Estadística descriptiva e inferencial y una introducción al método científico

Carlos de la Puente Viedma

UCM EDITORIAL COMPLUTENSE

Donoso Cortés, 63 – 4. planta (28015) Madrid
Tels.: 91 394 64 60/1 Fax: 91 394 64 58
e-mail: ecsa@rect.ucm.es
www.editorialcomplutense.com

Primera edición digital: octubre 2009

ISBN: 978-84-7491-992-9

Estadística descriptiva e inferencial y una introducción al método científico
Carlos de la Puente Viedma

A Rosa, Daniel y Jaime

Índice

1 Prólogo

Este manual quiere ser una ayuda o referente para quienes quieren acercarse a la Estadística. El empeño ha sido que fuese concreto en la exposición, práctico y útil, pero sin escatimar información. Pero serán quienes se acerquen a leerlo los que decidan si se ha cumplido el objetivo. La motivación para escribirlo ha sido que el enfoque está basado en la experiencia obtenida a través de los años que he impartido una asignatura de estas características, las preguntas realizadas por el alumnado y mi propio proceso teórico como sociólogo.

Se acompaña con una síntesis de los orígenes y evolución de *Homo sapiens sapiens*, hasta nuestros días. Se presenta la aparición de "lo social", en base a algunas hipótesis actuales. La aparición de la capacidad de "conocimiento" y "consciencia".[1] Sin entrar en las discusiones que plantean los diferentes paradigmas sobre la cuestión, aunque facilitaremos bibliografía para quienes deseen ampliar lo que se dice en estas páginas.

Se abordará el considerado Método Científico, la relación con la teoría, las técnicas, el objeto y el sujeto y la realidad en la que se insertan.

El resto de los capítulos están dedicados a la estadística como técnicas de descripción y análisis de datos.

Muchas gracias.[2]

[1] Por la dificultad en diferenciar los términos "conciencia" y "consciencia", se utilizan como sinónimos. Ambos proceden del latín *conscientia* literalmente "con conocimiento". No se hace referencia al concepto psicoanalítico de consciente que tendría otro tratamiento, terapéutico y teórico.

[2] Sugerencias cdelapuente@cps.ucm.es Poner en Asunto: Libro EDeIyMC.

2 Introducción

Siguiendo el criterio del considerado Método Científico, utilizado para que el conocimiento que se obtenga sea considerado como científico, antes de empezar a hablar de algo se debe definir ese *algo*. Según este criterio, se definen los primeros conceptos considerados clave que se van a utilizar: Método, Científico, Método Científico, Ciencia, Teoría y Estadística.

Una aclaración previa consiste en diferenciar los conceptos *descubrimiento* e *invento*. Según el Diccionario de la Real Academia Española (2008), "descubrir" es "Destapar lo que está tapado o cubierto", "inventar" es "Hallar o descubrir algo nuevo o no conocido", puede ocasionar alguna confusión al incorporar el verbo *descubrir* en su definición. Si no se diferencia *descubrir* de *inventar* se pierde la riqueza de matices de los dos conceptos.

En Oxford English Dictionary (2008), la definición nº 8 de *discover* es: "Ver o tener conocimiento de algo (previamente desconocido) por primera vez".[3] Utiliza una cita de Hugh Blair para ilustrar *discover* y diferenciarlo de *invent*, "Nosotros inventamos cosas que son nuevas; nosotros descubrimos lo que era antes oculto. Galileo inventó el telescopio; Harvey descubrió la circulación de la sangre" (ver nota 3). Otro ejemplo es la Fuerza de la Gravedad y Newton. Se puede considerar que Newton observó una *Fuerza* que emitían los cuerpos y que producía la atracción de otros cuerpos. En ese sentido, esta fuerza es un descubrimiento, pero el nombre *Fuerza de Gravedad* se puede considerar un invento.

En *Oxford English Dictionary* (*Ibíd.*), *a*l definir *invent,* introduce también el concepto *descubrir.* Para diferenciarlos, por *inventar* se propone considerar *hacer algo nuevo que no existía antes y por lo tanto no implica descubrimiento* y por descubrir se propone *destapar, ver, dar a conocer algo a los demás que existía previamente y que no era conocido por nadie o que nadie tenía conciencia de ello*. Aunque la definición y diferenciación puede ser clara, probablemente no lo sea tanto su aplicación.

Definiciones de: Método, Científico, Método Científico, Ciencia, Teoría y Estadística:

- ∉ Método: Formado por el prefijo *meta-* (fin al que se dirigen las acciones) y el sufijo *–odo* (camino) se puede considerar como "el camino hacia algo". En latín clásico *modo de proceder*, y en griego antiguo *persecución del conocimiento, modo de investigación* (ver nota 3) (*Ibíd.)*.
- ∉ Científico: Que tiene que ver con las exigencias de precisión y objetividad propias de la metodología de las ciencias (ver nota 3) (Oxford English Dictionary, *op. cit.*; Real Academia Española, *op. cit.*).
- ∉ Método Científico: Un modo de proceder que ha caracterizado a la ciencia natural desde el siglo XVII, que consiste en la observación sistemática, medida, y experimentación, y la formulación, comprobación, y modificación de hipótesis (ver nota 3).[4]

---

[3] Traducción propia.

[4] "scientific method noun" The Oxford Dictionary of English (revised edition). Ed. Catherine Soanes and Angus

- Ciencia: La definición de Ciencia es amplia y compleja y a veces el término hace referencia a un proceso y en otras ocasiones es el resultado de ese proceso (Tezanos, 2006: 393). Como definición breve, se considera ciencia "la actividad intelectual y práctica que abarca el estudio sistemático de la estructura y conducta del mundo físico y natural a través de la observación y la experimentación".[5] La ampliación de este punto se puede ver en Tezanos (2006: 393-396).
- Teoría: [Matemáticas] "La colección de teoremas y principios asociados con algún objeto o concepto matemático". [Ciencia y Tecnología] "Un intento de explicar cierta clase de fenómenos deduciéndolos como las consecuencias necesarias de otros fenómenos considerados más primitivos que son menos necesarios explicar" (McGraw-Hill, 2002). "Una suposición o un sistema de ideas con las que se pretende explicar algo, basado en principios generales independientes de la cosa a ser explicada o la teoría se usa para describir lo que se supone que pasa o puede pasar, con la implicación de que puede no pasar".[6] A finales del siglo XVI Teoría era: "Una concepción o esquema mental de algo que se iba a hacer, o del método de hacerlo; una declaración sistemática de reglas o principios que se debían seguir" (ver nota 3) (Oxford English Dictionary, *Op. cit.*).
- Estadística: "Estudio de los datos cuantitativos de la población, de los recursos naturales e industriales, del tráfico o de cualquier otra manifestación de las sociedades humanas" (Real Academia Española, *Op. cit.*). "La disciplina científica que trata de la recogida, análisis, y presentación de datos" (ver nota 3).[7]

Estos *términos* y las *operaciones* que ellos implican se consideran, lo primero un invento y lo segundo un descubrimiento. Las *operaciones* de teorización, metodología y análisis, en la forma en que se han definido se encuentran en la naturaleza y es el cerebro quien las realiza, aunque de forma automatizada, que consideramos inconsciente o que no es del todo consciente. Visto como un artefacto o como "una cosa" (Durkheim, 1895/1978; Aboitiz *et al.*, 2007; de la Puente, 2006), el cerebro es una máquina de elaborar explicaciones de la realidad que le rodea, de construir "Tipos Ideales".[8] Recoge información, la procesa y la analiza y como conclusión, toma decisiones que normalmente son adecuadas. Cuando las decisiones no son adecuadas, se puede producir un fatal desenlace. El análisis que realiza el cerebro no es sólo estadístico, sino que se puede considerar también matemático y balístico.

Stevenson. Oxford University Press, 2005. Oxford Reference Online. Oxford University Press. Universidad Complutense de Madrid. 2 June 2008 http://www.oxfordreference.com/views/ENTRY.html?subview=Main&entry=t140.e68940.

[5] "science *noun*" *The Oxford Dictionary of English* (revised edition). Ed. Catherine Soanes and Angus Stevenson. Oxford University Press, 2005. *Oxford Reference Online*. Oxford University Press. Universidad Complutense de Madrid. 13 January 2009 <http://www.oxfordreference.com/views/ENTRY.html?subview=Main&entry=t140.e68932>

[6] "theory noun" The Oxford Dictionary of English (revised edition). Ed. Catherine Soanes and Angus Stevenson. Oxford University Press, 2005. Oxford Reference Online. Oxford University Press. Universidad Complutense de Madrid. 2 June 2008 <http://www.oxfordreference.com/views/ENTRY.html?subview=Main&entry=t140.e79551>

[7] "statistics" A Dictionary of Genetics. Robert C. King, William D. Stansfield and Pamela K. Mulligan. Oxford University Press, 2007. Oxford Reference Online. Oxford University Press. Universidad Complutense de Madrid. 2 June 2008 <http://www.oxfordreference.com/views/ENTRY.html?subview=Main&entry=t224.e6186>

[8] Tiene las mismas características que el "Tipo Ideal" definido por Weber, ya que es subjetivo y está basado en la experiencia previa e histórica.

El cerebro de *Homo sapiens sapiens*, es un órgano que recoge información, la almacena y la analiza. La compara con la información previa y en base a ello toma una decisión. Se puede concluir que, en última instancia, en todo el proceso de adquisición de conocimiento, la herramienta analítica y de toma de decisiones final es el cerebro.

Ejemplo 1: Cuando se accede al Metro, se tiene una información previa de a qué lugar se entra (cuando es la primera vez y no se dispone de esta información el estado de vigilia y de atención son mayores y un cierto estado de angustia ante lo desconocido), al posicionarse en el anden y en base a la información previa se detecta cual es el lugar más adecuado, con menos público, para encontrar un asiento o quedarse cerca de la puerta si el lugar de destino está próximo. Existen otras muchas consideraciones que tiene en cuenta el cerebro, pero por su prolijidad se llenarían varias páginas con información que a veces es irrelevante o que sólo es ruido y que el cerebro sabe eliminar o filtrar "automáticamente". Los análisis realizados también se diferencian cualitativa y cuantitativamente en función de las características de la persona.

Ejemplo 2: Si una persona cruzase una gran avenida con el semáforo en rojo (cosa que no se debe hacer) cuando hay una gran afluencia de vehículos, el cerebro entra en un estado de procesamiento frenético para evitar ser atropellado y lo más probable es que llegue a la acera opuesta sin sufrir ningún percance. En esta operación, el cerebro de la persona tiene una información previa muy abundante que va desde el conocimiento de lo que le pasaría si es atropellado, hasta las características de los vehículos que tiene que esquivar.

Continuamente y de forma automática está realizando cálculos estadísticos, matemáticos y balísticos, comparándolos con la información anterior y en base a todo ello dando pasos adelante y atrás para tratar de no ser atropellado y además lo está haciendo en combinación con los cerebros de los conductores, previendo tanto el uno como los otros hacia qué lado se van a mover. Es una operación tan compleja que sólo es posible de manera automática o no-consciente, ya que realizar toda esa cantidad de cálculos y operaciones en intervalos tan cortos de tiempo de forma consciente sería muy difícil o imposible.

Si los cálculos se hiciesen con una computadora, que probablemente no exista, pero aunque existiese, el tiempo de introducción de datos y de proceso serían tan largos y tediosos que antes de empezar nos habría atropellado un coche y aunque todo esto fuese un proceso tan rápido como el que hace el cerebro, la computadora no se puede mover y si somos portadores de la computadora, entonces volvemos a empezar, porque el que toma la decisión en última instancia es el cerebro.

Ejemplo 3: Otro ejemplo puede ser cruzar la gran avenida con el semáforo en verde por el paso de peatones. El fallo del cerebro al realizar el análisis en esta ocasión tendrá consecuencias menos cruentas ya que los cálculos se realizan sobre los peatones que llevan el mismo sentido que nosotros y los de sentido opuesto. Un fallo en este caso no es tan dramático. Probablemente, en esta ocasión, lo más destacado serían las estimaciones de si se cumplen las reglas sociales de urbanidad de cómo y por dónde se transita y si los otros cumplen estas reglas.

La diferencia entre la aplicación consciente que ha desarrollado *Homo sapiens sapiens* de la teoría, el método y la estadística e incluso las matemáticas y la balística, y la aplicación que hace el cerebro, probablemente, está en que éste último lo hace de forma analógica o por lo menos la información que procesa es de este tipo y el ser humano lo ha digitalizado, entendiendo como tal, las fórmulas y protocolos desarrollados.

Para completar esta introducción, la estadística no sólo nos permite hacer tabulaciones y análisis, sino que aumenta la capacidad analítica y la manera de ver y analizar los hechos.

También es una gimnasia para el cerebro. Sucede como con los músculos, cuanto más se ejercita mejor preparado está. Los puntos más débiles de todo el proceso pueden ser: la adecuación del conocimiento almacenado, que la forma de recoger la información sea adecuada y que la pregunta sea correcta. Si la pregunta que se plantea no es correcta es casi completamente imposible que se llegue a la respuesta adecuada.

Como resumen de este Epígrafe, con la Estadística, el Método y las Técnicas de Investigación se realizan de forma sistemática operaciones que el cerebro las hace de forma habitual y cotidiana. Una diferencia es que en la aplicación del proceso científico se utilizan unas pocas variables, mientras que el cerebro utiliza una miríada. Además, en última instancia, la auténtica herramienta analítica y con la que se toman las decisiones es con el cerebro.

2.1 Sistema perceptivo: Los sentidos

Un aspecto fundamental en la captación de la realidad es el concepto que tenemos de los sentidos. La idea que tenemos de cómo percibimos la realidad a través de los sentidos y cómo funcionan realmente, condicionan las características atribuidas a la realidad, la realidad percibida y cómo la tratamos.

Ejemplo: se va a plantear un caso evidente de los problemas que ha ocasionado y ocasiona el concepto que tenemos de *ver*. El verbo *ver* se considera gramaticalmente transitivo "yo veo la casa" porque la acción de ver, gramaticalmente, cae en la casa. Probablemente, esta característica ha producido la gran confusión sobre la percepción y formación de la realidad. La acción de ver, fisiológicamente, es intransitiva o receptiva o se puede plantear que la acción del verbo *ver* es intransitiva.

Homo sapiens sapiens[9] no ve el exterior, sino que los objetos del exterior proyectan la luz (fotones) que reciben y ésta entra por el sentido de la vista de tal manera que se proyecta o se crea la imagen en la parte correspondiente del cerebro. Así es que lo que se "ve" es una representación (concepto fenomenológico de E. Husserl) del exterior que el cerebro recrea en el interior con la información de la que se dispone (concepto fenomenológico de A. Schutz). Entonces, las características que se asocian a las cosas son las que le asigna el cerebro, aunque probablemente, a veces, estas puede que no estén muy lejos de las características reales del objeto. Entonces, la realidad no es vista, sino que es recreada en el interior del cerebro. La pregunta podría ser entonces, "¿Cómo una imagen creada dentro de nosotros, en el cerebro, él mismo nos hace creer que nosotros estamos dentro de ella?" No obstante, se puede decir, que con el conocimiento del que dispone hoy día es imposible que se pueda saber como es, exactamente, la realidad, por ejemplo los colores (Ver referencias en: de la Puente, 2007 a).

El ser humano percibe la realidad por medio de los sentidos. Conocer su funcionamiento le da a los aspectos ontológicos y epistemológicos del *objeto* un sustrato físico y por lo tanto un sentido objetivo, lo que proporciona carácter de cientificidad. Desde esta base se puede decir que la formación (que es subjetiva) de la realidad percibida, es tratada de forma objetiva, esto es, la subjetividad es objetivamente subjetiva y es probable que no pueda ser de otra manera. La Tabla 1 es una clasificación de los sentidos.

[9] Probablemente se puede aplicar al sistema de visión de todos los seres vivos que poseen uno.

Tabla 1 Clasificación de los sentidos

	Sentido fisiológico	Percepción	Características	Elemento físico percibido	Característica física del elemento percibido	Característica del sentido
Sentidos Sujeto/objeto	Vista*	Color Movimiento 3D	Profundidad Distancia Dimensiones etc.	Ondas electromagnéticas	Ondas	Subjetivos
		Características	Peso Tamaño etc.			
	Oído*	Distancia Origen Movimiento Características	Peso Tamaño etc.	Ondas acústicas		
	Olfato*	Bondad Maldad		Moléculas	Químicos	
	Gusto*	Bondad Maldad				
	Tacto	Contacto Resistencia				Objetivos
		Dureza Blandura Suavidad Aspereza Elasticidad Flexibilidad Ductilidad Humedad Fluidez Temperatura etc.		Propiedades mecánicas de la materia		Subjetivos

Notas:
* Sentidos especiales.
Fuente: ver referencias en: C. de la Puente, 2007 a.

El sistema de la visión, probablemente, es el más complejo y sofisticado de todos los sistemas y órganos que se conocen del cuerpo humano y también el sentido más subjetivo porque no forma las imágenes por contacto directo con la realidad, sino a partir de ondas electromagnéticas reflejadas por la realidad, lo que supone que las imágenes se forman por el contacto de los fotones de la luz percibida con los sistemas neuronales correspondientes. En

esta escala de subjetividad, el siguiente en subjetividad puede ser el sistema auditivo. Las ondas sonoras que recibe el oído producen un trabajo mecánico en el sistema auditivo que convierte en una miríada de sonidos diferentes. Otro sistema sofisticado es el sistema vestibular que mantiene el equilibrio y la orientación.

Los sentidos del gusto y el olfato se pueden considerar los terceros en subjetividad, ya que la formación del olor o el sabor del exterior es a través de estímulos químicos a partir de las moléculas que perciben, lo que supone cierto contacto con la "cosa".

El sentido del tacto se puede considerar el más objetivo porque el cerebro entra en contacto directo con la realidad a través de la piel, que transmite las señales al sistema eferente y este a través de las canalizaciones correspondientes llega hasta las zonas correspondientes del neocórtex. Esta escala de subjetividad supone un cierto orden. La conclusión es que el cerebro construye una imagen de la realidad exterior y que con el conocimiento del que se dispone hoy día, probablemente no se puede saber objetivamente como es o qué características tiene la realidad externa material y objetiva: qué color, sabor, olor, tamaño, distancia, volumen, etc. tiene. Porque la realidad inmaterial (instintos, emociones, pensamientos, sentimientos, conducta, hechos sociales, etc.) del objeto, que se considera inmaterial y subjetiva, su construcción por parte del observador (sujeto) no puede ser otra cosa que subjetiva. Repitiendo la conclusión anterior, se asume que la subjetividad es objetivamente subjetiva.

Estos puntos pueden ser tratados y ampliados con manuales de neurociencia como por ejemplo, Kandel *et al.* (2001); Bear *et al*. (1998); Guyton (1994) y manuales de neurofisiología y neuroanatomía (Ferner & Staubesand, 1974; Sobotta *et al.*, 1996; Williams, 2001).

3 La Cultura el Lenguaje y lo Social

Probablemente, la aparición de la Cultura, en el sentido sociológico y antropológico (Tezanos, 2006: 253-276), aunque es un acontecimiento de difícil explicación en el proceso de la Evolución, se puede afirmar que es un *hecho* y lo mismo se puede decir del *lenguaje*. También se puede considerar otro *hecho* el Sistema Social humano. El *por qué* han aparecido es una respuesta que se da con el planteamiento de la Hipótesis del Residuo y para el *cómo*, se hace por referencia al proceso que plantean el conocimiento que se tiene al día de hoy de ciertos hechos y el planteamiento de otras Hipótesis.

Se puede asumir que la Cultura es una entidad de mayor amplitud que el Sistema Social humano y que este está configurado o constituido por Instituciones y es lo que estudia la Sociología. En la forma que lo plantea Durkheim, la Sociología es la ciencia que estudia las Instituciones.[10] Los tipos y clasificación de los hechos objeto de estudio de la Sociología se verán más adelante.

Plantear la aparición de la Cultura, el Lenguaje, el Sistema Social humano, el Conocimiento Científico y la Consciencia, puede servir de ayuda para comprender y estudiar la Realidad Social, y se resume en tres hechos: Cultura, Lenguaje y Consciencia. Asumiendo que la Cultura engloba al Sistema Social y al Conocimiento Científico y todos ellos deben ser producto de la Consciencia que en última instancia es un producto de ciertas partes del cerebro.

En otro lugar (de la Puente, 2007 a), la aparición de la Cultura y el Lenguaje se ha planteado como la Hipótesis del Residuo. La Cultura y el Lenguaje aparecen de forma accidental o aleatoria en el proceso evolutivo del ser humano, porque su cerebro reúne ciertas características, por lo tanto se considera que su aparición es de forma residual o accidental. No obstante, al hacer su aparición en un ser de la Evolución, se convierte en un producto de la misma y por tanto en objeto de estudio.

La Cultura y el Lenguaje deben haber favorecido la formación de *lo social* y aunque no se sepa exactamente qué es, su manifestación se puede ver en infinidad de situaciones de la vida diaria, por ejemplo, el que dos vehículos se puedan cruzar en una carretera de doble dirección o puedan circular en la misma dirección a más de 100 Km/h. Que un grupo de personas que no se conocen puedan estar encerradas en un ascensor un cierto período de tiempo. Que una masa de individuos pueda estar encerrada en un local a oscuras viendo una película durante dos o tres horas, son todos ellos acontecimientos que deben ser posibles por la existencia y manifestación de lo social y lo social se tiene que producir por la existencia de los hechos sociales (las Instituciones) coercitivos y punitivos que son, entre otros, las costumbres, la tradición, las normas de urbanidad, la ética, la moral, las señales de tráfico, el Código de la Circulación, el Código Civil y el Código Penal.

Entonces, haciendo un símil, se puede decir que "lo social" es como la electricidad que, dicen, no se sabe exactamente que es, pero sí se sabe que existe por sus manifestaciones y efectos, e incluso se puede producir, distribuir y controlar. Lo social podemos decir que no se sabe que es, y que no es algo material, pero se puede ver sus efectos y manifestaciones. Probablemente se puede decir también que se puede producir, distribuir y controlar.

[10] Las Instituciones, en definitiva, son los hechos sociales (Durkheim, 1895/1978: 30).

3.1 Proceso de *Homo sapiens sapiens*, la consciencia, el conocimiento y la ciencia

A *Homo* se le asignó la característica de *hábilis* por su capacidad o habilidad de hacer utensilios y herramientas. Posteriormente aparece *Homo erectus* y esta etiqueta es debida a su bipedestación y postura erguida. A *Homo sapiens* se le atribuye la *capacidad de conocer y a Homo sapiens sapiens*, la *capacidad de tener conciencia que puede conocer*. Siguiendo esta línea evolutiva, actualmente podemos decir que *conocemos que tenemos conciencia de que podemos conocer*, algo así como *Homo sapiens sapiens sapiens*, o en vez de esta forma cuatrinomial se podría abreviar diciendo *Homo meta-sapiens*. Este breve esquema pone de manifiesto la capacidad del humano de tener conciencia. Actualmente surge la pregunta de *cuándo* aparece la conciencia y *qué* y *cómo* es lo que la produce, pero independientemente de estas preguntas, se puede considerar un *hecho*.

Como resumen y para diferenciar las etapas de evolución de *Homo*, en un sentido estrictamente cultural y por el interés expositivo las fases, desde el período Neolítico hasta el presente, se establecen como: desde el Neolítico hasta la cultura Sumeria: *Homo pre-sapiens*; desde los sumerios y hasta el período griego: *Homo sapiens*; desde la época griega hasta el siglo XVII, considerado como el siglo de los genios: *Homo sapiens sapiens*, y desde el siglo XVII hasta la actualidad: *Homo meta-sapiens*. Esta clasificación más detallada nos ayuda a diferenciar las etapas, considerando que los cambios no son puntuales, sino que se corresponden con períodos de tiempo.

De manera formal se puede fijar el período de aparición de la conciencia y del hombre moderno que se ha llamado *pre-sapiens* (morfológica y conductualmente), desde la llamada Revolución del Neolítico,[11] y la llamada "Revolución de los Símbolos" o "Revolución Cultural" o "mutación mental" (Watkins, 2000 a, 2000 b; Bednarik, 2008; Asouti, 2006; Naccache, 2003; Bar-Yosef, 2002). Este período se considera el paso de las sociedades cazadoras-recolectoras a sociedades agrícolas-ganaderas, y las bases para la aparición posterior de: la sedentarización, los asentamientos urbanos estables, la cultura, el sistema social (base del actual), la ciencia (base de la actual), el conocimiento (base del actual), la escritura,[12] las instituciones, el comercio, los viajes, las instituciones educativas y las leyes. (Lara Peinado, 1988, 1998; Molina, 2000; Masó, 2007). Se considera un proceso largo que algunas facetas se inician hace un millón de años, otras desde hace 50.000 ó 40.000 y es desde hace 12.000 años que se puede considerar se empieza a consolidar.[13] Independientemente del proceso, el tiempo que tarda y los hechos que se producen, la cuestión que nos interesa es que el hecho "se produce" y nos llega hasta el presente.

Formalmente podemos asociar la culminación de *sapiens,* en el sentido de tener conciencia, al período de los Sumerios, fundamentado en el "darse cuenta de ..." que debió suponer el control sobre las cosechas (relación causa-efecto entre sembrar-florecer), la relación causa-efecto al observar las estaciones del año, relacionadas con las fases lunares que les permitió iniciar el desarrollo del calendario actual, el efecto de las *Instituciones* en la incipiente vida social, materializado en la losa de basalto en la que estaba escrito el Código de Hammurabi. Estos son algunos ejemplos de una actividad social incipiente que se puede ampliar en H. Crawford (2006).

Homo sapiens sapiens se asocia a la época que se origina en la Grecia Clásica y que alcanza hasta el siglo XVII. En este período tienen conciencia de que son conscientes y del

---

[11] Para ver un proceso relacionando la paleoantropología y la neurosociología ver C. De la Puente (2007 a) y su bibliografía. Desde un punto de vista neurocientífico ver F. Aboitiz *et al.* (2007).

[12] Que debía servir a un complejo lenguaje polisilábico para representar sonidos mejor que imágenes e ideas y por lo tanto un lenguaje más formal con sonidos fonéticos, aunque una forma de lenguaje que son más que ruidos guturales ya debía existir previamente, para comunicarse entre sí. Después de diferentes modos de escritura, en Occidente utilizamos el "izquierda-derecha arriba-abajo" (Kerckhove, 1987).

[13] Los periodos de tiempo son orientativos y pueden variar entre autores.

conocimiento. Por los descubrimientos y avances que se producen durante el siglo XVII y que están fundamentados en los siglos anteriores, desde este período, se le puede atribuir a *Homo* la característica de *meta-sapiens*, y es el que llega hasta el presente.

Probablemente el presente se corresponda con otra etapa que las generaciones futuras se ocuparán de dar nombre y que podría ser *Homo scientifícus*, que se caracterizaría por una tendencia al reconocimiento de los hechos y el abandono de las formas de conocimiento no científicas. Esta etapa sería el estadio científico de Comte.

Por los registros fósiles y arqueológicos de los que se dispone en la actualidad, este proceso se materializa en la zona del Golfo Pérsico, entre los ríos Éufrates y Tigris, asociado a los sumerios, y se puede considerar el origen de la Cultura Occidental. Otras ocurrencias serían la aparición de la cultura hindú, la asiática, la india americana y la arábiga, y en todas ellas, y considerado como un hecho, aparece la cultura, el leguaje, la escritura, la religión y un sistema social, que en todos los casos es en mayor o menor grado, más o menos evolucionado. La Cultura Occidental, y según sus patrones culturales, se puede considerar que es la más evolucionada.

En el sentido biológico, hay diferentes hipótesis sobre la línea evolutiva de la forma trinomial *Homo sapiens sapiens*. Si es la evolución desde el *sapiens Neandertalensis*, entonces es correcto porque somos más evolucionados que ellos. Si es a partir de *Cromañon* y se le considera *sapiens*, también sería correcta la forma trinomial. Pero si no se acepta la línea evolutiva de *Cromañon* y fuésemos biológicamente diferentes de *Neandertal*, entonces nos correspondería la forma binomial de *Homo sapiens*. Los análisis genéticos tampoco presentan un planteamiento claro para determinar una línea evolutiva (Bednarik, 2007). En cualquier caso, los sumerios y sus contemporáneos serían los continuadores de la tradición cultural de sus antecesores. La cultura griega debió tener influencias de la cultura sumeria, porque su alfabeto combina características de la escritura sumeria y egipcia. La influencia sumeria debió llegar directa o indirectamente a través de las invasiones de los pueblos indo-europeos y las tribus que descendieron desde las sierras próximas que extinguieron a los sumerios pero que debieron ser herederos de su cultura y llegarían hasta las costas del Mediterráneo (De Kerckhove, 1987).

Sobre el origen de *Homo*, una de las hipótesis pone en cuestión su "radiación africana". Pero si no ha habido radiación en ningún punto de la Evolución (*Eucariotas, Cordados, Terápsidos, Primates, Homínidos* y *Homo,* por ejemplo), entonces ha habido varios puntos de origen o la radiación se ha producido en alguna etapa y después ha seguido un proceso paralelo, porque existen restos fósiles de *Homo* en diversos puntos del planeta. Esta última hipótesis de "Evolución Paralela" sería otra explicación a las diferencias morfológicas entre razas, además de la debida al proceso de selección-adaptación del medio (Bednarik, 2007). Como se ha mencionado anteriormente, la herencia genética no permite resolver las dudas. Hace unos 200 millones de años la Tierra era un único continente (*Pangea*) y empezó a separarse hasta la forma actual, es la teoría de la *deriva continental* (Wegener, 1983).

Otra vez, independientemente de cómo haya sido el proceso, se puede considerar como un hecho que los sumerios han existido, que son los primeros a los que se les atribuye la creación de tablillas con textos grabados, los primeros a los que se les atribuye la creación de la primera ciudad, el primer código o conjunto de leyes y las Instituciones y por lo tanto "lo social". Que los griegos, entre otros, son los herederos de alguna parte de su tradición. Que aún siendo un hecho que todas las civilizaciones tienen cultura, lenguaje, sistema social y religión, en ningún otro lugar se ha detectado con la claridad que en los sumerios, porque hoy se pueden observar tribus que su nivel de desarrollo social, tecnológico y cultural, se puede considerar inferior al de los sumerios.

4 Aportaciones a la definición de persona como objeto de investigación[14]

Para dar una definición de *persona*, se va a seguir la regla fundamental de Durkheim, "considerar los hechos sociales como cosas" pero aplicado a la persona, considerarla como una *cosa* y no considerar aspectos filosóficos, teológicos ni psicológicos. La finalidad es buscar fundamentos materiales y objetivos desde los que respaldar las características y definiciones. Las preguntas que he consideran son: ¿Quién soy yo? ¿Cómo soy yo? y ¿Qué soy yo? Las respuestas a estas preguntas se consideran de interés para el estudio de *lo social* de *Homo sapiens sapiens*. En términos fisiológicos *Homo sapiens sapiens*, es un cuerpo, que se considera la base de la persona civil de tal manera que: la persona civil y sus contenidos son del cerebro y no al contrario, que el cerebro es de la persona civil.

Para hacer una aportación a la definición de las características de la *persona*, si se le quitan los conceptos aportados por la Psicología, la Filosofía y la Teología se prescinde, por ejemplo, de las Teorías de la Personalidad y las aportaciones psicoanalíticas de la estructura de la personalidad: *Yo*, *Superyo* y *Ello*. Desde la Teología se prescinde considerar el alma y desde la Filosofía se omiten términos como: óntico y ente. Estos conceptos presentan dificultades para darles una base material y objetiva.

De esta manera parece más apropiado aplicar el principio de Durkheim y considerar a la *persona* como "una cosa". Este mismo principio se aplica en otras Áreas de Conocimiento cuando hacen la siguiente observación "En este volumen hemos decidido tratar el estudio del origen y evolución del cerebro de los vertebrados, desde el sistema nervioso más simple, igual que los elementos que podemos observar en la Naturaleza" (Aboitiz, 2007).

Pero el significado de *persona* en latín, etrusco y griego es el mismo, "máscara". Entonces la máscara no es el *cuerpo*, sino la "persona" que se asigna al "cuerpo".

Entonces "la cosa" que nos queda es un cuerpo y buscando por "*body*" en siete diccionarios de habla inglesa se observa la misma definición: "La estructura física, incluso los huesos, carne y órganos, de una persona o animal". En el Diccionario de la Real Academia Española es: "Conjunto de los sistemas orgánicos que constituyen un ser vivo".

Las definiciones se orientan desde las preguntas ¿Quién soy yo? ¿Cómo soy yo? y ¿Qué soy yo? que pueden servir de referente. El *Yo* hace referencia a la *primera persona singular*, por lo que se considera una forma gramatical para referirse a la *persona*.

El *quién* hace referencia a *persona* y en el Diccionario María Moliner significa: *Pronombre interrogativo (con acento). —Es el único pronombre interrogativo aplicable a personas. Como todos los de esta clase, sirve para preguntar tanto en preguntas indirectas como en directas*. Entonces se pregunta por la persona, la mascara.

El *qué* hace referencia a *cosa* (el cuerpo) y en el María Moliner significa: *pronombre interrogativo. —«Que» representando una cosa, cualidad o determinación no expresadas o por las que se pregunta*. Entonces preguntamos por "el cuerpo".

El *cómo* hace referencia a características y en el María Moliner significa: *adverbio interrogativo que sirve para preguntar por el modo de ser o hacerse algo*. Entonces pueden ser características de la *persona* y del *cuerpo*. Las características de la persona son las del personaje que interpreta y asumimos que son del *carácter*, de la *persona-lidad*, y las características del *cuerpo* son físicas.

El Yo; *persona civil;* nombre y apellidos, y *persona* se considera que hacen referencia

[14] Este Epígrafe es una adaptación de una Comunicación presentada en las Jornadas de Sociología. Sociedad y Tecnología: ¿Qué futuro nos espera? Alcalá de Henares Madrid, 20 y 21 de noviembre de 2008.

a la misma cosa y el *cuerpo* es el soporte físico. Se expone el origen y el fin de la *persona*, el origen y el fin del *cuerpo* con una breve exposición de su evolución ontogenética o desarrollo y cuándo se produce la fusión de los dos: *persona* y *cuerpo*. Así como algunas consecuencias legales y sociales que tiene esta fusión y cuándo se produce la separación.

Homo sapiens sapiens es un *cuerpo* y la *persona* es una máscara asignada al cuerpo y ella y sus contenidos es una creación, si se quiere virtual, de la parte del cuerpo u órgano que tiene capacidad para ello, el cerebro, principalmente los lóbulos prefrontales y frontales del neocórtex. Entonces la *persona*, que es la persona civil y sus contenidos, son del cerebro y no al contrario, que el cerebro es de la persona civil.

Fernández Elías (1874/2005: 16) plantea la diferenciación entre persona y cuerpo,

> "Si el hombre natural respira el aire, digiere los alimentos, es sensible al calor, pesa, está enfermo o sano, muere, etc., el hombre social respira la comunicación con sus semejantes, digiere la enseñanza y el ejemplo, es sensible a la riqueza, es capaz de jerarquía, tiene plenitud de derechos, o está privado de algunos, tiene nombre que se trasmite, se trasforma en la sucesión a que da origen, etc."

4.1 Definición de cuerpo

El *cuerpo* es la base material y objetiva y se genera a partir de la inseminación de un óvulo por parte de un espermatozoide. El óvulo una vez inseminado cierra sus paredes de tal manera que no pueden entrar otros espermatozoides y empieza a dividirse durante cuatro días hasta formar la *mórula*.

Al final del proceso de *gastrulación* que se produce entre el día 13 y el 19, se han creado tres capas de células primitivas: *endodermo*, *mesodermo* y *ectodermo*, "Se cree que los primeros animales multicelulares poseían… una capa ectodérmica externa (de células), una capa media mesodérmica y una interna endodérmica" (Williams, 1998). "El embrión se inicia como un disco plano con tres capas diferentes de células que reciben el nombre de endodermo, mesodermo y ectodermo. El endodermo en último término da lugar al revestimiento de muchos de los órganos internos (vísceras). A partir del mesodermo surgen los huesos del esqueleto y los músculos. El sistema nervioso y la piel derivan completamente del ectodermo." (Bear, 1998). "A medida que las neuronas se diferencian, extienden axones que deben encontrar sus objetivos apropiados. … produciéndose en tres fases: la selección de la vía, la selección del objetivo y la selección del domicilio" (*Ibíd.*). "El resultado es la elaboración de una precisa red adulta de 100 mil millones de neuronas capaces de mover el cuerpo, percibir, emocionarse, y pensar" (Society for Neuroscience, 2008). El Sistema Nervioso extiende sus ramificaciones hasta contactar con los demás órganos y partes del cuerpo para llevar su control, a través de los cuales se va a sustentar/desplazar, percibir estímulos, alimentar y reproducir.

4.2 Definición de persona

La formación de la *persona* se inicia con el acto de dar nombre al cuerpo en el Registro Civil y se considera persona civil. Este acto es posterior a la aparición del *cuerpo*, y el tiempo que debe transcurrir varía dependiendo de la Administración. En España, el Código Civil establece en su artículo 30 que "Para los efectos civiles, sólo se reputará nacido el feto que tuviere figura humana y viviere veinticuatro horas enteramente desprendido del seno materno" y es a partir de este momento cuando puede inscribirse, darle nombre y apellidos, "poner" a la persona en el cuerpo.

La separación de la persona y el cuerpo aparece en el Artículo 32 del Código Civil que especifica "La personalidad civil se extingue por la muerte de las personas", y se separa del

cuerpo, aunque debería decir "por la muerte del cuerpo", ya que la *persona*, al ser una *máscara* no puede morir, sólo se *extingue*. De forma más precisa, la "separación" de la persona y el cuerpo se produce por la *muerte encefálica* porque se considera que el fin de la conciencia ocurre cuando termina la actividad eléctrica del cerebro y termina de forma irreversible.

Gazzaniga (1998) considera que el 98% de la actividad del cerebro está fuera del pensamiento consciente, incorporando como tales todas las funciones de control de los órganos que realiza el cerebro. Entonces ¿Se puede considerar que la parte del *Yo consciente* y sus *contenidos conscientes* son el 2% restante?

Entonces la *persona* es un nombre y sus contenidos. Durkheim y Mauss definen *persona* "como una categoría básica de pensamiento humano formada por las estructuras variables de las leyes y la moralidad"[15] y estos "contenidos" o "categoría básica" de pensamiento son una creación, si se quiere virtual, de la parte del cuerpo u órgano que tiene capacidad para ello, el cerebro, principalmente los lóbulos prefrontales y frontales del neocórtex.

4.3 Algunas consecuencias legales y sociales

En el Código Penal y en el Civil hay evidencias de la diferenciación entre *persona* y *cuerpo*. Entre las causas que eximen de la responsabilidad criminal están los Artículos,

> Artículo 20.1 El que al tiempo de cometer la infracción penal, a causa de cualquier anomalía o alteración psíquica, no pueda comprender la ilicitud del hecho o actuar conforme a esa comprensión.
>
> Artículo 20.2 El que al tiempo de cometer la infracción penal se halle en estado de intoxicación plena por el consumo de bebidas alcohólicas, drogas tóxicas, estupefacientes, sustancias psicotrópicas u otras que produzcan efectos análogos, siempre que no haya sido buscado con el propósito de cometerla o no se hubiese previsto o debido prever su comisión, o se halle bajo la influencia de un síndrome de abstinencia, a causa de su dependencia de tales sustancias, que le impida comprender la ilicitud del hecho o actuar conforme a esa comprensión.
>
> Artículo 20.3 El que, por sufrir alteraciones en la percepción desde el nacimiento o desde la infancia, tenga alterada gravemente la conciencia de la realidad.

En el Artículo 30 del Código Civil, se diferencia entre la persona y cuerpo, pero en ese acto se unen y en las eximentes se diferencia entre la persona y el cuerpo, porque si el cuerpo realiza un acto y la persona no es responsable, asumo que conceptualmente se les separa al eximir a la persona, pero se les junta porque son indisolubles y ello conlleva la exculpación del cuerpo. Pero entonces ¿No es cierto que el cuerpo ha realizado un acto? ¿Quién ha controlado entonces el cuerpo? ¿No se tendría que condenar al *cuerpo* y como la *persona* no se puede separar entonces también padece la condena?

Los Artículos 101, 102 y 103 del Código Penal hacen referencia a un período de educación de la persona equivalente a la pena que les hubiese correspondido cumplir si no hubiese habido eximente. ¿Entonces se requiere la reeducación de la *persona* para que controle el *cuerpo*?

---

[15] "person" *Dictionary of the Social Sciences*. Craig Calhoun, ed. Oxford University Press 2002. *Oxford Reference Online*. Oxford University Press. Universidad Complutense de Madrid. 9 November 2008 http://www.oxfordreference.com/views/ENTRY.html?subview=Main&entry=t104.e1249>.

Entonces en base a los planteamientos anteriores, la respuesta propuesta a cada una de las preguntas planteadas es:

¿Quién soy yo? La respuesta que se considera más adecuada es dar el nombre y apellidos, porque es la persona, la "mascara".

¿Cómo soy yo? La respuesta más adecuada se considera dar características físicas y de persona-lidad.

¿Qué soy yo? La respuesta más adecuada se considera la de "un cuerpo" según las características y definiciones dadas más arriba, *un conjunto de neuronas denominadas Sistema Nervioso que dispone y se conecta a una serie de órganos, tejidos y músculos que le permiten la percepción, el movimiento, el mantenimiento y la reproducción*. Aunque quizá sería más correcto sustituir la pregunta por ¿En dónde estoy yo? ¿En qué estoy yo? ¿Sobre qué estoy yo? ¿Qué re-produce el yo? ¿Qué soporta al yo?

4.4 Estudio empírico

Al inicio de algunas clases en la Universidad, y como introducción a cuestiones ontológicas, epistemológicas y metodológicas, se realizan las preguntas mencionadas, al alumnado. Posteriormente sirven para el desarrollo de diversos temas. La muestra se puede considerar intencional de 101 individuos que todos ellos son universitarios y en algunos casos ya tienen un título universitario. Las respuestas dadas a cada una de las preguntas son múltiples, pero para la tabulación se ha codificado siempre la primera respuesta, aunque se hacen comentarios sobre las demás respuestas dadas.

Según la Tabla 2, a la pregunta de ¿Quién soy yo? el 54,5% han contestado "Persona, ser humano, ciudadano, individuo" en primer lugar y el 17,8% han facilitado el nombre y apellidos. Esta pregunta puede tener un error de procedimiento, porque al ser el profesor de la signatura quien realiza la pregunta pueden haber evitado dar el nombre y apellidos por cuestiones de privacidad. No obstante, la mayoría de los que dan el nombre y apellidos son extranjeros. La respuesta suele ir siempre acompañada con respuestas de relación y parentesco con otras personas. Aproximadamente un 28,0% responde otras cosas.

Tabla 2 Tabla de frecuencias de la pregunta ¿Quién soy yo?

	Frecuencia	Porcentaje
Nombre y apellidos	18	17.8
Relación familia	2	2.0
Relación otros	1	1.0
Rol / estatus	10	9.9
Relación entorno	3	3.0
Persona, ser, humano, ciudadano, individuo	55	54.5
Depende del entorno	1	1.0
Características físicas	1	1.0
Características Personalidad	2	2.0
No sé definirme	1	1.0
Sé como me llamo pero no sé quien soy	1	1.0
No puedo responder	1	1.0
Cuerpo y mente	1	1.0
Sujeto social	2	2.0
Forma física con inteligencia	1	1.0
En quien me han convertido	1	1.0
Total	101	100.0

Las respuestas a la segunda pregunta (¿Cómo soy yo?) se muestran en la Tabla 3. El 77,2% contesta con características psicológicas o de personalidad y el 9,9% con características físicas. Esta pregunta presenta un acuerdo más generalizado. No obstante, hay aproximadamente un 13,0% que contesta otras cosas o no contesta.

Tabla 3 Tabla de frecuencias de la pregunta ¿Cómo soy yo?

	Frecuencia	Porcentaje
Características físicas	10	9.9
Características de personalidad, psicológicas, etc.	78	77.2
Relación otros	5	5.0
Persona, ser, humano	2	2.0
No quiere contestar	2	2.0
No puedo responder	1	1.0
Único e irrepetible	1	1.0
Total	99	98.0
Sistema	2	2.0
Total	101	100.0

En la Tabla 4, el 51,5% manifiesta que son “Persona, ser, humano, individuo, ciudadano”. Esta pregunta presenta un mayor nivel de confusión porque el porcentaje que contesta otra cosa (48,5%) es mayor que en las dos preguntas anteriores. Las opciones que más se aproximan a la respuesta que se ha dado como válida son: “Lo da la Naturaleza”, “Animal racional”, “Piel, órganos, vísceras y complementos artificiales”, “Cuerpo y mente”, “Animal social”, “Conjunto de partes”, “Un grupo de células”, “Forma física con inteligencia” y “Materia y espíritu” y estas respuestas son un 16,0%.

Tabla 4 Tabla de frecuencias de la pregunta ¿Qué soy yo?

	Frecuencia	Porcentaje
No se me ocurre nada	1	1.0
Persona, ser, humano, individuo, ciudadano	52	51.5
Características físicas	1	1.0
Características de personalidad	4	4.0
Lo que consideran los demás	1	1.0
Conjunto de acciones	1	1.0
Lo da la Naturaleza	1	1.0
Animal racional	2	2.0
Relación con otros	5	5.0
Igual que cómo soy yo	2	2.0
Piel, órganos, visceras y complementos artificiales	1	1.0
Proyecto de futuro profesional	1	1.0
Rol / estatus	9	8.9
Una proyección, posibilidades, realidades	2	2.0
No contesta	1	1.0
Una unidad	1	1.0
No puedo responder	1	1.0
Cuerpo y mente	3	3.0
Animal social	3	3.0
Conjunto de partes	2	2.0
Relación con el entorno	1	1.0
Un grupo de células	3	3.0
Forma física con inteligencia	1	1.0
La razón de las cosas o parte de ellas que hacen mis padres	1	1.0
Materia y espíritu	1	1.0
Total	101	100.0

El 33,7% coinciden en la categoría *persona* en las preguntas ¿Quién soy yo? y ¿Qué soy yo? Dando la misma respuesta en las dos lo que se interpreta como estado de confusión. En el resto de las celdas de la tabla de distribución conjunta de frecuencias, 355 tiene una frecuencia de 0,0%, y la frecuencia mayor es inferior al 8,0%.

Si se pregunta a un colectivo de personas universitarias y en algunos casos con un título universitario ¿Deberían saber la respuesta a las preguntas planteadas? ¿Deberían tener una idea clara de las respuestas adecuadas? ¿Son correctas las respuestas dadas en este escrito? En cualquiera de los casos, aunque existen posturas mayoritarias en alguna de las preguntas, no hay una respuesta de acuerdo total.

4.5 Conclusión

Es necesario asumir una definición de persona, de cuerpo y la relación entre ambos y considerarlo desde una base material y objetiva.

5 El Método Científico y el marco de la realidad

BREVE HISTORIA DEL MÉTODO CIENTÍFICO

Como se ha indicado, se puede considerar que la primera ciudad, los primeros textos escritos y los primeros códigos que regularon la vida social tienen su origen en Sumeria, son también los primeros que formalmente inician una forma de método científico y de la ciencia al observar los hechos que ocurren y tratar de elaborar explicaciones y construir modelos. Hicieron observaciones astronómicas, elaboraron un calendario con doce meses y corrigieron los desfases que se producían, desarrollaron las matemáticas (suma, resta, multiplicación, división, raíz cuadrada, ecuaciones), la geometría, las leyes. Su método estaba fundamentado en la observación de la realidad tal como lo entendemos hoy (Lara Peinado, *op. cit.,* 1988, 1998; Molina, *op. cit.,* 2000; Masó, *op. cit.,* 2007).

A los griegos les podemos atribuir la *conciencia de la conciencia,* empiezan a ser consciente de que tenemos conocimiento reflejado en los conceptos de inspiración platonianos de *doxa* y *episteme*, para diferenciar el conocimiento espontáneo o no científico del considerado científico. En este período, también se concibe al ser humano dentro del paradigma dualista de Aristóteles y Platón que supone la distinción: *materia-espíritu*; *cuerpo-alma*. Este paradigma se mantiene durante la Edad Media a través de la *Filosofía Escolástica* de Santo Tomás. Esta dualidad, probablemente, ha impedido hacer las preguntas adecuadas para buscar las respuestas.

La siguiente etapa considerada es el Renacimiento (s. XV), que supone la crítica del aristotelismo escolástico y el inicio del empirismo con Galilei (1632/1995), y Bacon (1620/1984) (método hipotético inductivo), desarrollado en el siglo XVII (Hobbes, Locke y Hume). Newton (1686/1987), sintetiza el método inductivo en el hipotético deductivo.

Simultáneamente, surge el racionalismo (Descartes, 1637/1986), como polo opuesto al empirismo. La razón (conciencia) humana, también es fuente de conocimiento. Introduce el concepto de mente o conciencia en lugar o además de alma o espíritu. La fusión de las corrientes empirista y racionalista se materializa en Kant (1787/2003).

En el Curso de Filosofía Positiva, Comte hace referencias directas al cerebro y conecta la *fisiología individual* con las *Físicas Sociales* (Comte, 1893/1968). Comte destaca el servicio que hizo Descartes al instituir un completo sistema de Filosofía Positiva que aplicó al mundo inorgánico y a las funciones físicas del mundo animal, pero estima que se detuvo cuando llegó al estudio del hombre, dejando éste al amparo de la Filosofía Metafísica y la Teología, e interrumpe la posibilidad de aplicarle los principios de la Filosofía Positiva (Martineau, 2000; Comte, 1893/1968).

Comte establece tres niveles para el estudio positivo de los elementos vivos: "Consideraciones filosóficas para el estudio general de la vida vegetativa u orgánica" (Lección 43); "Consideraciones filosóficas para el estudio general de la vida animal propiamente dicha" (Lección 44), y "Consideraciones generales para el estudio positivo de las funciones intelectuales y morales, o cerebrales" (Lección 45). Entendiendo que difiere menos el tercer apartado del segundo que el segundo del primero (Martineau, *ibíd.*; Comte, *ibíd.*).

Comte observa en los seres vivos, al menos en los animales superiores y en el hombre: actos afectivos e intelectuales. El estudio de estos actos pertenece al ámbito de la biología, la filosofía natural y las físicas sociales. La sociología, no sólo pertenecería a las ciencias sociales, sino también a las ciencias naturales, tomando como base las reglas de funcionamiento del órgano interprete/ productor de los actos: el cerebro; de esta manera quiebra la tradición de Descartes de estudiar al hombre sólo desde la filosofía metafísica y teológica (Martineau, *ibíd.*; Comte, *ibíd.*). Pero, probablemente, el primer sociólogo que sigue

esta línea es Warren D. TenHouten en 1972.

Popper (1994) inaugura una nueva etapa de la teoría del conocimiento, rechazando los puntos de apoyo absolutos de la razón pura y de los hechos puros, planteando la construcción de hipótesis interpretativas para falsarlas mediante el método de ensayo y error.

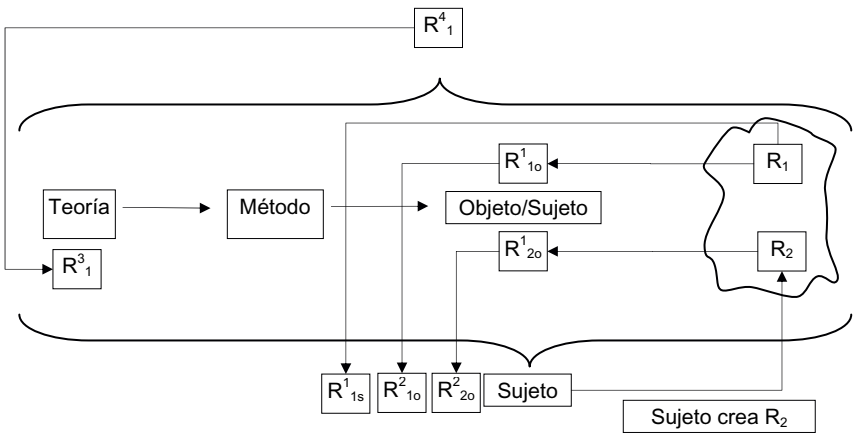
Thomas S. Kuhn (1977) y Paul K. Feyerabend (1989), representan las teorías postpopperianas de las "revoluciones científicas" y el anarquismo cognoscitivo de "contra el método", respectivamente.[16]

Frente al paradigma dualista de la tradición Platón-Aristóteles y Descartes diferenciando entre espíritu y materia, se presenta el paradigma monista-materialista. Los seres humanos tienen un cerebro que es material y objetivo, y de este emerge lo inmaterial y subjetivo: el comportamiento, o dualismo-rectificado al considerar una parte material: cuerpo y otra inmaterial: instintos, emociones y lo social (Biológico y Cultural).

Actualmente, relacionando el estudio del comportamiento subjetivo a partir de la materia, "una teoría que relacione ambas cosas, sostiene T. Nagel, puede suponer una transformación completa del pensamiento científico" (citado en, Kandel, 2001).

El denominado Método Científico consiste en una serie de pasos que difieren de un autor a otro, pero, básicamente, representan el mismo proceso. La aplicación de este método es lo que diferencia el conocimiento considerado científico (*episteme)* del considerado no científico (*doxa*). Pero es necesario tener el conocimiento teórico previo (Comte), tener un paradigma, para acercarse a conocer la Realidad. En Sociología el esquema que relaciona la Teoría, el Método, el Objeto, la realidad y el Sujeto, se puede representar según el Gráfico 1.

Gráfico 1 Esquema: Teoría - Método - Sujeto – Objeto - Realidad.



R_1	La realidad real que se asume que existe, aunque con el conocimiento que poseemos hoy día, probablemente, no es posible observarla tal como es.
R_2	La realidad creada por el sujeto derivada del propio proceso de la investigación (1ª hermenéutica).
R^1_{1o}	Representación que se hace el objeto, como sujeto, de la R_1 (1ª hermenéutica).
R^1_{2o}	Representación que se hace el objeto, como sujeto, de la R_2 (se puede considerar una 2ª hermenéutica).
R^1_{1s}	Representación que se hace el sujeto de la R_1 (1ª hermenéutica).
R^2_{1o}	Representación que se hace el sujeto de la R^1_{1o} del obeto (2ª hermenéutica).
R^2_{2o}	Representación que se hace el sujeto de la R^1_{2o} del obeto (sería una 3ª hermenéutica).
R^3_1	Representación teórica de la R_1.
R^4_1	Es la R final del proceso que puede confirmar o ser igual a R^3_1 o modificar a ésta. Si R^4_1 es igual a R^3_1 se mantiene y se confirma esta, pero si son distintas, entonces R^4_1 pasa a ser o midificar R^3_1.

[16] Una aplicación de la lógica de la ciencia en sociología se puede ver en W. Wallace (1980).

El proceso del denominado Método Científico es la aplicación del método sobre el objeto con el conocimiento previo sobre éste. El proceso, probablemente, esté muy cerca de ser que el Objeto se representa la Realidad, lo que convierte al objeto en objeto-sujeto. El sujeto se representa (observa) la representación de la realidad en el objeto-sujeto que éste manifiesta, Realidad que a veces es (re)creada por el sujeto[17]. El Sujeto trabaja con: la Realidad real, la que está "afuera" que con los conocimientos actuales, probablemente, no nos es posible saber como es; la Realidad real representada en el Objeto operando éste como sujeto; la Realidad re-creada por el Sujeto (cuestionarios, entrevistas, etc.); la representación en el Objeto de la Realidad re-creada por el Sujeto, y la Realidad del Marco Teórico que utiliza el Sujeto.

EL DENOMINADO MÉTODO CIENTÍFICO EN SOCIOLOGÍA

El esquema que se presenta se considera un resumen de fácil asimilación, comprensión y aplicación.[18]

1. Diseño Teórico

   1.1. Tema a Investigar

   Se expone el problema[19] (Miller, 1991: 13-20) o tema de investigación, poniendo de forma clara la definición de los conceptos que se van a investigar. También se debe hacer referencia a los motivos que han llevado a realizar la investigación.

   1.2. Marco Teórico (Documentación)

   El Marco Teórico recoge todo el conocimiento sobre el tema a investigar, sobre el objeto y también es una ayuda sobre la forma de investigarlos en base a las investigaciones anteriores. Se consideran tres grupos de paradigmas:[20] Paradigmas Teóricos, Paradigmas Técnicos y Paradigmas Epistemológicos.

   Los Paradigmas Teóricos dan los referentes para conocer-comprender la realidad. Los Paradigmas considerados son: Neurosociología (la línea francesa: Auguste Comte; la línea rusa: Lev S. Vygotsky, Alekxander R. Luria, Elkhonon Goldberg, la Línea Norteamericana: Warren D. TenHouten y la línea española Carlos de la Puente). Estructural-Funcionalismo (Emile Durkheim, Vilfredo Pareto, Talcott Parsons y Warren D. TenHouten). Sociología de las Emociones (Turner, TenHouten, etc.). Interaccionismo Simbólico (George Simmel, George H. Mead, Herbert Blumer). Institucionalismo (Emile Durkheim, Max Weber, Thorstein Veblen, Walter W. Powell y Paul J. DiMaggio). Teoría General de los Sistemas (Ludwig

[17] Paradigma fenomenológico y neuro-físico-químico.

[18] Basado en R. Hernández Sampieri (2007).

[19] Se considera "Problema" según la RAE "Planteamiento de una situación cuya respuesta desconocida debe obtenerse a través de métodos científicos". Pero no se considera como: "Conjunto de hechos o circunstancias que dificultan la consecución de algún fin" ni como: "Disgusto, preocupación".

[20] Paradigma se utiliza en el sentido de "ejemplos" o "ejemplares". En el caso de "Paradigma Teórico" se utiliza el sentido que le da T. S. Kuhn, (1977: 33) y que asocia con "ciencia normal" "Su logro carecía suficientemente de precedentes como para haber podido atraer a un grupo duradero de partidarios, alejándolos de los aspectos de competencia de la actividad científica. Simultáneamente, eran lo bastante incompletas para dejar muchos problemas para ser resueltos por el redelimitado grupo de científicos".

von Bertalanffy; Kenneth E. Boulding, y Daniel Katz y Robert C. Kahn). Teoría Sistémica (Gregory Bateson; Kenneth E. Boulding; Daniel Katz y Robert C. Kahn, y María T. Bollini). La Sociología Clínica, considerada más como una Ciencia Aplicada de los conocimientos anteriores (Louis Wirth; Ernest W. Burgess; Alfred McClung Lee, y Howard M. Rebach y John G. Bruhn). Otras corrientes teóricas se pueden ver en J. Félix Tezanos (2006: 371).

Los Paradigmas Técnicos facilitan los métodos y las técnicas para recoger y tratar la información y se consideran: Paradigma Cualitativo y Paradigma Cuantitativo. Los Paradigmas Ontológico, Epistemológico, Metodológico van a definir el objeto, sus características, la forma de relación del sujeto con el objeto (y éste a su vez como sujeto) y del objeto (como sujeto) con la realidad, y las características del Método, se consideran: Positivismo, Postpositivismo, Críticos, Constructivismo y Neuro-Cuántico.

Los cuatro primeros paradigmas se pueden considerar filosóficos en tanto que su fundamento sobre el sujeto, objeto y la realidad no tienen un fundamento de base material y objetiva.[21] El paradigma Neuro-Cuántico se puede considerar científico en cuanto que considera una base material y objetiva de las características del sujeto, objeto y la realidad.

Los hechos/objetos[22] (Tabla 5) que conforman la realidad que estudia la Sociología, (de la Puente, 2007 a) tienen una característica que les diferencia de los hechos que constituyen la realidad de otras ciencias. El ser humano, además de ser un elemento que está dentro de las reglas de la Evolución, ha creado una Cultura dentro de la cual también tiene su propia evolución. En C. de la Puente (*op. cit*. 2007 a) se ha propuesto que si la Cultura es un producto de un ser que está dentro de la Evolución, entonces esa Cultura forma parte a su vez de la Evolución. Pero hay una diferencia temporal, mientras la Evolución Biológica tiene procesos de cambio largos (millones a centenas de millones de años) la Evolución Cultural tiene procesos de cambio cortos (de años a decenas de años y, probablemente en menores ocasiones, a centenas de años). Este planteamiento supone que, probablemente, la Sociología debe tener teorías nomotéticas y teorías idiográficas, división acuñada por Wilhelm Windelband. Un ejemplo de Teorías nomotéticas se puede considerar la Teoría de la Evolución de las especies (por adaptación de éstas y) por selección del medio (Darwin, 1859/1977). La "Pirámide de las Necesidades Básicas" de A. H. Maslow (1954/1963). Un ejemplo de Teorías idiográficas pueden ser la formación del Capitalismo según M. Weber o la "Lucha de Clases" tal como la consideró K. F. Marx.

[21] Aunque en realidad este punto puede ser muy discutido y discutible. Más que proponer una verdad cierta pretende ser motivo de debate.

[22] "(a) una cosa externa a la mente del pensador o sujeto. (b) una cosa o ser del cual una persona piensa o tiene cognición" (traducción propia) "object *noun*" *The Canadian Oxford Dictionary*. Katherine Barber. Oxford University Press 2004. *Oxford Reference Online*. Oxford University Press. Universidad Complutense de Madrid. 8 July 2008 http://www.oxfordreference.com/views/ENTRY.html?subview=Main&entry=t150.e48017. Entonces por Objeto consideramos que es una persona, institución o cosa capaz de ser percibida por los sentidos. Siendo institución en el sentido durkheimiano, entonces son los hechos sociales.

Tabla 5 Clasificación de los hechos/sucesos/objetos.					
Hechos, sucesos, objetos	Naturales	Materiales	Mundo físico		
		Inmateriales	Mundo no físico	Instintos Emociones Social natural (efecto manada)	
	Culturales	Materiales	Mundo físico	Construcciones del humano	H. S. Materiales (Durkheim)
		Inmateriales	Mundo no físico	Hechos Sociales coercitivos	H. S. No Materiales (Durkheim)
				Hechos Sociales punitivos	H. S. Materiales[23] (Durkheim)

La definición del objeto (ver nota 22) y sus características se abordan también desde las posiciones de Comte, Marx, Durkheim, Weber, Parsons y las neurociencias (Tabla 6).

[23] Interpretamos como "hechos sociales materiales" en Durkheim aquellos hechos que son "consistentes" como por ejemplo el "Código Civil", "Código Penal", "Textos Bíblicos", "El Derecho", etc. No obstante, se recomienda a los lectores trabajar este tema. Así que los códigos de honor, normas, costumbres, buenos modales, aunque están escritos se consideran "Blandos" y por lo tanto "Inmateriales". De forma genérica se considera "hecho social" en el sentido dado pro Durkheim a todo lo que no es natural. Desde el aspecto social que puede parecer más insignificante: forma del saludo, de comer, de vestir, etc. hasta el más sofisticado y complejo, la Ciencia.

Tabla 6 Objeto y método de la Sociología.			
Autor	Objeto		Método
Auguste Comte	El individuo La familia Las combinaciones sociales		Observación Experimentación Histórico Comparativo
Karl Marx	Actores y Estructuras.		Materialismo Dialéctico.[24]
Emile Durkheim	Hechos Sociales		Comparación de las variaciones concomitantes.[25] Principio: Considera los hechos sociales como cosas.
Max Weber	Acción Social		Construcción de Tipos Ideales.
Talcott Parsons	El Acto Social El Status-Rol El Actor La Colectividad		Se considera el Sistema AGIL como un método.
Neurosociología	Multinivel: Nivel material: Celular (neurona, astrocito, etc.) Órgano (partes del sistema nervioso). Nivel inmaterial: Acciones del Individuo. Acciones del Grupo.	Hechos Sociales	Científico: Heurístico,[26] Holístico, de segunda hermenéutica y doble epistemología.

1.3. Definición de Objetivos e Hipótesis

A partir de los puntos anteriores y según los intereses y criterios de la investigación, se especificarán los Objetivos diferenciados en General y Específicos, y también la o las Hipótesis (Bunge, 1981: 248-333; Miller, 1991: 33-37). En función de la Técnica de Investigación utilizada y los intereses perseguidos, una investigación puede plantear: Objetivos, o Hipótesis, o ambos.

1.4. Definición de Variables (Ítems)

En los Objetivos y las Hipótesis se plantean variables (o ítems) y relaciones entre ellas. Los Objetivos se traducirán o implementarán en aquellas variables que permitirán comprobar su consecución o cumplimiento. Las Hipótesis, en su definición de proposiciones afirmativas, especifican también variables y establecen relación entre ellas. Para ampliar la información sobre las variables ver epígrafe 6.1.

1.5. Definición de Indicadores

Los indicadores son similares a las variables pero de construcción más elaborada. La clasificación o medición que realizan son una síntesis,

[24] A todo fenómeno social (Tesis), se opone otro (Antítesis) y surge otro nuevo (Síntesis).
[25] Define varios métodos, pero el de *Comparación de las variaciones concomitantes*, es el que considera más adecuado.
[26] (M. Bunge, 1981: 224-229).

normalmente de más de una variable o ítem, con algún criterio. Ejemplos de Indicadores son: distancia del hogar al centro de trabajo; equipamiento del hogar; características del hogar; densidad de un núcleo de población; características socio-económicas; etc. La construcción y tratamiento de los indicadores se puede seguir por Miller (1991: 323-582) y Morales (1988).

2. Diseño Técnico

2.1. Definición del Universo

La realización de una investigación en sociología, normalmente, precisa la definición de un Universo mediante su delimitación geográfica y la población que contiene.

Al definir los límites geográficos o administrativos de la Población y las características de la misma, se define el objeto o unidad de observación y análisis. Se asume que la Población es el conjunto de estas unidades (ver nota 22).

2.2. Definición de la Muestra (Ver Epígrafe 14)

Al ser limitados los recursos económicos y materiales para acceder a toda la Población, se opera sobre un conjunto limitado de objetos que denominado Muestra, con la misma delimitación geográfica que el Universo.

Los resultados obtenidos de la Muestra se pretenden inferir sobre la Población, por lo que aquella debe ser representativa de ésta. Para que la Muestra sea considerada representativa, es necesario aplicar Técnicas de Muestreo (Tabla 7) y Técnicas de Cálculo de Tamaño de Muestra (Tabla 8), según los requisitos o criterios de la Ficha Técnica. Con este proceso se define a *quién* (Técnicas de Muestreo) y *cuántos* (Técnica de Cálculo del Tamaño de la Muestra) se les va a aplicar el Instrumento de Obtención de Datos (Cochran, 1987; Lohr, 1999; Rodríguez Osuna, 1991, 1993; Cea D'Ancona, 2004; Scheaffer et al, 2007).

Tabla 7 Técnicas de Muestreo.

Técnicas de Muestreo	Probabilísticas	Muestreo Aleatorio Simple Muestreo Aleatorio Sistemático Muestreo Aleatorio Estratificado Muestreo por Conglomerados
	No probabilísticas	Muestreo Intencional Muestreo Accidental Muestreo Bola de Nieve Muestreo por Cuotas

Tabla 8 Fórmulas para el cálculo del tamaño de muestras asumiendo Muestreo Aleatorio Simple.	
Para estimación de %	
Población Finita.	Población Infinita.
$n \mid \frac{Z^2 \Delta p \Delta q \Delta N}{e^2 \Delta (N 4 1) 2 Z^2 \Delta p \Delta q}$	$n \mid \frac{Z^2 \Delta p \Delta q}{e^2}$
Error muestral	Error muestral
$e \mid Z \Delta \sqrt{\frac{p \Delta q}{n}} \Delta \sqrt{1 4 fm}$	$e \mid Z \Delta \sqrt{\frac{p \Delta q}{n}}$
Para estimación de $\overline{X}$	
Población finita.	Población Infinita.
$n \mid \frac{Z^2 \Delta \omega^2 \Delta N}{e^2 \Delta N 2 Z^2 \Delta \omega^2}$	$n \mid \frac{Z^2 \Delta \omega^2}{e^2}$
Error muestral	Error muestral
$e \mid Z \Delta \sqrt{\frac{\omega^2}{n}} \Delta \sqrt{1 4 fm}$	$e \mid Z \Delta \sqrt{\frac{\omega^2}{n}}$
Corrector por poblaciones finitas (cpf)[27]	
$\sqrt{\frac{N 4 n}{N}} \mid \sqrt{\frac{N}{N} 4 \frac{n}{N}} \mid \sqrt{1 4 fm}$	

2.3. Técnicas de Investigación

Las Técnicas de Investigación (Tabla 9) es la forma en como se va a proceder para recoger la información o datos de las unidades de observación.

La característica de la información implica dos Paradigmas: Cuantitativo (Festinger et al, 1953; Ander-Egg, 1982; D'Ancona, 1996; Gambara, 1998; León et al, 1998; Babbie, 1999; Baker, 1999; Corbetta, 2003; Losada et al, 2003; Hernández Sampieri, 2007) y Cualitativo (Festinger et al, 1953; Ander-Egg, 1982; Ruiz O. et al, 1989; Delgado, 1994; Denzin et al, 1994/2005; Valles, 1997; Baker, 1999; Corbetta, 2003; Flick, 2004a; Flick et al., 2004b; Seale, 2004;), cada uno con sus Técnicas de Investigación diferenciadas.

Tabla 9 Técnicas de Investigación.		
Paradigma	Técnica	
Cuantitativo	Encuesta Experimento Estudio de Caso	
Cualitativo	Técnicas Individuales	Dinámicas. Biográficas. Entrevistas. Observación.
	Técnicas de Grupo.	Dinámicas. Biográficas. Entrevistas. Observación.

[27] Para ampliar la información sobre el *cpf,* confrontar con W. G. Cochram (1974: 47-49).

2.4. Instrumento de Obtención de Datos (Ver Epígrafe 5.3)

El Instrumento de Obtención de Datos es el soporte estandarizado con el que se va a registrar la información de las unidades de observación.

Básicamente, el soporte es papel o magnético (audio, video o informático), sobre el que se diseña un formulario, si procede.

En el formulario se desarrollan, en forma de preguntas, las variables (o ítems) e indicadores de los Objetivos e Hipótesis, se incluyen otras preguntas relevantes o complementarias del tema de investigación, más las consideradas de clasificación: socio-político-económico-demográficas. La redacción del formulario se hace según reglas establecidas (Alvira, 2004; Converse et al, 1986; Payne, 1980).

2.5. Codificación, Grabación, Tabulación y Análisis

Terminado el trabajo de campo, se procede a estructurar la información en formato de matriz de datos (Ver epígrafe 6.2), para proceder a su tabulación y análisis (resto del texto).

El proceso, que no significa contigüidad inmediata, es: Codificación, Grabación, Tabulación y Análisis. En la Tabla 10 se muestra una clasificación de técnicas de análisis (Ver Bibliografía).

Tabla 10 Técnicas de Análisis.

Paradigma	Estadística	Grupo de Técnica	Técnica de Análisis
Cuantitativo	Descriptiva Univariable	Tendencia Central.	Moda Mediana Media Mínimo Máximo Sumatorio Percentiles Tabla de Frecuencias
		Dispersión.	Amplitud Varianza Desviación Típica Coeficiente de Variación.
		Forma.	Asimetría. Apuntamiento.
		Gráficos	Barras Histograma
	Dist. Prob.		Z t θ^2 F
	Descriptiva Bivariable	Tablas de Contingencia	Frecuencia Absoluta. Frecuencia Relativa Asociación Fuerza y Dirección Asociación de Celdas
		Tablas de Medias.	Tendencia Central, Dispersión y Forma.
		Asociación Lineal.	Gráfico de Dispersión. Covarianza. Correlación.
	Paramétrica	Una Muestra.	t-test
		Dos muestras Independientes.	t-test
		K muestras Independientes.	(M)ANOVA
		Dos muestras emparejadas.	t-test
		K Muestras Emparejadas.	Medidas Repetidas.
Cuantitativo	No Paramétrica	Una Muestra.	Kolmogorov-Smirnov; θ^2, Binomial, y Rachas.
		Dos muestras Independientes.	Mann-Whitney.
		K muestras Independientes.	Kruskal-Wallis.
		Dos muestras emparejadas.	Wilcoxon.
		K Muestras Emparejadas.	Friedman.
	Técnicas Multivariable	Modelos explicativos/ predictivos	Lineal simple Lineal múltiple Parciales lineal Simple no-lineal Múltiple no-lineal Parciales no-lineal Discriminante Logarítmico lineal jerárquico Logarítmico lineal no jerárquico Modelos binomiales Modelos polinomiales Series temporales (TRENDS) LISREL, AMOS. Red Bayesiana Perceptrón multicapa Función de base radial

Tabla 10 Técnicas de Análisis.

Paradigma	Estadística	Grupo de Técnica	Técnica de Análisis
Cuantitativo	Técnicas Multivariable		Red Kohonen
		Modelos de reducción de casos (grupos desconocidos)[1]	Conglomerados jerárquico Conglomerados no jerárquico
		Modelos de reducción de casos (grupos conocidos)	Discriminante
		Modelos de reducción de variables[11]	Componentes principales (ACP) Análisis Factorial. Análisis de Fiabilidad[***]
		Modelos de reducción de casos y variables	ANACOR. HOMALS. MDS PRINCALS OVERALS.
		Modelos de segmentación	CHAID CHAID Exhaustivo C&RT. QUEST
		Modelos de preferencia de mercado	CONJOINT. CONJOINT basado en elecciones (CBCA) BPTO (Brand Price Trace-Off)
		Otros entornos	CLEMENTINE CHURN QI Analist Power Sample Data warehouse
Cualitativo			Análisis de Discurso Análisis de Contenido

[1]: Usado también para variables.
[11]: Usado también para casos
[***]: El análisis de Fiabilidad no es una técnica específica de reducción de datos.

5.1 Paradigmas según los aspectos ontológicos, epistemológicos y metodológicos.

El conocimiento de la realidad social y los elementos que la componen se hace considerando los aspectos Ontológico, Epistemológico y Metodológico,[28] además de los Paradigmas Teóricos y Técnicos.

De las diferentes acepciones que puede tener el término ontología, se considera el que hace referencia a "¿Cuál es la forma y naturaleza de la realidad? y, por consiguiente, ¿Qué es lo que puede conocerse de ella? ... si se asume que existe un mundo real, entonces lo que puede ser conocido de él es ¿Cómo son las cosas realmente? y ¿Cómo operan las cosas realmente?" (Denzin et al, 1994: 108) (Ver Epígrafe 4).[29]

La cuestión epistemológica hace referencia a "¿Cuál es la naturaleza de la relación entre el conocedor o que puede conocer y lo conocido?" la respuesta a esta cuestión puede estar delimitada por la respuesta dada al nivel ontológico. Pero no vale cualquier relación, debe ser de separación objetiva y libre de valores por parte del conocedor, para poder descubrir las propuestas de la cuestión ontológica (*Ibíd*.).[30]

La cuestión metodológica. ¿Cómo puede el investigador (conocedor) hacer o proceder con lo que quiere conocer? (*Ibíd*.) La respuesta que se puede dar a esta cuestión está constreñida por la que se ha dado a las dos anteriores; esto es, no vale cualquier metodología, y además no se debe confundir las técnicas con la metodología (el método). Las técnicas deben estar ajustadas a una metodología (método) predeterminada[31].

Los Paradigmas considerados para ver los niveles ontológico, epistemológico y metodológico son: Positivismo, Postpositivismo, Teoría Crítica, Constructivismo y Neuro-Cuántico.

POSITIVISMO:

Ontología: Se considera un realismo ingenuo, se asume que existe una realidad aprehensible que es dirigida por leyes y mecanismos inmutables. El conocimiento de "cómo son las cosas" se considera que son generalizaciones independientes del contexto y el tiempo en el que ocurren. La postura básica del paradigma se asume reduccionista y determinista (*Ibíd*.).

Epistemología: Dualista y objetivista. El sujeto y el objeto se consideran entidades independientes y que el sujeto es capaz de no influir ni ser influido por el objeto (*Ibíd*.).

Metodología: Experimental y manipulativa. Las preguntas y las hipótesis se realizan en forma de proposición y sujetas a observación y test empíricos para verificar o refutar (*Ibíd*.).

POSTPOSITIVISMO:

Ontología: Considerado de realismo crítico. Se asume que existe una realidad, pero es aprehensible de forma imperfecta debido a los mecanismos intelectuales humanos que son imperfectos y la propia naturaleza compleja de las cosas. La ontología se etiqueta de realismo crítico porque la postura de los investigadores sobre la realidad debe estar sujeta a exámenes

---

[28] Para diferenciar *método* de *metodología*, ver las características de *método* en el Epígrafe 2. Según el el Diccionario de la RAE, *metodología* literalmente consiste en la *ciencia del método*. *–logía* "tratado", "estudio", "ciencia" del método. Cómo o qué vamos a considerar del método para la investigación.

[29] Para ver el tratamiento a nivel ontológico de *Homo sapiens sapiens* ver C. de la Puente (2007b).

[30] Considerando lo que puede ser la paradoja de la subjetividad, esto es, que "la subjetividad es objetivamente subjetiva".

[31] Pero manteniendo la disciplina científico-académica, no se debe olvidar que "lo importante" no son "los medios", sino "los hechos-sucesos" que se quiere conocer. Se podría proponer el "vale todo" de P. K. Feyerabend, que a veces ha sido mal interpretado.

críticos para facilitar su aprehensión de la forma más acertada (*Ibíd*.).

Epistemología: Se considera objetivista y dualismo modificado. El dualismo va siendo abandonado, pero el objetivismo permanece como un ideal que debe ser alcanzado por medio de su regulación, a través de la tradición crítica ¿Los resultados se ajustan al conocimiento preexistente? Y la comunidad crítica (editores, revisores y pares profesionales). Se introduce el concepto de "probables" en los resultados obtenidos (*Ibíd*.). Los resultados obtenidos en el proceso de investigación no se consideran verdaderos o falsos, sino que al introducir el concepto de probabilidad, los resultados tienen cierta probabilidad de ser ciertos, aunque si se cambia el criterio de probabilidad pueden pasar a ser no aceptados. El término preferido puede ser el de resultados "confirmables" (Ver Epígrafes correspondientes a contraste de Hipótesis).

Metodología: Experimental modificada y manipulativa. Se considera una versión restaurada de la triangulación, como vía para la falsación en vez de la verificación de hipótesis. La triangulación se consigue mediante la aplicación de diferentes técnicas de investigación y/o diferentes investigadores, introduciendo los resultados obtenidos en el proceso de investigación, de forma recursiva, para volver a reelaborarlos a partir de la realidad observada (*Ibíd*.).

TEORÍA CRÍTICA: (neo-marxismo, feminismo y materialismo)

Ontología: Realismo histórico. Se asume que la realidad (Social) es plástica y aprehensible, y que fue formada por un conglomerado de factores sociales, políticos, culturales, económicos, étnicos y de género y han cristalizado en una serie de estructuras que ahora se toman (de forma incorrecta) como "reales," esto es, natural e inmutables. Se considera que las estructuras son "reales," una realidad histórica virtual (*Ibíd*.).

Epistemología: Transaccional y subjetivista. Se asume que el sujeto investigador y el objeto investigado interactúan, tienen transacciones y además que los valores del investigador influyen en la investigación. En este paradigma se difumina la distinción entre ontología y epistemología (*Ibíd*.) al interactuar el sujeto y el objeto se influyen y construyen unos a otros y no son lo que "son" sino lo que "alter" dice que son.

Metodología: Dialéctica y dialógica.[32] La naturaleza transaccional de la investigación requiere un diálogo entre sujeto y objeto. La dialéctica debe transformar los defectos de observación y la ignorancia, en una representación más informada (*Ibíd*.).

CONSTRUCTIVISMO:

Ontología: Relativista. Las realidades son aprehensibles de forma múltiple, basado en el aprendizaje según la experiencia social, de naturaleza local y específica aunque pueden estar compartidos entre individuos y entre culturas, y dependiendo su forma y contenido de las personas o grupos que realizan la construcción. Las construcciones no son más o menos verdaderas en un sentido absoluto, sino que tienen un mayor o menor nivel de información y sofisticación. Tanto las construcciones como la realidad asociada se pueden alterar (*Ibíd*.).

Epistemología: Transaccional y subjetivista. Se asume que el sujeto investigador y el objeto investigado interactúan, así que los resultados son "literalmente" creados por el proceso de la investigación. La distinción entre ontología y epistemología desaparece como en el paradigma de la Teoría Crítica (*Ibíd*.).

Metodología: hermenéutica y dialéctica. La naturaleza personal y variable de las

[32] Concepto acuñado por el filósofo ruso Mikhail Bakhtin, que significaría diálogo continuo con otras técnicas y otros autores.

construcciones sociales (construcciones mentales) lo son a través de la relación entre el investigador y el investigado (Gráfico 1). Estas construcciones se hacen de forma hermenéutica y son comparadas a través del intercambio dialéctico. El resultado final es una construcción que tiene más información y es más sofisticada que cualquier construcción anterior (*Ibíd*.).

PROPUESTA DE PARADIGMA NEURO-CUÁNTICO:

Ontología: Se asume objetivamente una "realidad-fuera" (RF) y una "realidad-dentro" (RD). Existe una RF que sucede/ocurre de forma simple, pero de causación compleja y siempre dependiendo del estado inmediato anterior. Esta realidad proyecta la luz (fotones) que recibe, de forma discreta y en infinitas direcciones, por lo que hay un único hecho pero proyectado infinitas veces, y puede emitir ondas sonoras, entre otros estímulos. Estos estímulos externos son recibidos por el sujeto/objeto para constituir la RD (Guyton, 1994: 163-217; Bear et al. 1998. 210-308; Kandel et al. 2001: 492-624). Se asume el paradigma de la Teoría Crítica y el Constructivismo.

Epistemología: Constructivista. La RF no es "vista" por el sujeto/objeto, es percibida. Lo que "ven" es una representación virtual que realiza el cerebro a través de la energía (estímulos físicos y químicos) que recibe del exterior y que es filtrada por el cerebro mismo, ya que no muestra o procesa todo lo que recibe[33] sino que muestra o procesa exclusivamente aquello que debe ser necesario para el mantenimiento y supervivencia de la especie[34] y que en este escrito es lo que se considera la RD, esta RD es una representación psicológica (E. Husserl) con la información de la que se dispone (A. Schutz) que en el caso del sujeto cognoscente re-produce, además, la representación que se hace el objeto conocido de la realidad que percibe, como una segunda representación que A. Schutz denomina "segunda hermenéutica". Tampoco existe distinción entre el nivel ontológico y el epistemológico.

La RD se forma en base a los estímulos recibidos que son fotones, ondas sonoras y moléculas que se traducen en olores y sabores. Entonces asumimos que todos los significados y sentidos atribuidos a los hechos se crean en o los da el cerebro, entendiendo que los fotones, ondas sonoras y moléculas no contienen más información que lo que son. No transportan significados, ni sentido del hecho, ni color, ni olor, ni ruido, ni sabor. Se asume el paradigma de la Teoría Crítica y el Constructivismo.

Metodología: Multiparadigmática. La relación RF-RD se trata desde múltiples Áreas de Conocimiento: Matemáticas, Física, Química, Biología, Biología Molecular, Sociología, Psicología, Neurología, Fisiología, Antropología, Paleontología, Paleoantropología, Historia, Derecho, Economía, Filosofía, Filología, etc.

Entonces, el Positivismo, Postpositivismo, Teoría Crítica y Constructivismo, se pueden considerar paradigmas filosóficos, en tanto no hacen referencia a una base material y

---

[33] Lo que perciben el sujeto y el objeto, es lo que se considera la RF y que es el espectro electromagnético que está en el rango de una longitud de onda inferior a los 100 nm (nanometros) y superior al PHz (Peta Hz), longitud de onda corta-alta frecuencia (Rayos gamma y rayos cósmicos) y entre una longitud de onda superior al Mm y una frecuencia inferior al kHz (kilo Hz), longitud de onda larga-baja frecuencia (radio de muy baja frecuencia) y muy por debajo estarían las ondas electromagnéticas que se asume que emite el cerebro. La RD que muestra el cerebro es la que está en la longitud de onda de 380 nm (nanometros) a 780 nm y en un rango de frecuencia de 384 THz (Tera Hz) a 789 THz ("electromagnetic waves" A Dictionary of Space Exploration. Ed. E. Julius Dasch. Oxford University Press 2005. Oxford Reference Online. Oxford University Press. Universidad Complutense de Madrid. 6 December 2008 <http://www.oxfordreference.com/views/ENTRY.html?subview=Main&entry=t212.e577>). En el caso del sonido, las ondas que "muestra" el cerebro humano son las que están en el rango de 20 Hz a 20.000 Hz ("sound" A Dictionary of Physics. Ed. John Daintith. Oxford University Press, 2000. Oxford Reference Online. Oxford University Press. Universidad Complutense de Madrid. 6 December 2008 <http://0-www.oxfordreference.com.cisne.sim.ucm.es:80/views/ENTRY.html?subview=Main &entry=t83.e2840>).

[34] Expresado de forma proporcional y aproximada, respecto del total del espectro de longitud de onda, el cerebro humano muestra un 0,004 por mil millones y de la frecuencia es el 4,05 por mil millones.

objetiva, mientras que la propuesta de paradigma Neuro-cuántico se puede considerar científico en cuanto que hace referencia a una base material y objetiva.

5.2 Objeto, Objetivo, Técnica

Las palabras *objeto* y *objetivo* puede presentar confusión en su significado. Desde la Sociología se diferencia entre *objeto* y *objetivo* como:[35]

Objeto

- "Todo lo que puede ser materia de conocimiento o sensibilidad de parte del sujeto, incluso este mismo". "Aquello que sirve de materia o asunto al ejercicio de las facultades mentales" (Real Academia Española, *op. cit.*).
- "Cualquier cosa que se percibe por los sentidos". "Lo que sirve de materia al ejercicio del entendimiento". "Cosa" (Micronet S. A., abril de 1999).
- "Alguna cosa material que puede ser percibido por los sentidos". "Algo mental o físico hacia lo que el pensamiento, el sentimiento, o la acción se dirige" (Encyclopedia Britannica, Inc., 1994).

Es la unidad de observación, unidad de análisis o caso en el estudio sociológico. Puede ser persona, institución[36] o cosa capaz de ser percibida por los sentidos. En Sociología el objeto puede ser material o inmaterial como es el caso de los hechos sociales.

Objeto como *Objetivo*:

- "Fin o intento a que se dirige o encamina una acción u operación" (Real Academia Española, *op. cit.*).
- "Apuntar a, Dirigirse a, Perseguir; Tener por. Fin. Finalidad. Intención. Mira. Objetivo. Propósito. Cosa que se pretende al hacer algo". "Fin, finalidad, propósito con el que se realiza cualquier cosa o meta hacia la que tiende una acción" (Moliner, 1996).
- "Fin, intento, propósito" (Micronet S. A., *op. cit.*).
- "La meta o fin de un esfuerzo o actividad" (Encyclopedia Britannica, Inc., *op. cit.*).

Objeto como materia o asunto de una ciencia

- "Materia o asunto de que se ocupa una ciencia o estudio" (Real Academia Española, *op. cit.*).
- "Materia y asunto de que se ocupa una ciencia" (Micronet S. A., *op. cit.*).
- "Elemento o materia de una investigación o ciencia" (Encyclopedia Britannica, Inc., *op. cit.*).

Se debe considerar que la palabra *method* se puede traducir por *método* y *técnica*, pero en español tienen significados diferentes.

Método:

- "Orden que se sigue en las ciencias para investigar o enseñar la verdad" (Micronet S. A., *op. cit.*). Ver epígrafe 2.

Técnica:

---

[35] Para ampliar la definición de estos conceptos se puede ver Uña Juárez y Hernández Sánchez (2004).

[36] "Institución" en el sentido durkheimiano que se consideran "hechos sociales".

- "Conjunto de procedimientos y recursos de que se sirve una ciencia o un arte" (Real Academia Española, *op. cit.*).
- "Conjunto de procedimientos de un arte o ciencia" (Micronet S. A., *op. cit.*).
- "Un procedimiento sistemático, técnica o modo de investigación empleado como propio de una disciplina o arte" (Método como proceso y técnica). "Una vía, técnica o proceso de o para hacer algo" (Método como técnica y proceso) (Encyclopedia Britannica, Inc., *op. cit.*).

Ver Gráfico 1, Tabla 5 y Tabla 6

5.3 Comentarios al diseño del cuestionario[37]

El diseño de un cuestionario es una tarea que se puede considerar compleja y llena de dificultades. Se trata de la construcción de un instrumento de medida o, de forma más precisa en ciencias sociales, de clasificación de las unidades de observación, puesto que no existen instrumentos estandarizados de medida y observación de uso universal como en otras ciencias como pueden ser cintas métricas, balanzas, microscopios, telescopios, túneles aceleradores de partículas, instrumentos electroencefalográficos, etc.

García Ferrando (2005: 180) nos dice que "la función del cuestionario en el proceso de una investigación social, es doble. Por un lado, pretende situar a todos los entrevistados en la misma situación psicológica y por otro, mediante un sistema de notaciones simples, facilitar el examen y la comparabilidad de las respuestas".

La aplicación del cuestionario orientado principalmente a la modalidad mediante entrevista personal supone:

- La relación entre dos personas.
- Que no se conocen.
- Que una de ellas extrae información de la otra y además consume su tiempo.
- Que la persona entrevistada no obtiene ninguna contrapartida, excepto la de participar.
- Que entre las personas existe lo que llamamos "la primera impresión" que puede condicionar la entrevista.

Después de estas consideraciones, para que la entrevista mediante la aplicación del cuestionario transcurra de forma que se facilite la participación del entrevistado, es recomendable considerar los siguientes aspectos en su organización,

- Que tenga una introducción adecuada.
- Que tenga una transición fácil de un tema a otro.
- Formulación de un final adecuado.

Una estructura recomendable de cuestionario puede ser dividirlo en tres partes. La primera con temas introductorios, la segunda con el núcleo central del tema o temas de investigación y la tercera y última con las preguntas consideradas de clasificación que son las de tipo socio-político-económico-demográficas.

El cuestionario debe seguir un "hilo conductor" desde el principio hasta el final, de manera que la introducción sea gradual y sin preguntas comprometidas para facilitar la participación de la persona entrevistada, captando su interés.

---

[37] Este Epígrafe se incluye por el interés expositivo del autor.

La transición a la segunda parte también debe ser gradual. Este segundo bloque se considera que es el más importante o el que contiene la información que es el objetivo principal de la investigación (aquello que queremos saber o conocer). Las preguntas pueden ser desde temas aparentemente superficiales hasta de una gran trascendencia. Esta parte del cuestionario puede suponer que a la persona entrevistada se le someta a un gran esfuerzo de tipo emocional y/o memorístico. Además se pueden movilizar ciertos contenidos de la persona y puede llegar incluso a producir catarsis.

Preguntas aparentemente simples pueden resultar de una gran trascendencia para el entrevistado. Por ejemplo, un estudio de mercado de comidas de animales domésticos puede ocasionar que coincida con alguien a quien se le ha muerto su mascota y le puede movilizar recuerdos; un estudio de satisfacción con el automóvil nos puede llevar a alguien que hace poco ha perdido a un familiar en un accidente de tráfico; un estudio sobre la opinión del consumo de drogas, a alguien que tiene un problema familiar de este tipo; un estudio sobre la educación de los hijos ocasionaría un fuerte impacto a una persona que estuviese en un proceso de custodia de los hijos.

Cuando entrevistamos a alguien, entramos en un universo desconocido del que no sabemos en que situación se encuentra y al terminar la entrevista, al menos, debe quedar en las mismas circunstancias emocionales y anímicas en las que empezó, o por lo menos no debe quedar peor.

La terminación del segundo bloque debe ser de forma gradual para hacer la transición al tercer y último bloque en el que están las preguntas de clasificación. Este bloque debe ir al final. Si estas preguntas se realizasen al principio, la persona entrevistada podría sentirse identificada y mostrar rechazo a participar.

Todas las preguntas deben tener sentido en el marco de la investigación que se realiza. Preguntas que no aporten cierta utilidad van a romper el "hilo conductor" de la entrevista y aunque los entrevistados no son personas técnicas en la materia, la lógica y el sentido común les indica perfectamente que preguntas son útiles y cuales no y en estos casos pierden el interés en participar.

Un cuestionario puede tener más de tres bloque. La recomendación es proceder de la misma manera cuando sean más de tres bloques. La información final del entrevistado son sus datos: nombre y apellidos, dirección y teléfono. Esta información no se va a utilizar ni se va a grabar en la matriz de datos por ser información protegida por la Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal. Esta información se utiliza posteriormente en los procesos de supervisión y verificación del trabajo de campo. Para terminar, se especifica la información propia del cuestionario y del entrevistador,

Entrevistador	
Nombre:	
Apellidos:	
Municipio de la entrevista:	
Distrito:	
Barrio:	
Sección Censal:	
Comienzo de la entrevista (hora:minutos):	____:____
Fin de la entrevista (hora:minutos):	____:____
Fecha (día, mes, año):	__ __/__ __/200__

Firma del entrevistador

Cuestionario	
Revisado	
Codificado	
Supervisión telefónica	
Supervisión en campo	
Grabado	
Nulo	

La formulación de las preguntas del cuestionario puede ser una tarea propia o bien utilizar cuestionarios ya testados y utilizados en otras investigaciones. Para el proceso de la creación del cuestionario, su redacción, enmaquetación, pretest, etc., se recomiendan algunos manuales específicos que pueden consultar los lectores, ya que es una tarea que no está dentro del ámbito de este manual (García Ferrando, 2005: 167-201; Manzano, 1996; Payne, 1979; Converse y Preser, 1986; Bingham y Moore, 1973; Mayntz, 1976; Stoetzel y Girard, 1973; Scheuch, 1973).

No obstante, la realización de un estudio aplicando las diferentes Técnicas de Investigación Social y los Instrumentos de Obtención de Datos correspondientes sobre una población humana, no es una tarea simple, ni sencilla ni trivial, independientemente del tema o asunto tratado.

Una Investigación de Mercado de productos para mascotas tiene, según el Paradigma Sistémico, trascendencia e implicaciones en muchos niveles. Cualquier producto, por ejemplo, como los de mascotas, tiene una fase de producción, tratamiento y preparación, distribución, venta, compra y consumo. En todas estas etapas intervienen humanos y son interés de muchas Ciencias y entre ellas está la Sociología.

En la producción intervienen humanos y obtienen beneficios y sueldos con los que proporcionan, alimento, vestido, educación y ocio a su familia y con ello les dan ciertas dosis de felicidad. Básicamente todos ellos son compradores/consumidores.

En el tratamiento y preparación intervienen humanos y obtienen beneficios y sueldos con los que proporcionan, alimento, vestido, educación y ocio a su familia y con ello les dan ciertas dosis de felicidad. Básicamente todos ellos son compradores/consumidores.

En la distribución intervienen humanos y obtienen beneficios y sueldos con los que proporcionan, alimento, vestido, educación y ocio a su familia y con ello les dan ciertas dosis de felicidad. Básicamente todos ellos son compradores/consumidores.

En la venta intervienen humanos y obtienen beneficios y sueldos con los que proporcionan, alimento, vestido, educación y ocio a su familia y con ello les dan ciertas dosis de felicidad. Básicamente todos ellos son compradores/consumidores.

La compra es realizada por humanos para mantener en cierto estado de bienestar a sus mascotas y éstas forman parte de las relaciones e interacciones de la familia, proporcionando situaciones de felicidad y a veces cierto sufrimiento. Se convierten en un elemento más en el entorno del hogar y por lo tanto tienen su función dentro de la estructura familiar.

Un estudio sociológico, de mercado, etc. debe ayuda a mantener, mejorar y/o ampliar esta red.

6 Introducción a la Estadística

6.1 Estadística, preguntas y variables

Entendemos por Estadística "la disciplina científica que trata de la recolección, análisis, y presentación de datos" (ver nota 3).[38] Por el interés de la obra, la Estadística se divide en Estadística Descriptiva (Tabulación) y Estadística Inferencial (Análisis o contraste de hipótesis). Otro grupo sería la Estadística Multivariable, que no es objeto de este tratado.

Los datos se consideran de tres tipos: Tipo I, Tipo II y Tipo III. Los datos de Tipo I son los datos brutos, "raw data" o microdatos. Se dispone de los datos o valores que se tiene para todos y cada uno de los casos. En los datos Tipo II, se muestra la frecuencia, el número de casos que hay en cada categoría o valor distinto o el número de veces que se repite o aparece cada valor o categoría distinta (tabla de frecuencias). En los datos Tipo III, también se muestra la frecuencia o el número de casos, pero por intervalos (Tabla de frecuencias por intervalos). Los ejemplos se muestran en la Tabla 11.

Tabla 11 Tipos de estructura de los datos

Tipo I

Caso	Edad
1	18
2	19
3	20
4	18
5	19
6	20
7	21
8	22
9	23
10	21
11	22
12	23
13	21
14	22
15	23
16	24
17	25
18	26
19	24
20	25
21	26
22	24
23	25
24	26
25	27
26	28
27	29
28	27
29	28
30	30

Tipo II

Edad	Frecuencia
18	2
19	2
20	2
21	3
22	3
23	3
24	3
25	3
26	3
27	2
28	2
29	1
30	1

Tipo III

Edad_R	Frecuencia
18-21	9
21-24	9
24-27	8
27-30	4

Paso de Tipo III a Tipo II

Marca de Clase	Frecuencia
19,5	9
22,5	9
25,5	8
28,5	4

$$X_i' \mid \frac{L_{i41}\ 2\ L_{i21}}{2}$$

$$X_1' \mid \frac{18\ 2\ 21}{2} \mid 19{,}5$$

$$X_2' \mid \frac{21\ 2\ 24}{2} \mid 22{,}5$$

$$4$$

$$X_4' \mid \frac{27\ 2\ 30}{2} \mid 28{,}5$$

En donde:

L_{i-1}: Límite inferior del intervalo

L_{i+1}: Límite superior del intervalo

X_i': Marca de clase del intervalo *i-ésimo.*

La aplicación de los estadísticos se hace sobre los datos de Tipo I y Tipo II. Con los datos Tipo III se procede pasándolos a datos Tipo II, representando cada intervalo, estrato o categoría por el *valor medio* o *marca de clase* del intervalo. En este caso, a la variable se la

[38] "statistics" *A Dictionary of Genetics*. Robert C. King, William D. Stansfield and Pamela K. Mulligan. Oxford University Press, 2007. *Oxford Reference Online*. Oxford University Press. Universidad Complutense de Madrid. 16 July 2008 http://www.oxfordreference.com/views/ENTRY.html?subview=Main&entry=t224.e6186.

denomina como *prima* (*X'*), (ver Tabla 11).

Por el interés de este epígrafe se define *pregunta*, *variable*, *espacio muestral*, *suceso elemental*, *respuesta* y *categoría*.

Pregunta:

- RAE: "Interrogación que se hace para que alguien responda lo que sabe de un negocio u otra cosa".
- BDCD; "Un acto o instancia de pedir información en una investigación sistemática, a veces de interés público".

Ejemplos:

P-1	¿Podría indicar cual es su género en cuanto a sexo?	
	Varón	01
	Mujer	02

P-2	¿Podría indicar cual es su estado civil?	
	Soltero/a	01
	Casado/a	02
	Pareja	03
	Separado/a	04
	Divorciado/a	05
	Viudo/a	06

P-3	Puede indicar su,						
	Peso		Estatura		Edad		
		Kg.		m.		años	

Respuesta:

- RAE: "Satisfacción a una pregunta, duda o dificultad".
- BDCD: "Algo dicho o escrito en respuesta a una pregunta"

Ejemplos:

En las tres preguntas anteriores, la respuesta es marcar en la casilla correspondiente la respuesta dada a cada una. En la P-1, indicar cual es el sexo; en la P-2 el estado civil, y en cada una de las preguntas de la P-3, indicar el peso, la estatura y la edad, por este orden.

Categoría:

- RAE: "cualidad atribuida a un objeto".
- DMM: "Cada grupo de cosas o personas de la misma especie de los que resultan al ser clasificadas por su importancia, grado o jerarquía".
- BDCD: "Una de las clases distintas y fundamentales a la que pertenece una entidad o concepto". "Una división dentro de un sistema de clasificación".

Ejemplos:

De las tres preguntas anteriores, las categorías de la P-1 son: *varón* y *mujer*. En la P-2, las categorías son: *soltero/a, casado/a, pareja, separado/a, divorciado/a* y *viudo/a*. En la P-3,

al ser variables numéricas, las respuestas no se consideran categorías, sino valores. No obstante, se podría considerar categoría cada uno de los valores distintos que pueden contestar, ya que, por ejemplo, sería la categoría de las personas con "18 años".

Variable:

- OROP:[39] "En las ciencias sociales, el término se refiere a atributos que son fijos para cada persona u otra entidad social, el cual es observado a los diferentes niveles o cantidades de las muestras y otros grupos de agregados. Las variables miden una estructura social (como la clase social, edad, o tipo de albergue) y en cierto modo permite el análisis numérico. Así que el rasgo importante de una variable es que es capaz de reflejar la variación dentro de una población, y no es una constante".[40]
- RAE: "Que varía o puede variar".
- BDCD: "Capaz o apto para variar: sujeto a variación o cambio".

Ejemplos:

En el ejemplo considerado, las variables se corresponden con las preguntas y así, las variables serían: *sexo*, *estado civil*, *peso*, *estatura* y *edad*.

Espacio muestral:

- OROP: "Un conjunto completo de todos los posibles resultados de un experimento o procedimiento de observación. El concepto fue introducido por von Mises en 1931. El espacio muestral normalmente se representa por T , *S* o *E*.[41]
- DSTTMH:[42] "Un concepto o término en teoría de probabilidades que considera todos los posibles resultados de un experimento, juego o similar, como puntos en un espacio".

Ejemplos:

En la pregunta o variable sexo, el espacio muestral es: *varón* y *mujer*. En estado civil el espacio muestral está definido por: *soltero/a, casado/a, pareja, separado/a, divorciado/a* y *viudo/a*. Y en peso, estatura y edad, los espacios muestrales están definidos por todos los posibles valores de cada una de las preguntas o variables y que son finitos y conocidos. En el caso del peso y la estatura son los valores posibles de la población objetivo y la edad es la definida por los criterios de delimitación de la población.

Suceso elemental:

- OROP: "Un suceso elemental es uno de los resultados posibles del espacio

---

[39] Oxford Reference Online Premium.

[40] "variable" *A Dictionary of Sociology*. John Scott and Gordon Marshall. Oxford University Press 2005. *Oxford Reference Online*. Oxford University Press. Universidad Complutense de Madrid. 8 December 2008 http://www.oxfordreference.com/views/ENTRY.html?subview=Main&entry=t88.e2451.

[41] "sample space" *A Dictionary of Statistics*. Graham Upton and Ian Cook. Oxford University Press, 2008.*Oxford Reference Online*. Oxford University Press. Universidad Complutense de Madrid. 8 December 2008 http://www.oxfordreference.com/views/ENTRY.html?subview=Main&entry=t106.e1433.

[42] Diccionario de Términos Científicos y Técnicos. McGraw-Hill.

muestral".[43]

∉ DSTTMH: "Cada uno de los posibles resultados de un experimento aleatorio, es decir cada uno de los elementos del espacio muestral".

Ejemplos:

En la pregunta o variable sexo, los sucesos elementales del espacio muestral son: *varón* y *mujer*. En estado civil los sucesos elementales son: *soltero/a, casado/a, pareja, separado/a, divorciado/a* y *viudo/a*. Y en peso, estatura y edad, los sucesos elementales son todos los posibles valores de cada una de las preguntas o variables y que son finitos y conocidos. En el caso del peso y la estatura son los valores posibles de la población objetivo y de la edad los sucesos elementales están definidos por los criterios de delimitación de la población.

NIVEL DE MEDIDA DE LAS VARIABLES

Los niveles de medida se distinguen por propiedades de distancia y orden. Un ordenador no sabe las características de los valores que se le dan, por tanto, se deben determinar por el investigador los niveles de medida de los datos para poder aplicar las técnicas estadísticas apropiadas cuando se opera con programas estadísticos.

Las variables se clasifican en dos grupos: variables cualitativas, categóricas o de frecuencias y variables cuantitativas o numéricas. En el primer grupo se incluyen las variables de nivel de medida *nominal* y *ordinal*, y en el segundo las de *intervalo* o *escala* y *razón*.

Nivel de medida Nominal

Las variables de nivel de medida *nominal*, son aquellas que sus datos son valores numéricos o códigos que se asignan a las categorías de la variable, entre los que no existe ninguna relación y cada valor define una categoría distinta, es el nivel considerado inferior. La asignación de valores o códigos a las categorías se llama codificación (ver el apartado de codificación). Con estos valores no se pueden realizar operaciones aritméticas, pero sí se pueden aplicar operadores lógicos y operaciones de clasificación.

Son ejemplos de variables *nominales*: sexo, estado civil, carácter, religión, deportes practicados, productos comprados.

Un tipo especial de variables nominales son las dicotómicas, variables con dos categorías, pero también se pueden considerar variables dicotómicas a las binarias o falsas binarias. En la Tabla 12 se presenta la diferencia entre dicotómica, binaria y pseudobinaria.

Tabla 12 Variables dicotómicas, binarias y falsas binarias.

Dicotómica		Binaria		Falsa binaria	
	Código		Código		Código
Categoría 1	1	Verdadero	1	Categoría 1	1
Categoría 2	2	Falso	0	No categoría 1	0
Aplicación:					
Sexo	Código	Asistir a clase	Código	Sexo	Código
Varón	1	Verdadero	1	Varón	1
Mujer	2	Falso	0	No varón	0
		No asistir a clase	Código	Sexo	Código
		Verdadero	1	Mujer	1
		Falso	0	No mujer	0

[43] "sample space" *A Dictionary of Statistics*. Graham Upton and Ian Cook. Oxford University Press, 2008.*Oxford Reference Online*. Oxford University Press. Universidad Complutense de Madrid. 8 December 2008 http://www.oxfordreference.com/views/ENTRY.html?subview=Main&entry=t106.e1433.

Las variables dicotómicas pueden ser consideradas numéricas e independientes en el Análisis de Regresión. Las binarias y falsas binarias también se pueden considerar numéricas porque se pueden calcular funciones estadísticas.

Nivel de medida Ordinal

Las variables de nivel de medida *ordinal*, son aquellas que sus datos son valores numéricos o códigos que se asignan a las categorías de la variable, cada valor define una categoría distinta, lo que le asigna la característica de las variables *nominales*. Entre sus valores se puede establecer un criterio de orden. La asignación de valores o códigos a las categorías se llama codificación (ver el apartado de codificación). Con estos valores no se pueden realizar operaciones aritméticas, pero sí se pueden aplicar criterios de ordenación, operadores lógicos y operaciones de clasificación.

Son ejemplos de variables *ordinales* nivel de instrucción, categoría profesional, clase social.

Nivel de medida de intervalo o escalar

Las variables de nivel de medida de *intervalo*, son aquellas que sus datos son valores numéricos o códigos que se asignan a las categorías de la variable, cada valor define una categoría distinta, lo que le asigna la característica de las variables *nominales*. Entre sus valores se puede establecer un criterio de orden, lo que le asigna la característica de las variables *ordinales*. La característica que las diferencia es que se puede asumir distancia entre sus valores. La asignación de valores o códigos a las categorías se llama codificación (ver el apartado de codificación). La realización de operaciones aritméticas es compleja de determinar, pero se acepta la aplicación de funciones estadísticas. Se pueden aplicar criterios de ordenación, operadores lógicos y operaciones de clasificación.

Son ejemplos de variables de *intervalo* los ítems de las escalas y las propias escalas y las escalas termométricas, con las que se verá un ejemplo.

Un ejemplo típico es el termómetro, que mide temperatura en grados, entre los cuales existe la misma distancia entre dos puntos contiguos de la escala, pero no se pueden establecer magnitudes proporcionales. La diferencia entre 25º C y 26º C es la misma que entre 3º C y 4º C. Pero es incorrecto decir que 30º C sea el doble de calor que 15º C.

Nivel de medida de razón

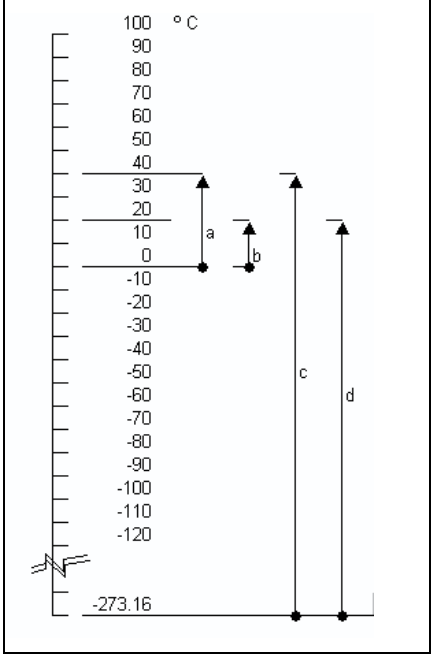
Las variables de nivel de medida de *razón*, son aquellas que sus datos son valores numéricos o códigos significativos. Cada valor define una categoría distinta, lo que le asigna la característica de las variables *nominales*. Entre sus valores se puede establecer un criterio de orden, lo que le asigna la característica de las variables *ordinales*. Existe distancia entre sus valores, lo que le asigna la característica de las variables *intervalares.* La característica que las diferencia es que el cero significa "ausencia de" "valor nulo". A los valores de estas variables se les puede aplicar operaciones aritméticas, criterios de ordenación, operadores lógicos y operaciones de clasificación.

Son ejemplos de medidas de RAZON: edad, peso, estatura, número de hijos, cantidad de productos comprados, salario.

No obstante esta clasificación, en la etapa de tabulación y análisis, la consideración del nivel de medida de las variables puede ajustarse en función de ciertas necesidades y consideraciones, todas ellas argumentadas, como es el caso de variables dicotómicas, binarias y ordinales.

La característica de ausencia de valor del cero, significa que se pueden comparar las magnitudes. Por ejemplo, es correcto decir que un adulto que mida 1,84 m. mide el doble que un niño de 0,92 m. o que una carrera de 300 m. es tres veces más larga que una de 100 m. Pero no es correcto decir que 40º C es el doble de calor que 20º C, sí se puede decir que 40º C es el doble del valor 20º C en la escala centígrada, en la que el 0º C es por convenio y es la posición en la que el agua se solidifica. Para que la temperatura se pueda comparar es necesario que esté referida a la *escala de temperatura termodinámica* o Kelvin en la que el cero tiene valor absoluto y se corresponde con los -273,16º C. El Gráfico 2 muestra que el segmento *a* con el valor 40 es el doble que el segmento *b* con el valor 20, según la escala Centígrada. Pero el segmento *c* no es el doble de calor que el segmento *d*, tomando como referencia el cero absoluto (0 K) que se corresponde con -273,16º C.

Gráfico 2 Comparación de escalas



Un ejemplo de las dificultades que se presentan en el momento de tomar la decisión de *clasificar* o *medir* a las unidades u objetos de observación, se puede ver al determinar la característica de si el objeto fuma o no. Dependiendo de cómo hagamos la pregunta, se considerará clasificación o medición, y determinará la implementación u operacionalización de la variable. La diferenciación entre *clasificación* y *medición*, lleva aparejada la consideración de *fiabilidad, validez* (del instrumento de medida) y *error de la medida*.[44]

La definición que se va a considerar de *medir* es la que facilita el Diccionario de la Real Academia Española (*op. cit.*) que es "Comparar una cantidad con su respectiva unidad, con el fin de averiguar cuántas veces la segunda está contenida en la primera".

La definición considerada de clasificación es: "Ordenar o disponer por clases" (Real Academia Española *op. cit.*), y de manera más amplia: "colocar (un grupo de personas o cosas) en clases o categorías según cualidades o características compartidas"[45] (ver nota 3).

---

[44] Una discusión detallada sobre el tema se puede ver en De la Puente (2007 b).

[45] "classify *verb*" *The Oxford Dictionary of English* (revised edition). Ed. Catherine Soanes and Angus Stevenson. Oxford University Press, 2005. *Oxford Reference Online*. Oxford University Press. Universidad Complutense de Madrid. 14 July 2008 <http://www.oxfordreference.com/views/ENTRY.html?subview=Main&entry=t140.e14265>

Estas definiciones se pueden considerar iguales a la utilizada en Ingeniería. “Ordenación o categorización de partículas u objetos por un criterio establecido, como el tamaño, función, o color” (McGraw-Hill, 2002).

Se considera *validez del instrumento de medida*: cuando el instrumento sirve para medir aquello que se quiere medir. Ejemplos de instrumento válidos son una balanza, una cinta métrica, un calibre. La balanza sirve para medir peso, la cinta métrica longitudes, etc.

Se considera *fiabilidad del instrumento de medida* cuando al aplicar el instrumento de medida por distintos investigadores, a iguales o distintas personas, en iguales o distintos momentos, pero en las mismas condiciones ambientales, producen los mismos resultados si los objetos medidos son iguales en la característica medida. Ejemplo: si diferentes investigadores con la misma balanza pesan a la misma persona, deben obtener el mismo resultado, entendiendo que el peso de la persona no ha variado.

La validez y fiabilidad del instrumento de medida son conceptos complejos ontológica y epistemológicamente y no se agotan con las definiciones dadas anteriormente, pero permiten saber de qué manera se usan en este texto, y se asume que es fácil dar la definición, pero puede ser compleja su aplicación.

El error de la medición en Ciencia y Tecnología sería “cualquier diferencia entre un cálculo, observación o cantidad medida y el verdadero, específico, o teórico valor correcto de esa cantidad” (McGraw-Hill, *op. cit.*).

Volviendo al caso mencionado antes, si se quiere saber si una persona, grupo de personas, muestra o universo fuma o no, se puede planificar la recogida de información de muchas maneras. Por ejemplo diseñando una pregunta con un espacio muestral exhaustivo, excluyente y dicotómico de tipo categórica, con dos sucesos elementales. La pregunta puede ser:

P-1	¿Fuma Ud.?	
	Sí	1
	No	2

Esta pregunta se implementaría o se operacionalizaría en una variable que tendría un espacio muestral exhaustivo, excluyente y dicotómico de tipo categórica, con dos sucesos elementales que al codificarla sería de nivel de medida nominal. El problema que presenta esta pregunta es de tipo epistemológico y ontológico combinado. El hecho o acto de fumar queda sometido al criterio de cada uno de los objetos, porque no fumar puede ser lo que entienda cada individuo: ningún cigarro al día, fumar sólo después de la comidas, algún cigarro al mes, etc. Por lo tanto, este instrumento de obtención de datos no sería ni fiable ni válido. Otra forma posible es hacer la pregunta de tipo categórica pero ordinal:

P-1	¿Considera Ud. que fuma ...?	
	Nada	1
	Regular	2
	Mucho	3

Pero plantea los mismos problemas que la anterior. Se puede optar por una pregunta de tipo escalar o intervalar: Escala de Intensidad de la siguiente manera:

P-1	¿Podría indicar cuánto fuma en una escala de 0 a 10 en la que el 0 significa nada y el 10 mucho?										
	0	1	2	3	4	5	6	7	8	9	10

En este tipo de pregunta se dan los mismos problemas que en las anteriores, además del problema indicado en las escalas termométricas. El criterio de subjetividad sería paradójico. Supongamos dos individuos A y B, siendo A que fuma 40 cigarrillos/día pero está en un grupo en el que cada individuo fuma 80 cigarrillos/día y el individuo B con 20 cigarrillos/día pero está en un grupo en el que cada individuo fuma 10 cigarrillos/día. En esta hipotética situación, el B podría situarse en la escala en el valor 8 mientras que el A podría situarse en el 4. Siendo que el A fuma el doble que el B, la escala mostraría que el B tiene el doble del valor de A. Probablemente este hecho no se producirá, pero si fuese así, no se podría controlar.

Por último, la pregunta de tipo de razón sería:

P-1	¿Podría indicar cuántos cigarrillos fuma al día?		
	Nº de cigarrillos	___	___

Este tipo de pregunta o instrumento de obtención de datos se puede considerar válido, fiable y medición, ya que el elemento base, el cero, es ajeno al sujeto y al objeto. Pero no han terminado los problemas, porque ahora que cumple esos requisitos aparece en escena el problema del error. ¿Cuál es la diferencia entre la respuesta y lo real? ¿Qué es lo que considera cada uno fumar un cigarrillo? ¿Quienes dan la misma respuesta han fumado lo mismo? Por fumar un cigarrillo se puede entender encenderlo y tirarlo; encenderlo fumar la mitad y tirarlo, o encenderlo y fumarlo hasta la boquilla. Evidentemente, estos tres individuos habiendo fumado el mismo número de cigarrillos no habrían fumado la misma cantidad de tabaco. Entonces la pregunta tendría que ser algo así:

P-1	¿De los cigarrillos que encendió "tal día", podría indicar la longitud total que fumó?		
	Longitud	___ , __	cm.

No obstante, seguiría existiendo el error, del instrumento de medida, el criterio de fallo humano, el redondeo utilizado. Se puede plantear la pregunta de diferentes maneras, pero todas ellas llevarían aparejado el problema del error. No obstante, se ha pasado de si el instrumento es válido y fiable a siendo válido y fiable cuál es el error que cometemos.

El acto de fumar es aparentemente simple, pero su clasificación o medición es compleja, igual que cualquier otro acto humano.

VARIABLES DISCRETAS Y CONTINUAS

Además del nivel de medida, otra diferencia es la que se da entre variables continuas y variables discretas. Una variable se considera continua si entre cualesquiera dos valores, puede tomar otros que se pueden considerar infinitos. Aunque en realidad las posiciones intermedias dependen de la precisión del instrumento de medida y el concepto infinito es más una cuestión filosófica que real. También se puede considerar como una variable continua la

que sus valores pertenecen a los números reales que se definen de manera axiomática como el conjunto de números que se encuentran en correspondencia biunívoca con los puntos de una recta infinita (*continuum*): la recta numérica. Ejemplos: salario, edad, estatura, peso.

Una variable discreta sería la que entre cualesquiera dos valores contiguos no existen posiciones intermedias y se corresponderían con los números enteros, siendo que los números enteros se representan gráficamente en la recta de números enteros como puntos a un mismo espacio entre sí, desde menos infinito, ..., -3 , -2, -1, 0, 1, 2, 3,... hasta más infinito. Ejemplos: número de hijos, número de cigarros fumados, veces que se ha ido al cine, número de días trabajados, edad.

En Sociología sería más apropiado hablar de números naturales, puesto que las variables utilizadas no pueden tener valores negativos. No se puede tener peso negativo, número de hijos negativo, etc. La excepción son las escalas construidas que pueden estar en el ámbito de los números enteros negativos.

A veces las variables tienen la doble consideración. Por ejemplo, la edad se trata siempre como variable discreta cuando se dice los años cumplidos, aunque en realidad es una variable continua. Sean consideradas continuas o discretas las variables, cuando se aplican funciones estadísticas (media [$\overline{X}$], varianza [S^2], desviación típica [S], etc.) éstas se consideran valores continuos y se presentarán con decimales.

LAS VARIABLES SEGUN SU RELACIÓN

En los procesos de análisis las variables se consideran según la relación entre ellas. Genéricamente se consideran variables *dependientes* o *independientes*.

El concepto de dependencia de una variable tiene varias definiciones. “En un estudio, análisis o modelo, una variable dependiente es el elemento social cuyas características o variaciones serán explicadas por la referencia a la influencia de otra, anterior, llamada variable independiente”[46] (ver nota 3).

En los métodos de investigación y estadísticos, “es una variable que potencialmente puede ser influida por una o más variables independientes. El propósito de un experimento es típicamente determinar si una o más variables independientes influyen en una o más variables dependientes de alguna manera”[47] (ver nota 3).

“En la regresión múltiple, un grupo de variables independientes o predictoras se combinan en un modelo lineal para proporcionar la mejor predicción de una variable dependiente que a veces se llama la variable criterio”[48] (ver nota 3).

Matemáticamente “si y es una función de x ($y = f(x)$), esto es, si la función asigna un solo valor a y por cada valor de x, entonces y es la variable dependiente” (McGraw-Hill, *op. cit.*) (Ver nota 3).

La variable independiente (o explicativa) es la que “en un estudio, análisis o modelo, (...) es el elemento social cuyas características o variaciones forman y determinan la variable dependiente: En una situación experimental, pueden manipularse las variables independientes sistemáticamente, para que se pueda observar el efecto producido en la variable dependiente.

---

[46] "dependent variable" A Dictionary of Sociology. John Scott and Gordon Marshall. Oxford University Press 2005. Oxford Reference Online. Oxford University Press. Universidad Complutense de Madrid. 11 July 2008 http://www.oxfordreference.com/views/ENTRY.html?subview=Main&entry=t88.e551.

[47] "dependent variable n." A Dictionary of Psychology. Andrew M. Colman. Oxford University Press, 2006. Oxford Reference Online. Oxford University Press. Universidad Complutense de Madrid. 11 July 2008 http://www.oxfordreference.com/views/ENTRY.html?subview=Main&entry=t87.e2193.

[48] *Ibíd.*

El que una variable sea tratada como dependiente o independiente esta determinado por el marco teórico y el enfoque del estudio, pero las variables independientes deben preceder a la variable dependiente, y debe ser la causa"[49] (ver nota 3).

En un diseño experimental la variable independiente es "una variable que es controlada/manipulada por el experimentador, independientemente de las variables extrañas, para examinar sus efectos en la variable dependiente"[50] (ver nota 3).

Matemáticamente la variable independiente es "en una ecuación $y = f(x)$, la variable de entrada x. También conocido como el argumento"[51] (ver nota 3).

Definir la variable dependiente (variable no controlada), asume la definición de la variable independiente (variable controlada). Los nombres que pueden recibir según los procedimientos estadísticos que se utilizan se muestran en la Tabla 13.

Tabla 13 Relación entre variables

Procedimiento Estadístico	Tabla de Contingencia	Diferencia de medias		Análisis de Varianza	Regresión
		Muestras Independientes	Muestras Emparejadas		
Variable Dependiente	Variable Dependiente	Agrupada y numérica	No procede relación y son numéricas	Agrupada y numérica	Explicada o Predicha (Variable Criterio)
Variable Independiente	Variable Independiente	Agrupamiento y categórica		Agrupamiento y categórica	Explicativa o Predictora

[49] "independent variable" A Dictionary of Sociology. John Scott and Gordon Marshall. Oxford University Press 2005. Oxford Reference Online. Oxford University Press. Universidad Complutense de Madrid. 11 July 2008 http://www.oxfordreference.com/views/ENTRY.html?subview=Main&entry=t88.e1083.

[50] "independent variable n." A Dictionary of Psychology. Andrew M. Colman. Oxford University Press, 2006. Oxford Reference Online. Oxford University Press. Universidad Complutense de Madrid. 11 July 2008 http://www.oxfordreference.com/views/ENTRY.html?subview=Main&entry=t87.e4110.

[51] McGraw-Hill (2003). Dictionary of Scientific and Technical Terms.

6.2 Matriz de datos

En Sociología y según el Paradigma Cuantitativo, una de las técnicas de investigación más utilizada es la Encuesta y el principal instrumento de obtención de datos es el Cuestionario en sus diferentes modalidades. El trabajo de campo consiste básicamente en aplicar el cuestionario a las unidades de observación, (Manzano, 1996; Alvira, 2004; García Ferrando, 2005: 167-202). A partir de los cuestionarios recogidos en campo y que se han rellenado con la información facilitada por los objetos o unidades de observación, se procede a crear la Matriz de Datos (Tabla 14) sobre la que posteriormente se aplicarán los procedimientos estadísticos y gráficos, a través de un programa estadístico.

Tabla 14 Matriz de datos.

			Columna 1	Columna 2	Columna 3	...	Columna c
			Variable 1	Variable 2	Variable 3	...	Variable c
Fila 1	Cuestionario 1	Caso 1	$Celda_{11}$	$Celda_{12}$	$Celda_{13}$		$Celda_{1c}$
Fila 2	Cuestionario 2	Caso 2	$Celda_{21}$	$Celda_{22}$	$Celda_{23}$		
Fila 3	Cuestionario 3	Caso 3	$Celda_{31}$	$Celda_{32}$	$Celda_{33}$		
.	.	.					
.	.	.					
.	.	.					
Fila f	Cuestionario f	Caso f	$Celda_{f1}$				$Celda_{fc}$

La matriz de datos es una matriz rectangular de dos dimensiones de *casos* por *variables*. Los casos definen las filas de la matriz y equivalen a las unidades de observación u objetos y cada una de las filas es un cuestionario de los que se recogió anteriormente (más adelante se tratan las matrices de más de dos dimensiones). Las columnas están definidas por las variables que se obtienen por la implementación u operacionalización de las preguntas, en una relación de *uno-a-uno* (a una pregunta le corresponde una variable) o de *uno-a-muchos* (a una pregunta le corresponde más de una variable). La cuadrícula o casilla que se define por el cruce de cada caso con cada variable se denomina *celda.* Cada celda contiene un valor, característica o atributo de la unidad de observación, que se denomina *dato,* y genéricamente, el dato se considera de dos tipos: *válido* y *no válido.*

Una variable toma un valor válido, cuando se corresponde con uno de los sucesos elementales de su espacio muestral. El no válido, es cualquier otro valor no contemplado en el espacio muestral de la variable. Son ejemplos de valores no válidos el no contestar o la respuesta "Ns/Nc" (No sabe/No contesta).

Una columna o variable es el conjunto de datos que se tiene para todos los casos, y deben ser de la misma unidad de medida y de la misma característica.[52] De todos los datos de una variable, al menos uno, debe tener un valor distinto a los demás, porque si no, se denomina *constante*. Una fila es el conjunto de datos que se tiene para cada caso en todas las variables. Los valores de los datos serán del tipo y unidad de medida de la variable correspondiente.

Una *variable* (Ver Epígrafe 6.1) "es la característica medida u observada cuando se realiza un experimento o una observación. Las variables pueden ser no-numéricas

[52] Si la variabl es el peso de las unidades de observación, la variable "peso" debe contener el peso de todas las unidades de observación y en la misma unidad de medida: kg, g, etc. No se puede, por ejemplo, grabar la estatura o el salario en la variable "peso".

(categóricas) o numéricas. Desde una observación no-numérica siempre puede codificarse numéricamente, por lo que una variable, normalmente, siempre es numérica"[53] (ver nota 3).

Los distintos valores, atributos o categorías de una variable constituyen su *espacio muestral* y los denominaremos *sucesos elementales* del espacio muestral de la variable. El espacio muestral es "el conjunto de todos los resultados posibles de un experimento u observación. El concepto se introdujo por von Mises en 1931"[54] (ver nota 3). El espacio muestral se representa con las letras: T , *S* o *E*, y los posibles eventos o sucesos elementales por letras minúsculas ($s_1, s_2, s_3, \ldots s_n$) (Ver Epígrafe 6.1).

Ejemplo 1:

El espacio muestral de tirar un dado de seis caras tiene seis elementos o sucesos elementales:

$$E = (s_1, s_2, s_3, s_4, s_5, s_6)$$

De tal manera que el $s_1 = 1$; el $s_2 = 2$; $s_3 = 3$; $s_4 = 4$; $s_5 = 5$, y $s_6 = 6$. Así que el espacio muestral de tirar un dado es:

$$E = (1, 2, 3, 4, 5, 6)$$

Los s_i de este *E* se consideran exhaustivos y excluyentes. Exhaustivos porque son todos los resultados posibles y son conocidos y excluyentes porque en cada ocasión sólo se puede obtener uno de los resultados posibles.

Ejemplo 2:

El *E* de género en cuanto a sexo tendrá dos elementos:

$$E = (s_1, s_2)$$

De tal manera que el s_1 = *Varón* y el s_2 = *Mujer*. Así que el *E* de *sexo* es:

$$E = (\text{Varón, Mujer})$$

Los s_i de este *E* se consideran exhaustivos y excluyentes. Exhaustivos porque son todos los resultados posibles y son conocidos, y excluyentes porque en cada ocasión sólo se puede obtener uno de los resultados posibles.

Ejemplo 3:

El *E* de Estado Civil, se puede considerar que tiene 6 elementos:

$$E = (s_1, s_2, s_3, s_4, s_5, s_6)$$

De tal manera que el s_1 = Soltero; el s_2 = Casado; s_3 = Pareja; s_4 = Separado; s_5 = Divorciado, y s_6 = Viudo. Así que el *E* de estado civil es:

$$E = (\text{Soltero, Casado, Pareja, Separado, Divorciado, Viudo})$$

Los s_i de este *E* se consideran exhaustivos y excluyentes. Exhaustivos porque son todos los resultados posibles y son conocidos y excluyentes porque en cada ocasión sólo se puede obtener uno de los resultados posibles.

[53] "variable" A Dictionary of Statistics. Graham Upton and Ian Cook. Oxford University Press, 2006. Oxford Reference Online. Oxford University Press. Universidad Complutense de Madrid. 17 July 2008 http://www.oxfordreference.com/views/ENTRY.html?subview=Main&entry=t106.e1703.

[54] "sample space" A Dictionary of Statistics. Graham Upton and Ian Cook. Oxford University Press, 2006. Oxford Reference Online. Oxford University Press. Universidad Complutense de Madrid. 17 July 2008 http://www.oxfordreference.com/views/ENTRY.html?subview=Main&entry=t106.e1433.

6.2.1 La codificación

Se denomina *codificación*, a la asignación de valores o códigos numéricos a las categorías, características o atributos de las variables categóricas (nominales y ordinales) y a las escalares o de intervalo. Esta asignación como no tiene ningún significado, es arbitraria y aleatoria. En las variables ordinales que indican orden, y en las escalares que indican orden y distancia, una vez establecido el origen, los códigos deben mantener un orden y en las escalares, además, distancia.

Ejemplo 1:

La variable “sexo” tiene dos características o atributos: Varón y Mujer.

La asignación de códigos puede ser: Varón = 1; Mujer = 2.

Ejemplo 2:

La variable “estado civil” tiene seis características o atributos: Soltero, Casado, Pareja, Separado, Divorciado y Viudo.

La asignación de códigos puede ser: Soltero = 1, Casado = 2, Pareja = 3, Separado = 4, Divorciado = 5 y Viudo = 6.

Al grabar o escribir en la matriz de datos, los datos que se ponen en cada celda son las características, atributos o valores de las variables que se corresponden con las respuestas a las preguntas. Con la codificación, todos los datos son estrictamente valores numéricos o códigos.

En la Tabla 15 se presenta un modelo de cuestionario, aplicado a un grupo de jóvenes, que servirá de ejemplo para la aplicación de los estadísticos posteriores. Este grupo se utiliza a modo de ejemplo y no tiene ninguna representatividad.

Tabla 15 Cuestionario.

Por favor, marque en la casilla de la derecha o debajo la respuesta a las siguientes preguntas

P1	Por favor, indique si Ud. es ...	
1.	Varón	
2.	Mujer	

P2	Su estado civil es ...	
1.	Soltero/a	
2.	Casado/a	
3.	Pareja	
4.	Separado/a	
5.	Divorciado/a	
6.	Viudo/a	
9	Ns/Nc	

P3	Su programa de TV favorito es de tipo ...	
1.	Cultural	
2.	Informativo	
3.	Recreativo	
4.	Otros: Cuál	

P4	Puede indicar su						
	Peso		Estatura		Edad		
		kg.		m.		años	

P5	Por favor, ¿Podría decir cuál es su interés por estudiar Sociología en una escala del 0 al 10 en la que el 0 es nada de interés y el 10 mucho?.											
	0	1	2	3	4	5	6	7	8	9	10	99
												Nc

P6	Por favor, ¿Podría decir cuánto conocimiento en Sociología considera que tiene en una escala del 0 al 10 en la que el 0 es nada de interés y el 10 mucho?.											
	0	1	2	3	4	5	6	7	8	9	10	99
												Nc

P7	Por favor, ¿Podría indicar cuánto espera aprender de Sociología en una escala del 0 al 10 en la que el 0 es nada de interés y el 10 mucho?.											
	0	1	2	3	4	5	6	7	8	9	10	99
												Nc

La Tabla 16 es la grabación de las respuestas a las preguntas del cuestionario, en las variables correspondientes y codificadas.

Tabla 16 Matriz de datos codificada.

id	p1	p2	p3	P4_1	p4_2	p4_3	p5	p6	p7	id	p1	p2	p3	p4_1	p4_2	p4_3	p5	p6	p7
1	1	1	1	63	1,63	21	7	7	9	50	2	2	3	55	1,74	27	8	6	10
2	1	1	1	63	1,63	21	7	7	9	51	2	1	3	67	1,7	20	5	5	9
3	1	1	1	68	1,75	23	8	5	9	52	1	1	3	77	1,87	19	7	3	8
4	1	1	1	80	1,75	19	7	4	7	53	1	1	3	77	1,87	19	7	3	8
5	1	1	3	73	1,82	24	8	4	9	54	2	1	2	52	1,67	19	8	3	8
6	1	1	3	73	1,82	24	8	4	9	55	1	1	3	78	1,85	21	8	3	10
7	2	1	3	45	1,6	19	5	0	5	56	2	3	3	50	1,67	20	7	5	10
8	2	1	.	60	1,6	20	7	3	8	57	1	1	3	66	1,78	18	5	4	6
9	2	1	.	60	1,72	22	7	5	10	58	1	1	.	65	1,73	19	0	5	6
10	2	1	3	55	1,63	18	9	5	10	59	2	3	3	58	1,63	21	2	1	6
11	1	6	1	85	1,85	20	10	3	9	60	2	1	3	70	1,68	21	7	3	8
12	1	6	1	75	1,75	19	5	5	5	61	1	3	1	70	1,6	20	9	1	9
13	1	6	1	75	1,75	19	5	5	5	62	2	1	2	65	1,77	18	7	5	9
14	2	3	2	53	1,66	18	3	1	99	63	2	1	3	73	1,71	26	8	7	9
15	2	1	2	.	.	.	5	3	6	64	2	1	3	58	1,75	19	8	10	7
16	2	1	1	52	1,66	17	8	6	9	65	2	1	3	75	1,58	18	6	6	8
17	2	2	3	55	1,74	27	8	6	10	66	1	1	3	76	1,9	28	10	5	2
18	2	1	3	67	1,7	20	5	5	9	67	1	1	1	63	1,63	21	7	7	9
19	1	1	3	77	1,87	19	7	3	8	68	2	1	1	52	1,63	25	9	6	9
20	1	1	3	77	1,87	19	7	3	8	69	1	1	1	68	1,75	23	8	5	9
21	2	1	2	52	1,67	19	8	3	8	70	1	1	1	80	1,75	19	7	4	7
22	1	1	3	78	1,85	21	8	3	10	71	1	1	3	73	1,82	24	8	4	9
23	1	1	3	78	1,85	21	8	3	10	72	2	1	1	55	1,6	24	8	6	9
24	1	1	3	66	1,78	18	5	4	6	73	2	1	3	45	1,6	19	5	0	5
25	1	1	.	65	1,73	19	0	5	6	74	2	1	.	60	1,6	20	7	3	8
26	1	1	.	65	1,73	19	0	5	6	75	2	1	.	60	1,72	22	7	5	10
27	2	1	3	70	1,68	21	7	3	8	76	2	1	3	55	1,63	18	9	5	10
28	1	3	1	70	1,6	20	9	1	9	77	1	6	1	85	1,85	20	10	3	9
29	1	3	1	70	1,6	20	9	1	9	78	1	6	1	75	1,75	19	5	5	5
30	2	1	3	73	1,71	26	8	7	9	79	2	1	3	58	1,63	19	6	7	5
31	2	1	3	58	1,75	19	8	10	7	80	2	3	2	53	1,66	18	3	1	99
32	2	1	3	75	1,58	18	6	6	8	81	2	1	2	.	.	.	5	3	6
33	1	1	3	76	1,9	28	10	5	2	82	2	1	1	52	1,66	17	8	6	9
34	1	1	1	63	1,63	21	7	7	9	83	2	2	3	55	1,74	27	8	6	10
35	1	1	1	63	1,63	21	7	7	9	84	1	1	3	66	1,78	18	5	4	6
36	1	1	1	68	1,75	23	8	5	9	85	1	1	3	77	1,87	19	7	3	8
37	1	1	1	80	1,75	19	7	4	7	86	2	1	3	.	1,65	20	6	3	8
38	1	1	3	73	1,82	24	8	4	9	87	2	1	2	52	1,67	19	8	3	8
39	2	1	1	55	1,6	24	8	6	9	88	1	1	3	78	1,85	21	8	3	10
40	2	1	3	45	1,6	19	5	0	5	89	2	3	3	50	1,67	20	7	5	10
41	2	1	.	60	1,6	20	7	3	8	90	1	1	3	66	1,78	18	5	4	6
42	2	1	.	60	1,72	22	7	5	10	91	1	1	.	65	1,73	19	0	5	6
43	2	1	3	55	1,63	18	9	5	10	92	1	3	1	70	1,6	20	9	1	9
44	1	6	1	85	1,85	20	10	3	9	93	2	1	3	70	1,68	21	7	3	8
45	1	6	1	75	1,75	19	5	5	5	94	1	3	1	70	1,6	20	9	1	9
46	1	6	1	75	1,75	19	5	5	5	95	1	1	3	76	1,9	28	10	5	2
47	2	3	2	53	1,66	18	3	1	99	96	2	1	3	73	1,71	26	8	7	9
48	2	1	2	.	.	.	5	3	6	97	2	1	3	58	1,75	19	8	10	7
49	2	1	1	52	1,66	17	8	6	9	98	2	1	3	75	1,58	18	6	6	8
										99	1	1	3	76	1,9	28	10	5	2

En la Tabla 17 se muestra como ejemplo la grabación de los cuestionarios: 1, 7 y 18 sin codificar.

Tabla 17 Matriz de datos sin codificar (tres casos).

id	p1	p2	p3	p4_1	p4_2	p4_3	p5	p6	p7
1	Varón	Soltero	Cultural	63	1,63	21	7	7	9
7	Mujer	Soltera	Recreativo	45	1,6	19	5	0	5
18	Mujer	Soltera	Recreativo	67	1,7	20	5	5	9

Las características o atributos de las variables categóricas (nominal y ordinal), generalmente, son datos de tipo “texto” y su grabación presenta diferencias respecto de las variables numéricas (escalas y razón). Para que todas las variables sean numéricas, es necesario aplicar la codificación, que consiste en asignar códigos o valores numéricos a las características o atributos de las variables categóricas de forma aleatoria y arbitraria, sin ningún significado. Entonces la codificación de la variable “sexo” podría ser: Varón = 12,36 y

Mujer: = 14,58. Aunque esta asignación puede ser válida, no cumple algunas de las reglas de la codificación. Para cumplir las reglas y de forma razonable, ya que es aleatorio y arbitrario, se codifica: Varón = 1 y Mujer = 2 ó Varón = 0 y Mujer = 1 ó Varón = 1 y Mujer = 3 ó Varón = 2 y Mujer = 4.

Las reglas que presenta la codificación son en parte obligatorias y en parte convencionales por opcionales, pero se van a tratar todas como obligatorias. Estas reglas se muestran en la Tabla 18.

Tabla 18 Reglas de la codificación.

∉ Evitan algunos errores.
Explicación: Los atributos o características se pueden escribir de diferentes maneras: con mayúsculas, minúsculas, ambas, con acentos, sin acentos, etc. Así que sería diferentes tipos de "varón" los siguientes. Varon ∏ varon ∏ Varón ∏ varón ∏ VARON ∏ VARÓN. Si se codifica con un valor, por ejemplo el 1, éste sólo puede ser escrito de una manera.
∉ Ahorran tiempo en la grabación.
Explicación: Esta regla se deriva de la anterior, ya que se tarda menos en escribir 1 que en poner Varón. El 1 tiene una única pulsación, mientras que Varón tiene 6 pulsaciones. En un celda el tiempo es imperceptible, pero si consideramos que en Sociología las matrices de datos pueden tener millones de casos y miles de variables, puede suponer muchas horas de trabajo/persona. Los lectores pueden hacer un cálculo de ejemplo con un millón de casos.
∉ Ahorran espacio en el soporte magnético.
Explicación: El sistema binario de almacenamiento de la información en un ordenador precisa para cada carácter un "byte", pero con ese mismo "byte" se pueden representar hasta 256 valores distintos (255 más el 0). La categoría Varón ocuparía 5 "byte", mientras que el código 1 ocuparía 1 "byte". Sugerimos a los lectores que realicen el mismo cálculo de antes para comprobar la diferencia de espacio requerido para el almacenamiento de un millón de casos. NOTA: es diferente el número 1 que el carácter "1", de la misma manera que es diferente el código o número 255 que los caracteres "255". El número 1 ocupa un "byte" el carácter "1" ocupa un "byte". El número 255 ocupa 1 "byte" pero los caracteres "255" ocupan 3 "byte".
∉ Ahorran tiempo de proceso.
Explicación: El procesador de un ordenador procesa más deprisa la información numérica que la información de caracteres. El programa estadístico (realmente es el microprocesador del ordenador) trata matemáticamente los valores numéricos, pero los caracteres tienen un proceso distinto y más elaborado que supone más tiempo.
∉ Algunos procedimientos estadísticos precisan que las variables categóricas estén codificadas con números enteros y más concretamente naturales.
Explicación: Los procedimientos de SPSS: T-test, Análisis de Varianza, Regresión binomial, regresión polinomial, tienen este requerimiento, y no es probable ni deseable que cambie en versiones futuras.

7 Estadística Descriptiva Univariable

La Estadística Descriptiva Univariable se agrupa según la Tabla 19.

Tabla 19 Estadísticos descriptivos univariables y gráficos.

Grupo	Estadístico
Tendencia Central	Moda (M_o)
	Mediana (M_e)
	Media ($\overline{X}$)
Dispersión	Rango o Amplitud (A)
	Varianza (S^2)
	Desviación Típica (S)
	Coeficiente de Variación (CV)
Forma	Asimetría (g_1)
	Apuntamiento (g_2)
Otros	Tabla de Frecuencias
	Percentiles
	Momentos
Gráficos	Diagrama de Barras
	Histograma
	Polígono de Frecuencias

La decisión de qué estadístico aplicar a cada variable, está en función del nivel de medida de la misma. A las variables cualitativas o categóricas, sólo se le puede aplicar la tabla de frecuencias y el diagrama de barras, siendo posible aunque no imprescindible la moda para las nominales y la mediana para las ordinales. El resto de los estadísticos no es estadísticamente apropiado aplicarlos, salvo algunas excepciones como son las variables dicotómicas, las binarias y las ordinales.

Todos los demás estadísticos, el histograma y el polígono de frecuencias se pueden aplicar a las variables cuantitativas o numéricas. La moda, mediana, tabla de frecuencias y el diagrama de barras es estadísticamente apropiado aplicarlos, aunque a veces no es conveniente por la cantidad de valores distintos que tienen los datos de las variables, no resumen lo suficiente, y es una de las finalidades de la Estadística.

7.1 Estadísticos de Tendencia Central

Los estadísticos de tendencia central son: la moda, mediana y media. Se utilizan para describir características de centralidad de las variables.

7.1.1 La moda

La moda es el valor o categoría de la variable que se repiten más veces o que tiene una frecuencia mayor. Esta es la moda considerada absoluta. Puede haber otras modas que se denominan relativas y su característica es que es un valor de la variable que tiene una frecuencia mayor que los valores anterior y posterior. Este estadístico se puede utilizar con las variables de nivel de medida: nominal, ordinal, intervalo y razón. Como el cálculo se realiza a partir de la Tabla de Frecuencias, el resultado puede variar en función del agrupamiento de los intervalos. En las variables categóricas, el valor de la moda se calcula por observación de la

Tabla de Frecuencias (Fórmula 1).

Fórmula 1 Moda.

La tabla de frecuencias de la variable peso recodificada se ha obtenido a partir de la variable *p4_1* (peso) de la matriz de datos de la Tabla 16 y que se corresponde con la pregunta *P4* del cuestionario de la Tabla 15. De las 99 entrevistas realizadas, 95 dieron respuesta válida y 4 no contestaron.

Proceso de cálculo:
1. Se ordena la tabla de frecuencias de menor a mayor.
2. La moda estará en aquella categoría que tenga el mayor número de casos.
3. Entonces se procede a calcular su valor.

En la tabla de frecuencias, el intervalo que tiene el mayor número de casos (32) es el intervalo de "65 a 75 Kg.", y se le llama intervalo crítico (*IC*).

Peso recodificada en 4 intervalos

		Frecuencia
Válidos	45-55 kg	23
	55-65 kg	21
	65-75 kg	32
	75-85 kg	19
	Total	95
Perdidos	Sistema	4
Total		99

$$M_o \mid L_{i41} 2 \frac{A_1}{A_1 2 A_2} \Delta a_i$$

Forma abreviada de la Fórmula 1: (resultado aproximado)

$$M_o \mid L_{i41} 2 \frac{n_{i21}}{n_{i21} 2 n_{i41}} \Delta a_i$$

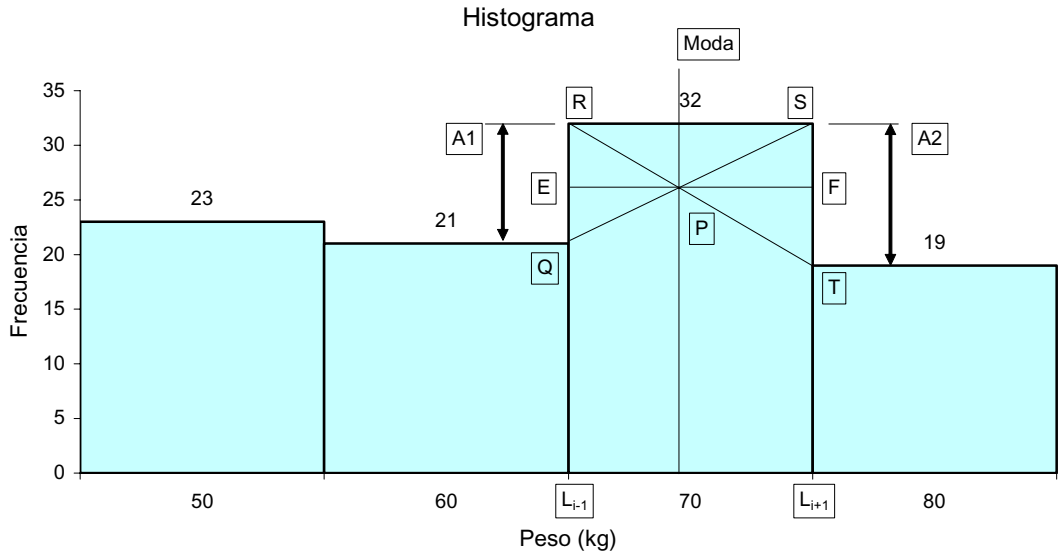
En donde:
M_o: Moda.
L_{i-1}: Límite inferior del *IC*.
A_1: $n_i - n_{i-1}$.
A_2: $n_i - n_{i+1}$.
a_i: Amplitud del *IC*.
n_i: Frecuencia del *IC*.
n_{i+1}: Frecuencia del intervalo posterior al *IC*.
n_{i-1}: Frecuencia del intervalo anterior al *IC*.

Ejemplo:

$$M_o \mid 65 2 \frac{32 4 21}{/32 4 21 0 2 /32 4 19 0} \Delta 10 \mid 69{,}58 \;,\; M_o \mid 65 2 \frac{19}{19 2 21} \Delta 10 \mid 69{,}75$$

La justificación geométrica de la Fórmula 1 se muestra en el Gráfico 3,

Gráfico 3 Desarrollo geométrico del cálculo de la moda.



Por triángulos alternos y según el Teorema de "tales" se cumple que,

$\frac{EP}{RQ} \mid \frac{PF}{ST}$ ó $\frac{M_o \; 4 \; L_{i41}}{A_1} \mid \frac{L_{i21} \; 4 \; M_o}{A_2}$	Fórmula 2

Entonces,

$A_2 / M_o \; 4 \; L_{i41} 0 \mid \; A_1 / L_{i21} \; 4 \; M_o 0$	Fórmula 3
$A_2 M_o \; 4 \; A_2 L_{i41} \mid \; A_1 L_{i21} \; 4 \; A_1 M_o$	Fórmula 4
$A_2 M_o \; 2 \; A_1 M_o \mid \; A_1 L_{i21} \; 2 \; A_2 L_{i41}$	Fórmula 5
$M_o / A_2 \; 2 \; A_1 0 \mid \; A_1 L_{i21} \; 2 \; A_2 L_{i41}$	Fórmula 6
$M_o \mid \frac{A_1 L_{i21} \; 2 \; A_2 L_{i41}}{A_1 \; 2 \; A_2}$	Fórmula 7

Como $L_{i+1} = L_{i-1} + a_i$, tenemos que,

$M_o \mid \frac{A_1 / L_{i41} \; 2 \; a_i 0 2 \; A_2 L_{i41}}{A_1 \; 2 \; A_2}$	Fórmula 8
$M_o \mid \frac{A_1 L_{i41} \; 2 \; A_1 a_i \; 2 \; A_2 L_{i41}}{A_1 \; 2 \; A_2}$	Fórmula 9
$M_o \mid \frac{L_{i41} / A_1 \; 2 \; A_2 0 2 \; A_1 a_i}{A_1 \; 2 \; A_2}$	Fórmula 10
$M_o \mid \frac{L_{i41} / \cancel{A_1 \; 2 \; A_2} 0}{\cancel{A_1 \; 2 \; A_2}} 2 \frac{A_1 a_i}{A_1 \; 2 \; A_2}$	Fórmula 11

Tachando en Fórmula 11, tenemos,

$M_o \mid \; L_{i41} \; 2 \frac{A_1}{A_1 \; 2 \; A_2} \Delta a_i$	Fórmula 12

Y el resultado de la Fórmula 12 coincide con la primera parte de la Fórmula 1.

7.1.2 La mediana

La mediana es el valor de la variable que deja por debajo el 50,0% de los casos. Por lo que por encima de su valor está el otro 50,0%. La mediana se puede utilizar con variables que al menos tengan el nivel de medida ordinal, pero su uso es más adecuado con las de intervalo y razón. Con las variables nominales no se puede utilizar ya que ni siquiera se pueden ordenar los casos.

La fórmula de la mediana es una derivación de la fórmula de los percentiles y se desarrolla con una regla de tres simple.

Fórmula 13 Mediana.

La tabla de frecuencias de la variable peso recodificada se ha obtenido a partir de la variable *p4_1* (peso) de la matriz de datos de la Tabla 16 y que se corresponde con la pregunta *P4* del cuestionario de la Tabla 15. De las 99 entrevistas realizadas, 95 dieron respuesta válida y 4 no contestaron.

Proceso de cálculo:

1. Se ordena la tabla de frecuencias de menor a mayor.
2. Se calculan las frecuencias absolutas acumuladas.
3. Se divide el número de casos por dos y el intervalo que los contiene se le denomina intervalo crítico (*IC*).
4. Entonces se procede a calcular el valor exacto de la mediana.

En la tabla de frecuencias, el intervalo que tiene la mitad de los casos acumulados (95/2 = 47,5) es el intervalo de “65 a 45 Kg.”, y se le considera el *IC*.

Peso recodificada en 4 intervalos

		Frecuencia	Frecuencia acumulada
Válidos	45-55 kg	23	23
	55-65 kg	21	44
	65-75 kg	32	76
	75-85 kg	19	95
	Total	95	
Perdidos	Sistema	4	
Total		99	

$$M_e \mid L_{i41} 2 \frac{\frac{N}{2} 4 N_{i41}}{n_i} \Delta a_i$$

En donde:
M_e: Mediana.
L_{i-1}: Límite inferior del *IC*.
$N/2$: La mitad de los casos.
N_{i-1}: Total de casos por debajo del *IC*.
a_i: Amplitud del *IC*.
n_i: Frecuencia del *IC*.

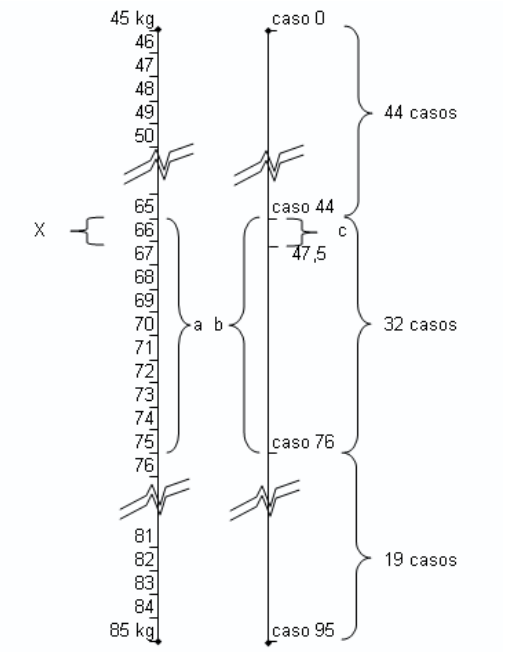
Ejemplo:

$$M_e \mid 65 2 \frac{\frac{95}{2} 4 44}{32} \Delta 10 \mid 65 2 \frac{47{,}5 4 44}{32} \Delta 10 \mid 65 2 0{,}11 \Delta 10 \mid 65 2 1{,}09 \mid 66{,}09$$

Lectura: El valor de la variable de 66,09 Kg. deja por debajo de sí el 50,0 % de los casos. No obstante, esta es una definición teórica, porque si se contabilizan los casos, puede que no coincidan exactamente.

En el Gráfico 4 se ve el razonamiento geométrico.

Gráfico 4 Cálculo gráfico de la mediana.



La mediana es el valor de la variable que le corresponde al caso 47,5, por lo tanto su valor es *65 Kg. + x*, siendo *x* el valor de la variable que le corresponde al segmento *c* de los 3,5 casos. El proceso consiste en calcular *x*.

b ——— *a*
c ——— *x*

32 ——— *10*
3,5 ——— *x*

$$x \mid \frac{c\Delta a}{b} \;,\; x \mid \frac{c}{b}\Delta a \;,\; x \mid \frac{3{,}5}{32}\Delta 10 \mid 1{,}09$$

$$M_e \mid 65 \,2\, x \mid 65 \,2\, 1{,}09 \mid 66{,}09$$

La relación con la Fórmula 13 es:

$$c \mid \frac{N}{2} 4\, N_{i41} \;,\; b \mid n_i \;,\; a \mid a_i \;,\; x \mid \frac{\frac{N}{2} 4\, N_{i41}}{n_i}\Delta a_i$$

Sustituyendo tenemos la Fórmula 13:

$$M_e \mid L_{i41} \,2\, x \;,\; M_e \mid L_{i41} \,2\, \frac{\frac{N}{2} 4\, N_{i41}}{n_i}\Delta a_i$$

7.1.3 La media

La media es el valor que tendrían todos los casos, si todos los casos tuviesen el mismo valor o también se puede considerar como el centro de gravedad de la variable o el punto de apoyo que la mantiene en equilibrio, esto es, que la suma de los valores de los casos que hay a la izquierda "pesan" lo mismo que la suma de los valores de los casos que hay a la derecha y es la suma de los valores de todos los casos dividida por el número de casos. El nivel de medida de las variables debe ser de intervalo o razón.

Fórmula 14 Media.[55]

Tabla de datos Tipo I	$\overline{X} \mid \frac{\sum_{i \mid 1}^{n} x_i}{n}$	En donde: $\overline{X}$: Media. x_i : Valor de la variable para el caso *i*-ésimo. n: Total de los casos.
Tabla de datos Tipo II	$\overline{X} \mid \frac{\sum_{i \mid 1}^{n} x_i n_i}{\sum_{i \mid 1}^{n} n_i}$	En donde: $\overline{X}$: Media. x_i : Valor de la variable en la categoría i. n_i: Número de casos en la categoría i.

Ejemplo T-I: Cálculo de la media de la variable *peso* desde la matriz de datos de la Tabla 16

$$\overline{X} \mid \frac{\sum_{i \mid 1}^{n} x_i}{n} \;,\; \overline{X} \mid \frac{x_1 2 x_2 2 x_3 2 \ldots 2 x_{n41} 2 x_n}{n} \;,\; \overline{X} \mid \frac{63 2 63 2 68 2 \ldots 2 75 2 76}{95} \mid 65{,}86$$

Lectura: La media de la variable *peso*, esto es, el valor que es el centro de gravedad de la variable o el valor que tendrían todos los casos si todos tuviesen el mismo valor es 65,86 Kg. (Datos Tipo I).

Ejemplo T-II: Cálculo de la media de la variable *p4_1* (peso) desde la tabla de datos Tipo II, obtenida de la Tabla 16.

$$\overline{X} \mid \frac{\sum_{i \mid 1}^{20} x_i n_i}{\sum_{i \mid 1}^{20} n_i}$$

$$\overline{X} \mid \frac{x_1 n_1 2 x_2 n_2 2 x_3 n_3 2 \ldots 2 x_{19} n_{19} 2 x_{20} n_{20}}{n_1 2 n_2 2 n_3 2 \ldots 2 n_{19} 2 n_{20}}$$

$$\overline{X} \mid \frac{45\Delta 3 2 50\Delta 2 2 52\Delta 7 2 \ldots 2 80\Delta 3 2 85\Delta 3}{3 2 2 2 7 2 \ldots 2 3 2 3} \mid \frac{6.257}{95} \mid 65{,}86$$

Peso (kg)

		Frecuencia
Válidos	45	3
	50	2
	52	7
	53	3
	55	8
	58	5
	60	6
	63	5
	65	5
	66	4
	67	2
	68	3
	70	8
	73	7
	75	8
	76	4
	77	5
	78	4
	80	3
	85	3
	Total	95
Perdidos	Sistema	4
Total		99

Lectura: La media de la variable *p4_1*, esto es, el valor que es el centro de gravedad de la variable o el valor que tendrían todos los casos si todos tuviesen el mismo valor es 65,86 Kg. (Datos Tipo II).

Ejemplo T-III: Cálculo de la media de la variable *p4_1* (peso) desde la tabla de datos Tipo II obtenida a partir de la tabla de datos Tipo III que se muestra en la Fórmula 13.

$$\overline{X} \mid \frac{\sum_{i \mid 1}^{4} x_i' \Delta n_i}{\sum_{i \mid 1}^{4} n_i}$$

$$\overline{X} \mid \frac{x_1' n_1 2 x_2' n_2 2 x_3' n_3 2 x_4' n_4}{n_1 2 n_2 2 n_3 2 n_4}$$

$$\overline{X} \mid \frac{50\Delta 23 2 60\Delta 21 2 70\Delta 32 2 80\Delta 19}{3 2 2 2 7 2 \ldots 2 3 2 3} \mid \frac{6.257}{95} \mid 64{,}95$$

Peso recodificada en 4 intervalos

		Frecuencia
Válidos	50 kg	23
	60 kg	21
	70 kg	32
	80 kg	19
	Total	95
Perdidos	Sistema	4
Total		99

Lectura: La media de la variable *p4_1*, esto es, el valor que es el centro de gravedad de la variable o el valor que tendrían todos los casos si todos tuviesen el mismo valor es 64,95 Kg. (Datos Tipo III).[56]

[55] Los resultados de los estadísticos calculados sobre datos Tipo I y Tipo II coinciden, pero no coinciden los calculados con los datos Tipo III, por la sustitución que se hace de los intervalos por la Marca de Clase.

[56] Los resultados de los estadísticos calculados sobre datos Tipo I y Tipo II coinciden, pero no coinciden los calculados con los datos Tipo III, por la sustitución que se hace de los intervalos por la Marca de Clase.

En las variables categóricas no se pueden calcular funciones estadísticas como la media, porque no tienen significado los valores al ser asignados de forma arbitraria y aleatoria. Un caso especial es el de las variables dicotómicas codificadas como 1 y 0 y las binarias (ver Tabla 12). En estos casos la media es la proporción de unos.

7.1.3.1 Propiedades de la media

Propiedad 1 Media 1.	X		Y	Ejemplo	X		Y
Si a los valores x_i de una variable *X* le sumamos una constante *A* (x_i + *A*), obtenemos una nueva variable *Y*, de tal forma que $\overline{Y} = \overline{X} + A$.	$x_1 + A$	=	y_1	Sea la *cte*. = 2, si sumamos a todos los valores de *X* la *cte*., la media de la nueva variable obtenida es igual a la media de *X* más la *cte*.	2 + 2	=	4
	$x_2 + A$	=	y_2		3 + 2	=	5
	$x_3 + A$	=	y_3		5 + 2	=	7
	$x_4 + A$	=	y_4		8 + 2	=	10
	$x_5 + A$	=	y_5		1 + 2	=	3
	$x_6 + A$	=	y_6		3 + 2	=	5
	$x_7 + A$	=	y_7		5 + 2	=	7
	$x_8 + A$	=	y_8		7 + 2	=	9
	$x_9 + A$	=	y_9		9 + 2	=	11
	$x_{10} + A$	=	y_{10}		1 + 2	=	3

Demostración:

$$\overline{Y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{\sum_{i=1}^{n} x_i + A}{n} = \frac{(x_1 + A) + (x_2 + A) + (x_3 + A) + \ldots + (x_{n-1} + A) + (x_n + A)}{n} =$$

$$\frac{(x_1 + x_2 + x_3 + \ldots + x_{n-1} + x_n) + (A + A + \ldots + A)}{n} = \frac{(x_1 + x_2 + x_3 + \ldots + x_{n-1} + x_n)}{n} + 3$$

$$3 \frac{(A + A + \ldots + A)}{n} = \frac{\sum_{i=1}^{n} x_i}{n} + \frac{n \cdot A}{n} = \overline{X} + A$$

$$\overline{Y} = \overline{X} + A$$

Ejemplo:

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + x_3 + \ldots + x_{n-1} + x_n}{n} = \frac{2+2+3+2+5+2+8+2+1+2+3+2+5+2+7+2+9+2+1}{10} = 4{,}40$$

$$\overline{Y} = \frac{\sum_{i=1}^{10} y_i}{n} = \frac{y_1 + y_2 + y_3 + y_4 + y_5 + y_6 + y_7 + y_8 + y_9 + y_{10}}{n} = \frac{4+2+5+2+7+2+10+2+3+2+5+2+7+2+9+2}{10} \ldots$$

$$\ldots \frac{11+2+3}{} = 6{,}40$$

$$\overline{Y} = \frac{\sum_{i=1}^{10} y_i}{n} = \frac{\sum_{i=1}^{10} x_i + A}{n} = \frac{(x_1 + A) + (x_2 + A) + (x_3 + A) + (x_4 + A) + (x_5 + A) + (x_6 + A) + 2}{n} \ldots$$

$$\ldots\frac{(x_7 - A) + (x_8 - A) + (x_9 - A) + (x_{10} - A)}{} \mid \frac{(2 + 2) + (3 + 2) + (5 + 2) + (8 + 2) + (1 + 2) +}{10}\ldots$$

$$\ldots\frac{(3 + 2) + (5 + 2) + (7 + 2) + (9 + 2) + (1 + 2)}{} \mid \frac{(2 + 3 + 5 + 8 + 1 + 3 + 5 + 7 + 9 + 1) + (2 + 2 +}{10}\ldots$$

$$\ldots\frac{2 + 2 + 2 + 2 + 2 + 2 + 2 + 2 + 2 + 2)}{} \mid \frac{(2 + 3 + 5 + 8 + 1 + 3 + 5 + 7 + 9 + 1)}{10} \cdot \frac{10 \Delta 2}{10} \mid 4{,}40 + 2 \mid 6{,}40$$

$$\overline{Y} \mid \overline{X} + A \mid 4{,}40 + 2 \mid 6{,}40$$

Propiedad 2 Media 2.	X		Y	Ejemplo	X		Y
Si a los valores x_i de una variable X le restamos una constante A $(x_i - A)$, obtenemos una nueva variable Y, de tal forma que $\overline{Y} \mid \overline{X} - A$.	$x_1 - A$	=	Y_1	Sea la *cte*. = 2, si restamos a todos los valores de X la *cte*., la media de la nueva variable obtenida es igual a la media de X menos la *cte*.	2 - 2	=	0
	$x_2 - A$	=	Y_2		3 - 2	=	1
	$x_3 - A$	=	Y_3		5 - 2	=	3
	$x_4 - A$	=	Y_4		8 - 2	=	6
	$x_5 - A$	=	Y_5		1 - 2	=	-1
	$x_6 - A$	=	Y_6		3 - 2	=	1
	$x_7 - A$	=	Y_7		5 - 2	=	3
	$x_8 - A$	=	Y_8		7 - 2	=	5
	$x_9 - A$	=	Y_9		9 - 2	=	7
	$x_{10} - A$	=	y_{10}		1 - 2	=	-1

Demostración:

$$\overline{Y} \mid \frac{\sum_{i=1}^{n} y_i}{n} \mid \frac{\sum_{i=1}^{n} x_i - A}{n} \mid \frac{(x_1 - A) + (x_2 - A) + (x_3 - A) + \ldots + (x_{n-1} - A) + (x_n - A)}{n} \mid$$

$$\frac{(x_1 + x_2 + x_3 + \ldots + x_{n-1} + x_n) - (A + A + \ldots + A)}{n} \mid \frac{(x_1 + x_2 + x_3 + \ldots + x_{n-1} + x_n)}{n} - \ldots$$

$$\ldots\frac{(A + A + \ldots + A)}{} \mid \frac{\sum_{i=1}^{n} x_i}{n} - \frac{n \Delta A}{n} \mid \overline{X} - A \;,\; \overline{Y} \mid \overline{X} - A$$

Ejemplo:

$$\overline{X} \mid \frac{\sum_{i=1}^{n} x_i}{n} \mid \frac{x_1 + x_2 + x_3 + \ldots + x_{n-1} + x_n}{n} \mid \frac{2 + 2 + 3 + 2 + 5 + 2 + 8 + 2 + 1 + 2 + 3 + 2 + 5 + 2 + 7 + 2 + 9 + 2 + 1}{10} \mid 4{,}40$$

$$\overline{Y} \mid \frac{\sum_{i=1}^{10} y_i}{n} \mid \frac{y_1 + y_2 + y_3 + y_4 + y_5 + y_6 + y_7 + y_8 + y_9 + y_{10}}{n} \mid \frac{0 + 2 + 1 + 2 + 3 + 2 + 6 + 2 / 4\,10 + 2 + 1 + 2 + 3 + 2 + 5 + 2}{10}\ldots$$

$$\ldots\frac{7 + 2 / 4\,10}{} \mid 2{,}40$$

$$\overline{Y} \mid \frac{\sum_{i=1}^{10} y_i}{n} \mid \frac{\sum_{i=1}^{10} x_i - A}{n} \mid \frac{(x_1 - A) + (x_2 - A) + (x_3 - A) + (x_4 - A) + (x_5 - A) + (x_6 - A) +}{n}\ldots$$

$$\ldots\frac{(x_7 \; 4 \; A) \; 2 \; (x_8 \; 4 \; A) \; 2 \; (x_9 \; 4 \; A) \; 2 \; (x_{10} \; 4 \; A)}{} \; | \; \frac{(2 \; 4 \; 2) \; 2 \; (3 \; 4 \; 2) \; 2 \; (5 \; 4 \; 2) \; 2 \; (8 \; 4 \; 2) \; 2 \; (1 \; 4 \; 2) \; 2 \; (3 \; 4 \; 2) \; 2}{10}\ldots$$

$$\ldots\frac{(5 \; 4 \; 2) \; 2 \; (7 \; 4 \; 2) \; 2 \; (9 \; 4 \; 2) \; 2 \; (1 \; 4 \; 2)}{10} \; | \; \frac{(2 \; 2 \; 3 \; 2 \; 5 \; 2 \; 8 \; 2 \; 1 \; 2 \; 3 \; 2 \; 5 \; 2 \; 7 \; 2 \; 9 \; 2 \; 1) \; 4 \; (2 \; 2 \; 2 \; 2 \; 2 \; 2 \; 2 \; 2 \; 2 \; 2}{10}\ldots$$

$$\ldots\frac{2 \; 2 \; 2 \; 2 \; 2 \; 2 \; 2 \; 2 \; 2)}{} \; | \; \frac{(2 \; 2 \; 3 \; 2 \; 5 \; 2 \; 8 \; 2 \; 1 \; 2 \; 3 \; 2 \; 5 \; 2 \; 7 \; 2 \; 9 \; 2 \; 1)}{10} \; 4 \; \frac{10 \; \Delta \; 2}{10} \; | \; 4{,}40 \; 4 \; 2 \; | \; 2{,}40$$

$$\overline{Y} \; | \; \overline{X} \; 4 \; A \; | \; 4{,}40 \; 4 \; 2 \; | \; 2{,}40$$

Propiedad 3 Media 3.	X		Y	Ejemplo	X		Y
Si los valores x_i de una variable X se multiplican por una constante A (x_i x A), obtenemos una nueva variable Y, de tal forma que $\overline{Y} \; | \; \overline{X} \; \Delta \; A$.	x_1 x A	=	y_1	Sea la *cte*. = 2, si multiplicamos a todos los valores de X por la *cte*., la media de la nueva variable obtenida es igual a la media de X por la *cte*.	2 x 2	=	4
	x_2 x A	=	y_2		3 x 2	=	6
	x_3 x A	=	y_3		5 x 2	=	10
	x_4 x A	=	y_4		8 x 2	=	16
	x_5 x A	=	y_5		1 x 2	=	2
	x_6 x A	=	y_6		3 x 2	=	6
	x_7 x A	=	y_7		5 x 2	=	10
	x_8 x A	=	y_8		7 x 2	=	14
	x_9 x A	=	y_9		9 x 2	=	18
	x_{10} x A	=	y_{10}		1 x 2	=	2

Demostración:

$$\overline{Y} \; | \; \frac{\overset{n}{\underset{i \; | \; 1}{\longrightarrow}} y_i}{n} \; | \; \frac{\overset{n}{\underset{i \; | \; 1}{\longrightarrow}} x_i \; \Delta \; A}{n} \; | \; \frac{(x_1 \; \Delta \; A) \; 2 \; (x_2 \; \Delta \; A) \; 2 \; (x_3 \; \Delta \; A) \; 2 \ldots 2 \; (x_{n41} \; \Delta \; A) \; 2 \; (x_n \; \Delta \; A)}{n} \; |$$

$$\frac{A \; \Delta \; (x_1 \; 2 \; x_2 \; 2 \; x_3 \; 2 \ldots 2 \; x_{n41} \; 2 \; x_n)}{n} \; | \; \frac{A \; \Delta \; \overset{n}{\underset{i \; | \; 1}{\longrightarrow}} x_i}{n} \; | \; A \; \Delta \; \frac{\overset{n}{\underset{i \; | \; 1}{\longrightarrow}} x_i}{n} \; | \; \overline{X} \; \Delta \; A$$

Ejemplo:

$$\overline{X} \; | \; \frac{\overset{n}{\underset{i \; | \; 1}{\longrightarrow}} x_i}{n} \; | \; \frac{x_1 \; 2 \; x_2 \; 2 \; x_3 \; 2 \ldots 2 \; x_{n41} \; 2 \; x_n}{n} \; | \; \frac{2 \; 2 \; 3 \; 2 \; 5 \; 2 \; 8 \; 2 \; 1 \; 2 \; 3 \; 2 \; 5 \; 2 \; 7 \; 2 \; 9 \; 2 \; 1}{10} \; | \; 4{,}40$$

$$\overline{Y} \; | \; \frac{\overset{10}{\underset{i \; | \; 1}{\longrightarrow}} y_i}{n} \; | \; \frac{y_1 \; 2 \; y_2 \; 2 \; y_3 \; 2 \; y_4 \; 2 \; y_5 \; 2 \; y_6 \; 2 \; y_7 \; 2 \; y_8 \; 2 \; y_9 \; 2 \; y_{10}}{n} \; | \; \frac{4 \; 2 \; 6 \; 2 \; 10 \; 2 \; 16 \; 2 \; 2 \; 2 \; 6 \; 2 \; 10 \; 2 \; 14 \; 2}{10}\ldots$$

$$\ldots\frac{18 \; 2 \; 2}{} \; | \; 8{,}80$$

$$\overline{Y} \; | \; \frac{\overset{10}{\underset{i \; | \; 1}{\longrightarrow}} y_i}{n} \; | \; \frac{\overset{10}{\underset{i \; | \; 1}{\longrightarrow}} x_i \; \Delta \; A}{n} \; | \; \frac{(x_1 \; \Delta \; A) \; 2 \; (x_2 \; \Delta \; A) \; 2 \; (x_3 \; \Delta \; A) \; 2 \; (x_4 \; \Delta \; A) \; 2 \; (x_5 \; \Delta \; A) \; 2 \; (x_6 \; \Delta \; A) \; 2}{n}\ldots$$

$$\ldots\frac{(x_7 \; \Delta \; A) \; 2 \; (x_8 \; \Delta \; A) \; 2 \; (x_9 \; \Delta \; A) \; 2 \; (x_{10} \; \Delta \; A)}{} \; | \; \frac{/2 \; \Delta \; 20 2 \; /3 \; \Delta \; 20 2 \; /5 \; \Delta \; 20 2 \; /8 \; \Delta \; 20 2 \; /1 \; \Delta \; 20 2 \; /3 \; \Delta \; 20 2}{10}\ldots$$

$$\ldots\frac{/5\Delta 202\ /7\Delta 202\ /9\Delta 202\ /1\Delta 20}{} \mid \frac{2\Delta(2\,2\,3\,2\,5\,2\,8\,2\,1\,2\,3\,2\,5\,2\,7\,2\,9\,2\,1)}{10} \mid 2\Delta\frac{2\,2\,3\,2\,5\,2\,8\,2}{10}\ldots$$

$$\ldots\frac{1\,2\,3\,2\,5\,2\,7\,2\,9\,2\,1}{} \mid 2\Delta 4{,}40 \mid 8{,}80$$

$$\overline{Y} \mid \overline{X}\Delta A \mid 4{,}40\Delta 2 \mid 8{,}80$$

Propiedad 4 Media 4.	X		Y	Ejemplo	X		Y
Si los valores x_i de una variable *X* se dividen por una constante *A* (x_i / *A*), obtenemos una nueva variable *Y*, de tal forma que $\overline{Y} \mid \frac{\overline{X}}{A}$ ó $\overline{Y} \mid \overline{X}\Delta\frac{1}{A}$	$x_1 : A$	=	y_1	Sea la *cte*. = 2, si dividimos a todos los valores de *X* por a *cte*., la media de la nueva variable obtenida es igual a la media de *X* dividida por la *cte*. [nótese que dividir por *A* es igual que multiplicar por el inverso de *A (1/A)*].	2 : 2	=	1,0
	$x_2 : A$	=	y_2		3 : 2	=	1,5
	$x_3 : A$	=	y_3		5 : 2	=	2,5
	$x_4 : A$	=	y_4		8 : 2	=	4,0
	$x_5 : A$	=	y_5		1 : 2	=	0,5
	$x_6 : A$	=	y_6		3 : 2	=	1,5
	$x_7 : A$	=	y_7		5 : 2	=	2,5
	$x_8 : A$	=	y_8		7 : 2	=	3,5
	$x_9 : A$	=	y_9		9 : 2	=	4,5
	$x_{10} : A$	=	y_{10}		1 : 2	=	0,5

Demostración:

$$\overline{Y} \mid \frac{\overline{\overline{y_i}}^{\,n}_{i\mid 1}}{n} \mid \frac{\overline{\frac{x_i}{A}}^{\,n}_{i\mid 1}}{n} \mid \frac{\left(\frac{x_1}{A}\right)\left(2\left(\frac{x_2}{A}\right)\right)\left(2\left(\frac{x_3}{A}\right)\right)\left(2\ldots 2\left(\frac{x_{n41}}{A}\right)\right)\left(2\left(\frac{x_n}{A}\right)\right)}{n} \mid$$

$$\frac{\frac{1}{A}\Delta(x_1\,2\,x_2\,2\,x_3\,2\ldots 2\,x_{n41}\,2\,x_n)}{n} \mid \frac{\frac{1}{A}\Delta\overline{\overline{x_i}}^{\,n}_{i\mid 1}}{n} \mid \frac{1}{A}\Delta\frac{\overline{\overline{x_i}}^{\,n}_{i\mid 1}}{n} \mid \overline{X}\Delta\frac{1}{A}$$

Ejemplo:

$$\overline{X} \mid \frac{\overline{\overline{x_i}}^{\,n}_{i\mid 1}}{n} \mid \frac{x_1\,2\,x_2\,2\,x_3\,2\ldots 2\,x_{n41}\,2\,x_n}{n} \mid \frac{2\,2\,3\,2\,5\,2\,8\,2\,1\,2\,3\,2\,5\,2\,7\,2\,9\,2\,1}{10} \mid 4{,}40$$

$$\overline{Y} \mid \frac{\overline{\overline{y_i}}^{\,10}_{i\mid 1}}{10} \mid \frac{y_1\,2\,y_2\,2\,y_3\,2\,y_4\,2\,y_5\,2\,y_6\,2\,y_7\,2\,y_8\,2\,y_9\,2\,y_{10}}{10} \mid \frac{1{,}0\,2\,1{,}5\,2\,2{,}5\,2\,4{,}0\,2\,0{,}5\,2\,1{,}5\,2}{10}\ldots$$

$$\ldots\frac{2{,}5\,2\,3{,}5\,2\,4{,}5\,2\,0{,}5}{} \mid 2{,}20$$

$$\overline{Y} \mid \frac{\overline{\overline{y_i}}^{\,10}_{i\mid 1}}{n} \mid \frac{\overline{\frac{x_i}{A}}^{\,10}_{i\mid 1}}{n} \mid \frac{\frac{x_1}{A}\,2\,\frac{x_2}{A}\,2\,\frac{x_3}{A}\,2\,\frac{x_4}{A}\,2\,\frac{x_5}{A}\,2\,\frac{x_6}{A}\,2\,\frac{x_7}{A}\,2\,\frac{x_8}{A}\,2\,\frac{x_9}{A}\,2\,\frac{x_{10}}{A}}{n} \mid \frac{\frac{2}{2}\,2\,\frac{3}{2}\,2\,\frac{5}{2}\,2\,\frac{8}{2}\,2\,\frac{1}{2}\,2\,\frac{3}{2}}{10}\ldots$$

$$\ldots\frac{\frac{5}{2}\,2\,\frac{7}{2}\,2\,\frac{9}{2}\,2\,\frac{1}{2}}{} \mid \frac{\frac{1}{2}\Delta(2\,2\,3\,2\,5\,2\,8\,2\,1\,2\,3\,2\,5\,2\,7\,2\,9\,2\,1)}{10} \mid \frac{1}{2}\Delta\frac{2\,2\,3\,2\,5\,2\,8\,2\,1\,2\,3\,2\,5\,2\,7\,2}{10}\ldots$$

$$\ldots\frac{9\,2\,1}{} \mid \frac{1}{2}\Delta 4{,}40 \mid 2{,}20$$

$$\overline{Y} \mid \overline{X}\Delta\frac{1}{A} \mid 4{,}40\Delta\frac{1}{2} \mid 2{,}20$$

Propiedad 5 Media 5.		Ejemplo	X	$\overline{X}$	=
El sumatorio de la distancia de todos los casos respecto de la media es igual a cero.	$\sum_{i\mid 1}^{n} x_i 4 \overline{X}0 \mid 0$	Si a todos les restamos la media, el sumatorio de los resultados obtenidos es iguala cero.	2 -	4,40	-2,40
			3 -	4,40	-1,40
			5 -	4,40	0,60
			8 -	4,40	3,60
			1 -	4,40	-3,40
			3 -	4,40	-1,40
			5 -	4,40	0,60
			7 -	4,40	2,60
			9 -	4,40	4,60
			1 -	4,40	-3,40
			Total		0

Demostración:

$$\sum_{i\mid 1}^{n} x_i 4 \overline{X}0 \mid \sum_{i\mid 1}^{n} x_i 4 \overline{X}0 \mid /x_1 4 \overline{X}02 /x_2 4 \overline{X}02 /x_3 4 \overline{X}02 \ldots 2 /x_{n41} 4 \overline{X}02 /x_n 4 \overline{X}0 \mid$$

$$/\overline{X} 2 \overline{X} 2 \overline{X} 2 \ldots 2 \overline{X} 2 \overline{X}(4 (x_1 2 x_2 2 x_3 2 \ldots 2 x_{n41} 2 x_n)$$

$$n/\overline{X}04\sum_{i\mid 1}^{n} x_i \mid \not{n}\frac{\sum_{i\mid 1}^{n} x_i}{\not{n}} 4\sum_{i\mid 1}^{n} x_i \mid \sum_{i\mid 1}^{n} x_i 4\sum_{i\mid 1}^{n} x_i \mid 0$$

Ejemplo:

$$\sum_{i\mid 1}^{n} x_i 4 \overline{X}0 \mid \sum_{i\mid 1}^{10} x_i 4 \overline{X}0 \mid /x_1 4 \overline{X}02 /x_2 4 \overline{X}02 /x_3 4 \overline{X}02 /x_4 4 \overline{X}02 /x_5 4 \overline{X}02 /x_6 4 \overline{X}02 /x_7 4 \overline{X}02 \ldots$$

$$\ldots/x_8 4 \overline{X}(2 /x_9 4 \overline{X}(2 /x_{10} 4 \overline{X}(\mid /2 4 4{,}4002 /3 4 4{,}4002 /5 4 4{,}4002 /8 4 4{,}4002 /1 4 4{,}4002 \ldots$$

$$\ldots/3 4 4{,}4002 /5 4 4{,}4002 /7 4 4{,}4002 /9 4 4{,}4002 /1 4 4{,}400 \mid (4{,}40 2 4{,}40 2 4{,}40 2 4{,}40 2 \ldots$$

$$2 4{,}40 2 4{,}40 2 4{,}40 2 4{,}40 2 4{,}40 2 4{,}402) 4 (2 2 3 2 5 2 8 2 1 2 3 2 5 2 7 2 9 2 1) \mid$$

$$10\Delta/4{,}4004\sum_{i\mid 1}^{10} x_i \mid 1\acute{0}\frac{\sum_{i\mid 1}^{10} x_i}{1\acute{0}} 4\sum_{i\mid 1}^{10} x_i \mid \sum_{i\mid 1}^{10} x_i 4\sum_{i\mid 1}^{10} x_i \mid 44 4 44 \mid 0$$

7.2 Estadísticos de Dispersión

Los estadísticos de Tendencia Central representan a la variable a través de un único valor. El riesgo es que este valor sea representativo o no de todos los casos y esta característica afecta a la media. Según la definición de la media, no facilita información de cómo están situados todos los casos respecto de ella, pueden coincidir todos con la media y entonces no sería una variable, sería una constante, pueden estar próximos a la media y entonces ésta se consideraría representativa o pueden estar muy alejados. Con las medidas de dispersión se obtiene información de cómo están situados los casos respecto a un estadístico de tendencia central, normalmente, la media.

7.2.1 Rango o Amplitud de la variable

Se define *rango* o *amplitud* de una variable, y se denominará por *A*, a la diferencia entre el valor mayor y el menor de la variable, o sea, los valores más extremos de la variable.

Fórmula 15 Rango o Amplitud.		
Tabla de datos Tipo I	$A \mid V_M 4 V_m$	En donde: *A*: Amplitud. V_M: Valor máximo de la variable. V_m: Valor menor de la variable.
Ejemplo T-I: Cálculo de la amplitud de las variables peso, estatura y edad desde la matriz de datos de la Tabla 16.		
$A_{peso} \mid V_M 4 V_m \mid 85 4 45 \mid 40\,\text{kg}$ $A_{estatura} \mid V_M 4 V_m \mid 1{,}90 4 1{,}58 \mid 0{,}32\,\text{m}$ $A_{edad} \mid V_M 4 V_m \mid 28 4 17 \mid 11\,\text{años}$		

7.2.2 La varianza

El concepto de dispersión es medir la distancia de todos los casos respecto a algún estadístico de tendencia central, normalmente la media. La dispersión de un caso respecto de la media se puede ver por la distancia que hay entre ellos a través de la diferencia.

$dispersión \mid x_i 4 \overline{X}$	Fórmula 16

Para obtener la dispersión de todos los casos se puede aplicar el sumatorio.

$dispersión\_total \mid \sum_{i \mid 1}^{n} (x_i 4 \overline{X})$	Fórmula 17

Y dividiéndolo por el total de casos se obtiene la dispersión media.

$dispersión\_media \mid \frac{\sum_{i \mid 1}^{n} (x_i 4 \overline{X})}{n}$	Fórmula 18

Pero este proceso de cálculo de la dispersión media es importante a efectos conceptuales, pero estadísticamente da siempre como resultado cero, porque el sumatorio de los valores positivos y los negativos, por los casos que quedan por debajo y por encima de la media dan siempre cero (Ver Propiedad 5 de la media, pág. 66), por lo que no resulta útil.

Elevando la diferencia $/x_i 4 \overline{X}0$ al cuadrado, tanto las diferencias positivas como las negativas, se hacen positivas y se obtiene la *varianza*.

<table>
<tr><td colspan="3">Fórmula 19 Varianza (Ver nota 56).[57]</td></tr>
<tr><td>Tabla de datos Tipo I</td><td>$S^2 | \frac{\sum_{i|1}^{n} /x_i 4 \overline{X}0}{n}$ ó
$S^2 | \frac{\sum_{i|1}^{n} x_i^2}{n} 4 \overline{X}^2$</td><td>En donde:
S^2: Varianza.
$\overline{X}$: Media de X.
x_i : Valor de la variable X para el caso i-ésimo.
n: Total de los casos.</td></tr>
<tr><td>Tabla de datos Tipo II</td><td>$S^2 | \frac{\sum_{i|1}^{n} /x_i 4 \overline{X}0 \Delta n_i}{\sum_{i|1}^{n} n_i}$</td><td>En donde:
S^2: Varianza.
$\overline{X}$: Media de X.
x_i : Valor de la variable X en la categoría i-ésima.
n_i: Número de casos en la categoría i.</td></tr>
<tr><td>Tabla de datos Tipo III</td><td>$S^2 | \frac{\sum_{i|1}^{n} /x_i' 4 \overline{X}0 \Delta n_i}{\sum_{i|1}^{n} n_i}$</td><td>En donde:
S^2: Varianza.
$\overline{X}$: Media X.
x_i' : Marca de clase de la variable X en la categoría i-ésima.
n_i: Número de casos en la categoría i.</td></tr>
<tr><td colspan="3">Ejemplo T-I: Cálculo de la varianza de la variable p4_1 (peso) desde la matriz de datos de la Tabla 16
$S^2 | \frac{\sum_{i|1}^{95} /x_i 4 \overline{X}0}{95}$
$\frac{/x_1 4 \overline{X}0 2 /x_2 4 \overline{X}0 2 ... 2 /x_{95} 4 \overline{X}0}{95} | \frac{/63 4 65{,}860 2 /63 4 65{,}860 2 ... 2 /76 4 65{,}860}{95} | \frac{9.651{,}22}{95} | 101{,}59 kg^2$</td></tr>
<tr><td colspan="2">Ejemplo T-II: Cálculo de la media de la variable p4_1 (peso) desde la tabla de datos Tipo II, obtenida desde la matriz de datos de la Tabla 16.
$S^2 | \frac{\sum_{i|1}^{20} /x_i 4 \overline{X}0 \Delta n_i}{\sum_{i|1}^{20} n_i} |$
$\frac{/x_1 4 \overline{X}0 \Delta n_1 2 /x_2 4 \overline{X}0 \Delta n_2 2 ... 2 /x_{20} 4 \overline{X}0 \Delta n_{20}}{n_1 2 n_2 2 ... 2 n_{20}} |$
$\frac{/45 4 65{,}860 \Delta 3 2 /50 4 65{,}860 \Delta 2 2 ... 2 /85 4 65{,}860 \Delta 3}{3 2 2 2 ... 2 3} | \frac{9.651{,}22}{95} | 101{,}59 kg^2$</td><td>
Peso (kg)
<table>
<tr><th></th><th></th><th>Frecuencia</th></tr>
<tr><td>Válidos</td><td>45</td><td>3</td></tr>
<tr><td></td><td>50</td><td>2</td></tr>
<tr><td></td><td>52</td><td>7</td></tr>
<tr><td></td><td>53</td><td>3</td></tr>
<tr><td></td><td>55</td><td>8</td></tr>
<tr><td></td><td>58</td><td>5</td></tr>
<tr><td></td><td>60</td><td>6</td></tr>
<tr><td></td><td>63</td><td>5</td></tr>
<tr><td></td><td>65</td><td>5</td></tr>
<tr><td></td><td>66</td><td>4</td></tr>
<tr><td></td><td>67</td><td>2</td></tr>
<tr><td></td><td>68</td><td>3</td></tr>
<tr><td></td><td>70</td><td>8</td></tr>
<tr><td></td><td>73</td><td>7</td></tr>
<tr><td></td><td>75</td><td>8</td></tr>
<tr><td></td><td>76</td><td>4</td></tr>
<tr><td></td><td>77</td><td>5</td></tr>
<tr><td></td><td>78</td><td>4</td></tr>
<tr><td></td><td>80</td><td>3</td></tr>
<tr><td></td><td>85</td><td>3</td></tr>
<tr><td></td><td>Total</td><td>95</td></tr>
<tr><td>Perdidos</td><td>Sistema</td><td>4</td></tr>
<tr><td>Total</td><td></td><td>99</td></tr>
</table>
</td></tr>
</table>

[57] Todos los cálculos estadísticos se realizan con SPSS y EXCEL. Los resultados expuestos están de acuerdo a las fórmulas expresadas. Si se comparan los resultados, se comprobará que en la *varianza* no coinciden. En la fórmula de la *varianza*, SPSS utiliza el concepto de *cuasi-varianza*, que en el denominador en vez de *n* utiliza *(n-1)*. La *cuasi-varianza* es un corrector para *n* pequeñas, ya que a medida que la *n* se hace grande, la diferencia entre la *varianza* y la *cuasi-varianza* se reduce, llegando incluso a anularse.

Fórmula 19 Continuación.

Ejemplo T-III: Cálculo de la media de la variable *p4_1* (peso) desde la tabla de datos Tipo II obtenida de la tabla de datos Tipo III que se muestra en la Fórmula 13.

Peso recodificada en 4 intervalos

		Frecuencia
Válidos	50 kg	23
	60 kg	21
	70 kg	32
	80 kg	19
	Total	95
Perdidos	Sistema	4
Total		99

$$S^2 = \frac{\sum_{i=1}^{4} x_i' \cdot \bar{X} \cdot \Delta n_i}{\sum_{i=1}^{4} n_i}$$

$$\frac{x_1' \cdot \bar{X} \cdot \Delta n_1 + x_2' \cdot \bar{X} \cdot \Delta n_2 + x_3' \cdot \bar{X} \cdot \Delta n_3 + x_4' \cdot \bar{X} \cdot \Delta n_4}{n_1 + n_2 + n_3 + n_4}$$

$$\frac{50 \cdot 64{,}95 \cdot \Delta 23 + 60 \cdot 64{,}95 \cdot \Delta 21 + 70 \cdot 64{,}95 \cdot \Delta 32 + 80 \cdot 64{,}95 \cdot \Delta 19}{23 + 21 + 32 + 19} = \frac{10.774{,}74}{95} = 113{,}42 kg^2$$

7.2.2.1 Propiedades de la varianza

Propiedad 6 Varianza 1.	X		Y	Ejemplo	X		Y
Si a los valores x_i de una variable *X* le sumamos una constante *A* (x_i + *A*), obtenemos una nueva variable Y, de tal forma que $S_Y^2 = S_X^2$.	$x_1 + A$	=	y_1	Sea la *cte*. = 2, si sumamos a todos los valores de *X* la *cte*.,la varianza de la nueva variable *Y* obtenida es igual a la varianza de *X*.	2 + 2	=	4
	$x_2 + A$	=	y_2		3 + 2	=	5
	$x_3 + A$	=	y_3		5 + 2	=	7
	$x_4 + A$	=	y_4		8 + 2	=	10
	$x_5 + A$	=	y_5		1 + 2	=	3
	$x_6 + A$	=	y_6		3 + 2	=	5
	$x_7 + A$	=	y_7		5 + 2	=	7
	$x_8 + A$	=	y_8		7 + 2	=	9
	$x_9 + A$	=	y_9		9 + 2	=	11
	$x_{10} + A$	=	y_{10}		1 + 2	=	3

Demostración:

$$S_Y^2 = \frac{\sum_{i=1}^{n} y_i \cdot \bar{Y}}{n} = \frac{\sum_{i=1}^{n} (x_i + A) \cdot (\bar{X} + A)^2}{n} = \frac{\sum_{i=1}^{n} (x_i \cdot \bar{X} + x_i \cdot A + A \cdot A)^2}{n} = \frac{\sum_{i=1}^{n} x_i \cdot \bar{X}}{n} = S_X^2$$

Ejemplo:

$$S_Y^2 = \frac{\sum_{i=1}^{n} y_i \cdot \bar{Y}}{n} = \frac{y_1 \cdot \bar{Y} + y_1 \cdot \bar{Y} + \ldots + y_n \cdot \bar{Y}}{n}$$

$$S_Y^2 = \frac{\sum_{i=1}^{10} y_i \cdot \bar{Y}}{10} = \frac{y_1 \cdot \bar{Y} + y_1 \cdot \bar{Y} + \ldots + y_{10} \cdot \bar{Y}}{10} = \frac{4 \cdot 6{,}40 + 5 \cdot 6{,}40 + \ldots}{10} \ldots$$

$$\ldots \frac{3 \cdot 6{,}40}{10} = \frac{74{,}40}{10} = 7{,}44$$

$$S_X^2 = \frac{\sum_{i=1}^{n} x_i \cdot \bar{X}}{n} = \frac{x_1 \cdot \bar{X} + x_2 \cdot \bar{X} + \ldots + x_n \cdot \bar{X}}{n}$$

$$S_X^2 = \frac{\sum_{i=1}^{10}(x_i - \bar{X})^2}{10} = \frac{(x_1 - \bar{X})^2 + (x_2 - \bar{X})^2 + \ldots + (x_{10} - \bar{X})^2}{10} = \frac{(2 - 4{,}40)^2 + (3 - 4{,}40)^2 + \ldots}{10} \ldots$$

$$\ldots \frac{\ldots + (1 - 4{,}40)^2}{10} = \frac{74{,}40}{10} = 7{,}44$$

$$S_Y^2 = S_X^2 \Rightarrow 7{,}44 = 7{,}44$$

Propiedad 7 Varianza 2.	X		Y	Ejemplo	X		Y
Si a los valores x_i de una variable X le restamos una constante A $(x_i - A)$, obtenemos una nueva variable Y, de tal forma que $S_Y^2 = S_X^2$.	$x_1 - A$	=	Y_1	Sea la *cte.* = 2, si restamos a todos los valores de X la *cte.*,la varianza de la nueva variable Y obtenida es igual a la varianza de X.	2 - 2	=	0
	$x_2 - A$	=	Y_2		3 - 2	=	1
	$x_3 - A$	=	Y_3		5 - 2	=	3
	$x_4 - A$	=	Y_4		8 - 2	=	6
	$x_5 - A$	=	Y_5		1 - 2	=	-1
	$x_6 - A$	=	Y_6		3 - 2	=	1
	$x_7 - A$	=	Y_7		5 - 2	=	3
	$x_8 - A$	=	Y_8		7 - 2	=	5
	$X_9 - A$	=	Y_9		9 - 2	=	7
	$x_{10} - A$	=	y_{10}		1 - 2	=	-1

Demostración:

$$S_Y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{Y})^2}{n} = \frac{\sum_{i=1}^{n}[x_i - A - (\bar{Y} - A)]^2}{n} = \frac{\sum_{i=1}^{n}[x_i - \bar{X} - (A - A)]^2}{n} = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n} = S_X^2$$

Ejemplo:

$$S_Y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{Y})^2}{n} = \frac{(y_1 - \bar{Y})^2 + (y_1 - \bar{Y})^2 + \ldots + (y_n - \bar{Y})^2}{n}$$

$$S_Y^2 = \frac{\sum_{i=1}^{10}(y_i - \bar{Y})^2}{10} = \frac{(y_1 - \bar{Y})^2 + (y_2 - \bar{Y})^2 + \ldots + (y_{10} - \bar{Y})^2}{10} = \frac{(0 - 2{,}40)^2 + (1 - 2{,}40)^2 + \ldots}{10} \ldots$$

$$\ldots \frac{\ldots + (1 - 2{,}40)^2}{10} = \frac{74{,}40}{10} = 7{,}44$$

$$S_X^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n} = \frac{(x_1 - \bar{X})^2 + (x_2 - \bar{X})^2 + \ldots + (x_n - \bar{X})^2}{n}$$

$$S_X^2 = \frac{\sum_{i=1}^{10}(x_i - \bar{X})^2}{10} = \frac{(x_1 - \bar{X})^2 + (x_2 - \bar{X})^2 + \ldots + (x_{10} - \bar{X})^2}{10} = \frac{(2 - 4{,}40)^2 + (3 - 4{,}40)^2 + \ldots}{10} \ldots$$

$$\ldots \frac{\ldots + (1 - 4{,}40)^2}{10} = \frac{74{,}40}{10} = 7{,}44$$

$$S_Y^2 = S_X^2 \Rightarrow 7{,}44 = 7{,}44$$

Propiedad 8 Varianza 3.	X		Y	Ejemplo	X		Y
Si los valores x_i de una variable *X* se multiplican por una constante *A* (x_i x *A*), obtenemos una nueva variable *Y*, de tal forma que $S_Y^2 \mid S_X^2 \Delta A^2$.	x_1 x *A*	=	y_1	Sea la *cte*. = 2, si multiplicamos a todos los valores de *X* por la *cte*., la varianza de la nueva variable *Y* es igual a la varianza de *X* por la *cte*. elevada al cuadrado.	2 x 2	=	4
	x_2 x *A*	=	y_2		3 x 2	=	6
	x_3 x *A*	=	y_3		5 x 2	=	10
	x_4 x *A*	=	y_4		8 x 2	=	16
	x_5 x *A*	=	y_5		1 x 2	=	2
	x_6 x *A*	=	y_6		3 x 2	=	6
	x_7 x *A*	=	y_7		5 x 2	=	10
	x_8 x *A*	=	y_8		7 x 2	=	14
	x_9 x *A*	=	y_9		9 x 2	=	18
	x_{10} x *A*	=	y_{10}		1 x 2	=	2

Demostración:

$$S_Y^2 \mid \frac{\sum_{i=1}^{n}(y_i - \overline{Y})}{n} \mid \frac{\sum_{i=1}^{n}(x_i \Delta A - \overline{X} \Delta A)}{n} \mid \frac{\sum_{i=1}^{n}(A \Delta x_i - \overline{X} \Delta A)}{n} \mid \frac{\sum_{i=1}^{n} A^2 \Delta (x_i - \overline{X})}{n} \mid$$

$$\frac{A^2 \Delta (x_1 - \overline{X}) + A^2 \Delta (x_2 - \overline{X}) + \ldots + A^2 \Delta (x_n - \overline{X})}{n} \mid \frac{A^2 \Delta [(x_1 - \overline{X}) + (x_2 - \overline{X}) + \ldots + (x_n - \overline{X})]}{n} \mid$$

$$\frac{A^2 \Delta \sum_{i=1}^{n}(x_i - \overline{X})}{n} \mid S_X^2 \Delta A^2$$

Ejemplo:

$$S_Y^2 \mid \frac{\sum_{i=1}^{n}(y_i - \overline{Y})}{n} \mid \frac{\sum_{i=1}^{10}(y_i - \overline{Y})}{10} \mid \frac{(y_1 - \overline{Y}) + (y_2 - \overline{Y}) + (y_3 - \overline{Y}) + \ldots + (y_{10} - \overline{Y})}{10} \mid$$

$$\frac{(4 - 8{,}80)^2 + (6 - 8{,}80)^2 + (10 - 8{,}80)^2 + \ldots + (2 - 8{,}80)^2}{10} \mid \frac{297{,}60}{10} \mid 29{,}76$$

$$S_X^2 \mid \frac{\sum_{i=1}^{n}(x_i - \overline{X})}{n} \mid \frac{\sum_{i=1}^{10}(x_i - \overline{X})}{10} \mid \frac{(x_1 - \overline{X}) + (x_2 - \overline{X}) + (x_3 - \overline{X}) + \ldots + (x_{10} - \overline{X})}{10} \mid$$

$$\frac{(2 - 4{,}40)^2 + (3 - 4{,}40)^2 + (5 - 4{,}40)^2 + \ldots + (1 - 4{,}40)^2}{10} \mid \frac{74{,}10}{10} \mid 7{,}44$$

$$S_Y^2 \mid S_X^2 \Delta A^2 \mid 7{,}44 \Delta 2^2 \mid 29{,}76$$

Propiedad 9 Varianza 4.	X		Y	Ejemplo	X		Y
Si los valores x_i de una variable *X* se dividen por una constante *A* (x_i : *A*), obtenemos una nueva variable *Y*, de tal forma que $S_Y^2 \mid \frac{S_X^2}{A}$ ó $S_Y^2 \mid S_X^2 \Delta \frac{1}{A}$	x_1 : *A*	=	y_1	Sea la *cte*. = 2, si dividimos a todos los valores de *X* por a *cte*. La varianza de la nueva variable obtenida es igual a la varianza de *X* dividida por la *cte*. [nótese que dividir por *A* es igual que multiplicar por el inverso de *A* (*1/A*)].	2 : 2	=	1,0
	x_2 : *A*	=	y_2		3 : 2	=	1,5
	x_3 : *A*	=	y_3		5 : 2	=	2,5
	x_4 : *A*	=	y_4		8 : 2	=	4,0
	x_5 : *A*	=	y_5		1 : 2	=	0,5
	x_6 : *A*	=	y_6		3 : 2	=	1,5
	x_7 : *A*	=	y_7		5 : 2	=	2,5
	x_8 : *A*	=	y_8		7 : 2	=	3,5
	x_9 : *A*	=	y_9		9 : 2	=	4,5
	x_{10} : *A*	=	y_{10}		1 : 2	=	0,5

Demostración:

$$S_Y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{Y})^2}{n} = \frac{\sum_{i=1}^{n}\left(x_i - \frac{1}{A}\right)\left(\bar{X} - \frac{1}{A}\right)^2}{n} = \frac{\sum_{i=1}^{n}\frac{1}{A^2}(x_i - \bar{X})^2}{n} = \frac{\frac{1}{A^2}(x_1 - \bar{X})^2 + \frac{1}{A^2}(x_2 - \bar{X})^2 + \ldots + \frac{1}{A^2}(x_n - \bar{X})^2}{n} = \frac{\frac{1}{A^2}\left[(x_1 - \bar{X})^2 + (x_2 - \bar{X})^2 + \ldots\right]}{n} = \ldots$$

$$\ldots = \frac{\frac{1}{A^2}\sum_{i=1}^{n}(x_i - \bar{X})^2}{n} = S_X^2\left(\frac{1}{A^2}\right)^2$$

Ejemplo:

$$S_Y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{Y})^2}{n} = \frac{\sum_{i=1}^{10}(y_i - \bar{Y})^2}{10} = \frac{(y_i - \bar{Y})^2 + (y_2 - \bar{Y})^2 + (y_3 - \bar{Y})^2 + \ldots + (y_{10} - \bar{Y})^2}{10} =$$

$$\frac{(0{,}50 - 2{,}20)^2 + (1{,}50 - 2{,}20)^2 + (2{,}50 - 2{,}20)^2 + \ldots + (0{,}50 - 2{,}20)^2}{10} = \frac{18{,}60}{10} = 1{,}86$$

$$S_X^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n} = \frac{\sum_{i=1}^{10}(x_i - \bar{X})^2}{10} = \frac{(x_i - \bar{X})^2 + (x_2 - \bar{X})^2 + (x_3 - \bar{X})^2 + \ldots + (x_{10} - \bar{X})^2}{10} =$$

$$\frac{(2 - 4{,}40)^2 + (3 - 4{,}40)^2 + (5 - 4{,}40)^2 + \ldots + (1 - 4{,}40)^2}{10} = \frac{74{,}40}{10} = 7{,}44$$

$$S_Y^2 = S_X^2\left(\frac{1}{A^2}\right)^2 = 7{,}44\left(\frac{1}{2^2}\right)^2 = 1{,}86$$

7.2.3 La desviación típica

Al tener que elevar al cuadrado la diferencia $(\bar{X} - x_i)$ para eliminar la igualdad a cero, la varianza es un valor elevado al cuadrado y en unidades de la variable elevadas al cuadrado. Si se aplica la raíz cuadrada, se le quita el cuadrado al valor y a las unidades y el valor obtenido se le llama *desviación típica.*

Fórmula 20 Desviación típica.	
$S = \sqrt{S^2}$	En donde: *S*: Desviación típica. S^2: Varianza.
Ejemplo: Cálculo realizado a partir de la varianza de la Fórmula 19	
$S = \sqrt{S^2} = \sqrt{101{,}59\ kg^2} = 10{,}08\ kg$	

7.2.3.1 Propiedades de la desviación típica

Propiedad 10 Desviación Típica 1.	X		Y	Ejemplo	X		Y
Si a los valores x_i de una variable *X* le sumamos una constante *A* (x_i + *A*), obtenemos una nueva variable *Y*, de tal forma que $S_Y \mid S_X$.	$x_1 + A$	=	y_1	Sea la *cte*. = 2, si sumamos a todos los valores de *X* la *cte*., la desviación típica de la nueva variable *Y* obtenida es igual a la desviación típica de *X*.	2 + 2	=	4
	$x_2 + A$	=	y_2		3 + 2	=	5
	$x_3 + A$	=	y_3		5 + 2	=	7
	$x_4 + A$	=	y_4		8 + 2	=	10
	$x_5 + A$	=	y_5		1 + 2	=	3
	$x_6 + A$	=	y_6		3 + 2	=	5
	$x_7 + A$	=	y_7		5 + 2	=	7
	$x_8 + A$	=	y_8		7 + 2	=	9
	$x_9 + A$	=	y_9		9 + 2	=	11
	$x_{10} + A$	=	y_{10}		1 + 2	=	3

Demostración:

Según la demostración de la Propiedad 6, $S_Y^2 \mid S_X^2$, entonces:

$S_Y \mid \sqrt{S_Y^2}$ y $S_X \mid \sqrt{S_X^2}$ entonces $S_Y \mid S_X$

Ejemplo:

$S_Y \mid \sqrt{S_Y^2} \mid \sqrt{7{,}44} \mid 2{,}73$

$S_X \mid \sqrt{S_X^2} \mid \sqrt{7{,}44} \mid 2{,}73$

$S_Y \mid S_X \blacktriangledown 2{,}73 \mid 2{,}73$

Propiedad 11 Desviación Típica 2.	X		Y	Ejemplo	X		Y
Si a los valores x_i de una variable *X* le restamos una constante *A* (x_i - *A*), obtenemos una nueva variable *Y*, de tal forma que $S_Y \mid S_X$.	$x_1 - A$	=	Y_1	Sea la *cte*. = 2, si restamos a todos los valores de *X* la *cte*., la desviación típica de la nueva variable *Y* obtenida es igual a la desviación típica de *X*.	2 - 2	=	2
	$x_2 - A$	=	Y_2		3 - 2	=	1
	$x_3 - A$	=	Y_3		5 - 2	=	3
	$x_4 - A$	=	Y_4		8 - 2	=	6
	$x_5 - A$	=	Y_5		1 - 2	=	-1
	$x_6 - A$	=	Y_6		3 - 2	=	1
	$x_7 - A$	=	Y_7		5 - 2	=	3
	$x_8 - A$	=	Y_8		7 - 2	=	5
	$x_9 - A$	=	Y_9		9 - 2	=	7
	$x_{10} - A$	=	y_{10}		1 - 2	=	-1

Demostración:

Según la demostración de la Propiedad 7, $S_Y^2 \mid S_X^2$, entonces:

si $S_Y \mid \sqrt{S_Y^2}$ y $S_X \mid \sqrt{S_X^2}$ entonces $S_Y \mid S_X$

Ejemplo:

$S_Y \mid \sqrt{S_Y^2} \mid \sqrt{7{,}44} \mid 2{,}73$

$S_X \mid \sqrt{S_X^2} \mid \sqrt{7{,}44} \mid 2{,}73$

$S_Y \mid S_X \blacktriangledown 2{,}73 \mid 2{,}73$

Propiedad 12 Desviación Típica 3.	X		Y	Ejemplo	X		Y
Si los valores x_i de una variable X se multiplican por una constante A (x_i x A), obtenemos una nueva variable Y, de tal forma que $S_Y \mid S_X \Delta A$.	x_1 x A	=	y_1	Sea la *cte*. = 2, si multiplicamos a todos los valores de X por la *cte*., la desviación típica de la nueva variable Y es igual a la desviación típica de X por la *cte*.	2 x 2	=	4
	x_2 x A	=	y_2		3 x 2	=	6
	x_3 x A	=	y_3		5 x 2	=	10
	x_4 x A	=	y_4		8 x 2	=	16
	x_5 x A	=	y_5		1 x 2	=	2
	x_6 x A	=	y_6		3 x 2	=	6
	x_7 x A	=	y_7		5 x 2	=	10
	x_8 x A	=	y_8		7 x 2	=	14
	x_9 x A	=	y_9		9 x 2	=	18
	x_{10} x A	=	y_{10}		1 x 2	=	2

Demostración:

Según la demostración de la Propiedad 8, $S_Y^2 \mid S_X^2 \Delta A^2$, entonces:

$$si\ S_Y \mid \sqrt{S_Y^2}\ \text{y}\ \text{S}_\text{Y}^2 \mid S_X^2 \Delta A^2\ \text{entonces}\ S_Y \mid \sqrt{S_X^2 \Delta A^2}\ \text{y}\ S_Y \mid S_X \Delta A$$

Ejemplo:

$$S_Y \mid \sqrt{S_Y^2} \mid \sqrt{29{,}76} \mid 5{,}46$$
$$S_X \mid \sqrt{S_X^2} \mid \sqrt{7{,}44} \mid 2{,}73$$
$$S_Y \mid S_X \Delta A \mid 2{,}73 \Delta 2 \mid 5{,}46$$

Propiedad 13 Desviación Típica 4.	X		Y	Ejemplo	X		Y
Si los valores x_i de una variable X se dividen por una constante A (x_i : A), obtenemos una nueva variable Y, de tal forma que $S_Y \mid \frac{S_X}{A}$ ó $S_Y \mid S_X \Delta \frac{1}{A}$	$X_1 : A$	=	y_1	Sea la *cte*. = 2, si dividimos a todos los valores de X por la *cte*., la desviación típica de la nueva variable obtenida es igual a la desviación típica de X dividida por la *cte*. [nótese que dividir por A es igual que multiplicar por el inverso de A ($1/A$)].	2 : 2	=	1,0
	$X_2 : A$	=	y_2		3 : 2	=	1,5
	$X_3 : A$	=	y_3		5 : 2	=	2,5
	$X_4 : A$	=	y_4		8 : 2	=	4,0
	$X_5 : A$	=	y_5		1 : 2	=	0,5
	$X_6 : A$	=	y_6		3 : 2	=	1,5
	$X_7 : A$	=	y_7		5 : 2	=	2,5
	$X_8 : A$	=	y_8		7 : 2	=	3,5
	$x_9 : A$	=	y_9		9 : 2	=	4,5
	$x_{10} : A$	=	y_{10}		1 : 2	=	0,5

Demostración:

Según la demostración de la Propiedad 9, $S_Y^2 \mid S_X^2 \Delta \left| \frac{1}{A} \right|^2$, entonces:

$$si\ S_Y \mid \sqrt{S_Y^2}\ \text{y}\ \text{S}_\text{Y}^2 \mid S_X^2 \Delta \left| \frac{1}{A} \right|^2\ \text{entonces}\ S_Y \mid \sqrt{S_X^2 \Delta \left| \frac{1}{A} \right|^2}\ \text{y}\ S_Y \mid S_X \Delta \frac{1}{A}$$

Ejemplo:

$$S_Y \mid \sqrt{S_Y^2} \mid \sqrt{1{,}86} \mid 1{,}36$$
$$S_X \mid \sqrt{S_X^2} \mid \sqrt{7{,}44} \mid 2{,}73$$
$$S_Y \mid S_X \Delta \frac{1}{A} \mid 2{,}73 \Delta \frac{1}{2} \mid 1{,}36$$

7.2.4 El coeficiente de variación

La *varianza* es un estadístico que se puede considerar abstracto porque el resultado es un valor de la variable elevado al cuadrado, al hallar la raíz cuadrada, se elimina la abstracción, pero el valor está influido por la unidad de medida de la variable. Según la Propiedad 12 pág. 74 y Propiedad 13 pág. 74 de la desviación típica, al multiplicar o dividir los valores de la variable por una constante, la desviación típica queda multiplicada o dividida por esa constante. Si se cambia la unidad de medida de una variable, se multiplica o divide por una constante y la desviación típica queda multiplicada o dividida por esa constante, por lo que es un estadístico que no permite interpretar la dispersión de la variable ni compararla con la de otras variables.

El *coeficiente de variación* es la estandarización de la *desviación típica* al eliminar la unidad de medida de la variable.

<table>
<tr><td colspan="3">Fórmula 21 Coeficiente de variación.</td></tr>
<tr><td colspan="2">$CV \mid \frac{S_X}{\overline{X}_X}$</td><td>En donde:
S_X: Desviación típica de X.
$\overline{X}_X$:Media de X.</td></tr>
<tr><td colspan="3">Ejemplo: Cálculo realizado a partir de la varianza de la Fórmula 19</td></tr>
<tr><td colspan="3">$S \mid \sqrt{S^2} \mid \sqrt{101{,}59\ kg^2} \mid 10{,}08\ kg$
$\overline{X} \mid 65{,}86\ \text{kg}$
$CV \mid \frac{S_X}{\overline{X}_X} \mid \frac{10{,}08\cancel{kg}}{65{,}86\cancel{kg}} \mid 0{,}1530 \Delta 100 \mid 15{,}30\%$</td></tr>
<tr><td colspan="3">Interpretación:
Valores que puede tomar el coeficiente de variación.</td></tr>
<tr><td rowspan="3">$CV \mid \frac{S_X}{\overline{X}_X}$</td><td>$S_X \} \overline{X}_X$</td><td>Si se cumple esta condición, entonces $CV > 1$.</td></tr>
<tr><td>$S_X \mid \overline{X}_X$</td><td>Si se cumple esta condición, entonces $CV = 1$.</td></tr>
<tr><td>$S_X \{ \overline{X}_X$</td><td>Si se cumple esta condición, entonces $CV < 1$.</td></tr>
<tr><td colspan="3">El CV establece la relación entre la desviación típica y la media. Cuanto más se aproxime a cero, menor será la dispersión de la variable. No tiene unidad de medida porque las unidades del numerador (desviación típica) al ser las mismas que las del denominador (media) se eliminan.</td></tr>
</table>

Según el Teorema de Tchebycheff, si la desviación típica es mayor o igual que la media, se puede dar la probabilidad de que haya casos con valores negativos en la variable, esta circunstancia es imposible en la mayoría de las variables que se estudian en Sociología, o que la variable tuviese un comportamiento muy anómalo. Por lo tanto un CV igual o mayor que la unidad se interpretará como un valor que indica una dispersión anómala. Incluso por debajo de 1 se considerará desproporcionada. Teóricamente, sólo se puede considerar una dispersión aceptable cuando *CV* sea igual o incluso inferior a 0,5 ó 50,0 %, y cuanto más se aproxime a cero menor será la dispersión.

El *CV* no varía aunque se cambie la unidad de medida de la variable. Esta característica permite que se pueda relacionar la dispersión de variables de diferente magnitud y unidad de medida. La dispersión sólo no dice nada de la distribución de la variable. La dispersión no es indicativa de la representatividad de la variable sobre la población, aunque siempre es estadísticamente más agradable que la dispersión sea baja.

Tabla 20 Ejemplo de cálculo de *coeficiente de variación*.	
Variable *peso* de la Fórmula 14 en kg.	Variable *peso* de la Fórmula 14 en g.
$S_X \mid 10{,}08\,\text{kg}$ $\overline{X}_X \mid 65{,}86\,\text{kg}$ $CV \mid \frac{S_X}{\overline{X}_X} \mid \frac{10{,}08\,\text{kg}}{65{,}86\,\text{kg}} \mid 0{,}1530 \Delta 100 \mid 15{,}30\%$	Según la Propiedad 3 de la media y la Propiedad 12 de la desviación típica: $S_X \mid 10{,}08\,\text{kg} \mid 10.080\,\text{g}$ $\overline{X}_X \mid 65{,}86\,\text{kg} \mid 65.860\,\text{g}$ $CV \mid \frac{S_X}{\overline{X}_X} \mid \frac{10.080\,\text{g}}{65.860\,\text{g}} \mid 0{,}1530 \Delta 100 \mid 15{,}30\%$
Al multiplicar el numerador y el denominador por una constante, el cociente no varía.	

Ejemplo general:

Dos distribuciones pueden tener la misma media y ser diferentes en otros aspectos. Por ejemplo, supuestas las siguientes variables:

Tabla 21 Ejemplo de dispersión.		
	X	*Y*
Caso 1	0	18
Caso 2	10	19
Caso 3	20	20
Caso 4	30	21
Caso 5	40	22

$$\overline{X} \mid \frac{\overline{\sum_{i \mid 1}^{n} x_i}}{n} \mid \frac{\overline{\sum_{i \mid 1}^{5} x_i}}{5} \mid \frac{x_1 2 x 2 x_3 2 x_4 2 x_5}{5} \mid \frac{0 2 10 2 20 2 30 2 40}{5} \mid \frac{100}{5} \mid 20$$

$$\overline{Y} \mid \frac{\overline{\sum_{i \mid 1}^{n} y_i}}{n} \mid \frac{\overline{\sum_{i \mid 1}^{5} y_i}}{5} \mid \frac{y_1 2 y 2 y_3 2 y_4 2 y_5}{5} \mid \frac{18 2 19 2 20 2 21 2 22}{5} \mid \frac{100}{5} \mid 20$$

$$S_X^2 \mid \frac{\overline{\sum_{i \mid 1}^{n} /\overline{X} 4 x_i 0}}{n} \mid \frac{\overline{\sum_{i \mid 1}^{5} /\overline{X} 4 x_i 0}}{5} \mid \frac{/\overline{X} 4 x_1 0 2 /\overline{X} 4 x_2 0 2 /\overline{X} 4 x_3 0 2 /\overline{X} 4 x_4 0 2 /\overline{X} 4 x_5 0}{5} \mid$$

$$\frac{/20 4 0 0 2 /20 4 10 0 2 /20 4 20 0 2 /20 4 30 0 2 /20 4 40 0}{5} \mid 200$$

$$S_Y^2 \mid \frac{\overline{\sum_{i \mid 1}^{n} /\overline{Y} 4 y_i 0}}{n} \mid \frac{\overline{\sum_{i \mid 1}^{5} /\overline{Y} 4 y_i 0}}{5} \mid \frac{/\overline{Y} 4 y_1 0 2 /\overline{Y} 4 y_2 0 2 /\overline{Y} 4 y_3 0 2 /\overline{Y} 4 y_4 0 2 /\overline{Y} 4 y_5 0}{5} \mid$$

$$\frac{/20 4 18 0 2 /20 4 19 0 2 /20 4 20 0 2 /20 4 21 0 2 /20 4 22 0}{5} \mid 2$$

$$S_X \mid \sqrt{S_X^2} \mid \sqrt{200} \mid 14{,}14$$

$$CV_X \mid \frac{S_X^2}{\overline{X}} \mid \frac{14{,}14}{20} \mid 0{,}7071 \Delta 100 \mid 70{,}7\%$$

$$S_Y \mid \sqrt{S_Y^2} \mid \sqrt{2} \mid 1{,}41$$

$$CV_Y \mid \frac{S_Y^2}{\overline{Y}} \mid \frac{1{,}41}{20} \mid 0{,}0707 \Delta 100 \mid 7{,}1\%$$

Tabla 22 Resumen.		
Estadístico	X	Y
V_M	40	22
V_m	0	18
$\overline{X}$	20	20
S^2	200	2
S	14,14	1,41
CV	0,71	0,07

La media de las dos variables es 20. No obstante, un examen detenido de los datos y los demás estadísticos, muestran que las dos variables tienen características distintas. Este ejemplo muestra que la Estadística Descriptiva Univariable no consiste en aplicar un estadístico (la media) a una variable, sino aplicar todos los estadísticos adecuados a cada variable. La media de las dos variables tienen el mismo valor, pero la dispersión de la variable X (CV = 0,71) es mayor que el de la variable Y (CV = 0,07).

7.3 Estadísticos de Forma

Estos estadísticos permiten decir algo sobre la característica de la forma de la distribución de la variable. La forma de la distribución se establece comparándola con la Normal, pero no significa contrastar con la Normal. El significado es que se va a comparar si la forma de la distribución de una variable tiene características similares a la Normal. Pero no se contrasta si la distribución de la variable es normal, supuestamente normal o marcadamente normal. La comparación es descriptiva, el contraste implica cálculo de probabilidades y contraste de Hipótesis.

Los estadísticos de forma miden la *asimetría, oblicuidad* o "*skewness"* (g_1) y el *apuntamiento, curtosis* o *"kurtosis"* (g_2) de la distribución de una variable.

7.3.1 Momentos

Ya se han medido algunas características de las variables (tendencia central y dispersión), ahora se busca dar un tratamiento unificado a estos estadísticos y que sea la base para el cálculo de los estadísticos de *forma.*

Los *momentos* describen características de un conjunto de datos que componen una o más variables. En esta ocasión se tratan solo los momentos de una variable.

Los momentos se clasifican como: momentos *respecto al origen* de una variable y momentos *respecto a un estadístico de tendencia central,* en este caso se considera respecto de la media de la variable.

El momento *a* de orden *r,* de la variable *X,* respecto al origen se representa como a_r y es, por definición:

Tabla de datos Tipo I	
$a_r = \frac{\sum_{i=1}^{n} x_i^r}{n}$ para $r = 0, 1, 2, ...$	Fórmula 22
Tabla de datos Tipo II	
$a_r = \frac{\sum_{i=1}^{n} x_i^r n_i}{\sum_{i=1}^{n} n_i}$ para $r = 0, 1, 2, ...$	Fórmula 23
Tabla de datos Tipo III	
$a_r = \frac{\sum_{i=1}^{n} x_i'^r n_i}{\sum_{i=1}^{n} n_i}$ para $r = 0, 1, 2, ...$	Fórmula 24

Dando valores a *r* se tiene:

Momento de orden $r = 0$	$a_0 = \frac{\sum_{i=1}^{n} x_i^0}{n} = 1$
Momento de orden $r = 1$	$a_1 = \frac{\sum_{i=1}^{n} x_i^1}{n} = media$
Momento de orden $r = 2$	$a_2 = \frac{\sum_{i=1}^{n} x_i^2}{n}$
Momento de orden $r = 3$	$a_3 = \frac{\sum_{i=1}^{n} x_i^3}{n}$

El momento *m* de orden *r,* de la variable *X,* respecto a la media se representa como m_r y es, por definición:

Tabla de datos Tipo I	
$m_r = \frac{\sum_{i=1}^{n} (x_i - \overline{X})^r}{n}$ para $r = 0, 1, 2, ...$	Fórmula 25
Tabla de datos Tipo II	
$m_r = \frac{\sum_{i=1}^{n} (x_i - \overline{X})^r \Delta n_i}{\sum_{i=1}^{n} n_i}$ para $r = 0, 1, 2, ...$	Fórmula 26

Tabla de datos Tipo III	
$m_r = \frac{\sum_{i=1}^{n} (x_i' - \overline{X}')\,\Delta n_i}{\sum_{i=1}^{n} n_i}$ para $r = 0, 1, 2, \ldots$	Fórmula 27

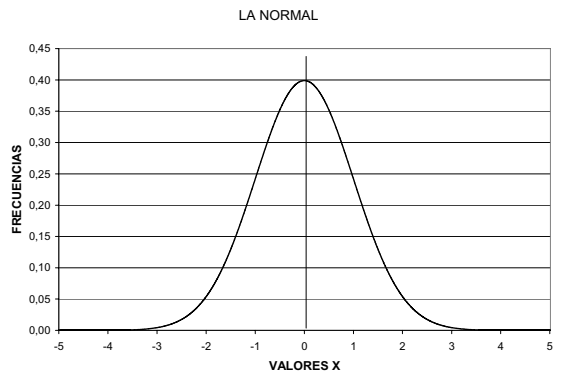
Dando valores a *r* se tiene:

Momento de orden $r = 0$	$m_0 = \frac{\sum_{i=1}^{n} (x_i - \overline{X})^0}{n} = 1$
Momento de orden $r = 1$	$m_1 = \frac{\sum_{i=1}^{n} (x_i - \overline{X})^0}{n} = 0$
Momento de orden $r = 2$	$m_2 = \frac{\sum_{i=1}^{n} (x_i - \overline{X})^2}{n} = \text{varianza}$
Momento de orden $r = 3$	$m_3 = \frac{\sum_{i=1}^{n} (x_i - \overline{X})^3}{n}$
Momento de orden $r = 4$	$m_4 = \frac{\sum_{i=1}^{n} (x_i - \overline{X})^4}{n}$

7.3.2 Asimetría y apuntamiento

La medida de la forma de la distribución de una variable se hace respecto a la Normal. Los estadísticos son *asimetría* (g_1) y *apuntamiento* (g_2). En una distribución normal o *campana de Gauss,* los estadísticos de tendencia central (M_o, M_e y $\overline{X}$) tienen el mismo o similar valor y es el eje que divide la distribución en dos partes iguales y simétricas. En los valores extremos de la variable se dan frecuencias bajas y éstas aumentan a medida que los valores se acercan a los valores medios de la misma y g_1 y g_2 toman el valor cero (Gráfico 5). Pero no significa que una distribución que tenga g_1 y g_2 igual a cero, sea normal. Lo que se pretende es comparar la forma de una distribución con la normal, no contrastar si la distribución es normal o marcadamente normal.

Gráfico 5 La normal.

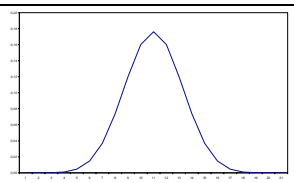
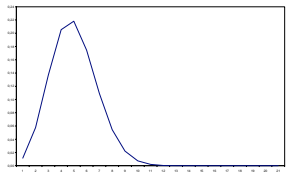
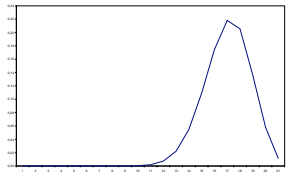


$M_o = M_e = \overline{X}$

Valores extremos -> frecuencias bajas.
Valores medios -> frecuencias altas.

$g_1 = 0$
$g_2 = 0$

La normal no presenta asimetría ni apuntamiento.

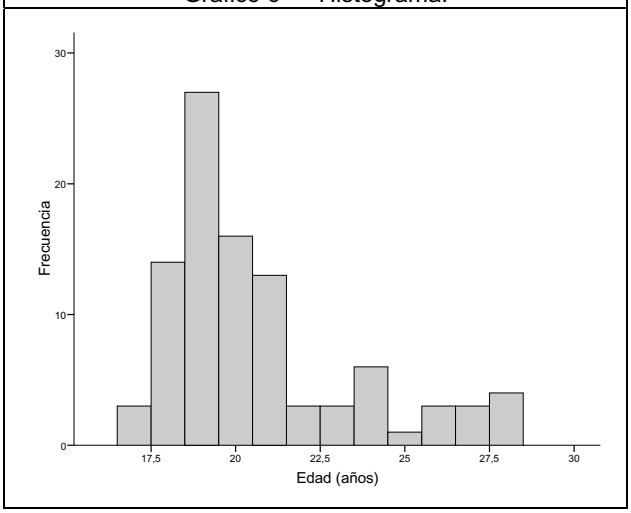
Fórmula 28 Asimetría o g_1	
Datos Tipo I	
$g_1 = \frac{m_3}{S^3} = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^3}{\frac{n}{S^3}} = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^3}{nS^3}$	
Datos Tipo II	
$g_1 = \frac{m_3}{S^3} = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^3 n_i}{S^3 \sum_{i=1}^{n} n_i}$	
Datos Tipo III	
$g_1 = \frac{m_3}{S^3} = \frac{\sum_{i=1}^{n}(x_i' - \bar{X})^3 n_i}{S^3 \sum_{i=1}^{n} n_i}$	
Interpretación:	
$g_1 = 0$ no presenta asimetría	
$g_1 > 0$ presenta asimetría positiva u oblicua a la derecha. La asimetría se presenta hacia los valores altos de la variable, por lo tanto los casos se concentran hacia los valores bajos de la variable.	
$g_1 < 0$ presenta asimetría negativa u oblicua a la izquierda. La asimetría se presenta hacia los valores bajos de la variable, por lo tanto se concentran los casos hacia los valores altos de la variable.	

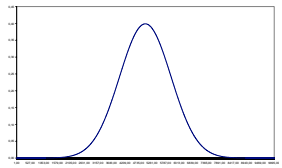
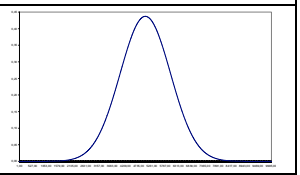
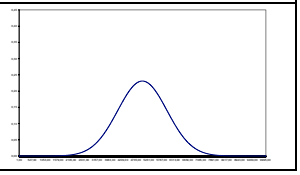
Ejemplo: Cálculo de la asimetría de la variable *p4_3* (edad) de la matriz de datos de la Tabla 16. Se recomienda seguir el proceso de cálculo de la Tabla 23.

Tabla 23 Cálculo de la asimetría de la variable edad.

Caso	Edad	$(x_i - \bar{X})$	$(x_i - \bar{X})^2$	$(x_i - \bar{X})^3$	Caso	Edad	$(x_i - \bar{X})$	$(x_i - \bar{X})^2$	$(x_i - \bar{X})^3$
1	21	-0,33	0,11	-0,04	55	21	-0,33	0,11	-0,04
2	21	-0,33	0,11	-0,04	56	20	0,67	0,44	0,30
3	23	-2,33	5,44	-12,70	57	18	2,67	7,11	18,96
4	19	1,67	2,78	4,63	58	19	1,67	2,78	4,63
5	24	-3,33	11,11	-37,04	59	21	-0,33	0,11	-0,04
6	24	-3,33	11,11	-37,04	60	21	-0,33	0,11	-0,04
7	19	1,67	2,78	4,63	61	20	0,67	0,44	0,30
8	20	0,67	0,44	0,30	62	18	2,67	7,11	18,96
9	22	-1,33	1,78	-2,37	63	26	-5,33	28,44	-151,70
10	18	2,67	7,11	18,96	64	19	1,67	2,78	4,63
11	20	0,67	0,44	0,30	65	18	2,67	7,11	18,96
12	19	1,67	2,78	4,63	66	28	-7,33	53,78	-394,37
13	19	1,67	2,78	4,63	67	21	-0,33	0,11	-0,04
14	18	2,67	7,11	18,96	68	25	-4,33	18,78	-81,37
15	.				69	23	-2,33	5,44	-12,70
16	17	3,67	13,44	49,30	70	19	1,67	2,78	4,63
17	27	-6,33	40,11	-254,04	71	24	-3,33	11,11	-37,04
18	20	0,67	0,44	0,30	72	24	-3,33	11,11	-37,04
19	19	1,67	2,78	4,63	73	19	1,67	2,78	4,63
20	19	1,67	2,78	4,63	74	20	0,67	0,44	0,30
21	19	1,67	2,78	4,63	75	22	-1,33	1,78	-2,37
22	21	-0,33	0,11	-0,04	76	18	2,67	7,11	18,96
23	21	-0,33	0,11	-0,04	77	20	0,67	0,44	0,30
24	18	2,67	7,11	18,96	78	19	1,67	2,78	4,63
25	19	1,67	2,78	4,63	79	19	1,67	2,78	4,63
26	19	1,67	2,78	4,63	80	18	2,67	7,11	18,96
27	21	-0,33	0,11	-0,04	81	.			
28	20	0,67	0,44	0,30	82	17	3,67	13,44	49,30
29	20	0,67	0,44	0,30	83	27	-6,33	40,11	-254,04
30	26	-5,33	28,44	-151,70	84	18	2,67	7,11	18,96
31	19	1,67	2,78	4,63	85	19	1,67	2,78	4,63
32	18	2,67	7,11	18,96	86	20	0,67	0,44	0,30
33	28	-7,33	53,78	-394,37	87	19	1,67	2,78	4,63
34	21	-0,33	0,11	-0,04	88	21	-0,33	0,11	-0,04
35	21	-0,33	0,11	-0,04	89	20	0,67	0,44	0,30
36	23	-2,33	5,44	-12,70	90	18	2,67	7,11	18,96
37	19	1,67	2,78	4,63	91	19	1,67	2,78	4,63
38	24	-3,33	11,11	-37,04	92	20	0,67	0,44	0,30
39	24	-3,33	11,11	-37,04	93	21	-0,33	0,11	-0,04
40	19	1,67	2,78	4,63	94	20	0,67	0,44	0,30
41	20	0,67	0,44	0,30	95	28	-7,33	53,78	-394,37
42	22	-1,33	1,78	-2,37	96	26	-5,33	28,44	-151,70
43	18	2,67	7,11	18,96	97	19	1,67	2,78	4,63
44	20	0,67	0,44	0,30	98	18	2,67	7,11	18,96
45	19	1,67	2,78	4,63	99	28	-7,33	53,78	-394,37
46	19	1,67	2,78	4,63					
47	18	2,67	7,11	18,96	Suma	1.984		751,33	2.600,89
48	.				n	96			
49	17	3,67	13,44	49,30	$\bar{X}$	20,67			
50	27	-6,33	40,11	-254,04	S^2			7,83	
51	20	0,67	0,44	0,30	S			2,80	
52	19	1,67	2,78	4,63	S^3			21,89	
53	19	1,67	2,78	4,63	$n\Delta S^3$	2.101,91			
54	19	1,67	2,78	4,63	g_1				1,24

Gráfico 6 Histograma.



Fórmula 29 Apuntamiento ó g_2	
Datos Tipo I	
$g_2 = \frac{m_4}{S^4} - 3 = \frac{\frac{\sum_{i=1}^{n}(x_i - \bar{X})^4}{n}}{S^4} - 3 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^4}{nS^4} - 3$	
Datos Tipo II	
$g_2 = \frac{m_4}{S^4} - 3 = \frac{\frac{\sum_{i=1}^{n}(x_i - \bar{X})^4 n_i}{n}}{S^4 - \sum_{i=1}^{n} n_i} - 3$	
Datos Tipo III	
$g_2 = \frac{m_4}{S^4} = \frac{\frac{\sum_{i=1}^{n}(x_i' - \bar{X})^4 n_i}{n}}{S^4 - \sum_{i=1}^{n} n_i} - 3$	
Como el estadístico de apuntamiento g_2 tiene un valor de 3 en la distribución normal, se le resta esta cantidad para que presente el valor 0.	
Interpretación:	
$g_2 = 0$ no presenta apuntamiento (mesocúrtica).	
$g_2 > 0$ presenta apuntamiento positivo (leptocúrtica). La curva es más apuntada que una supuesta normal. Los casos tienden a estar más concentrados.	
$g_2 < 0$ presenta apuntamiento negativo (platicúrtica). La curva es más plana que una supuesta normal. Los casos tienden a estar más dispersos.	

Ejemplo: Cálculo de la asimetría de la variable *p4_3* (edad) de la matriz de datos de la Tabla 16. Se recomienda seguir el proceso de cálculo de la Tabla 24.

Tabla 24 Cálculo del apuntamiento.

Caso	Edad	$(x_i - \overline{X})$	$(x_i - \overline{X})^2$	$(x_i - \overline{X})^4$	Caso	Edad	$(x_i - \overline{X})$	$(x_i - \overline{X})^2$	$(x_i - \overline{X})^4$
1	21	-0,33	0,11	0,01	55	21	-0,33	0,11	7,72
2	21	-0,33	0,11	0,01	56	20	0,67	0,44	0,01
3	23	-2,33	5,44	29,64	57	18	2,67	7,11	0,20
4	19	1,67	2,78	7,72	58	19	1,67	2,78	50,57
5	24	-3,33	11,11	123,46	59	21	-0,33	0,11	7,72
6	24	-3,33	11,11	123,46	60	21	-0,33	0,11	0,01
7	19	1,67	2,78	7,72	61	20	0,67	0,44	0,01
8	20	0,67	0,44	0,20	62	18	2,67	7,11	0,20
9	22	-1,33	1,78	3,16	63	26	-5,33	28,44	50,57
10	18	2,67	7,11	50,57	64	19	1,67	2,78	809,09
11	20	0,67	0,44	0,20	65	18	2,67	7,11	7,72
12	19	1,67	2,78	7,72	66	28	-7,33	53,78	50,57
13	19	1,67	2,78	7,72	67	21	-0,33	0,11	2.892,05
14	18	2,67	7,11	50,57	68	25	-4,33	18,78	0,01
15	.				69	23	-2,33	5,44	352,60
16	17	3,67	13,44	180,75	70	19	1,67	2,78	29,64
17	27	-6,33	40,11	1.608,90	71	24	-3,33	11,11	7,72
18	20	0,67	0,44	0,20	72	24	-3,33	11,11	123,46
19	19	1,67	2,78	7,72	73	19	1,67	2,78	123,46
20	19	1,67	2,78	7,72	74	20	0,67	0,44	7,72
21	19	1,67	2,78	7,72	75	22	-1,33	1,78	0,20
22	21	-0,33	0,11	0,01	76	18	2,67	7,11	3,16
23	21	-0,33	0,11	0,01	77	20	0,67	0,44	50,57
24	18	2,67	7,11	50,57	78	19	1,67	2,78	0,20
25	19	1,67	2,78	7,72	79	19	1,67	2,78	7,72
26	19	1,67	2,78	7,72	80	18	2,67	7,11	7,72
27	21	-0,33	0,11	0,01	81	.			50,57
28	20	0,67	0,44	0,20	82	17	3,67	13,44	
29	20	0,67	0,44	0,20	83	27	-6,33	40,11	180,75
30	26	-5,33	28,44	809,09	84	18	2,67	7,11	1.608,90
31	19	1,67	2,78	7,72	85	19	1,67	2,78	50,57
32	18	2,67	7,11	50,57	86	20	0,67	0,44	7,72
33	28	-7,33	53,78	2.892,05	87	19	1,67	2,78	0,20
34	21	-0,33	0,11	0,01	88	21	-0,33	0,11	7,72
35	21	-0,33	0,11	0,01	89	20	0,67	0,44	0,01
36	23	-2,33	5,44	29,64	90	18	2,67	7,11	0,20
37	19	1,67	2,78	7,72	91	19	1,67	2,78	50,57
38	24	-3,33	11,11	123,46	92	20	0,67	0,44	7,72
39	24	-3,33	11,11	123,46	93	21	-0,33	0,11	0,20
40	19	1,67	2,78	7,72	94	20	0,67	0,44	0,01
41	20	0,67	0,44	0,20	95	28	-7,33	53,78	0,20
42	22	-1,33	1,78	3,16	96	26	-5,33	28,44	2.892,05
43	18	2,67	7,11	50,57	97	19	1,67	2,78	809,09
44	20	0,67	0,44	0,20	98	18	2,67	7,11	7,72
45	19	1,67	2,78	7,72	99	28	-7,33	53,78	50,57
46	19	1,67	2,78	7,72					
47	18	2,67	7,11	50,57	Suma	1.984		751,33	21.475,78
48	.				n	96			
49	17	3,67	13,44	180,75	$\overline{X}$	20,67			
50	27	-6,33	40,11	1.608,90	S^2			7,83	
51	20	0,67	0,44	0,20	S			2,80	
52	19	1,67	2,78	7,72	S^4			61,25	
53	19	1,67	2,78	7,72	$n\Delta S^4$	5.880,23			
54	19	1,67	2,78	0,01	g_2				0,65

Gráfico 7 Histograma.

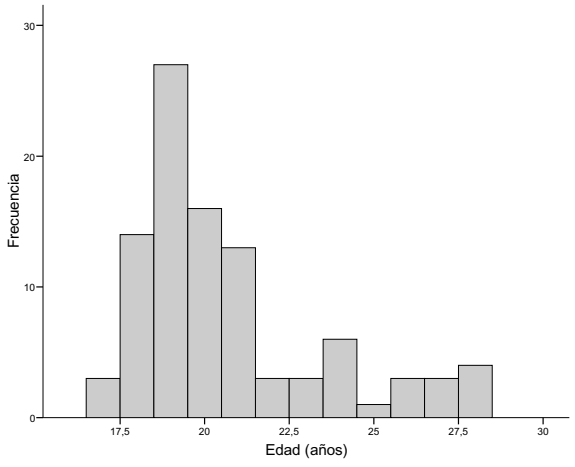


Tabla 25 Resumen.	
Estadístico	Edad
V_M	28
V_m	17
$\overline{X}$	20,67
S^2	7,83
S	2,80
CV	0,14
g_1	1,24
g_2	0,65

Tabla 26 Recomendaciones de lectura de la estadística descriptiva (orientativo).		
1. Para la presentación en los informes, se recomienda utilizar: en los porcentajes 1 decimal (redondeo por defecto o exceso con un error menor que 0,1); en los demás estadísticos 2 decimales (redondeo por defecto o exceso con un error menor que 0,01), y en las probabilidades, al presentarse también como porcentajes se deben conocer cuatro decimales.		
Ejemplos: Para porcentajes: 2,45 a 2,49 ♥ 2,5 2,40 a 2,44 ♥ 2,4	Resto de estadísticos: 2,545 a 2,549 ♥ 2,55 2,540 a 2,544 ♥ 2,54	Probabilidades: 0,54545 a 0,54549 ♥ 0,5455 ♥ 54,6% 0,54540 a 0,54544 ♥ 0,5454 ♥ 54,5%
2. Cuando se hacen referencias a Censos, se pueden expresar valores absolutos de población, valores relativos y otros estadísticos.		
Ejemplos: La población de los jóvenes (18 a 29 años) es de 15.486.387 habitantes, el 17,5% de la población total.		
3. Siempre que se hable de muestras o conjuntos que no sean Censos no se deben reflejar los valores absolutos y sólo se deben poner valores relativos u otros estadísticos (a). Una excepción puede ser cuando un porcentaje muy alto haga referencia a un valor absoluto muy bajo y entonces se puede reflejar el valor absoluto entre paréntesis (b).		
Ejemplos: (a) Según la Encuesta realizada, la población de estudiantes universitarios sigue siendo minoritaria (17,0%), aunque ... (b) Los clientes de Estrella Polar que tienen cuenta corriente son el 90% (n = 10).		
4. Siempre se debe poner el valor numérico en el texto, por relación directa o por relación indirecta entre paréntesis.		
Ejemplos: Es una referencia indirecta la edad de los jóvenes en el punto 2 (18 a 29 años) y el % de estudiantes universitarios (17,0%) en el punto 3a. Es una referencia directa el % de usuarios de cuenta corriente del punto 3b (90%) y en este caso es una referencia indirecta.		
5. La lectura se puede hacer en tres niveles diferentes: valor numérico, concepto estadístico y concepto coloquial culto.		
Ejemplos de lectura sobre la Tabla 25:[58] Valor numérico: El *CV* de la variable edad es 0,14, con una media de 20,67 años, la asimetría -1,24 y el apuntamiento es 0,65. (Es un comentario aséptico que no aporta nada de información, sólo dice lo que se ve). Concepto estadístico: Los individuos presentan baja dispersión (*CV* = 0,14) respecto de la media de edad (20,67 años). Su distribución está sesgada a la derecha (g_1 = 1,24) con un cierto apuntamiento (g_2 = 0,65). (Es un comentario que aporta información técnica, por lo tanto sólo sería útil para personas técnicas). Nivel coloquial culto o aclaratorio para la mayoría de las personas: La edad media de los individuos es de 20,67 años con un CV = 0,14, que por su proximidad a cero, es indicativo de que los casos tienen poca dispersión. La concentración de los casos tiende hacia los valores bajos de la edad, como indica el coeficiente positivo de asimetría (1,24), con cierta concentración, como indica el apuntamiento positivo (0,65). La estadística descriptiva es nada más, pero nada menos, que descriptiva, por lo que no se puede buscar una estadística de precisión.		

[58] Una lectura puede ser empezar por el *CV* que indica la dispersión, continuar con la media y después hablar de la forma. La lectura está fuera de contexto, quiere decir que en un informe hay que dar las referencias necesarias de tiempo, espacio y población de referencia.

7.4 Tabla de frecuencias.

La *tabla de frecuencias* o *distribución de frecuencias*, es apropiada para variables categóricas y numéricas discretas, y numéricas continuas cuando las categorías se presentan por intervalos, aunque se aplica preferentemente a las categóricas. Es un resumen de la variable de tal manera que presenta de forma ordenada, normalmente de menor a mayor, las categorías o valores distintos de la variable, indicando para cada uno de ellos cuantas veces se repite, o lo que es lo mismo, cuantos casos hay en cada categoría o que tienen un determinado valor, característica o atributo. La forma de presentación es según la Tabla 27.

Tabla 27 Tabla de frecuencias o distribución de frecuencias.

X	n	N	f	F	X: Es la variable.
x_1	n_1	N_1	f_1	F_1	x_i: Valor de la variable en la categoría *i-ésima*.
x_2	n_2	N_2	f_2	F_2	n_i: Frecuencia absoluta en la categoría *i-ésima*.
x_3	n_3	N_3	f_3	F_3	N_i: Frecuencia absoluta acumulada hasta la categoría *i-ésima*.
…	…	…	…	…	f_i: Frecuencia relativa expresada en base 1 o 100.
x_{n-1}	n_{n-1}	N_{n-1}	f_{n-1}	F_{n-1}	F_i: Frecuencia relativa acumulada expresada en base 1 o 100.
x_n	n_n	N_n	f_n	F_n	N: Total de casos.
Total	N		100		

n_i |	Número de casos en la categoría o valor x_i o veces que se repite la categoría o valor x_i o número de casos que tienen la categoría o valor x_i.
$N_i \mid \sum_{j\mid 1}^{i} n_j$ ♥	Total de casos acumulados hasta la categoría o valor *i-ésimo*.
$f_i \mid \frac{n_i}{N}$	Proporción de casos en la categoría *i-ésima*.
$f_i \mid \frac{n_i}{N}\Delta 100$	Porcentaje de casos en la categoría *i-ésima*
$F_i \mid \frac{N_i}{N}$	Proporción acumulada de casos hasta la categoría *i-ésima*.
$F_i \mid \frac{N_i}{N}\Delta 100$	Porcentaje acumulado de casos hasta la categoría *i-ésima*

Ejemplo: Tabla de frecuencias de la variable *p2* (estado civil) (Tabla 28), de la matriz de datos de la Tabla 16. Esta variable tiene un espacio muestral de seis categorías, características o sucesos elementales: *Soltero/a, Casado/a, Pareja, Separado/a, Divorciado/a* y *Viudo/a*. Como las unidades de observación que han participado son jóvenes, el espacio muestral queda reducido a: *Soltero/a, Casado/a,* y *Pareja*.

Tabla 28 Cálculo de tabla de frecuencias.

i	*Atributo*	*X*	*n*	*N*	*f (%)*	*F (%)*
1	Soltero/a	1	77	77	77,8	77,8
2	Casado/a	2	9	86	9,1	86,9
3	Pareja	3	13	99	13,1	100,0
	Total		99			

Frecuencias absolutas (n)

$n_{i|1} \mid 77$ Se obtiene por recuento de los casos que tienen la característica o atributo de *soltero/a,* que es el código 1 en la matriz de datos de la Tabla 16.

$n_{i|2} \mid 9$ Se obtiene por recuento de los casos que tienen la característica o atributo de *casado/a,* que es el código 2.

$n_{i|3} \mid 13$ Se obtiene por recuento de los casos que tienen la característica o atributo de *pareja,* que es el código 3.

Frecuencias absolutas acumuladas (N)

$$N_{i|1} \mid \sum_{j|1}^{1} n_j \mid n_1 \mid 77$$

$$N_{i|2} \mid \sum_{j|1}^{2} n_j \mid n_1 2\, n_2 \mid 77 2\, 9 \mid 86$$

$$N_{i|3} \mid \sum_{j|1}^{3} n_j \mid n_1 2\, n_2 2\, n_3 \mid 77 2\, 9 2\, 13 \mid 99$$

Frecuencias relativas en % (f)

$$f_{i|1} \mid \frac{n_1}{N}\Delta 100 \mid \frac{77}{99}\Delta 100 \mid 77{,}8\%$$

$$f_{i|2} \mid \frac{n_2}{N}\Delta 100 \mid \frac{9}{99}\Delta 100 \mid 9{,}1\%$$

$$f_{i|3} \mid \frac{n_3}{N}\Delta 100 \mid \frac{13}{99}\Delta 100 \mid 13{,}1\%$$

Frecuencias relativas acumuladas en % (F)

$$F_{i|1} \mid \frac{N_1}{N}\Delta 100 \mid \frac{77}{99}\Delta 100 \mid 77{,}8\%$$

$$F_{i|2} \mid \frac{N_2}{N}\Delta 100 \mid \frac{86}{99}\Delta 100 \mid 86{,}9\%$$

$$F_{i|3} \mid \frac{N_3}{N}\Delta 100 \mid \frac{99}{99}\Delta 100 \mid 100{,}0\%$$

Las frecuencias acumuladas absoluta y relativa, tienen más sentido cuando se aplica con variables que al menos tienen nivel de medida ordinal. La denominación de *distribución de frecuencias* se debe a que la suma de los porcentajes es 100, o sea, la suma de todas las frecuencias absolutas, coincide con el total de la tabla.

Ejemplo: Tabla de frecuencias (Tabla 29) de la variable *p4_1* (peso), de la matriz de datos de la Tabla 16. Esta variable tiene un espacio muestral de 20 sucesos elementales: 45, 50, 52, 53, 55, 58, 60, 63, 65, 66, 67, 68, 70, 73, 75, 76, 77, 78, 80 y 85.

Tabla 29 Tabla de frecuencias de la variable *p4_1* (peso).

i	*X*	*n*	*N*	*f (%)*	*F (%)*
1	45	3	3	3,2	3,2
2	50	2	5	2,1	5,3
3	52	7	12	7,4	12,6
4	53	3	15	3,2	15,8
5	55	8	23	8,4	24,2
6	58	5	28	5,3	29,5
7	60	6	34	6,3	35,8
8	63	5	39	5,3	41,1
9	65	5	44	5,3	46,3
10	66	4	48	4,2	50,5
11	67	2	50	2,1	52,6
12	68	3	53	3,2	55,8
13	70	8	61	8,4	64,2
14	73	7	68	7,4	71,6
15	75	8	76	8,4	80,0
16	76	4	80	4,2	84,2
17	77	5	85	5,3	89,5
18	78	4	89	4,2	93,7
19	80	3	92	3,2	96,8
20	85	3	95	3,2	100,0
	Total	95		100,0	

$n_{i|1} | 3$ Se obtiene por recuento de los casos que tienen el peso de 45 kg. en la matriz de datos de la Tabla 16.

$n_{i|2} | 2$ Se obtiene por recuento de los casos que tienen el peso de 50 kg.

$n_{i|3} | 7$ Se obtiene por recuento de los casos que tienen el peso de 52 kg.

$n_{i|19} | 3$, Se obtiene por recuento de los casos que tienen el peso de 80 kg.

$n_{i|20} | 3$ Se obtiene por recuento de los casos que tienen el peso de 85 kg.

$$N_{i|1} | \overset{1}{\underset{j|1}{\text{——}}} n_j | n_1 | 3$$

$$N_{i|2} | \overset{2}{\underset{j|1}{\text{——}}} n_j | n_1\, 2\, n_2 | 3\, 2\, 2 | 5$$

$$N_{i|3} | \overset{3}{\underset{j|1}{\text{——}}} n_j | n_1\, 2\, n_2\, 2\, n_3 | 3\, 2\, 2\, 2\, 7 | 12$$

$$N_{i|20} | \overset{20}{\underset{j|1}{\text{——}}} n_j | n_1\, 2\, n_2\, 2 \ldots 2\, n_{20} | 3\, 2\, 2\, 2 \ldots 2\, 3 | 95$$

$$f_{i|1} | \frac{n_1}{N}\Delta 100 | \frac{3}{95}\Delta 100 | 3{,}2\% \qquad f_{i|2} | \frac{n_2}{N}\Delta 100 | \frac{3}{95}\Delta 100 | 2{,}1\%$$

$$f_{i|3} | \frac{n_3}{N}\Delta 100 | \frac{7}{95}\Delta 100 | 7{,}4\% \qquad f_{i|20} | \frac{n_3}{N}\Delta 100 | \frac{3}{95}\Delta 100 | 3{,}2\%$$

$$F_{i|1} | \frac{N_1}{N}\Delta 100 | \frac{3}{95}\Delta 100 | 3{,}2\% \qquad F_{i|2} | \frac{N_2}{N}\Delta 100 | \frac{5}{95}\Delta 100 | 5{,}3\%$$

$$F_{i|3} | \frac{N_3}{N}\Delta 100 | \frac{12}{95}\Delta 100 | 12{,}6\% \qquad F_{i|20} | \frac{N_3}{N}\Delta 100 | \frac{95}{95}\Delta 100 | 100{,}0\%$$

7.4.1 Tabla de frecuencias por intervalos.

La tabla de frecuencias o distribución de frecuencias por intervalos es la representación de la tabla de datos Tipo I ó Tipo II agrupada en intervalos. Las categorías o estratos son intervalos definidos por un valor mínimo (límite inferior del intervalo) y un valor máximo (límite superior del intervalo) que suma o reúne los casos que tienen los valores o datos comprendidos dentro de cada intervalo. La amplitud del intervalo está definida por la diferencia entre el valor máximo y el mínimo y se denomina *amplitud del intervalo* (a_i), y el punto medio del intervalo se denomina *marca de clase*.

$a_i \mid v_{Mi} \; 4 \; v_{mi}$	en donde: a_i: Amplitud del intervalo *i-ésimo*. v_{Mi}: Límite superior del intervalo *i-ésimo*. v_{mi}: Límite inferior del intervalo *i-ésimo*.	Fórmula 30

$x_i^{'} \mid \frac{v_{Mi} \; 2 \; v_{mi}}{2}$	En donde: x'_i: Marca de clase del intervalo *i-ésimo*. v_{Mi}: Límite superior del intervalo *i-ésimo*. v_{mi}: Límite inferior del intervalo *i-ésimo*.	Fórmula 31

Para crear una tabla de frecuencias por intervalos hay que definir el número de intervalos y averiguar la amplitud del intervalo o definir la amplitud de los intervalos y hallar el número de intervalos. Las posibilidades se muestran en la Tabla 30.

Tabla 30 Definición de intervalos.

Intervalos de igual amplitud	Conocida la amplitud, calcular el nº de intervalos.
	Conocido el nº de intervalos, calcular la amplitud.
Intervalos de distinta amplitud	Definición de Intervalos por percentiles.
	Definición de intervalos por valores críticos.

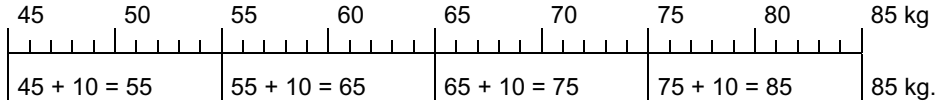
Ejemplo: Paso de la tabla de frecuencias (Tabla 31) de la variable *p4_1* (peso), de la matriz de datos de la Tabla 16, de datos Tipo II a datos Tipo III, en intervalos de igual amplitud de 10 kg. cada uno. Esta variable tiene un espacio muestral de 20 sucesos elementales: 45, 50, 52, 53, 55, 58, 60, 63, 65, 66, 67, 68, 70, 73, 75, 76, 77, 78, 80 y 85.

Tabla 31 Paso de tabla de datos T-II a T-III en intervalos de 10 kg.

$A_{peso} \mid V_M\ 4\ V_m \mid 85\ 4\ 45 \mid 40kg$

$\frac{40kg}{10kg} \mid 4$ intervalos

Límite inferior	Límite superior
45	45 + 10 = 55
55	55 + 10 = 65
65	65 + 10 = 75
75	75 + 10 = 85



	Tipo II			Tipo III			Tipo II	
i	*X*	*n*		*X*	*n*		*X'*	*n*
1	45	3		45-55	23	→	50	23
2	50	2		55-65	25	→	60	25
3	52	7		65-75	28	→	70	28
4	53	3		75-85	19	→	80	19
5	55	8						
6	58	5		Total	95			95
7	60	6						
8	63	5						
9	65	5						
10	66	4						
11	67	2						
12	68	3						
13	70	8						
14	73	7						
15	75	8						
16	76	4						
17	77	5						
18	78	4						
19	80	3						
20	85	3						
	Total	95						

$$x_1' \mid \frac{45\ 2\ 55}{2} \mid 50$$

$$x_2' \mid \frac{55\ 2\ 65}{2} \mid 60$$

$$x_3' \mid \frac{65\ 2\ 75}{2} \mid 70$$

$$x_4' \mid \frac{75\ 2\ 85}{2} \mid 80$$

Ejemplo: Paso de la tabla de frecuencias (Tabla 32) de la variable *p4_1* (peso), de la matriz de datos de la Tabla 16, de datos Tipo II a datos Tipo III, en 3 intervalos de igual amplitud. Esta variable tiene un espacio muestral de 20 sucesos elementales: 45, 50, 52, 53, 55, 58, 60, 63, 65, 66, 67, 68, 70, 73, 75, 76, 77, 78, 80 y 85.

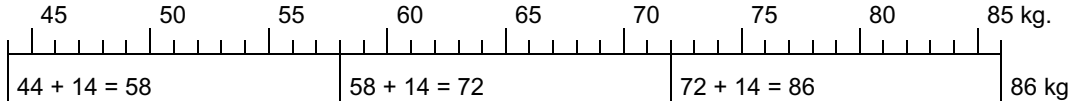
Tabla 32 Paso de tabla de datos T-II a T-III en tres intervalos.

$A_{peso} \mid V_M \; 4 \, V_m \mid 85 \, 4 \, 45 \mid 40kg$

$\frac{40kg}{3} \mid 13{,}3 \text{ kg amplitud del intervalo} \; \heartsuit \; 14\text{kg}$

Límite inferior	Límite superior
44	44+14=58
58	58+14=72
72	72+14=86

Como la amplitud del intervalo no es un número entero, se procede a redondear por exceso al entero superior: 14. Pero al multiplicar 14 por el número de intervalos, la amplitud de la variable se amplia a 42 kg. La diferencia entre la amplitud original y la actual se reparte entre los dos extremos de la variable. Si los valores de límite de intervalos de la nueva tabla de datos T-III no coincidiesen con los valores de límite de la T-II, se tendría que proceder desde la T-I para poder sumar los casos que hay en cada intervalo. En la T-II, los valores 71 y 72 no se consideran porque no existen en el espacio muestral de la variable *peso* (zona de las celdas 70 y 73, sombreadas).



	Tipo II	
i	*X*	*n*
1	45	3
2	50	2
3	52	7
4	53	3
5	55	8
6	58	5
7	60	6
8	63	5
9	65	5
10	66	4
11	67	2
12	68	3
13	70	8
14	73	7
15	75	8
16	76	4
17	77	5
18	78	4
19	80	3
20	85	3
	Total	95

Tipo III			Tipo II	
X	*n*		*X'*	*n*
44-58	28	→	51	28
58-72	33	→	65	33
72-86	34	→	79	34
Total	95			95

$x_1' \mid \frac{44 \, 2 \, 58}{2} \mid 51$

$x_2' \mid \frac{58 \, 2 \, 72}{2} \mid 65$

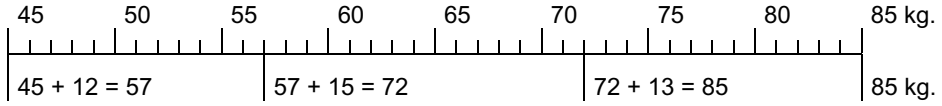
$x_3' \mid \frac{72 \, 2 \, 86}{2} \mid 79$

Ejemplo: Paso de la tabla de frecuencias (Tabla 33) de la variable *p4_1* (peso), de la matriz de datos de la Tabla 16, de datos Tipo II a datos Tipo III, en intervalos de diferente amplitud, considerando valores críticos. Al estar trabajando con la variable *p4_1*, se va a utilizar, como ejemplo, los límites del peso de las categorías de los boxeadores, sin diferenciar por sexo. Esta variable tiene un espacio muestral de 20 sucesos elementales: 45, 50, 52, 53, 55, 58, 60, 63, 65, 66, 67, 68, 70, 73, 75, 76, 77, 78, 80 y 85.

Tabla 33 Paso de tabla de datos T-II a T-III en intervalos por valores críticos.

Los límites considerados son:	Límite inferior	Límite superior
Peso Pluma: inferior a 57 kg.	45	57
Peso Ligero: inferior a 72 kg.	57	72
El resto superior a: 72 kg.	72	85

El límite superior del primer intervalo se marca 55 kg en la tabla T-II, porque en el espacio muestral de la variable no existen 56 kg y 57 kg, pero el valor se mantiene en la tabla T-III a efectos de cálculos.



	Tipo II		Tipo III			Tipo II	
i	*X*	*n*	*X*	*n*		*X'*	*n*
1	45	3	45-57	23	→	51,00	23
2	50	2	57-72	38	→	64,50	38
3	52	7	72-85	34	→	78,50	34
4	53	3					
5	55	8	Total	95			95
6	58	5					
7	60	6					
8	63	5					
9	65	5					
10	66	4					
11	67	2					
12	68	3					
13	70	8					
14	73	7					
15	75	8					
16	76	4					
17	77	5					
18	78	4					
19	80	3					
20	85	3					
	Total	95					

$$x_1' \mid \frac{45\,2\,57}{2} \mid 51{,}00$$

$$x_2' \mid \frac{57\,2\,72}{2} \mid 64{,}50$$

$$x_3' \mid \frac{72\,2\,85}{2} \mid 78{,}50$$

Ejemplo: Paso de la tabla de frecuencias de la variable *p4_1* (peso), de la matriz de datos de la Tabla 16, de datos Tipo II a datos Tipo III, en intervalos de diferente amplitud, considerando los percentiles cuartiles. Esta variable tiene un espacio muestral de 20 sucesos elementales: 45, 50, 52, 53, 55, 58, 60, 63, 65, 66, 67, 68, 70, 73, 75, 76, 77, 78, 80 y 85.

El cálculo de los percentiles cuartiles se realiza sobre la tabla de datos Tipo II de la variable *p4_1*. A efectos de cálculo, se considera que cada intervalo tiene la amplitud de una unidad y los intervalos que no aparecen es porque no hay casos. Matemáticamente es cierto, ya que en el primer cuartil, por ejemplo, los 5 casos que tienen 58 kg, en realidad se puede considerar que tienen entre 58 kg y 59 kg, pero sin llegar a tener los 59, porque entonces se habría generado esa categoría y la realidad es que no existe.

Tabla 34 Percentiles cuartiles de la variable *peso.*

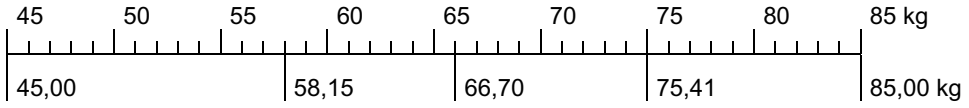
i	*X*	*n*	*N*	*Intervalo Crítico*	*Cuartil*
1	45(-46)	3	3		
2	50(-51)	2	5		
3	52(-53)	7	12		$Q_1 = P_{25}$
4	53(-54)	3	15		
5	55(-56)	8	23		
6	58(-59)	5	28	$IC_{25} \mid 25\Delta\frac{95}{100} \mid 23{,}75$	$P_{25} \mid 58\,2\,\frac{25\Delta\frac{95}{100}\,4\,23}{5}\Delta 1 \mid 58\,2\,\frac{23{,}75\,4\,23}{5}\Delta 1 \mid 58\,2\,0{,}15 \mid 58{,}15$
7	60(-61)	6	34		
8	63(-64)	5	39		$Q_2 = P_{50}$
9	65(-66)	5	44		
10	66(-67)	4	48	$IC_{50} \mid 50\Delta\frac{95}{100} \mid 47{,}50$	$P_{50} \mid 66\,2\,\frac{50\Delta\frac{95}{100}\,4\,44}{4}\Delta 1 \mid 66\,2\,\frac{47{,}50\,4\,44}{5}\Delta 1 \mid 66\,2\,0{,}70 \mid 66{,}70$
11	67(-68)	2	50		
12	68(-69)	3	53		$Q_3 = P_{75}$
13	70(-71)	8	61		
14	73(-74)	7	68		
15	75(-76)	8	76	$IC_{75} \mid 75\Delta\frac{95}{100} \mid 71{,}25$	$P_{75} \mid 75\,2\,\frac{75\Delta\frac{95}{100}\,4\,68}{8}\Delta 1 \mid 75\,2\,\frac{71{,}25\,4\,68}{8}\Delta 1 \mid 75\,2\,0{,}41 \mid 75{,}41$
16	76(-77)	4	80		
17	77(-78)	5	85		
18	78(-79)	4	89		
19	80(-81)	3	92		
20	85(-86)	3	95		
	Total	95			

Por coherencia expositiva, se mantienen los valores obtenidos de los cuartiles, pero como las unidades de observación no tienen peso con fracciones de kg, se podría haber truncado[59] a los valores enteros: $58{,}15 \blacktriangledown 58$, $66{,}70 \blacktriangledown 66$ y $75{,}41 \blacktriangledown 75$.

Tabla 35 Paso de tabla de datos T-II a T-III en intervalos por cuartiles.

Los límites son:	En donde:	Límite inferior	Límite superior
v_m - Q_1	v_m: Valor mínimo.	45,00	58,15
Q_1 - Q_2	Q_1: Cuartil 1.	58,15	66,70
Q_2 - Q_3	Q_2: Cuartil 2.	66,70	75,41
Q_3 - v_M	Q_3: Cuartil 3.	75,41	85,00
	v_M: Valor máximo.		

Los límites de los intervalos se marcan por valores enteros (58, 66 y 75), pero en la tabla Tipo III se mantiene el valor exacto a efectos de cálculo.



	Tipo II	
i	*X*	*n*
1	45	3
2	50	2
3	52	7
4	53	3
5	55	8
6	58	5
7	60	6
8	63	5
9	65	5
10	66	4
11	67	2
12	68	3
13	70	8
14	73	7
15	75	8
16	76	4
17	77	5
18	78	4
19	80	3
20	85	3
	Total	95

Tipo III		Tipo II	
X	*n*	*X'*	*n*
45,00-58,15	28	51,58	28
58,15-66,70	20	62,43	20
66,70-75,41	28	71,06	28
75,41-85,00	19	80,21	19
Total	95		95

$$x_1' = \frac{45{,}00 + 58{,}15}{2} = 51{,}58$$

$$x_2' = \frac{58{,}15 + 66{,}70}{2} = 62{,}43$$

$$x_3' = \frac{66{,}70 + 75{,}41}{2} = 71{,}06$$

$$x_4' = \frac{75{,}41 + 85{,}00}{2} = 80{,}21$$

Teóricamente, cada uno de los estratos o categorías, debe tener 23,75 casos, según los cuartiles, que son el 25,0% del total de los casos. Pero al calcularlos a partir de la tabla de datos Tipo II, la teoría difiere de la realidad. Si el cálculo se hubiese realizado sobre la tabla de datos Tipo I, la situación no habría mejorado mucho, porque las frecuencias absolutas en cada uno de los estratos o categorías, respectivamente, hubiesen sido de: 25, 22, 25 y 27 casos. Otra dificultad para la coincidencia de los valores teóricos y empíricos es que las unidades de observación no se pueden dividir como es el caso del valor 23,75.

Un problema no tratado hasta ahora es el de los límites de los intervalos, esto es, en qué intervalo se deben considerar a aquellos casos que se encuentran justo en los límites de los intervalos. Cuando un caso coincide con el límite de un intervalo y se asigna a ese intervalo, entonces se considera que es un límite cerrado, y cuando un caso que coincide con el límite de un intervalo no es asignado a ese intervalo, entonces se considera que es un límite abierto. Hasta ahora se ha considerado el criterio de SPSS, que consiste en asignar cada caso

[59] Truncar es quitar la parte decimal y dejar la parte entera, mientras que redondear implica redondeo por defecto o por exceso.

al intervalo en el que aparece primero su valor, procediendo de arriba a bajo de la tabla.

La tabla puede estar ordenada de forma ascendente o descendente (Tabla 36). Si está ordenada ascendentemente, que es el caso seguido hasta ahora, entonces el primer intervalo tiene los dos límites cerrados y el resto de los intervalos, el límite inferior es abierto y el superior es cerrado. Si la tabla está en orden descendente, al seguir el mismo criterio, entonces el primer intervalo tiene los dos límites cerrados y el resto de los intervalos, el límite inferior es cerrado y el superior es abierto, pero considerándolo sobre la tabla ordenada ascendentemente, el resultado es que el último intervalo tiene los dos límites cerrados y el resto de intervalos tienen el límite inferior cerrado y el superior abierto, que es la opción tradicional de la Estadística. Procediendo sobre el ejemplo de la Tabla 31

Tabla 36 Definición de los límites en la tabla de frecuencias.

Tabla ordenada ascendentemente		
	X	
(45) Límite inferior cerrado	45-55	(55) Límite superior cerrado
(55) Límite inferior abierto	55-65	(65) Límite superior cerrado
(65) Límite inferior abierto	65-75	(75) Límite superior cerrado
(75) Límite inferior abierto	75-85	(85) Límite superior cerrado
Tabla ordenada descendentemente		
	X	
(75) Límite inferior cerrado	75-85	(85) Límite superior cerrado
(65) Límite inferior cerrado	65-75	(75) Límite superior abierto
(55) Límite inferior cerrado	55-65	(65) Límite superior abierto
(45) Límite inferior cerrado	45-55	(55) Límite superior abierto

Tabla 37 Asignación de casos por intervalos según el orden de la tabla de frecuencias.

	Tipo II						Tipo II	
i	*X*	*n*	*X*	*n*		*X'*	*n*
1	45	3	45-55	23	→	50	23
2	50	2	55-65	25	→	60	25
3	52	7	65-75	28	→	70	28
4	53	3	75-85	19	→	80	19
5	55	8					
6	58	5	Total	95			95
7	60	6					
8	63	5					
9	65	5					
10	66	4					
11	67	2					
12	68	3					
13	70	8					
14	73	7					
15	75	8					
16	76	4	*X*	*n*		*X'*	*n*
17	77	5	75-85	27	⇢	80	27
18	78	4	65-75	29	⇢	70	29
19	80	3	55-65	24	⇢	60	24
20	85	3	45-55	15	⇢	50	15
	Total	95	Total	95			95

7.5 Percentiles

Los percentiles se pueden considerar un estadístico de tendencia central, pero se ha optado presentarlos aparte. El *percentil* es un valor de la variable que deja por debajo de sí un determinado porcentaje de casos, por lo tanto, el complemento a 100% es el porcentaje de casos que quedará por encima del mencionado valor. Entonces, el *percentil k*, es el valor *x* de la variable que deja por debajo de sí el *k%* de los casos, y por encima de *x*, deja el (100 - *k*) % de los casos. Esta cuestión plantea proponer un convenio, o que en el valor *x* de la variable no existen casos o que el valor de *x* está contemplado como limite abierto en un intervalo y cerrado en el complementario. En las variables numéricas supuestas continuas y con las integrales definidas está resuelto matemáticamente, ya que la integral entre un valor y él mismo es igual a cero.

La *mediana* es un percentil tipo, ya que es el valor de la variable que deja por debajo y por encima de sí el 50% de los casos.

La fórmula de los percentiles es una derivación de la fórmula de la mediana (Fórmula 13).

Fórmula 32 Percentiles.

La tabla de frecuencias de la variable peso recodificada se ha obtenido a partir de la variable *p4_1* de la matriz de datos de la Tabla 16 y que se corresponde con la pregunta P4 del cuestionario de la Tabla 15. De las 99 entrevistas realizadas, 95 dieron respuesta válida y 4 no contestaron.

Proceso de cálculo:
1. Se ordena la tabla de frecuencias de menor a mayor.
2. Se calculan las frecuencias absolutas acumuladas.
3. Se calcula el % de casos que corresponden al percentil *k*, y el intervalo que los contiene se le denomina intervalo crítico (*IC*).
4. Entonces se procede a calcular el valor exacto del percentil.

Peso recodificada en 4 intervalos

		Frecuencia	Frecuencia acumulada
Válidos	45-55 kg	23	23
	55-65 kg	21	44
	65-75 kg	32	76
	75-85 kg	19	95
	Total	95	
Perdidos	Sistema	4	
Total		99	

$$M_e \mid L_{i41} 2 \frac{\frac{N}{2} 4 N_{i41}}{n_i} \Delta a_i$$, pero como N/2 es el 50% de N:

$$para\ k \mid 50 \heartsuit 50 \Delta \frac{N}{100} \mid \frac{N}{2}$$, entonces:

$$M_e \mid L_{i41} 2 \frac{50 \Delta \frac{N}{100} 4 N_{i41}}{n_i} \Delta a_i$$, y generalizando:

$$P_k \mid L_{i41} 2 \frac{k \Delta \frac{N}{100} 4 N_{i41}}{n_i} \Delta a_i$$

En donde:
M_e: Mediana.
N/2: La mitad de los casos.
50 x N/100: La mitad de los casos.
k x N/100: El *k%* de los casos.
P_k: Percentil *k*.
L_{i-1}: Límite inferior del *IC*.
N_{i-1}: Total de casos por debajo del *IC*.
a_i: Amplitud del *IC*.
n_i: Frecuencia absoluta del *IC*.

Ejemplo: Calcular el P_{45} (Percentil 45). En la tabla de frecuencias el intervalo que tiene el 45% de los casos (45 x 95/100 = 42,75) acumulados, es el intervalo de "55 a 65 Kg." y es el intervalo crítico.

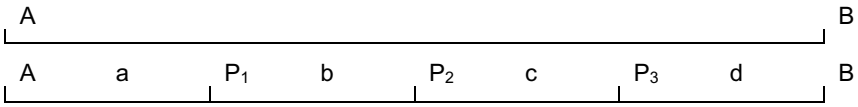
$$P_{k \mid 45} \mid 55 2 \frac{45 \frac{95}{100} 4 23}{21} \Delta 10 \mid 55 2 \frac{42{,}75 4 23}{21} \Delta 10 \mid 55 2 0{,}94 \Delta 10 \mid 55 2 9{,}4 \mid 64{,}40 kg$$

Lectura: El valor de la variable *peso* que deja por debajo de sí el 45 % de los casos es 64,40 kg.

Los percentiles denominados tipo o típicos son los cuartiles, deciles y centiles. Antes de dar su definición, diremos que un segmento se divide en tantas partes como puntos de corte

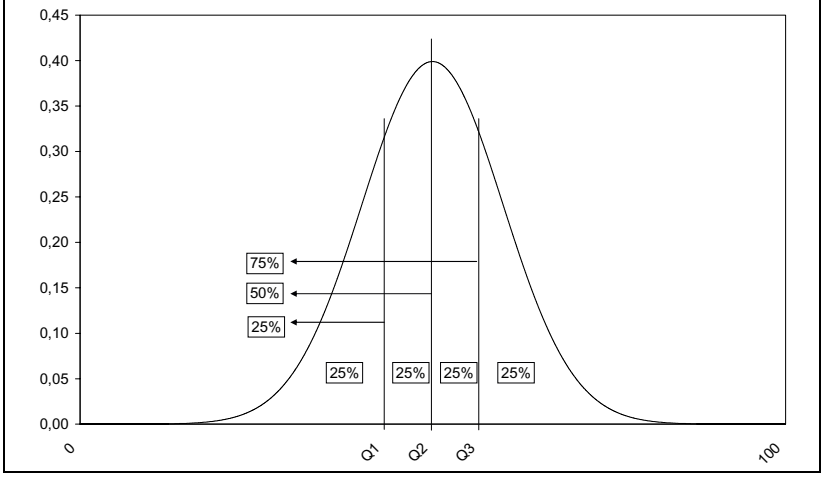
tiene más uno. Por ejemplo si a un segmento le damos tres cortes, se divide en cuatro partes. Si al segmento AB de la Tabla 38 le damos tres puntos de corte: P1, P2 y P3, entonces lo dividimos en cuatro partes: a, b, c y d.

Tabla 38 División de un segmento



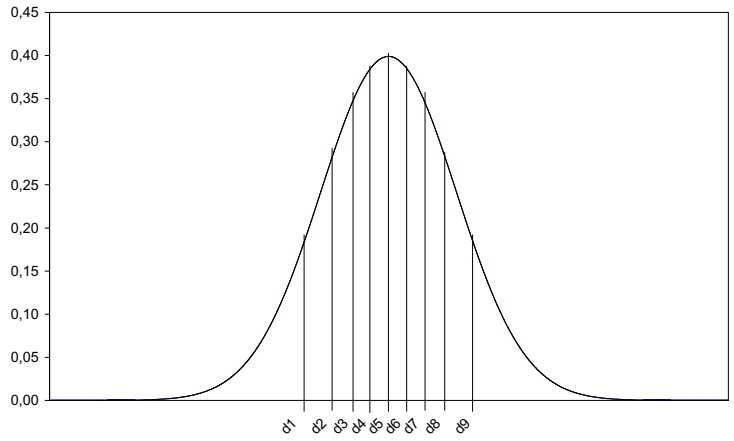
Entonces definimos a los *percentiles cuartiles* como los tres puntos de corte (Q_1, Q_2 y Q_3) que divide a la variable en cuatro partes iguales en cuanto al número de casos se refiere y cada una de ellas tiene el 25% de los casos. Por debajo de Q_1 quedan el 25% de los casos. Entre el Q_1 y el Q_2 hay otro 25% de los casos. Entre el Q_2 y el Q_3 otro 25% de los casos. Y por encima del Q_3 se encuentran el restante 25% de los casos. Por lo tanto por debajo del Q_2 están el 50% de los casos, que es la mediana. Por debajo del Q_3 el 75% de los casos y por debajo del valor máximo de la variable estarían el total de los casos (100%) (Gráfico 8).

Gráfico 8 Representación gráfica de los cuartiles.



Los *percentiles deciles* son los nueve puntos de corte (d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8 y d_9) que divide a la variable en diez partes iguales en cuanto al número de casos se refiere y cada una de ellas tiene el 10% de los casos. Por debajo de d_1 quedan el 10% de los casos. Entre el d_1 y el d_2 hay otro 10% de los casos. Entre el d_2 y el d_3 el 10% de los casos, y así sucesivamente. Y por encima del d_9 se encuentran el último 10% de los casos. Por lo tanto por debajo del d_2 están el 20% de los casos, por debajo del d_3 están el 30% de los casos. Así sucesivamente hasta el d_5 que son el 50%, que es la mediana. Por debajo del valor máximo de la variable estarían el total de los casos (100%). El Gráfico 9 es la representación gráfica de los deciles.

Gráfico 9 Representación gráfica de los deciles.



Los *percentiles centiles* son los 99 puntos de corte (c_1, c_2, c_3, … c_{50}, … c_{97}, c_{98} y c_{99}) que divide a la variable en 100 partes iguales en cuanto al número de casos se refiere y cada una de ellas tiene el 1% de los casos. Por debajo de c_1 queda el 1% de los casos. Entre el c_1 y el c_2 hay otro 1% de los casos. Entre el c_2 y el c_3 el 1% de los casos, y así sucesivamente. Y por encima del c_{99} se encuentran el último 1% de los casos. Por lo tanto por debajo del c_2 están el 2% de los casos, por debajo del c_3 están el 3% de los casos. Así sucesivamente hasta el c_{50} que son el 50%, que es la mediana. Por debajo del valor máximo de la variable estarían el total de los casos (100%).

De la misma manera, la variable se podría dividir en: 5, 6, 7, 8, … partes iguales.

7.6 Gráficos

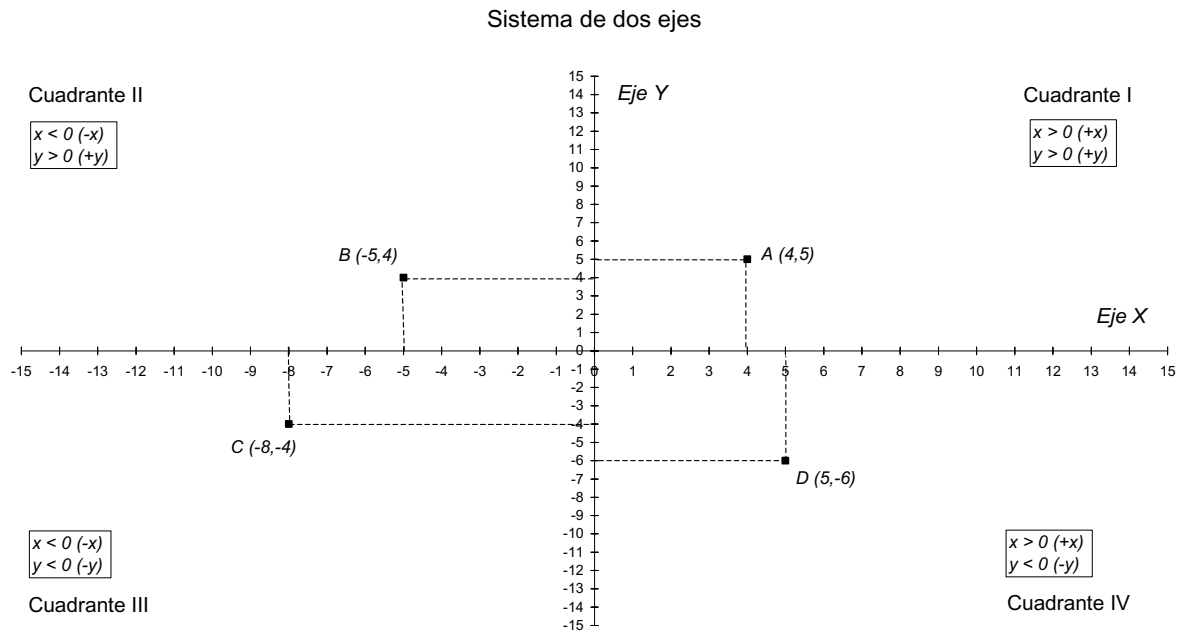
La representación gráfica de los datos se hace con el *diagrama de barras* y el *histograma,* y el *polígono de frecuencias* como derivación del histograma. Son los gráficos que se pueden considerar básicos de la Estadística. No obstante, cierto software, como *SPSS*, *Excel*, *Harvard Graphics*, MathCAD, Matlab, así como otros programas estadísticos, hojas de cálculo y matemáticos, pueden facilitar la creación de otro tipo de gráficos.

7.6.1 Introducción a los sistemas de representación gráfica

Para la representación gráfica de los datos se consideran sistemas de coordenadas cartesianas de dos y tres dimensiones.

El sistema de coordenadas cartesianas de dos dimensiones es un sistema de coordenadas de dos ejes ortogonales (perpendiculares entre sí) que dividen el plano en cuatro partes que llamamos cuadrantes: *I, II, III* y *IV*. El eje horizontal es el de *abscisas* o eje *X* y el eje vertical es el de *ordenadas* o eje *Y*. El punto en el que se cruzan se dice que tiene coordenadas (*x,y)* (0,0) y se le considera el origen del sistema (Gráfico 10).

Gráfico 10 Sistema de coordenadas cartesianas de dos dimensiones.

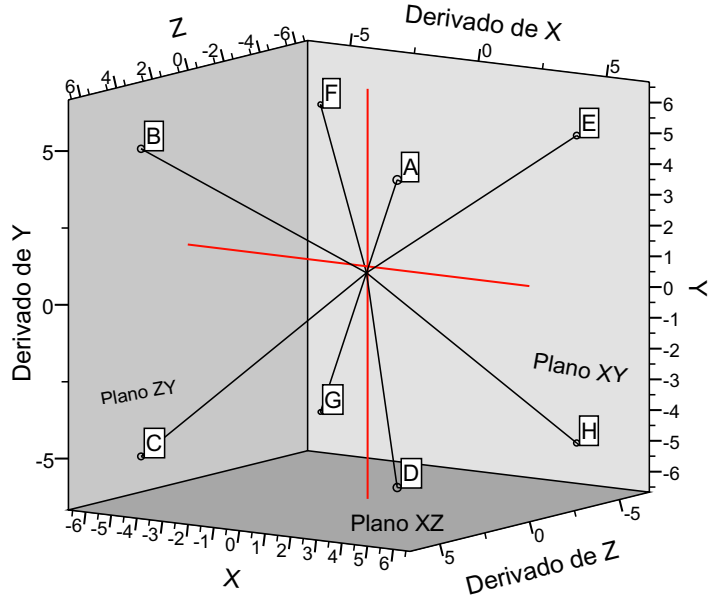


Desde el origen del sistema, coordenadas (0,0), a la derecha, el eje *X* tiene valores positivos y hacia la izquierda valores negativos. El eje *Y* tiene valores positivos por encima del eje *X* y negativos por debajo. Cualquier punto en el plano se puede representar por un par de coordenadas *(x,y)*. Cualquier punto en el cuadrante *I* tiene coordenadas *x* e *y* positivas; en el *II* la *x* es negativa y la *y* positiva; en el cuadrante *III* *x* e *y* son negativas, y en el último cuadrante, la coordenada *x* es positiva y la *y* negativa. Cualquier punto en el eje *X* tiene coordenada $y = 0$ y cualquier punto en el eje *Y* tiene coordenada $x = 0$.

En el Gráfico 10, las coordenadas de los puntos *A, B, C* y *D* son: (4,5), (-5,4), (-8,-4) y (5,-6), respectivamente.

El sistema de coordenadas cartesianas de tres dimensiones es un sistema de coordenadas de tres ejes ortogonales (perpendiculares entre sí) que dividen el espacio en ocho partes que llamamos octantes: *I, II, III, IV, V, VI, VII* y *VIII*. El eje horizontal es el *X,* el eje vertical es el *Y* y el eje que saldría hacia el exterior de este papel es el *Z* . El punto en el que se cruzan se dice que tiene coordenadas *(x,y,z)* (0,0,0) y se le considera el origen del sistema. En el Gráfico 11 ese punto está representado por el punto donde se cruzan los dos ejes de color rojo y es el centro del cubo formado por el sistema de tres ejes. Llamamos *plano XY*, al plano definido por los ejes *X-Y,* el *plano ZY*, el formado por los ejes *Z-Y*, y el formado por los ejes *X-Z* es el *plano XZ*.

Gráfico 11 Sistema de coordenadas cartesianas de tres dimensiones.



Desde el origen del sistema, coordenadas (0,0,0), a la derecha, el eje *X* tiene valores positivos y hacia la izquierda valores negativos; el eje *Y* tiene valores positivos por encima del origen (0,0,0) y negativos por debajo, y el eje *Z* tiene valores positivos por delante del origen (0,0,0) y negativos hacia atrás. Cualquier punto en el espacio se puede representar por las coordenadas *(x,y,z)*. En los octantes *I, II, III* y *IV,* las coordenadas *x,y* tienen los mismos signos que en el Gráfico 10 y la *z* es positiva, y en los octantes *V, VI, VII* y *VIII, x* e *y* tienen los mismos signos que antes, pero *z* es negativa.

Se ha representado un punto en cada octante para clarificar el sistema de signos (todos tienen el valor 5), y para dar sensación de perspectiva se representa la traza desde el origen (0,0,0) hasta el punto correspondiente. El punto *A* tiene coordenadas (5,5,5) y está en el octante *I*. El punto *B* tiene coordenadas (-5,5,5) y está en el octante *II*. El punto *C* tiene coordenadas (-5,-5,5) y está en el octante *III*. El punto *D* tiene coordenadas (5,-5,5) y está en el octante *IV*. El punto *E* tiene coordenadas (5,5,-5) y está en el octante *V*. El punto *F* tiene coordenadas (-5,5,-5) y está en el octante *VI*. El punto *G* tiene coordenadas (-5,-5,-5) y está en el octante *VII*. Y el punto *H* tiene coordenadas (5,-5,-5) y está en el octante *VIII*. Los puntos *A* y *D* son los que están más cerca de los ojos del lector y el *F* y *G* los más alejados.

Cualquier punto en el eje *X* tiene coordenada cero en los ejes *Y* y *Z*. Los puntos del eje *Y* el valor cero es en los ejes *X* y *Z*. Y los puntos del eje *Z* presentan valor cero en los ejes *X* e *Y*. De la misma manera, cualquier punto en el plano *XZ,* tiene valor cero en el eje *Y;* los del plano *XY,* tienen coordenada cero en el *Z,* y los del plano *ZY,* el valor cero es en el *X*. Considerando que los ejes y los planos pasan por el punto de coordenadas (0,0,0).

7.6.2 Diagrama de barras

El *diagrama de barras* es la representación gráfica en un sistema de coordenadas cartesianas de dos dimensiones de la *tabla de frecuencias* de una variable cualitativa o categórica, nominal u ordinal. En el eje horizontal, *X* o de abscisas se representan los valores o categorías de la variable y en el eje vertical, *Y* o de ordenadas se representan las frecuencias absolutas o relativas de cada valor o categoría de la variable.

Como los valores de las variables de nivel de medida *nominal* u *ordinal* no tienen ninguna relación entre sí, si acaso de orden, entonces la escala o división del eje *X* son marcas sin ninguna relación entre ellas, colocadas de forma aleatoria y razonable dentro del marco del gráfico y sin solución de continuidad entre ellas.

El ancho de las barras es estético. Las frecuencias representadas en el eje *Y* indican el número de casos o la proporción de casos que se da en cada valor o categoría de la variable y este eje se escala en función del valor de la categoría con frecuencia mayor (Tabla 39 y Tabla 40).

Cada segmento del eje *Y* representa a *n* casos. En el gráfico de la variable *sexo* de la Tabla 39, cada segmento representa dos casos. En el gráfico de la variable *estado civil* de la Tabla 39, cada segmento representa cinco casos. En el gráfico de la variable *sexo* de la Tabla 40, cada segmento representa dos puntos porcentuales. En el gráfico de la variable *estado civil* de la Tabla 40, cada segmento representa cinco puntos porcentuales.

Tabla 39 Diagrama de Barras (frecuencias absolutas).

Tabla de frecuencias de la variable *sexo*.

X	n
Varón	49
Mujer	50
Total	99

Diagrama de barras de la variable *sexo*.

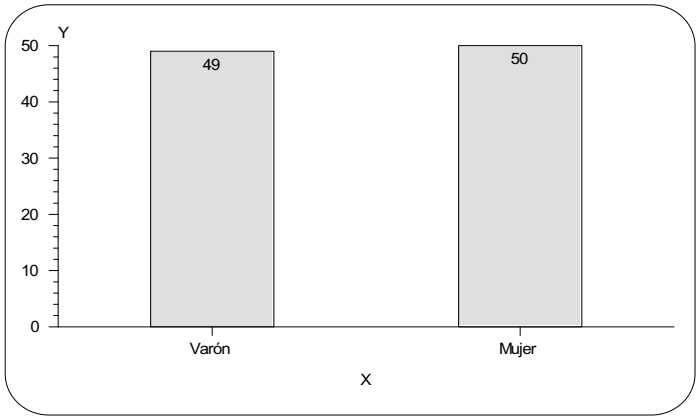
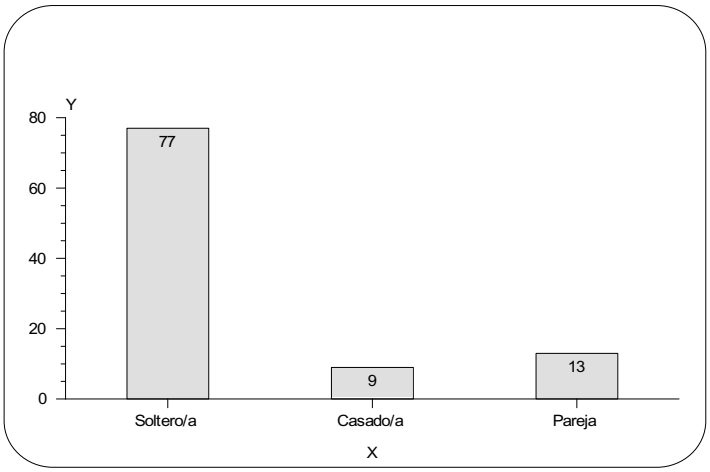


Tabla de frecuencias de la variable *estado civil*.

X	n
Soltero/a	77
Casado/a	9
Pareja	13
Total	99

Diagrama de barras de la variable *estado civil*.



Nota: Se ha utilizado la matriz de datos de la Tabla 16.

Tabla 40 Diagrama de Barras (frecuencias relativas).

Tabla de frecuencias de la variable *sexo*.

X	f (%)
Varón	49,5
Mujer	50,5
Total	100,0

Diagrama de barras de la variable *sexo*.

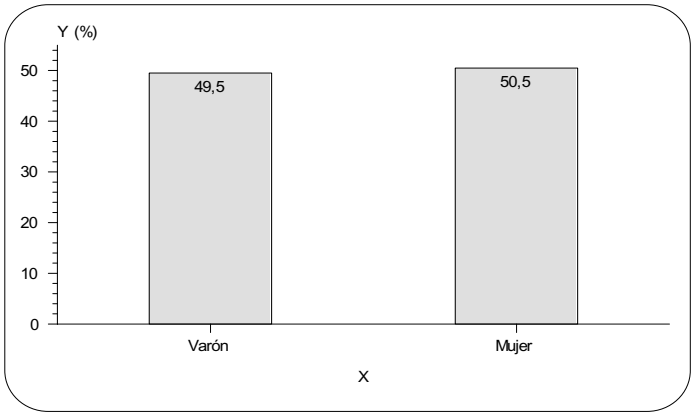
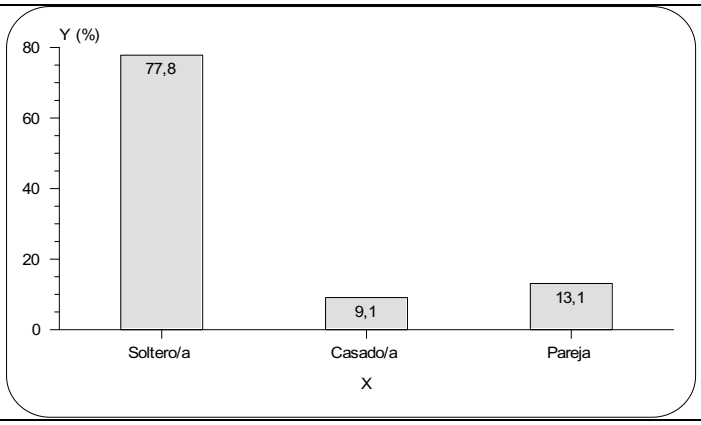


Tabla de frecuencias de la variable *estado civil*.

X	f (%)
Soltero/a	77,8
Casado/a	9,1
Pareja	13,1
Total	100,0

Diagrama de barras de la variable *estado civil*.



Nota: Se ha utilizado la matriz de datos de la Tabla 16.

7.6.3 Histograma de intervalos de igual amplitud

El *histograma* es la representación gráfica en un sistema de coordenadas cartesianas de dos dimensiones de la *tabla de frecuencias* de una variable cuantitativa o numérica, intervalar o de razón. En el eje horizontal, *X* o de abscisas se representan los valores de la variable y en el eje vertical, *Y* o de ordenadas se representan las frecuencias absolutas o relativas de cada valor de la variable.

Como entre los valores de las variables de nivel de medida *intervalar* o de *razón* existe el concepto distancia, entonces el eje *X* se escala en función del valor máximo de la variable y cada segmento representa *n* unidades de la variable. Las frecuencias representadas en el eje *Y* indican el número de casos o la proporción de casos que se da en cada valor de la variable y este eje se escala en función del valor con frecuencia mayor (Tabla 41).

Tabla 41 Histograma (frecuencias absolutas).

Proceso:

1º La *tabla de frecuencias* muestra la frecuencia o el número de casos por cada valor de la variable *X*. Pero sólo aparecen aquellos valores que tienen *frecuencia.* Como los valores de la variable son significativos y hay distancia entre ellos, en el sistema de coordenadas cartesianas hay que representar todos los valores en el eje *X* empezando en el valor 0 kg. En cada valor *x,* se representa el número de casos que lo tienen con un segmento. El primer valor es 45 kg. que lo tienen 3 unidades de observación, y así sucesivamente.

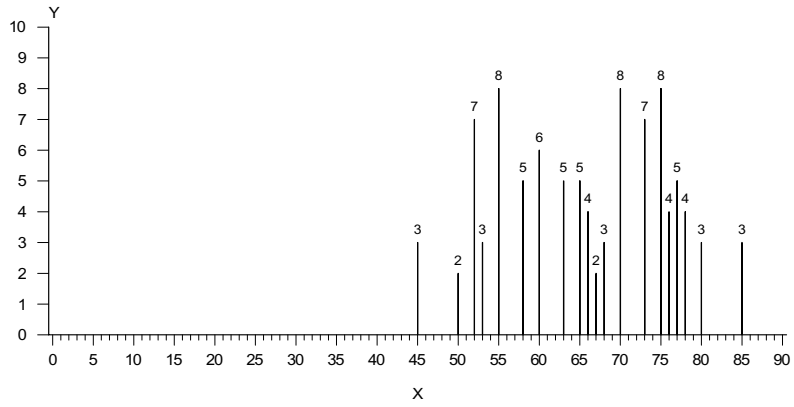
2º En realidad los casos que aparecen en cada valor de peso no tienen exactamente el peso indicado, sino que se asume que esos casos tienen el peso comprendido entre el valor en el que aparecen y el valor a continuación. Así los 3 individuos que aparecen en el valor de 45 kg. se considera que tienen entre 45 kg. y 46 kg., sin llegar a 46 kg., porque si fuese así, tendrían el valor de 46 kg. Entonces, los 3 primeros individuos están en el intervalo de 45 a 46 kg. pero el límite superior se considera abierto. La tabla de frecuencias ha pasado a ser considerada por intervalos de igual amplitud.

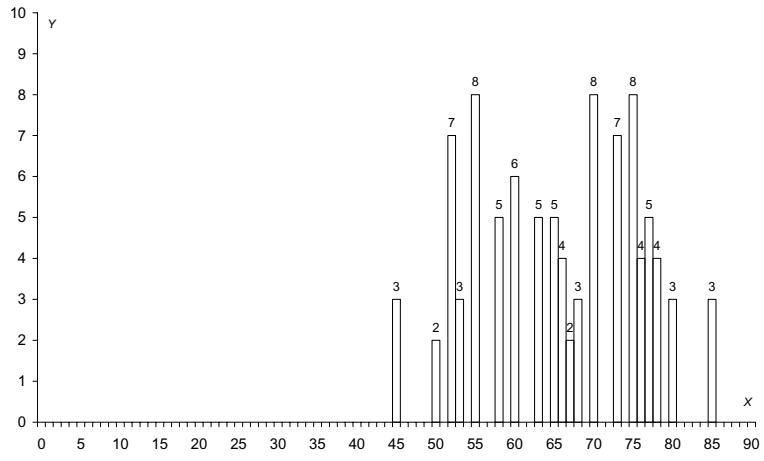
3º Desde 0 kg. a 45 kg. no hay casos. En los histogramas se elimina la zona inicial que no tienen frecuencia para mejorar la representación gráfica. Pero se deben mantener los valores intercalados que no tienen frecuencia. Entonces los valores hasta 40 kg. no se representarán pero se mantienen los valores: 46, 47, 48 y 49 kg. así como el resto de valores que no tienen casos o que la frecuencia es cero.

Tabla de frecuencias de la variable *peso*.

X	n
45	3
50	2
52	7
53	3
55	8
58	5
60	6
63	5
65	5
66	4
67	2
68	3
70	8
73	7
75	8
76	4
77	5
78	4
80	3
85	3
Total	95

Histograma de la variable *peso*.





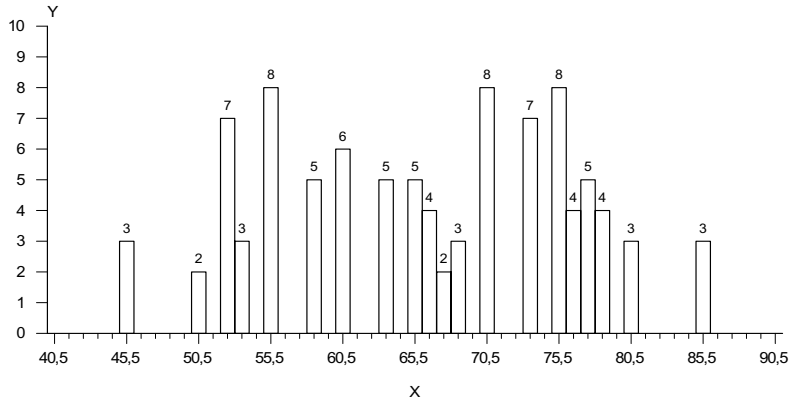
Continuación Tabla 41.

Observación:

Según las tres observaciones hechas anteriormente, la tabla de frecuencias y el histograma quedan de la siguiente forma. Cada barra se representa en el eje *X* por la *marca de clase* o *punto medio del intervalo*. Los tres individuos que tienen 45 kg o más, pero que no llegan a tener 46 kg estarán representados por su *marca de clase* 45,5 kg.

Si la amplitud de los intervalos es igual, se considera la base unitaria, y entonces la altura y la superficie de la barra se igualan a la frecuencia y se hace la asociación de que la superficie representa a los casos y es igual a la altura. La barra de 45,5 kg que representa a tres casos, si se considera que la base, al ser igual en todos los intervalos, es unitaria, la superficie tiene el mismo valor de tres, que representa a los tres casos, y así sucesivamente. Por lo tanto, la superficie de todas las barras representa al total de los casos (n=95).

X	n
45 - 46	3
50 - 51	2
52 - 53	7
53 - 54	3
55 - 56	8
58 - 59	5
60 - 61	6
63 - 64	5
65 - 66	5
66 - 67	4
67 - 68	2
68 - 69	3
70 - 71	8
73 - 74	7
75 - 76	8
76 - 77	4
77 - 78	5
78 - 79	4
80 - 81	3
85 - 86	3
Total	95

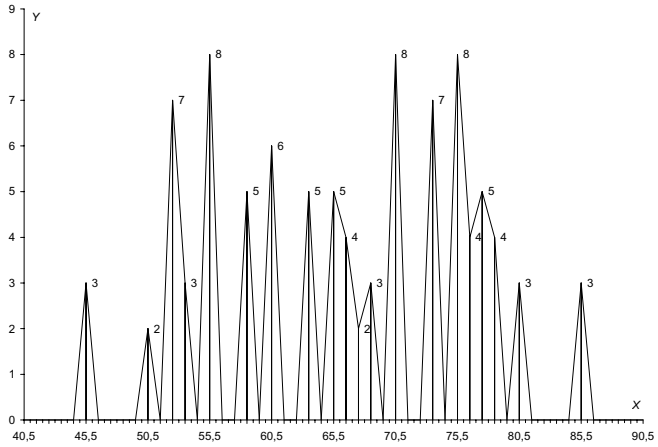


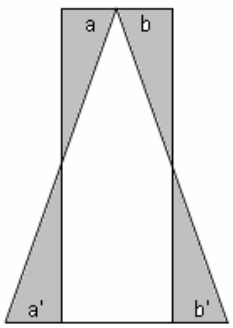
Polígono de frecuencias

Observación:

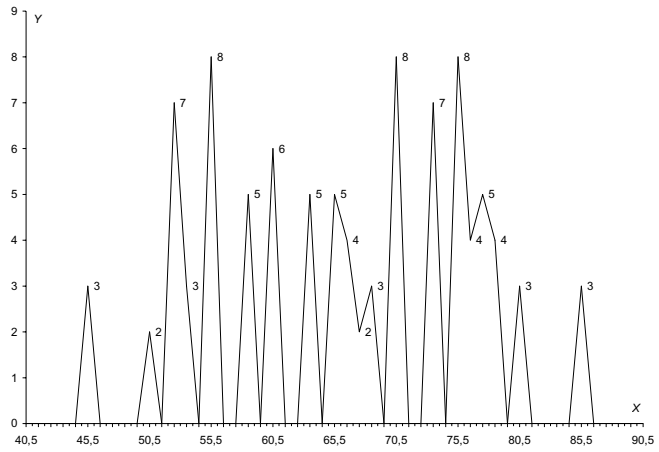
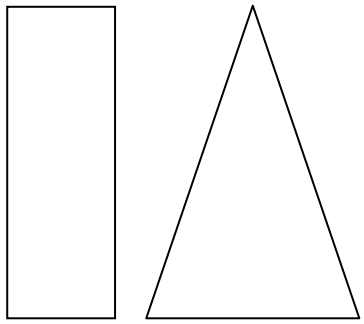
El *polígono de frecuencias* resulta de unir los puntos medios de la cara opuesta a la base, de cada intervalo, que representamos con y sin las verticales. Si el intervalo anterior y/o el posterior no tienen frecuencia, se considera cero y el polígono se cierra con el eje *X*. De la misma manera, la superficie por debajo del polígono de frecuencias y por encima del eje de abscisas representa al total de los casos (n = 95). La equivalencia de la superficie de los rectángulos con los triángulos, se debe a que de las superficies marcadas de gris a = a' y b = b' al ser triángulos opuestos por el vértice con lados y ángulos iguales. La superficie *a* que se deja fuera al trazar el triángulo del polígono de frecuencias, es igual a *a'* que se coge de fuera y lo mismo sucede con *b* y *b'* (Gráfico 12). La superficie del rectángulo y del triángulo son iguales (Gráfico 13).

Gráfico 12





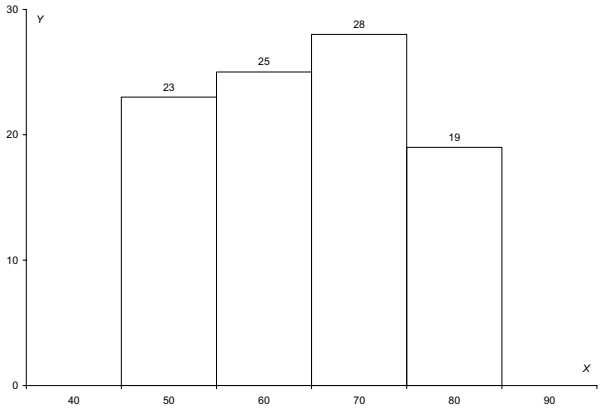
Continuación Tabla 41.

	Gráfico 13
	

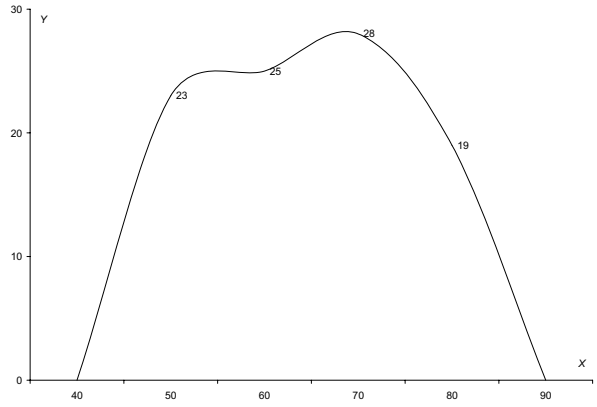
Nota: Se ha utilizado la matriz de datos de la Tabla 16.

Una de las posibles definiciones de la Estadística es que se utiliza para resumir o sintetizar datos y en los gráficos anteriores se puede reducir el número de intervalos haciendo mayor su amplitud o recorrido como en la Tabla 31.

Tabla 42 Histograma con datos agrupados* (frecuencias absolutas)

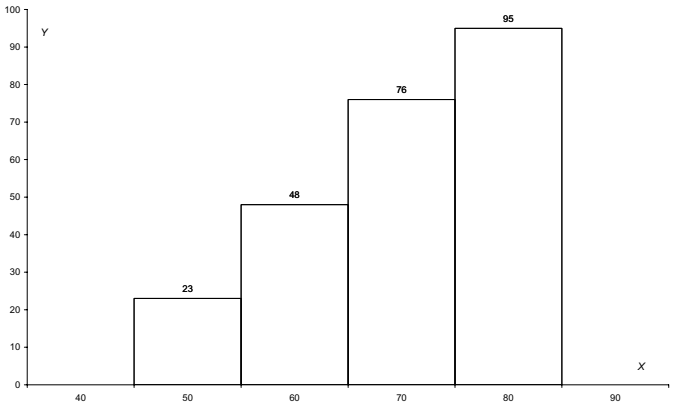


Polígono de frecuencias**

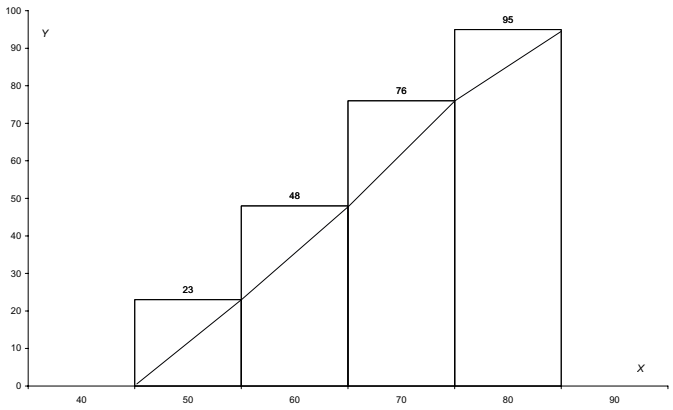


Notas:
* Los gráficos se han realizado a partir de la tabla de frecuencias de la Tabla 31.
** El polígono de frecuencias se puede representar con la línea suavizada, por estética.

Tabla 43 Histograma de frecuencias acumuladas* (frecuencias absolutas).



Polígono de frecuencias acumuladas

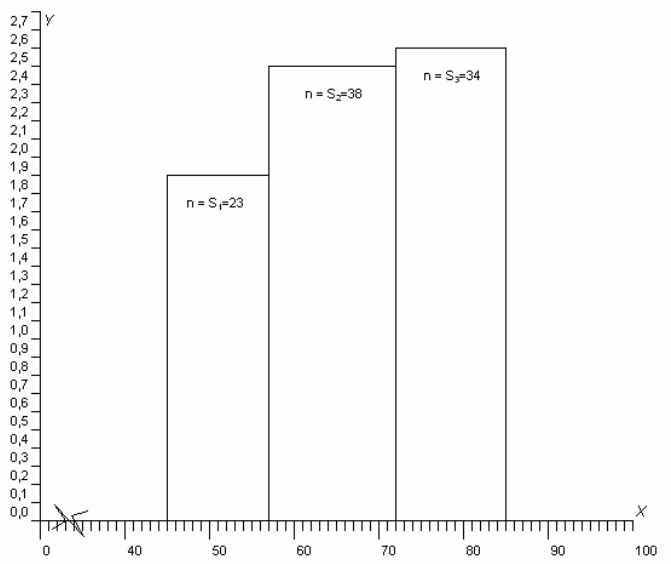


Notas:
* Los gráficos se han realizado a partir de la tabla de frecuencias, acumuladas, de la Tabla 31.

Según se dijo, en el eje *Y* del histograma se representa la frecuencia absoluta y ésta determina el número de observaciones a las que representa la superficie de la barra, al ser todos los intervalos de la misma amplitud y considerarlos unitarios. Entonces la superficie total del histograma representa al total de los casos Si los intervalos son de distinta amplitud, entonces las alturas y las superficies no son equivalentes a las frecuencias y para que las superficies representen a los casos, se tiene que hacer una transformación. No obstante, este caso es poco habitual.

Como ejemplo se representa el histograma de intervalos de distinta amplitud (Tabla 44) de la tabla de frecuencias de la Tabla 33.

Tabla 44 Histograma de intervalos de distinta amplitud.



Se Parte del cálculo de la superficie del rectángulo	$S \mid b\Delta h$, siendo *h* la altura, *b* la base y *S* la superficie.
Para el cálculo de la superficie de cada uno de los rectángulos del histograma se considera que:	$b \mid a \mid$ amplitud del intervalo $h \mid$ frecuencia (eje Y) $S \mid$ número de casos
Intervalo de 45 kg. a 57 kg.	$S_1 \mid n_1 \mid 23 \mid b\Delta h \mid a_1\Delta h_1 \mid (57\,4\,45)\Delta h_1$, entonces $h_1 \mid \frac{23}{(57\,4\,45)} \mid 1{,}92$
Intervalo de 57 kg. a 72 kg.	$S_2 \mid n_2 \mid 38 \mid b\Delta h \mid a_2\Delta h_2 \mid (72\,4\,57)\Delta h_2$, entonces $h_2 \mid \frac{38}{(72\,4\,57)} \mid 2{,}53$
Intervalo de 72 kg. a 85 kg.	$S_3 \mid n_3 \mid 34 \mid b\Delta h \mid a_3\Delta h_3 \mid (85\,4\,72)\Delta h_3$, entonces $h_3 \mid \frac{34}{(85\,4\,72)} \mid 2{,}62$

8 Estadística Descriptiva Bivariable

La *Estadística Descriptiva Bivariable* es la estadística que describe o tabula las variables de dos en dos. Ofrece tablas que son el resultado del cruce de dos variables. Esto no significa que no se puedan especificar más de dos variables simultáneamente. En el cruce se pueden definir más de dos variables, pero las terceras y sucesivas variables se consideran intervinientes, de control o de capa.

Al cruzar dos variables, como el nivel de medida puede ser cualitativas o categóricas y cuantitativas o numéricas, entonces los cruces posibles son: categórica por categórica, numérica pro categórica y numérica por numérica. Cada una de estas vías tiene sus estadísticos propios y es la entrada a las tres ramas de la que se considera la Estadística Analítica (Tabla 45), y esta tabla, más los estadísticos anteriores, se puede considerar la base de la Estadística Multivariable.

Tabla 45 Estadística Bivariable.

Combinación	Estadísticos	Análisis (Contraste de Hipótesis).	Estadísticos	Nomenclatura	Base de las Técnicas Multivariable.
Categórica por categórica	Tabla de doble entrada o distribución conjunta de frecuencias	Tablas de Contingencia.	θ^2	Chi cuadrado	
Numérica por categórica	Tabla de medias.	Diferencia de medias y Análisis de Varianza	Z *t-Student* F_S	Diferencia de medias. Diferencia de medias. ANOVA.	
Numérica por numérica	Covarianza y correlación	Asociación lineal.	S_{XY} r_{XY}	Covarianza. Correlación de Pearson	

8.1 Variable categórica por categórica

Si se cruzan dos variables categóricas: X e Y, cada una de ellas con tres sucesos elementales: x_1, x_2 y x_3 e y_1, y_2 e y_3, se obtiene una tabla de X por Y, de tres por tres categorías. Terceras y sucesivas variables se consideran variables intervinientes o de control, esto es, se obtendría una tabla de X por Y por cada una de las categorías o combinación de categorías de las variables intervinientes o de control. En la Tabla 46 se muestra el esquema correspondiente.

Tabla 46 Cruce de dos variables sin y con variables de control.

Cruce	Tabla
Cruce de dos variables (*X, Y*). *X* (x_1, x_2, x_3) por *Y* (y_1, y_2, y_3)	Variable *Y* por *X** *X* |	x_1 | x_2 | x_3 | *Y* y_1 | | | | y_2 | | | | y_3 | | | |
Cruce de dos variables (*X, Y*) con variable de control (*Z*). *X* (x_1, x_2, x_3) por *Y* (y_1, y_2, y_3) por *Z* (z_1, z_2)	Variable *Y* por *X*, según z_1 z_1 *X* |	x_1 | x_2 | x_3 | *Y* y_1 | | | | y_2 | | | | y_3 | | | | Variable *Y* por *X*, según z_2 z_2 *X* |	x_1 | x_2 | x_3 | *Y* y_1 | | | | y_2 | | | | y_3 | | | |
Cruce de dos variables (*X, Y*) con variables de control (*Z, W*). *X* (x_1, x_2, x_3) por *Y* (y_1, y_2, y_3) por *Z* (z_1, z_2) por *W* (w_1, w_2).	Variable *Y* por *X*, según w_1 y z_1 w_1, z_1 *X* |	x_1 | x_2 | x_3 | *y* y_1 | | | | y_2 | | | | y_3 | | | | Variable *Y* por *X*, según w_1 y z_2 w_1, z_2 *X* |	x_1 | x_2 | x_3 | *y* y_1 | | | | y_2 | | | | y_3 | | | | Variable *Y* por *X*, según w_2 y z_1 w_2, z_1 *X* |	x_1 | x_2 | x_3 | *y* y_1 | | | | y_2 | | | | y_3 | | | | Variable *Y* por *X*, según w_2 y z_2 w_2, z_2 *X* |	x_1 | x_2 | x_3 | *Y* y_1 | | | | y_2 | | | | y_3 | | | |

Nota:

* El título de las tablas se construye expresando primero la variable de filas, después la variable de columnas y a continuación, si hay variables de control, se expresan las categorías a las que corresponden la tabla.

8.2 Tabla de doble entrada

Una *tabla de doble entrada* o *distribución conjunta de frecuencias* es una matriz rectangular o cuadrada que en la entrada de filas se representan las categorías, códigos, valores o sucesos elementales del espacio muestral de una de las variables y en la entrada de columnas se representan las categorías, códigos, valores o sucesos elementales del espacio muestral de la otra variable. En esta tabla no se plantea asociación entre las variables, pero como es la antesala del contraste de hipótesis de asociación, entonces la variable considerada o propuesta como dependiente se pone en las filas y la considerada o propuesta como independiente en las columnas. De esta manera se orienta el proceso hacia la tabla de contingencia. No obstante, este criterio es por convenio, quiere decir que la colocación de las variables no va a influir en el resultado de los estadísticos aplicados. En cualquier caso, es habitual que las variables consideradas de clasificación (socio-político-económico-demográficas: sexo, edad, estudios, estado civil, etc.) se pongan siempre en las columnas.

En la Tabla 46 la variable Y sería la considerada como posible dependiente, la X la considerada como posible independiente y la Z y W las de control o intervinientes. El cuadro definido por el cruce de cada dos categorías de X e Y se llama celda, y contiene las frecuencias absolutas y frecuencias relativas.

Las frecuencias absolutas indican el número de veces que se dan conjuntamente dos categorías o el número de casos o unidades de observación que pertenecen a esas dos categorías. La última columna es el sumatorio de las frecuencias absolutas de filas y se llama marginal de filas, y la última fila se le llama marginal de columnas y es el sumatorio de las frecuencias absolutas de las columnas. Por lo tanto hay tres totales, el total de las filas, el total de las columnas y el total del marginal de filas o el total del marginal de columnas que es igual al total de tabla (Tabla 47).

Tabla 47 Elementos de la tabla de doble entrada.

		X			
		x_1	x_2	x_3	TF
Y	y_1	n_{11}	n_{12}	n_{13}	TF_1
	y_2	n_{21}	n_{22}	n_{23}	TF_2
	y_3	n_{31}	n_{32}	n_{33}	TF_3
	TC	TC_1	TC_2	TC_3	TT

En donde:
Y: Variable de filas.
X: Variable de columnas.
TF: Marginal o total de filas.
TC: Marginal o total de columnas.
TF_i: Total de fila i-ésima.
TC_j: Total de columna j-ésima.
TT: Total de tabla.
n_{ij}: Frecuencia absoluta de la celda ij-ésima.
y_i: Valor, código, categoría o suceso elmental i-ésimo de Y.
x_i: Valor, código, categoría o suceso elmental j-ésimo de X.

Las frecuencias relativas, expresadas en proporciones o en porcentajes, son la relación entre la frecuencia absoluta de la celda y los totales de fila, de columna y de tabla, como se muestra en la Tabla 48 y su cálculo.

Tabla 48 Frecuencias absolutas y relativas de la tabla de doble entrada.

Título de tabla

		X			
		x_1	x_2	x_3	*TF*
Y	y_1	n_{11} $f_{11}TF_1$ $f_{11}TC_1$ $f_{11}TT$	n_{12} $f_{12}TF_1$ $f_{12}TC_2$ $f_{12}TT$	n_{13} $f_{13}TF_1$ $f_{13}TC_3$ $f_{13}TT$	TF_1
	y_2	n_{21} $f_{21}TF_2$ $f_{21}TC_1$ $f_{21}TT$	n_{22} $f_{22}TF_2$ $f_{22}TC_2$ $f_{22}TT$	n_{23} $F_{23}TF_2$ $f_{23}TC_3$ $f_{23}TT$	TF_2
	y_3	n_{31} $f_{31}TF_3$ $f_{31}TC_1$ $f_{31}TT$	n_{32} $f_{32}TF_3$ $f_{32}TC_2$ $f_{32}TT$	n_{33} $f_{33}TF_3$ $f_{33}TC_3$ $f_{33}TT$	TF_3
	TC	TC_1	TC_2	TC_3	*TT*

En donde:

Y: Variable de filas.
X: Variable de columnas.
TF: Marginal o total de filas.
TC: Marginal o total de columnas.
TF_i: Total de fila i-ésima.
TC_j: Total de columna j-ésima.
n_{ij}: Frecuencia absoluta de la celda ij-ésima.
y_i: Valor, código, categoría o suceso elmental i-ésimo de *Y*.
x_j: Valor, código, categoría o suceso elmental j-ésimo de *X*.
$f_{ij}TF_i$: Frecuencia relativa de la celda ij-ésima sobre el total de fila i-ésima.
$f_{ij}TC_j$: Frecuencia relativa de la celda ij-ésima sobre el total de columna j-ésima.
$f_{ij}TT$: Frecuencia relativa de la celda ij-ésima sobre el total de tabla.

$$f_{ij}TF_i \mid \frac{n_{ij}}{\overset{C}{\overline{\overline{n_{.j}}}}} \mid \frac{n_{ij}}{TF_i} \text{ ó } f_{ij}TF_i \mid \frac{n_{ij}}{\overset{C}{\overline{\overline{n_{.j}}}}}\Delta 100 \mid \frac{n_{ij}}{TF_i}\Delta 100$$

$$f_{ij}TC_j \mid \frac{n_{ij}}{\overset{F}{\overline{\overline{n_{i.}}}}} \mid \frac{n_{ij}}{TC_j} \text{ ó } f_{ij}TC_j \mid \frac{n_{ij}}{\overset{F}{\overline{\overline{n_{i.}}}}}\Delta 100 \mid \frac{n_{ij}}{TC_j}\Delta 100$$

$$f_{ij}TT \mid \frac{n_{ij}}{\overset{F\ C}{\overline{\overline{n_{ij}}}}} \mid \frac{n_{ij}}{TT} \text{ ó } f_{ij}TT \mid \frac{n_{ij}}{\overset{C\ F}{\overline{\overline{n_{ij}}}}}\Delta 100 \mid \frac{n_{ij}}{TT}\Delta 100$$

Ejemplo de tabla de doble entrada (se ha utilizado la matriz de datos de la Tabla 16), con las frecuencias absolutas y el cálculo de las frecuencias relativas expresadas en porcentajes. La frecuencia absoluta de la primera celda (36), es el número de casos que son varones y que están solteros, o el número de veces que se repite conjuntamente el suceso elemental "varón" y "soltero", o el número de casos que cumplen la condición "varón" y "soltero" (Tabla 49).

Tabla 49 Frecuencias absolutas y cálculo de frecuencias relativas de la tabla de doble entrada: Estado civil según el sexo.

Estado civil según el sexo.

			Sexo		
			Varón	*Mujer*	*Total fila*
Estado civil	*Soltero/a*	*n*	36	41	77
		%TF	46,8%	53,2%	100,0%
		%TC	73,5%	82,0%	77,8%
		%TT	36,4%	41,4%	77,8%
	Casado/a	*n*	6	3	9
		%TF	66,7%	33,3%	100,0%
		%TC	12,2%	6,0%	9,1%
		%TT	6,1%	3,0%	9,1%
	Pareja	*n*	7	6	13
		%TF	53,8%	46,2%	100,0%
		%TC	14,3%	12,0%	13,1%
		%TT	7,1%	6,1%	13,1%
	Total columna	*n*	49	50	99
		%TF	49,5%	50,5%	100,0%
		%TC	100,0%	100,0%	100,0%
		%TT	49,5%	50,5%	100,0%

Cálculo de porcentaje para una celda.

$$f_{11}TF_1 \mid \frac{n_{11}}{\sum_{j \mid 1}^{2} n_{1j}} \Delta 100 \mid \frac{36}{36 + 41} \Delta 100 \mid \frac{36}{77} \Delta 100 \mid 46{,}8\%$$

$$f_{11}TC_1 \mid \frac{n_{11}}{\sum_{i \mid 1}^{3} n_{i1}} \Delta 100 \mid \frac{36}{36 + 6 + 7} \Delta 100 \mid \frac{36}{49} \Delta 100 \mid 73{,}5\%$$

$$f_{11}TT \mid \frac{n_{11}}{\sum_{i \mid 1}^{3} \sum_{j \mid 1}^{2} n_{ij}} \Delta 100 \mid \frac{36}{36 + 41 + 6 + 3 + 7 + 6} \Delta 100 \mid \frac{36}{99} \Delta 100 \mid 36{,}4\%$$

Cálculo de porcentaje para el total de fila.

$$fTF_1TF_1 \mid \frac{TF_1}{\sum_{j \mid 1}^{2} n_{1j}} \Delta 100 \mid \frac{77}{36 + 41} \Delta 100 \mid \frac{77}{77} \Delta 100 \mid 100{,}0\%$$

$$fTF_1TC \mid \frac{TF_1}{\sum_{i \mid 1}^{3} TF_i} \Delta 100 \mid \frac{77}{77 + 9 + 13} \Delta 100 \mid \frac{77}{99} \Delta 100 \mid 77{,}8\%$$

$$fTF_1TT \mid \frac{TF_1}{\sum_{i \mid 1}^{3} \sum_{j \mid 1}^{2} n_{ij}} \Delta 100 \mid \frac{77}{36 + 41 + 6 + 3 + 7 + 6} \Delta 100 \mid \frac{77}{99} \Delta 100 \mid 77{,}8\%$$

Cálculo de porcentaje para el total de columna.

$$fTC_1TF \mid \frac{TC_1}{\sum_{j \mid 1}^{2} TC_j} \Delta 100 \mid \frac{49}{49 + 50} \Delta 100 \mid \frac{49}{99} \Delta 100 \mid 49{,}5\%$$

$$fTC_1TC_1 \mid \frac{TC_1}{\sum_{i \mid 1}^{3} n_{i1}} \Delta 100 \mid \frac{49}{36 + 6 + 7} \Delta 100 \mid \frac{49}{49} \Delta 100 \mid 100{,}0\%$$

$$fTC_1TT \mid \frac{TC_1}{\sum_{i \mid 1}^{3} \sum_{j \mid 1}^{2} n_{ij}} \Delta 100 \mid \frac{49}{36 + 41 + 6 + 3 + 7 + 6} \Delta 100 \mid \frac{49}{99} \Delta 100 \mid 49{,}5\%$$

Tabla 50 Continuación.

Cálculo de porcentaje para el total de tabla.

$$fTFTF \mid \frac{TF}{\sum_{j \mid 1}^{2} TC_j} \Delta 100 \mid \frac{99}{49\ 2\ 50} \Delta 100 \mid \frac{99}{99} \Delta 100 \mid 100{,}0\%$$

$$fTCTC \mid \frac{TC}{\sum_{i \mid 1}^{3} TF_i} \Delta 100 \mid \frac{99}{77\ 2\ 9\ 2\ 13} \Delta 100 \mid \frac{99}{99} \Delta 100 \mid 100{,}0\%$$

$$fTTTT \mid \frac{TT}{\sum_{i \mid 1}^{3} \sum_{j \mid 1}^{2} n_{ij}} \Delta 100 \mid \frac{99}{36\ 2\ 41\ 2\ 6\ 2\ 3\ 2\ 7\ 2\ 6} \Delta 100 \mid \frac{99}{99} \Delta 100 \mid 100{,}0\%$$

Símbolos:
TF: Total de Fila.
TC. Total de columna.
TT: Total de tabla.

LECTURA DE PORCENTAJES

Al haber tres porcentajes, son tres las lecturas posibles de porcentajes. La cuestión entonces es qué información da cada uno ellos y si existe alguno porcentaje mejor que el otro. A estas lecturas se añade la denominada Regla de Zeisel (Tabla 51).

Tabla 51 Lectura de porcentajes de la tabla de doble entrada: estado civil según el sexo.

Estado civil según el sexo

			Sexo		
			Varón	*Mujer*	*Total fila*
Estado civil	*Soltero/a*	*n*	36	41	77
		%TF	46,8%	53,2%	100,0%
		%TC	73,5%	82,0%	77,8%
		%TT	36,4%	41,4%	77,8%
	Casado/a	*n*	6	3	9
		%TF	66,7%	33,3%	100,0%
		%TC	12,2%	6,0%	9,1%
		%TT	6,1%	3,0%	9,1%
	Pareja	*n*	7	6	13
		%TF	53,8%	46,2%	100,0%
		%TC	14,3%	12,0%	13,1%
		%TT	7,1%	6,1%	13,1%
	Total columna	*n*	49	50	99
		%TF	49,5%	50,5%	100,0%
		%TC	100,0%	100,0%	100,0%
		%TT	49,5%	50,5%	100,0%

Lectura de frecuencias absolutas. Primera fila.*

De los 77 “soltero/a” de la primera fila, 36 dicen ser varones y 41 mujeres.

Lectura de frecuencias absolutas. Primera columna.

De los 49 “varones”, 36 dicen estar solteros, 6 casados y 7 viven en pareja.

Lectura de frecuencias absolutas. Total de tabla.

De las 99 unidades de observación, 36 dicen ser varones y estar solteros; 41 mujeres y solteras; 6 varones y casados; 3 mujeres y casadas; 7 varones y vivir en pareja, y 6 mujeres y en pareja.

Lectura de porcentajes sobre el total de fila.

De los 77 solteros, el 46,8% dicen ser varones y el 53,2% mujeres.

Lectura de porcentajes sobre el total de columna.

De los 49 varones, el 73,5% dicen estar solteros, el 12,2% casados y el 14,3% vivir en pareja.

Lectura de porcentajes sobre el total de tabla.

De las 99 unidades de observación, 36,4% dicen ser varones y estar solteros; 41,4% mujeres y solteras; 6,1% varones y casados; 3,0% mujeres y casadas; 7,1% varones y vivir en pareja, y 6,1% mujeres y en pareja.

Notas:

* La distinción entre el “ser”, “estar” y “decir” no es trivial en el caso de la aplicación de los cuestionarios y depende de si el cuestionario es autoadministrado o mediante entrevista. Es una cuestión ontológica y epistemológica fundamental. Por ejemplo, si el cuestionario es autoadministrado, el entrevistado “dice” lo que es, pero el investigador no “sabe” lo que es, por lo tanto debe creer lo que dice, pero decir “dice”. Si el cuestionario es mediante entrevista, en el caso del sexo, el entrevistador “ve” lo que “es” aunque la realidad fisiológica pueda ser distinta. En el caso del “estado civil” la realidad va a ser siempre lo que diga el entrevistado, y por lo tanto decir “dice”. Exceptuando las circunstancias en las que el entrevistado adjunte algún documento oficial en el que demuestre oficialmente su estado civil y entonces no es “dice” sino “está”. Investigando la complejidad de la realidad humana, debido a que se trabaja con lo que “dice”, la “realidad” es lo que “dice” sin cuestionar si es verdad o mentira, porque en asuntos generales de opiniones, actitudes, aptitudes, valores, normas, etc. incluso el propio entrevistado puede desconocer (o no ser completamente consciente) cual es su realidad “real”. Cuestionarse continuamente lo que se “dice” no llevaría a ningún lugar o a cerrar la investigación. Se deben aceptar las cosas como las dicen puesto que es el único instrumento del que se dispone actualmente para saber algo, hasta que aparezcan otros instrumentos. Con los procedimientos estadísticos se pueden detectar ciertas incongruencias.

Aunque la lectura de porcentajes anterior es correcta, no todos dan la misma información, ni una información completa, sin ser falsa. Si se leen los porcentajes de la fila sobre su total, no se sabe nada de lo que ocurre verticalmente, sobre el total de las columnas. Si se leen sobre el total de la columna, se observa lo que ocurre verticalmente, pero no se sabe que pasa horizontalmente. La lectura de porcentajes sobre el total de la tabla es demasiado generalista. Entonces una forma que resuelve estas cuestiones es la aplicación de la regla de Zeisel. Zeisel estableció que la lectura de porcentajes se debía hacer de la siguiente forma: "Calcular los porcentajes en el sentido de la variable (considerada) independiente y leerlos en el sentido de la variable (considerada) dependiente". Según el convenio de situar la variable dependiente en filas y la variable independiente en columnas, entonces la regla de Zeisel también se puede expresar: "Calcular los porcentajes en el sentido de las columnas y leerlos (compararlos) en el sentido de las filas".

Tabla 52 Lectura de porcentajes según la regla de Zeisel.

Estado civil según el sexo

			Sexo Varón	Mujer	Total fila
Estado civil	*Soltero/a*	*n*	36	41	77
		%TF	46,8%	53,2%	100,0%
		%TC	73,5%	82,0%	77,8%
		%TT	36,4%	41,4%	77,8%
	Casado/a	*n*	6	3	9
		%TF	66,7%	33,3%	100,0%
		%TC	12,2%	6,0%	9,1%
		%TT	6,1%	3,0%	9,1%
	Pareja	*n*	7	6	13
		%TF	53,8%	46,2%	100,0%
		%TC	14,3%	12,0%	13,1%
		%TT	7,1%	6,1%	13,1%
	Total columna	*n*	49	50	99
		%TF	49,5%	50,5%	100,0%
		%TC	100,0%	100,0%	100,0%
		%TT	49,5%	50,5%	100,0%

Lectura de Zeisel.

Aplicación de la regla de Zeisel a la fila de "soltero/a".

Se calculan los porcentajes en el sentido de las columnas y son: el 73,5% de los varones y el 82,0% de las mujeres, dicen estar solteros. Entonces la lectura sería:

De los 49 varones, el 73,5% dicen estar solteros, y de las 50 mujeres el 82,0% dicen estar solteras.

De forma abreviada: del total de varones, el 73,5% dice estar soltero, frente al 82,0% de las mujeres.

Por lo tanto se puede decir que, de forma relativa, hay una diferencia de 9,5 puntos porcentuales. Como es descriptivo, no se puede decir nada más.

Para la segunda fila sería: del total de varones, el 12,2% dice estar casados, frente al 6,0% de las mujeres.

Y para la tercera: del total de varones, el 14,3% dice vivir en pareja, frente al 12,0% de las mujeres.

En la Tabla 53 se muestran algunos casos de las tablas de doble entrada que se pueden presentar como raros. Para otras observaciones ver la Tabla 26.

Tabla 53 Simulación de posibles casos extremos de las tablas de doble entrada.

		X			
		x_1	X_2	x_3	*TF*
Y	y_1	1	1	1	3
	%TF	33,3%	33,3%	33,3%	
	%TC	10,0%	10,0%	10,0%	
	y_2	1	1	2	4
	%TF	25,0%	25,0%	50,0%	
	%TC	10,0%	10,0%	20,0%	
	y_3	8	8	7	23
	TC	10	10	10	30

Cuando las frecuencias absolutas son bajas, se pueden presentar frecuencias relativas altas.

En la primera fila, con sólo tres casos, en cada columna hay un 33,3% de los casos.

En la segunda fila, con sólo 4 casos, en las dos primeras filas hay un 25,0% respectivamente, mientras que en la tercera hay un 50,0%. Sólo con un caso más se presenta una diferencia de 25 puntos porcentuales.

La regla de Zeisel es una forma de dar respuesta a estas situaciones. Si para el suceso elemental y_1 se calculan los porcentajes sobre el total de las columnas, los porcentajes se reducen al 10,0% y en la fila del suceso elemental y_2, la diferencia ya no es de 25 puntos porcentuales, sino de 10 puntos porcentuales.

Estos casos extremos, no se consideran errores, sino información incompleta.

		X			
		x_1	x_2	x_3	*TF*
Y	y_1	8	100	10	118
	%TF	6,8%	84,7%	8,5%	
	%TC	80,0%	10,0%	10,0%	
	y_2	1	100	50	151
	y_3	1	800	40	841
	TC	10	1.000	100	1.110

Cuando los totales de columnas presentan grandes diferencias, se puede dar la aparente paradoja de que frecuencias absolutas menores, presenten frecuencias relativas mayores.

En la fila del suceso elemental y_1, los porcentajes sobre el total de filas son 6,8%, 84,7% y 8,5% a cada categoría respectiva de *X*. pero si se aplica la regla de Zeisel, los 8 casos son el 80,0% sobre su total de columna (TC=10), mientras que los 100 casos de la segunda categoría de *X* representan el 10,0% (TC=1.000). Una posible lectura significaría que si la base de la columna de la categoría x_1 hubiese sido mayor, la frecuencia absoluta también lo habría sido, pero esto se asume.

Otra forma de presentar la tabla de doble entrada o distribución conjunta de frecuencias es en forma de columna o matriz de datos Tipo I, así la presentación de las tablas anteriores sería como en la Tabla 54.

Tabla 54 Presentación de la tabla de doble entrada en formato columna.

Tabla de doble entrada

Variable Y por X*

		X	
		x_1	x_2
Y	y_1	n_{11}	n_{12}
	y_2	n_{21}	n_{22}

Variable Y por X, según z_1

z_1		X	
		x_1	x_2
Y	y_1	n_{11}	n_{12}
	y_2	n_{21}	n_{22}

Variable Y por X, según z_2

z_2		X	
		x_1	x_2
Y	y_1	n_{11}	n_{12}
	y_2	n_{21}	n_{22}

Variable Y por X, según w_1 y z_1

w_1, z_1		X	
		x_1	x_2
y	y_1	n_{11}	n_{12}
	y_2	n_{21}	n_{22}

Variable Y por X, según w_1 y z_2

w_1, z_2		X	
		x_1	x_2
y	y_1	n_{11}	n_{12}
	y_2	n_{21}	n_{22}

Variable Y por X, según w_2 y z_1

w_2, z_1		X	
		x_1	x_2
y	y_1	n_{11}	n_{12}
	y_2	n_{21}	n_{22}

Variable Y por X, según w_2 y z_2

w_2, z_2		X	
		x_1	x_2
Y	y_1	n_{11}	n_{12}
	y_2	n_{21}	n_{22}

Formato columna

Y	X	n_{yx}
1	1	n_{11}
1	2	n_{12}
2	1	n_{21}
2	2	n_{22}

Z	Y	X	n_{yxz}
1	1	1	N_{111}
1	1	2	N_{121}
1	2	1	N_{211}
1	2	2	N_{221}
2	1	1	N_{112}
2	1	2	N_{122}
2	2	1	N_{212}
2	2	2	N_{222}

W	Z	Y	X	n_{yxwz}
1	1	1	1	n_{1111}
1	1	1	2	n_{1211}
1	1	2	1	n_{2111}
1	1	2	2	n_{2211}
1	2	1	1	n_{1112}
1	2	1	2	n_{1212}
1	2	2	1	n_{2112}
1	2	2	2	n_{2212}
2	1	1	1	n_{1121}
2	1	1	2	n_{1221}
2	1	2	1	n_{2121}
2	1	2	2	n_{2221}
2	2	1	1	n_{1122}
2	2	1	2	n_{1222}
2	2	2	1	n_{2122}
2	2	2	2	n_{2222}

9 Concepto de probabilidad y probabilidad condicionada (variables discretas)

Antes de empezar con los contrastes de Hipótesis es necesario ver previamente los conceptos de probabilidad, probabilidad condicionada y distribución de densidad de probabilidad de las variables z, t, θ^2 y F.

Para introducir el concepto de probabilidad se distingue entre espacio muestral discreto y espacio muestral continuo. En el primero hay un enfoque *interpretativo* y otro *formal*. En el enfoque interpretativo se intenta definir explícitamente qué es probabilidad y ofrecer normas apropiadas que permitan atribuir probabilidades a diferentes sucesos. El enfoque formal propone leyes formales que deben ser respetadas por ciertos valores numéricos para que puedan ser llamados probabilidades.

A su vez el enfoque interpretativo se clasifica en: punto de vista objetivo y el subjetivo. Y a su vez el primero se clasifica en: punto de vista objetivo clásico (o "*a priori*") y punto de vista objetivo frecuencista (o "*a posteriori*"). Por el interés de los temas desarrollados en este manual, en lo sucesivo se va a considerar sólo el punto de vista frecuencista o "*a posteriori*" (Tabla 55).

Tabla 55 Cuadro de las probabilidades.

	Espacio muestral	Enfoque	Punto de vista	
Probabilidad	Discreto	Interpretativo	Objetivo	Clásico ("*a priori*")
				Frecuencista ("*a posteriori*")
			Subjetivo	(No se trata).
		Formal		(No se trata).
	Continuo	Z T θ^2 F		

9.1 Punto de vista objetivo clásico ("*a priori*")

Consideremos el experimento aleatorio de "lanzar un dado al aire" con el espacio muestral *E* asociado compuesto por los seis sucesos elementales: s_1 (cara con un punto), s_2 (cara con dos puntos), … s_6 (cara con seis puntos). Suponiendo que el dado no está trucado, atribuimos que la probabilidad de cada suceso $s_1, s_2, .. s_6$ de aparecer es igual a 1/6. Este valor se interpreta como el cociente entre s_i (el suceso favorable s_i) que es 1 y el número de elementos de *E* (sucesos posibles) que es 6 (Ejemplos en Tabla 56).

Tabla 56 Ejemplos de probabilidades "*a priori*" I.

Sea un dado de seis caras que su espacio muestral *E* es $s_1 = 1; s_2 = 2; s_3 = 3; \ldots s_6 = 6$. Todos con la misma oportunidad de salir.	Probabilidad de salir un 1	Probabilidad de salir un 2
	$P_{(S_1)} \mid \frac{1}{6}$	$P_{(S_2)} \mid \frac{1}{6}$
	Probabilidad de salir un 1 ó 2	
	$P_{(S_1 \cong S_2)} \mid P_{(S_1)} \; 2 \; P_{(S_2)} \mid \frac{1}{6} 2 \frac{1}{6} \mid \frac{2}{6} \mid \frac{1}{3}$	
Sea un dado de seis caras que su espacio muestral *E* es $s_1 = 1; s_2 = 2; s_3 = 3; \ldots s_6 = 6$. Todos con la misma oportunidad de salir. Se realiza el experimento de tirar dos veces el dado. El espacio muestral *E* tiene 36 sucesos elementales compuestos: 9 de dos caras pares; 9 de dos caras impares, y 18 con una cara par y la otra impar.	Probabilidad de salir dos caras pares	Probabilidad de salir dos caras impares
	$P_{(S_1)} \mid \frac{9}{36} \mid \frac{1}{4}$	$P_{(S_2)} \mid \frac{9}{36} \mid \frac{1}{4}$
	Probabilidad de salir dos caras pares o dos impares	
	$P_{(S_1 \cong S_2)} \mid P_{(S_1)} \; 2 \; P_{(S_2)} \mid \frac{9}{36} 2 \frac{9}{36} \mid \frac{18}{36} \mid \frac{2}{4}$	

En general, sea un espacio muestral *E*, con *n* resultados posibles, todos ellos con la misma oportunidad de aparecer y de los cuales n_a constituye el suceso o subconjunto *a*, n_b constituye el suceso o subconjunto *b*, n_i constituye el suceso o subconjunto *i*, y $n_a + n_b + \ldots + n_i = N$. Entonces, diremos que n_a/N, es la probabilidad del suceso *a*, que n_b/N, es la probabilidad del suceso *b*, … n_i/N, es la probabilidad del suceso *i*. Esto es, dado un experimento aleatorio con un espacio muestral *E*, cuyos sucesos posibles tienen todos la misma oportunidad de aparecer, la probabilidad de cualquier suceso *s* de *E*, es igual al cociente entre el número de elementos de *s* (número de resultados favorables) y el número de elementos de *E* (número total de resultados, los posibles).

Esta definición dada se dice que es clásica, por ser una de las primeras que se propusieron y se le atribuye a Laplace (1749-1827).

Siendo N el número de resultados posibles del espacio muestral *E* en un experimento aleatorio, las probabilidades que se le atribuyen a los sucesos cumplen estas tres condiciones:

1. La probabilidad de la unión de dos sucesos mutuamente excluyentes (la aparición de uno implica la imposibilidad de la aparición del otro) es igual a la suma de las dos probabilidades (Tabla 56).
2. La probabilidad del suceso seguro (o sea obtener uno de los resultados del espacio muestral *E*) vale 1. En efecto si *E* tiene seis sucesos elementales la probabilidad que salga uno cualquiera de los seis es 6/6 = 1. Si se repite dos veces el experimento de tirar el dado, la probabilidad de que las dos caras sean pares, impares o una par y otra impar es 36/36 = 1.
3. La probabilidad de cualquier suceso *s* (compuesto de n_i elementos de *E*) será no negativa. Dicha probabilidad vale n_i/N y todos ellos son positivos.

La Tabla 57 muestra los ejemplos de la primera y la segunda condición.

<table>
<tr><td colspan="4">Tabla 57 Ejemplos de probabilidades “a priori” II.</td></tr>
<tr><td rowspan="6">Sea un dado de seis caras que su espacio muestral E es s1 = 1; s2 = 2; s3 = 3; … s6 = 6. Todos con la misma oportunidad de salir.</td><td>Probabilidad de salir un 1</td><td>Probabilidad de salir un 2</td><td>Probabilidad de salir un 3</td></tr>
<tr><td>$P_{(S_1)} \mid \frac{1}{6}$</td><td>$P_{(S_2)} \mid \frac{1}{6}$</td><td>$P_{(S_3)} \mid \frac{1}{6}$</td></tr>
<tr><td>Probabilidad de salir un 4</td><td>Probabilidad de salir un 5</td><td>Probabilidad de salir un 6</td></tr>
<tr><td>$P_{(S_4)} \mid \frac{1}{6}$</td><td>$P_{(S_5)} \mid \frac{1}{6}$</td><td>$P_{(S_6)} \mid \frac{1}{6}$</td></tr>
<tr><td colspan="3">Probabilidad de salir un 1 ó 2 ó 3 ó 4 ó 5 ó 6</td></tr>
<tr><td colspan="3">$P_{(S_1 \cong S_2 \cong S_3 \cong S_4 \cong S_5 \cong S_6)} \mid P_{(S_1)} 2 P_{(S_2)} 2 P_{(S_3)} 2 P_{(S_4)} 2 P_{(S_5)} 2 P_{(S_6)} \mid \frac{1}{6} 2 \frac{1}{6} 2 \frac{1}{6} 2 \frac{1}{6} 2 \frac{1}{6} 2 \frac{1}{6} \mid \frac{6}{6} \mid 1$</td></tr>
<tr><td rowspan="4">Sean dados de seis caras que su espacio muestral E es s1 = 1; s2 = 2; s3 = 3; … s6 = 6. Todos con la misma oportunidad de salir. Se realiza el experimento de tirar dos dados. El espacio muestral E tiene 36 sucesos elementales compuestos: 9 de dos caras pares; 9 de dos caras impares, y 18 con una cara par y la otra impar.</td><td>Probabilidad de salir dos caras pares.</td><td>Probabilidad de salir dos caras impares.</td><td>Probabilidad de salir una cara par y otra impar</td></tr>
<tr><td>$P_{(S_1)} \mid \frac{9}{36} \mid \frac{1}{4}$</td><td>$P_{(S_2)} \mid \frac{9}{36} \mid \frac{1}{4}$</td><td>$P_{(S_3)} \mid \frac{18}{36} \mid \frac{2}{4}$</td></tr>
<tr><td colspan="3">Probabilidad de que salgan dos pares o dos impares o una par y la otra impar.</td></tr>
<tr><td colspan="3">$P_{(S_1 \cong S_2 \cong S_3)} \mid P_{(S_1)} 2 P_{(S_2)} 2 P_{(S_3)} \mid \frac{9}{36} 2 \frac{9}{36} 2 \frac{18}{36} \mid \frac{1}{4} 2 \frac{1}{4} 2 \frac{2}{4} \mid \frac{4}{4} \mid 1$</td></tr>
</table>

9.2 Punto de vista objetivo frecuencista (“*a posteriori*”).

Si se lanza un dado N veces consecutivas. Sean $n_1, n_2, \ldots n_6$ el número de veces que aparece la cara con un punto, la de dos puntos, … la de seis puntos y además $n_1+n_2+\ldots+n_6 = N$, diremos que la probabilidad de aparición del suceso *cara con i puntos* siendo $i = 1, 2, \ldots 6$, es el límite al que tiende n_i/N cuando N tiende al infinito.

Generalizando el ejemplo del dado, si se realiza el experimento aleatorio de lanzarlo N veces consecutivas. Sea n_a el número de veces que aparece el suceso a, n_b el número de veces que aparece el suceso b, n_i el número de veces que aparece el suceso i, y $n_a + n_b + \ldots + n_i = N$. Entonces, diremos que las probabilidades de los sucesos $a, b, \ldots i$ son los límites hacia los que tienden, respectivamente, los cocientes n_a/N, n_b/N, … n_i/N, cuando N tiende a infinito, es decir, la probabilidad es definida como el límite de una proporción o frecuencia relativa y por eso la denominación de *frecuencista.* Cualquier n_i son los hechos favorables y N son los hechos posibles. Si se lanza un dado un número muy grande de veces, simulado con un ordenador, con un espacio muestral E de seis sucesos elementales: s_1 (cara con un punto), s_2 (cara con dos puntos), … s_6 (cara con seis puntos), y sea n_1 el número de veces que aparece el suceso elemental s_1 (cara con un punto), n_2 el número de veces que aparece el suceso elemental s_2 (cara con dos puntos), … n_6 el número de veces que aparece el suceso elemental s_6 (cara con seis puntos), el resultado para sucesivas N’s se muestra en la Tabla 58 (Simulación con ordenador).

Tabla 58 Tendencia de P cuando N tiende a infinito.

s_i	$P_{(S_i)} \mid \frac{n_i}{N}$	$P_{(S_i)} \mid \frac{n_i}{N}$	$P_{(S_i)} \mid \frac{n_i}{N}$	$P_{(S_i)} \mid \frac{n_i}{N}$	$P_{(S_i)} \mid \frac{n_i}{N}$
1	0,1630	0,1668	0,1666	0,1665	0,1667
2	0,1610	0,1668	0,1665	0,1666	0,1666
3	0,1570	0,1664	0,1668	0,1668	0,1666
4	0,1690	0,1664	0,1667	0,1667	0,1667
5	0,1770	0,1667	0,1667	0,1667	0,1667
6	0,1730	0,1669	0,1667	0,1666	0,1667
N	1.000	1.000.000	10.000.000	20.000.000	40.000.000

Siendo:

- s_i Suceso elemental i; i = 1, 2, 3, 4, 5, 6.
- $P_{(Si)}$ Probabilidad del suceso elemental *i-ésimo*.
- n_i Hechos favorables o frecuencia del suceso elemental *i-ésimo*.
- N Hechos posibles o frecuencia total.

Se acepta como probabilidad de un suceso la frecuencia relativa del mismo. La probabilidad se acercará más a su valor cuanto mayor sea el número de veces que se repite el experimento.

Por las características de los hechos tratados en Sociología, es difícil considerar el punto de vista objetivo clásico y se va a considerar en lo sucesivo el punto de vista frecuencista. El comportamiento humano difiere del comportamiento de los dados, los naipes, las monedas y las bolas de colores. El género en cuanto a sexo de una persona puede ser *varón* o *mujer*, pero no quiere decir que la probabilidad clásica o "*a priori*" de tener o encontrar uno u otro sea de 0,50. El estado civil de una sociedad considerada Occidental, puede ser: *soltero/a, casado/a, en pareja, separado/a, divorciado/a y viudo/a*, la probabilidad de cada uno de estos sucesos elementales no tiene porqué ser de 1/6. Entonces se opera con las probabilidades después de conocer la frecuencia relativa, la probabilidad "*a posteriori"* (Tabla 59).

Tabla 59 Ejemplos de probabilidades "*a posteriori*".

<table>
<tr><td rowspan="10">Sea un grupo de personas del que se sabe el estado civil: s_1 = solteros, s_2 = casados, s_3 = en pareja, s_4 = divorciados, s_5 = separados y s_6 = viudos.</td><td colspan="2">Distribución del grupo de personas</td></tr>
<tr><td colspan="2">s_1 aparece n_1 veces = 30; s_2 aparece n_2 veces = 55; s_3 aparece n_3 veces = 15; s_4 aparece n_4 veces = 20, s_5 aparece n_5 veces = 15; s_6 aparece n_6 veces = 15
Entonces:
$N = n_1 + n_2 + n_3 + n_4 + n_5 + n_6 = 30 + 55 + 15 + 20 + 15 + 15 = 150$</td></tr>
<tr><td>Probabilidad de ser soltero:</td><td>Probabilidad de ser casado:</td></tr>
<tr><td>$P_{(S_1)} \mid \frac{n_1}{N} \mid \frac{30}{150} \mid \frac{1}{5}$</td><td>$P_{(S_2)} \mid \frac{n_2}{N} \mid \frac{55}{150} \mid \frac{11}{30}$</td></tr>
<tr><td>Probabilidad de ser pareja:</td><td>Probabilidad de ser divorciado:</td></tr>
<tr><td>$P_{(S_3)} \mid \frac{n_3}{N} \mid \frac{15}{150} \mid \frac{1}{10}$</td><td>$P_{(S_4)} \mid \frac{n_4}{N} \mid \frac{20}{150} \mid \frac{2}{15}$</td></tr>
<tr><td>Probabilidad de ser separado:</td><td>Probabilidad de ser viudo:</td></tr>
<tr><td>$P_{(S_5)} \mid \frac{n_5}{N} \mid \frac{15}{150} \mid \frac{1}{10}$</td><td>$P_{(S_6)} \mid \frac{n_6}{N} \mid \frac{15}{150} \mid \frac{1}{10}$</td></tr>
<tr><td colspan="2">Probabilidad de ser soltero ó casado ó en pareja ó divorciado ó separado ó viudo</td></tr>
<tr><td colspan="2">$P_{(S_1 \equiv S_2 \equiv S_3 \equiv S_4 \equiv S_5 \equiv S_6)} \mid P_{(S_1)} \; 2 \; P_{(S_2)} \; 2 \; P_{(S_3)} \; 2 \; P_{(S_4)} \; 2 \; P_{(S_5)} \; 2 \; P_{(S_6)} \mid$
$\frac{1}{5} 2 \frac{11}{30} 2 \frac{1}{10} 2 \frac{2}{15} 2 \frac{1}{10} 2 \frac{1}{10} \mid \frac{6}{30} 2 \frac{11}{30} 2 \frac{3}{30} 2 \frac{4}{30} 2 \frac{3}{30} 2 \frac{3}{30} \mid \frac{30}{30} \mid 1$</td></tr>
</table>

En el punto de vista "*a posteriori*" que se define la probabilidad como el límite de una frecuencia relativa, se cumplen las tres propiedades que se vieron en el punto de vista "*a priori*".

1. La probabilidad de la unión de dos sucesos mutuamente excluyentes (la aparición de uno implica la imposibilidad de la aparición del otro) es igual a la suma de las dos probabilidades.

En el ejemplo del dado, simulado con un ordenador, con un espacio muestral *E* de seis sucesos elementales: s_1 (cara con un punto), s_2 (cara con dos puntos), … s_6 (cara con seis puntos). Suponiendo que el dado no está trucado, después de lanzarlo 39.243 veces el resultado es el que se ve en la Tabla 60. Si el experimento es tirar dos dados, después de

lanzarlos 23.679 veces el resultado se muestra también en la Tabla 60 y Tabla 61.

Tabla 60 Experimento con dados I.

Tirar un dado.			Tirar dos dados.									
s_i	n_i	$P_{(si)}$ ó f	s_i	Dado 1	Dado 2	n_i	$P_{(si)}$ ó f	s_i	Dado 1	Dado 2	n_i	$P_{(si)}$ ó f
1	6.549	0,167	1	1	1	650	0,0275	19	4	1	672	0,0284
2	6.556	0,167	2	1	2	678	0,0286	20	4	2	696	0,0294
3	6.387	0,163	3	1	3	621	0,0262	21	4	3	664	0,0280
4	6.628	0,169	4	1	4	664	0,0280	22	4	4	645	0,0272
5	6.666	0,170	5	1	5	636	0,0269	23	4	5	704	0,0297
6	6.457	0,165	6	1	6	628	0,0265	24	4	6	637	0,0269
N	39.243	1,000	7	2	1	641	0,0271	25	5	1	629	0,0266
Siendo:			8	2	2	643	0,0272	26	5	2	668	0,0282
s_i	Suceso *i*-ésimo. Una tirada o dos tiradas.		9	2	3	648	0,0274	27	5	3	685	0,0289
$P_{(si)}$	Probabilidad del suceso i-ésimo.		10	2	4	654	0,0276	28	5	4	646	0,0273
f_i	Frecuencia relativa del suceso *i-ésimo*.		11	2	5	626	0,0264	29	5	5	637	0,0269
n_i	Hechos favorables o frecuencia del suceso elemental i.		12	2	6	680	0,0287	30	5	6	653	0,0276
N	Hechos posibles o frecuencia total.		13	3	1	622	0,0263	31	6	1	663	0,0280
			14	3	2	661	0,0279	32	6	2	670	0,0283
			15	3	3	646	0,0273	33	6	3	682	0,0288
			16	3	4	686	0,0290	34	6	4	633	0,0267
			17	3	5	711	0,0300	35	6	5	656	0,0277
			18	3	6	662	0,0280	36	6	6	682	0,0288
								N			23.679	1,0000

Tabla 61 Ejemplos con dados II.

Sea un dado de seis caras que su espacio muestral *E* es $s_1 = 1; s_2 = 2; s_3 = 3; \ldots s_6 = 6$. Todos con la misma oportunidad de salir, después de 39.243 lanzamientos (Tabla 60).	Prob. de salir un 1	Prob. de salir un 2
	$P_{(S_1)} \mid \frac{6.549}{39.243} \mid 0{,}167$	$P_{(S_2)} \mid \frac{6.556}{39.243} \mid 0{,}167$
	Probabilidad de que salga un 1 ó 2	
	$P_{(S_1 \cong S_2)} \mid P_{(S_1)} \; 2 \; P_{(S_2)} \mid \frac{6.549}{39.243} \; 2 \; \frac{6.556}{39.243} \mid \frac{13.105}{39.243} \mid 0{,}334 \mid 0{,}167 \; 2 \; 0{,}167$	
Sean dados de seis caras que su espacio muestral *E* es $s_1 = 1; s_2 = 2; s_3 = 3; \ldots s_6 = 6$. Todos con la misma oportunidad de salir. Se realiza el experimento de tirar dos dados. El espacio muestral *E* tiene 36 sucesos elementales compuestos: 9 de dos caras pares; 9 de dos caras impares, y 18 con una cara par y la otra impar (Tabla 60).	Probabilidad de salir el 2 y el 4).	Probabilidad de salir el 1 y el 3).
	$P_{(S_{10})} \mid \frac{654}{23.679} \mid 0{,}0276$	$P_{(S_3)} \mid \frac{621}{23.679} \mid 0{,}0262$
	Probabilidad de que salgan 2 y 4 ó 1 y 3.	
	$P_{(S_{10} \cong S_3)} \mid P_{(S_{10})} \; 2 \; P_{(S_3)} \mid \frac{654}{23.679} \; 2 \; \frac{621}{23.679} \mid \frac{1.275}{23.679} \mid 0{,}0538 \mid 0{,}0276 \; 2 \; 0{,}0262$	

2. La probabilidad del suceso seguro (o sea obtener uno de los resultados del espacio muestral *E*) vale 1. En efecto si *E* tiene seis sucesos y realizamos 39.243 lanzamientos del dado, la probabilidad que salga uno cualquiera de los resultados posibles es 39.243/39.243 = 1. Si se hace otro experimento de tirar 23.679 veces dos dados, la probabilidad de que las dos caras sean pares, impares o una par y otra impar es 23.679/23.679 = 1. Según la primera condición,

Tabla 62 Ejemplo con dados III.

Sea un dado de seis caras que su espacio muestral *E* es $s_1 = 1;\ s_2 = 2;\ s_3 = 3;\ \ldots\ s_6 = 6$. Todos con la misma oportunidad de salir. Se realiza el experimento de tirar el dado 39.243 veces. El 1 ha salido 6.549 veces; el 2 aparece 6.546 veces, el 3, 6.387; el 4, 6.628; el 5, 6.666, y el 6, 6.457 (Tabla 60).

Probabilidad de salir un 1	Probabilidad de salir un 2	Probabilidad de salir un 3
$P_{(S_1)} \mid \frac{6.549}{39.243}$	$P_{(S_2)} \mid \frac{6.546}{39.243}$	$P_{(S_3)} \mid \frac{6.387}{39.243}$
Probabilidad de salir un 4	**Probabilidad de salir un 5**	**Probabilidad de salir un 6**
$P_{(S_4)} \mid \frac{6.628}{39.243}$	$P_{(S_5)} \mid \frac{6.666}{39.243}$	$P_{(S_6)} \mid \frac{6.457}{39.243}$

Probabilidad de salir un 1 ó 2 ó 3 ó 4 ó 5 ó 6

$$P_{(S_1 \cong S_2 \cong S_3 \cong S_4 \cong S_5 \cong S_6)} \mid P_{(S_1)} 2\, P_{(S_2)} 2\, P_{(S_3)} 2\, P_{(S_4)} 2\, P_{(S_5)} 2\, P_{(S_6)} \mid \frac{6.549}{39.243} 2 \frac{6.546}{39.243} 2 \frac{6.387}{39.243} 2 \frac{6.628}{39.243} 2 \frac{6.666}{39.243} 2 \frac{6.457}{39.243} \mid \frac{39.243}{39.243} \mid 1$$

Sean dados de seis caras que su espacio muestral *E* es $s_1 = 1;\ s_2 = 2;\ s_3 = 3;\ \ldots\ s_6 = 6$. Todos con la misma oportunidad de salir. Se realiza el experimento de tirar dos dados. El espacio muestral *E* tiene 36 sucesos elementales compuestos: 9 de dos caras pares; 9 de dos caras impares, y 18 con una cara par y la otra impar (Tabla 60).

Probabilidad de salir dos caras pares	Probabilidad de salir dos caras impares	Probabilidad de salir una cara par y otra impar.
$P_{(S_1)} \mid \frac{5.940}{23.679}$	$P_{(S_2)} \mid \frac{5.837}{23.679}$	$P_{(S_3)} \mid \frac{11.902}{23.679}$

Probabilidad de salir dos caras pares, dos impares o una par y otra impar.

$$P_{(S_1 \cong S_2 \cong S_3)} \mid P_{(S_1)} 2\, P_{(S_2)} 2\, P_{(S_3)} \mid \frac{5.940}{23.679} 2 \frac{5.837}{23.679} 2 \frac{11.902}{23.679} \mid \frac{23.679}{23.679} \mid 1$$

3. La probabilidad de cualquier suceso *s* (compuesto de n_i elementos de *E*) será no negativa. Dicha probabilidad vale n_i/N y todas ellas son positivas.

Tabla 63 Ejemplos de probabilidad frecuencista de la Tabla 51.

Sea los sucesos elementales:

s_s: *soltero/a,*
s_c: *casado/a,*
s_p: *pareja,*

s_v: varón
s_m: mujer

$$P_{(S_s)} \mid \frac{\text{hechos favorables}}{\text{hechos posibles}} \mid \frac{77}{99} \mid 0{,}7778$$

$$P_{(S_c)} \mid \frac{\text{hechos favorables}}{\text{hechos posibles}} \mid \frac{9}{99} \mid 0{,}0909$$

$$P_{(S_p)} \mid \frac{\text{hechos favorables}}{\text{hechos posibles}} \mid \frac{13}{99} \mid 0{,}1313$$

$$P_{(S_v)} \mid \frac{\text{hechos favorables}}{\text{hechos posibles}} \mid \frac{49}{99} \mid 0{,}4949$$

$$P_{(S_m)} \mid \frac{\text{hechos favorables}}{\text{hechos posibles}} \mid \frac{50}{99} \mid 0{,}5050$$

Tabla 63 Ejemplos de probabilidad frecuencista de la Tabla 51.

Probabilidad de la intersección de dos sucesos,

$$P_{(S_s \sim S_v)} \mid \frac{\text{hechos favorables}}{\text{hechos posibles}} \mid \frac{36}{99} \mid 0{,}3636$$

$$P_{(S_s \sim S_m)} \mid \frac{\text{hechos favorables}}{\text{hechos posibles}} \mid \frac{41}{99} \mid 0{,}4141$$

$$P_{(S_c \sim S_v)} \mid \frac{\text{hechos favorables}}{\text{hechos posibles}} \mid \frac{6}{99} \mid 0{,}0606$$

$$P_{(S_c \sim S_m)} \mid \frac{\text{hechos favorables}}{\text{hechos posibles}} \mid \frac{3}{99} \mid 0{,}0303$$

$$P_{(S_p \sim S_v)} \mid \frac{\text{hechos favorables}}{\text{hechos posibles}} \mid \frac{7}{99} \mid 0{,}0707$$

$$P_{(S_s \sim S_m)} \mid \frac{\text{hechos favorables}}{\text{hechos posibles}} \mid \frac{6}{99} \mid 0{,}0606$$

9.3 Probabilidad condicionada.

Este epígrafe es la base del análisis de las *tablas de contingencia.* En estas tablas se trata de ver la relación, asociación o dependencia entre variables a través de la relación, asociación o dependencia entre los sucesos elementales. Este análisis consiste en ver la independencia mejor que la dependencia y este aspecto es el que se desarrolla. El proceso se desarrolla a partir de la Tabla 51 que se reproduce en la Tabla 64.

Tabla 64 Tabla de doble entrada.

Estado civil según el sexo

		Sexo		
		Varón	*Mujer*	*T. fila*
Estado civil	*Soltero/a*	36	41	77
	Casado/a	6	3	9
	Pareja	7	6	13
	T. columna	49	50	99

En esta tabla, de 99 estudiantes distribuidos según el estado civil y el sexo, la probabilidad de que un estudiante seleccionado al azar sea soltero es

$$P_{(S_s)} \mid \frac{77}{99} \mid 0{,}7778$$

Si se impone ahora la condición de que sea varón, la probabilidad de ser *soltero* es

$$\frac{36}{49} \mid 0{,}7347$$

Los 36 elementos cumplen la condición *ser soltero y varón,* esto es, de la intersección de los sucesos $s_s \sim s_v$, siendo s_s el suceso estar soltero y s_v el suceso ser varón, y que 49 son los elementos del suceso s_v ser *varón*. Si ahora proponemos los sucesos $s_s \sim s_v$ y s_v como subconjuntos del conjunto de los 99 estudiantes, tenemos que

$$P_{(S_s \sim S_v)} \mid \frac{36}{99} \mid 0{,}3636$$

Y que

$$P_{(S_v)} \mid \frac{49}{99} \mid 0{,}4949$$

Por lo tanto, la probabilidad condicional de estar *soltero*, supuesto que se es *varón,* que hemos visto que valía 36/49, se puede expresar por

$$\frac{36/99}{49/99} \mid \frac{P_{(S_s \sim S_v)}}{P_{(S_v)}} \mid \frac{36}{49}$$

Es el porcentaje sobre el total de la columna y lo llamamos la probabilidad condicional de estar soltero supuesto que se es varón. Y de forma genérica se define como: la probabilidad condicional (o condicionada) de *A* supuesto *B*, y se designa por $P_{(A/B)}$, y con la expresión:

$P_{(A\|B)} \mid \frac{P_{(A \sim B)}}{P_{(B)}}$	Fórmula 33

De la misma manera, se define la probabilidad de B supuesto A por:

$P_{(B\|A)} \mid \frac{P_{(B \sim A)}}{P_{(A)}} \mid \frac{P_{(A \sim B)}}{P_{(A)}}$	Fórmula 34

Entonces:

$P_{(A\|B)} \mid \frac{P_{(A \sim B)}}{P_{(B)}},\ P_{(B)} \Delta P_{(A\|B)} \mid P_{(A \sim B)}$	Fórmula 35

Y

$P_{(B\|A)} \mid \frac{P_{(A \sim B)}}{P_{(A)}},\ P_{(A)} \Delta P_{(B\|A)} \mid P_{(A \sim B)}$	Fórmula 36

Entonces, según Fórmula 35 y Fórmula 36 como $P_{(A \sim B)} \mid P_{(A \sim B)}$, entonces,

$P_{(B)} \Delta P_{(A\|B)} \mid P_{(A)} \Delta P_{(B\|A)}$	Fórmula 37

9.4 Sucesos independientes.

Dos sucesos, *A* y *B,* son estadísticamente independientes (o de forma abreviada, independientes) sí, y sólo si, se verifica

$P_{(A \sim B)} \mid P_{(A)} \Delta P_{(B)}$	Fórmula 38

Teorema 1 Si dos sucesos A y B, verifican la Fórmula 38, entonces

$P_{(A\|B)} \mid P_{(A)}$	Fórmula 39

$P_{(B\|A)} \mid P_{(B)}$	Fórmula 40

En efecto,

$P_{(A\|B)} \mid \frac{P_{(A\sim B)}}{P_{(B)}} \mid \frac{P_{(A)}\Delta P_{(B)}}{P_{(B)}} \mid P_{(A)}$	Fórmula 41

Según Fórmula 33 y Fórmula 38, y

$P_{(B\|A)} \mid \frac{P_{(A\sim B)}}{P_{(A)}} \mid \frac{P_{(A)}\Delta P_{(B)}}{P_{(A)}} \mid P_{(B)}$	Fórmula 42

Según Fórmula 34 y Fórmula 38,

Teorema 2 Si dos sucesos A y B, verifican las relaciones Fórmula 39 y Fórmula 40, entonces necesariamente verifican la relación Fórmula 38

En efecto,

$P_{(A\sim B)} \mid P_{(A\|B)}\Delta P_{(B)} \mid P_{(A)}\Delta P_{(B)}$	Fórmula 43

Según Fórmula 33, Fórmula 35 y Fórmula 39

$P_{(A\sim B)} \mid P_{(B\|A)}\Delta P_{(A)} \mid P_{(B)}\Delta P_{(A)}$	Fórmula 44

Según Fórmula 34, Fórmula 36 y Fórmula 40

Según el Teorema 1 y el Teorema 2, como *corolario,* se puede decir que dos sucesos *A* y *B* son independientes si, y sólo si, $P_{(A|B)} = P_{(A)}$ o si $P_{(B|A)} = P_{(B)}$. Como ejemplo se presenta una simulación sobre la Tabla 64 en la Tabla 65.

Tabla 65 Ejemplo de sucesos independientes.

Estado civil según el sexo

		Sexo		
		Varón	*Mujer*	*T. F.*
Estado civil	*Soltero/a*	20	20	40
	Casado/a	20	20	40
	Pareja	20	20	40
	T. C.	60	60	120

Como se ha planteado anteriormente,

$$P_{(S_s|S_v)} \mid \frac{P_{(S_s \sim S_v)}}{P_{(S_v)}} \mid \frac{20/120}{60/120} \mid \frac{20}{60} \mid \frac{1}{3}$$

$$P_{(S_v|S_s)} \mid \frac{P_{(S_s \sim S_v)}}{P_{(S_s)}} \mid \frac{20/120}{40/120} \mid \frac{20}{40} \mid \frac{1}{2}$$

Si los sucesos son independientes, se debe cumplir que:

$$P_{(S_s \sim S_v)} \mid P_{(S_s)} \Delta P_{(S_v)}, \frac{20}{120} \mid \frac{40}{120} \Delta \frac{60}{120},$$

$$\frac{20}{120} \mid \frac{2.400}{14.400} \text{ reduciendo } \frac{1}{6} \mid \frac{1}{6}$$

Entonces como: $P_{(S_s|S_v)} \Delta P_{(S_v)} \mid P_{(S_s \sim S_v)}$ y $P_{(S_v|S_s)} \Delta P_{(S_s)} \mid P_{(S_s \sim S_v)}$

Entonces se cumple la igualdad,

$$P_{(S_s|S_v)} \Delta P_{(S_v)} \mid P_{(S_v|S_s)} \Delta P_{(S_s)}, \frac{1}{3} \Delta \frac{60}{120} \mid \frac{1}{2} \Delta \frac{40}{120}$$

$$\frac{1}{3} \Delta \frac{60}{120} \mid \frac{1}{2} \Delta \frac{40}{120}, \frac{1}{3} \Delta \frac{1}{2} \mid \frac{1}{2} \Delta \frac{1}{3}, \frac{1}{6} \mid \frac{1}{6}$$

Se comprueba que los sucesos son independientes, pero no excluyentes (las celdas no están vacías), como se puede observar a simple vista en la tabla.

Si se comprueba la independencia sobre los valores originales de la Tabla 64 se obtiene la Tabla 66,

Tabla 66 Ejemplo de sucesos no independientes.

Estado civil según el sexo

		Sexo		
		Varón	*Mujer*	*T. F.*
Estado civil	*Soltero/a*	36	41	77
	Casado/a	6	3	9
	Pareja	7	6	13
	T. C.	49	50	99

Como se ha planteado anteriormente,

$$P_{(S_s|S_v)} \mid \frac{P_{(S_s \sim S_v)}}{P_{(S_v)}} \mid \frac{36/99}{49/99} \mid \frac{36}{49} \mid 0{,}7347$$

$$P_{(S_v|S_s)} \mid \frac{P_{(S_s \sim S_v)}}{P_{(S_s)}} \mid \frac{36/99}{77/99} \mid \frac{36}{77} \mid 0{,}4675$$

Como los sucesos no son independientes, no se cumple la igualdad:

$$P_{(S_s \sim S_v)} \prod P_{(S_s)} \Delta P_{(S_v)}, \frac{36}{99} \prod \frac{77}{99} \Delta \frac{49}{99},$$

$$\frac{36}{99} \prod \frac{3.773}{9.801}, 0{,}3636 \prod 0{,}3850$$

Entonces como: $P_{(S_s|S_v)} \Delta P_{(S_v)} \mid P_{(S_s \sim S_v)}$ y $P_{(S_v|S_s)} \Delta P_{(S_s)} \mid P_{(S_s \sim S_v)}$

Entonces se cumple la igualdad,

$$P_{(S_s|S_v)} \Delta P_{(S_v)} \mid P_{(S_v|S_s)} \Delta P_{(S_s)}, \frac{36}{49} \Delta \frac{49}{99} \mid \frac{36}{77} \Delta \frac{77}{99}$$

$$\frac{36 \Delta 49}{49 \Delta 99} \mid \frac{36 \Delta 77}{77 \Delta 99}, \frac{36}{99} \mid \frac{36}{99}$$

Lo que se comprueba es que no se puede confirmar que los sucesos son independientes, ni se puede confirmar que sean dependientes. Esta confirmación queda pendiente al contraste de Hipótesis de la tabla de contingencia.

9.5 Prueba de Bernoulli y distribución binomial

Se llama prueba de Bernoulli a toda realización de un experimento aleatorio en el que sólo son posibles dos resultados, arbitraria, pero tradicionalmente, llamados *éxito* y *fracaso* y que son mutuamente excluyentes. Son ejemplos:

- Lanzar una moneda y observar el resultado: cara "éxito" o cruz "fracaso" o viceversa.
- Que al elegir una pieza fabricada no sea defectuosa "éxito" o defectuosa "fracaso".
- Que la respuesta a una pregunta de examen sea correcta "éxito" o incorrecta "fracaso".
- Que al seleccionar una persona de una población vote a un partido político "éxito" o no "fracaso".
- Que al seleccionar una persona de una población tenga cierta enfermedad "éxito" o no "fracaso".
- Que al seleccionar una persona de una población compre un producto "éxito" o no "fracaso".

Se insiste que indistintamente se puede llamar éxito o fracaso a uno de los resultados, normalmente se llama éxito a aquel que se quiere conocer.

Sobre este espacio muestral de dos sucesos elementales: "éxito" "fracaso" , se define una variable aleatoria X que les atribuye un valor o código distinto a cada uno, que por sencillez y tradición y como se vio en la *codificación* serán 1 "éxito" y 0 "fracaso". Sea, además, p la probabilidad de obtener "éxito" en la prueba a la que se asigna el valor o código 1 y q (=(1-p)) la probabilidad de obtener "fracaso" en la prueba a la que se asigna un 0 (Tabla 67).

Tabla 67

Espacio muestral E	Variable X	Probabilidad $f(x) = P(X=x)$
Éxito	1	p
Fracaso	0	q

$f_{(x)} \mid P_{(X \mid x)} \mid p^x \Delta(1 4\ p)^{(14x)} \mid p^x \Delta q^{(14x)}$ $(x = 0, 1)$	Fórmula 45

Entonces la función de densidad o probabilidad de Bernoulli es

$$P_{(éxito)} \mid f_{(1)} \mid P_{(X \mid 1)} \mid p^1 \Delta q^{(141)} \mid p^1 \Delta q^0 \mid p$$

$$P_{(fracaso)} \mid f_{(0)} \mid P_{(X \mid 0)} \mid p^0 \Delta q^{(140)} \mid p^0 \Delta q^1 \mid q$$

Y la función de distribución o acumulada

$F_{(k)} \mid P_{(X \Omega k)} \mid \underset{x \mid 0}{\overset{k}{\text{——}}} p^x \Delta q^{(14x)}$	Fórmula 46

Para $k = 1$, tenemos,

$$F_{(1)} \mid P_{(X\Omega 1)} \mid \sum_{x\mid 0}^{1} p^{x}\,\Delta q^{(14x)} \mid p^{0}\,\Delta q^{(140)}\ 2\ p^{1}\,\Delta q^{(141)} \mid 1\Delta q\ 2\ p\Delta 1 \mid q\ 2\ p \mid 1$$

Este resultado era de esperar, ya que la suma de las probabilidades de los dos únicos resultados posibles (éxito / fracaso) en la prueba de Bernoulli tiene que ser 1.

Suponiendo ahora que se repite *n* veces la prueba independiente de Bernoulli tales que la probabilidad de éxito se mantiene constante en todas las pruebas. Es decir, suponiendo *N* variables aleatorias (v.a.) de Bernoulli, $X_1, X_2, X_3, \ldots X_n$, tales que $p_1 = p_2 = p_3 = \ldots = p_n = p$; y $q_1 = q_2 = q_3, = \ldots q_n = q$, y con $x_1, x_2, x_3, \ldots x_n = 0$ ó 1. Se considera la v.a. $X = x_1 + x_2 + \ldots + x_n$ (Tabla 68).

Tabla 68

v.a. de Bernoulli	Resultado	p	(1 - *p* = *q*)	Valor de v.a. *X*
X_1	$x_1, x_2, \ldots x_n$	*p*	*q*	0
X_2	$x_1, x_2, \ldots x_n$	*p*	*q*	1
X_3	$x_1, x_2, \ldots x_n$	*p*	*q*	2
...	…	…	…	…
X_N	$x_1, x_2, \ldots x_n$	*p*	*q*	*n*

Los resultados posibles de la v.a. *X* serán: 0 (fracaso en las *n* pruebas de Bernoulli, entonces todas las $x_i = 0$); 1 (éxito en una de las *n* pruebas de Bernoulli, entonces una $x_i = 1$); 2 (éxito en dos de las *n* pruebas de Bernoulli, entonces dos $x_i = 1$); n (éxito en todas las *n* pruebas de Bernoulli, entonces todas las $x_i = 1$). Entonces la variable aleatoria *X* contiene el número de éxitos en las *n* pruebas. Ahora se calcula su distribución.

La probabilidad de éxito *p,* se mantiene constante a lo largo de las *n* pruebas y, por lo tanto, se mantiene constante la probabilidad de fracaso *q* (= (1-*p*)). Como estas pruebas son independientes, la probabilidad conjunta de *x* éxitos y *n-x* fracasos, será igual al producto de las probabilidades de *x* éxitos y *n-x* fracasos,

$\Psi p0\Delta/p0\Delta..x$ veces$..\Delta/p0\beta\Delta\Psi q0\Delta/q0\Delta..(n\ 4\ x)$ veces$..\Delta/q0\beta \mid p^{x}\,\Delta q^{(n4x)}$	Fórmula 47

Los *x* éxitos pueden aparecer en las *n* pruebas de maneras distintas. Esto es, dados *n* elementos tomados de *x* en *x* se obtienen combinaciones que se llaman de orden *x*, entendiendo que dos combinaciones son distintas si difieren, al menos, en uno de sus elementos y vale,

$C_{n,x} \mid \dfrac{n!}{x!\Delta(n\ 4\ x)!}$	Fórmula 48

Se recuerda que: $C_{n,\,n} = 1$, $0! = 1$; $C_{n,\,0} = 1$; $C_{n,\,1} = n$.

$C_{n,\,n} = 1$	$\binom{n}{n} \mid C_{n,n} \mid \frac{n!}{n!\Delta(n-n)!} \mid \frac{n!}{n!\Delta(0)!} \mid \frac{n!}{n!\Delta 1} \mid \frac{n!}{n!} \mid 1$
$0! = 1$	Se acepta por convenio para poder mantener las igualdades anterior y posterior.
$C_{n,\,0} = 1$	$\binom{n}{0} \mid C_{n,0} \mid \frac{n!}{0!\Delta(n-0)!} \mid \frac{n!}{1\Delta(n)!} \mid \frac{n!}{n!} \mid 1$
$C_{n,\,1} = n$	$\binom{n}{1} \mid C_{n,1} \mid \frac{n!}{1!\Delta(n-1)!} \mid \frac{(n-1)!\Delta n}{(n-1)!} \mid n$

Entonces los *x* éxitos pueden aparecer en las *n* pruebas de

$\binom{n}{x} \mid C_{n,x} \mid \frac{n!}{x!\Delta(n-x)!}$	Fórmula 49

Maneras distintas, los modos distintos según los cuales pueden ser colocados los *x* éxitos en *n* posiciones distintas, que son las combinaciones de orden *x* a las que dan lugar *n* elementos según Fórmula 48. Entonces, la probabilidad de *x* éxitos en *n* pruebas vale:

$f(x) \mid P_{(X \mid x)} \mid \binom{n}{x} \Delta p^{x} \Delta q^{(n-x)} \mid \frac{n!}{x!\Delta(n-x)!} \Delta p^{x} \Delta q^{(n-x)}$, $(q = 1-p)$[60]	Fórmula 50

Y es la distribución o función de probabilidad binomial, o variable aleatoria binomial.

La función de distribución está dada por:

$F(k) \mid P_{(X \le k)} \mid \sum_{x=0}^{k} \binom{n}{x} p^{x} \Delta q^{(n-x)}$	Fórmula 51

Y para $k = n$, tenemos

$$F(n) \mid P_{(X \le n)} \mid \sum_{x=0}^{n} \binom{n}{x} p^{x} \Delta q^{(n-x)} \mid \binom{n}{0} p^{0} \Delta q^{(n-0)} + \binom{n}{1} p^{1} \Delta q^{(n-1)} + \cdots + \binom{n}{n-1} p^{(n-1)} \Delta q^{(n-(n-1))} + \binom{n}{n} p^{n} \Delta q^{(n-n)} \mid$$

$$\frac{n!}{0!\Delta(n-0)!} \Delta p^{0} \Delta q^{(n-0)} + \frac{n!}{1!\Delta(n-1)!} \Delta p^{1} \Delta q^{(n-1)} + \cdots + \frac{n!}{(n-1)!\Delta(n-(n-1))!} \Delta p^{(n-1)} \Delta q^{(n-(n-1))} + \frac{n!}{n!\Delta(n-n)!} \Delta p^{n} \Delta q^{(n-n)} \mid 1$$

La probabilidad suma de las probabilidades de los *n* resultados posibles: *0* éxitos, *1* éxito, … *n* éxitos, del experimento aleatorio, tiene que valer *1*.

Ejemplo: se lanza una moneda tres veces y admitimos que son independientes ya que un lanzamiento no influye en los demás lanzamientos. Consideraremos la probabilidad de

[60] Se denomina binomial, porque es el término general del binomio $(p+q)^n$.

obtener cara, por lo que se considera "éxito" el obtener "cara" (c) y "fracaso" el obtener "cruz" (z). Entonces, ¿Cuál es la probabilidad de no obtener ninguna cara? ¿Cuál es la probabilidad de obtener una cara? ¿Cuál es la probabilidad de obtener dos caras? ¿Cuál es la probabilidad de que las tres sean cara? y ¿Cuál es la probabilidad de obtener tres caras o menos? Consideramos que la moneda no está trucada y por lo tanto la probabilidad *p* de obtener cara es 0,50 (Tabla 69).

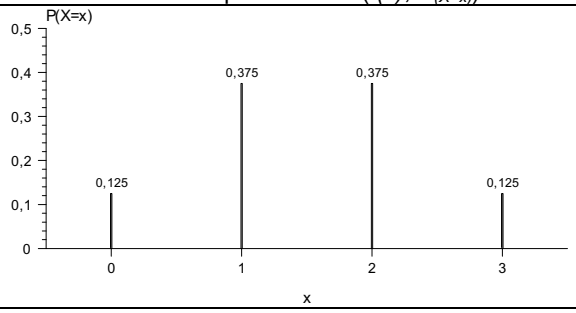
Tabla 69 Aplicación de la binomial.

Probabilidad de obtener cero caras (Fórmula 50).	Lanzamientos
$f(0) = P(X = 0) = \binom{3}{0} \Delta 0{,}5^0 \Delta 0{,}5^{(3-0)} = \frac{3!}{0!\Delta(3-0)!} \Delta(0{,}5^0 \Delta 0{,}5^{(3-0)}) = 1\Delta(0{,}5^0 \Delta 0{,}5^3) = 0{,}1250$	z, z, z (p^0 x q^3) *1* x (p^0 x q^3)
Cuando ninguno de los tres lanzamientos es cara, los tres son cruz. La probabilidad de que no salga cara (0,1250) es la misma que la de salir tres cruces.	
Probabilidad de obtener una cara (Fórmula 50).	c, z, z (p^1 x q^2) z, c, z (p^1 x q^2) z, z, c (p^1 x q^2) *3* x (p^1 x q^2)
$f(1) = P(X = 1) = \binom{3}{1} \Delta 0{,}5^1 \Delta 0{,}5^{(3-1)} = \frac{3!}{1!\Delta(3-1)!} \Delta(0{,}5^1 \Delta 0{,}5^{(3-1)}) = 3\Delta(0{,}5^1 \Delta 0{,}5^2) = 0{,}3750$	
La probabilidad de obtener en alguno de los lanzamientos cara es 0,3750, es la misma que la de obtener dos cruces.	
Probabilidad de obtener dos caras (Fórmula 50).	c, c, z (p^2 x q^1) z, c, c (p^2 x q^1) c, z, c (p^2 x q^1) *3* x (p^2 x q^1)
$f(2) = P(X = 2) = \binom{3}{2} \Delta 0{,}5^2 \Delta 0{,}5^{(3-2)} = \frac{3!}{2!\Delta(3-2)!} \Delta(0{,}5^2 \Delta 0{,}5^{(3-2)}) = 3\Delta(0{,}5^2 \Delta 0{,}5^1) = 0{,}3750$	
La probabilidad de obtener dos caras en los tres lanzamientos es 0,3750, es la misma que la de obtener una cruz. Cuando $p = q$, entonces, *3* x (p^1 x q^2) es igual que *3* x (p^2 x q^1).	
Probabilidad de obtener tres caras (Fórmula 50).	c, c, c (p^3 x q^0) *1* x (p^3 x q^0)
$f(3) = P(X = 3) = \binom{3}{3} \Delta 0{,}5^3 \Delta 0{,}5^{(3-3)} = \frac{3!}{3!\Delta(3-3)!} \Delta(0{,}5^3 \Delta 0{,}5^{(3-3)}) = 1\Delta(0{,}5^3 \Delta 0{,}5^0) = 0{,}1250$	
Cuando los tres lanzamientos son cara, ninguno de los tres es cruz. La probabilidad de que salgan tres caras (0,1250) es la misma que la de no salir cruz. Cuando $p = q$, entonces, *1* x (p^0 x q^3) es igual que *1* x (p^3 x q^0).	
Probabilidad de obtener tres caras o menos (Fórmula 51).	
$F(3) = P(X \le 3) = \sum_{x=0}^{3} \binom{n}{x} p^x \Delta q^{(n-x)} = \binom{3}{0} 0{,}5^0 \Delta 0{,}5^{(3-0)} + \binom{3}{1} 0{,}5^1 \Delta 0{,}5^{(3-1)} + \binom{3}{2} 0{,}5^2 \Delta 0{,}5^{(3-2)} + \binom{3}{3} 0{,}5^3 \Delta 0{,}5^{(3-3)} =$ $\frac{3!}{0!\Delta(3-0)!} \Delta 0{,}5^0 \Delta 0{,}5^{(3-0)} + \frac{3!}{1!\Delta(3-1)!} \Delta 0{,}5^1 \Delta 0{,}5^{(3-1)} + \frac{3!}{2!\Delta(3-2)!} \Delta 0{,}5^2 \Delta 0{,}5^{(3-2)} + \frac{3!}{3!\Delta(3-3)!} \Delta 0{,}5^3 \Delta 0{,}5^{(3-3)} =$ $0{,}1250 + 0{,}3750 + 0{,}3750 + 0{,}1250 = 1$	
La probabilidad de obtener tres o menos caras, es la probabilidad del suceso seguro y tiene que valer *1*. Ya que es la suma de la probabilidad de obtener *0* caras (*0,1250*), más la de obtener *1* cara (*0,3750*) más la de obtener *2* caras (*0,3750*) más la probabilidad de obtener las *3* caras (*0,1250*).	

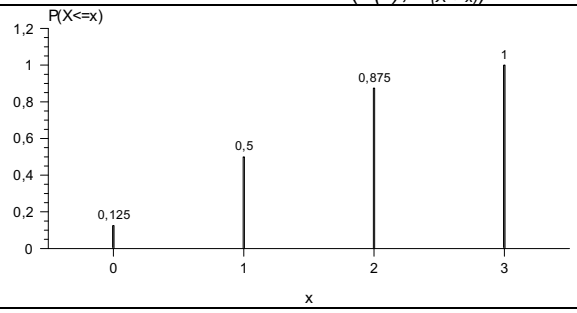
La representación gráfica de la función de probabilidad ($f(x)$, $P_{(X=x)}$) y de la función de distribución ($F(x)$, $P_{(X \le x)}$), en un sistema de coordenadas cartesianas de dos dimensiones en el que en el eje abscisas (X) se representa el valor de x y en el eje de ordenadas (Y) se representa la probabilidad ($f(x)$) o la probabilidad acumulada ($F(x)$), es la del Gráfico 14.

Gráfico 14 Gráficos de la binomial ($n = 3$ y $p = q$).

Función de probabilidad ($f(x)$, $P_{(X=x)}$)



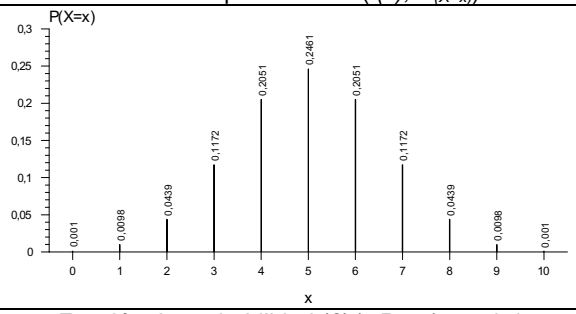
Función de distribución ($F(x)$, $P_{(X<=x)}$)



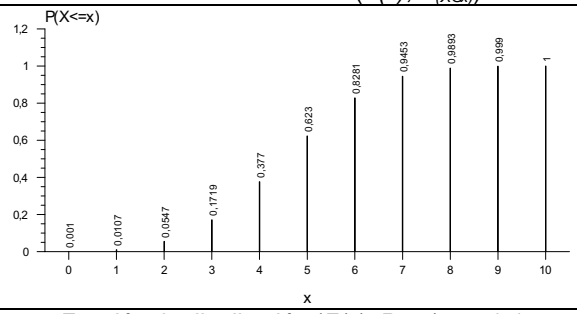
La función de probabilidad de la binomial se puede considerar normal cuando p *å* q. Esta característica permite, posteriormente, ver la aproximación a la binomial por la normal (Ver Epígrafe 10.1). Cuando p es menor que q el gráfico presenta asimetría a la derecha y a la izquierda cuando p es mayor que q. En el Gráfico 15 se muestran los ejemplos con $n = 10$.

Gráfico 15 Gráficos de la binomial ($n = 10$ y $p = q$).

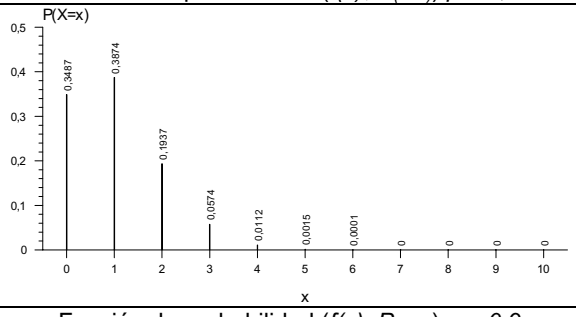
Función de probabilidad ($f(x)$, $P_{(X=x)}$)



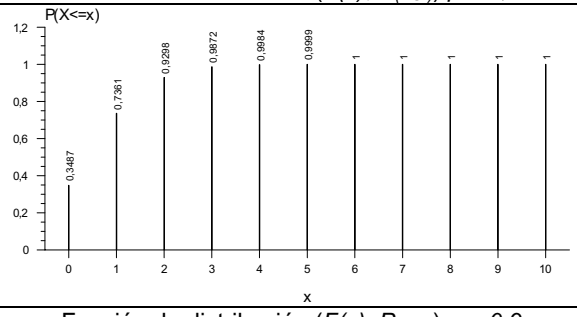
Función de distribución ($F(x)$, $P_{(X\grave{O}x)}$)



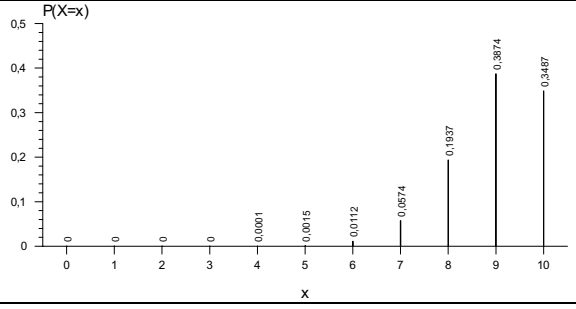
Función de probabilidad ($f(x)$, $P_{(X=x)}$) $p = 0,1$



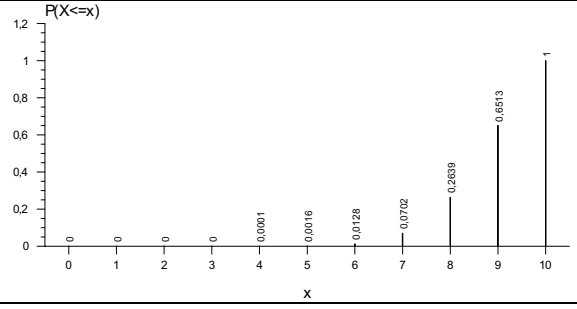
Función de distribución ($F(x)$, $P_{(X\grave{O}x)}$) $p = 0,1$



Función de probabilidad ($f(x)$, $P_{(X=x)}$) $p = 0,9$



Función de distribución ($F(x)$, $P_{(X\grave{O}x)}$) $p = 0,9$



Las propiedades de la distribución binomial se presentan en la Tabla 70.

Tabla 70 Propiedades de la distribución binomial	
Media	$\overline{X} \mid n \Delta p$
Varianza	$S^2 \mid n \Delta p \Delta q$
Desviación típica	$S \mid \sqrt{n \Delta p \Delta q}$
Coeficiente de sesgo	$g_1 \mid \frac{q 4 p}{\sqrt{n \Delta p \Delta q}}$
Coeficiente de apuntamiento	$g_2 \mid 3 2 \frac{4 6 \Delta p \Delta q}{n \Delta p \Delta q}$

10 Puntuación directa, diferencial y típica.

Se llama *puntuación directa* al valor que obtiene el individuo, caso o unidad de observación *i-ésimo* en una variable, y se representa por x_i. La *puntuación diferencial* (*pd*), es la distancia que tiene un individuo desde su puntuación directa hasta un estadístico de tendencia central, que habitualmente es la media y se representa por $(x_i - \overline{X})$. La *puntuación típica*, es la relación entre la puntuación diferencial y un estadístico de dispersión, habitualmente la desviación típica, y se representa por $(x_i - \overline{X})/S$. A la relación llamada *puntuación típica*, se la representa con la letra minúscula z_i. En la Tabla 71 se presenta el resumen y los ejemplos, utilizando la matriz de la Tabla 16, y los estadísticos se pueden ver en la Fórmula 14 y Fórmula 20.

Tabla 71 Resumen de los tipos de puntuaciones.

Puntuación	Símbolo	Caso 1 variable *p4_1* (peso)	Caso 4 variable *p4_1* (peso)
Directa	x_i	63 kg.	80 kg.
Diferencial	$(x_i - \overline{X})$	63 – 65,86 = - 2,86 kg	80 – 65,86 = 14,14 kg
Típica	$z_i = \frac{(x_i - \overline{X})}{S}$	$z_1 = \frac{(63 - 65{,}86)}{10{,}08} = \frac{-2{,}86}{10{,}08} = -0{,}2837$	$z_4 = \frac{(80 - 65{,}86)}{10{,}08} = \frac{14{,}04}{10{,}08} = 1{,}3929$

La diferencia entre los tres tipos de puntuación es la información que da cada uno. La puntuación directa dice el peso de cada uno de los individuos y por experiencia se puede saber si el peso es mucho o poco pero sin saber la estatura, por ejemplo, no se puede decir mucho más.

La puntuación diferencial amplía la información al indicar la distancia que tiene cada uno de los casos respecto a la media del grupo. Así el individuo 1 está próximo y por debajo de la media al ser negativa la puntuación (-2,86), y que el individuo 4 tiene un peso superior a la media (14,14). Aunque esta información es mayor, sigue siendo imprecisa porque no da idea de la magnitud de la diferencia.

La puntuación z_i o tipificada dice si el individuo está por encima o por debajo de la media y a que distancia de la media en unidades de desviación típica. También permite comparar valores del mismo individuo o distintos individuos en distintas variables. Otra opción es que a partir de la distribución o función de densidad de probabilidad de z_i se puede saber el porcentaje o probabilidad de individuos por debajo, por encima o entre dos valores. La Tabla 72 muestra el cálculo e interpretación de z_i y posteriormente una aplicación.

<table>
<tr><td colspan="5">Tabla 72 Interpretación de z_i.</td></tr>
<tr><td colspan="5">Caso 1 (matriz de datos de la Tabla 16).</td></tr>
<tr><td>x_i</td><td>$\overline{X}$</td><td>S</td><td colspan="2">$z_i = \frac{(x_i - \overline{X})}{S}$</td></tr>
<tr><td>63 kg</td><td>65,86 kg</td><td>10,08 kg</td><td colspan="2">$z_1 = \frac{(63\,\text{kg} - 65{,}86\,\text{kg})}{10{,}08\,\text{kg}} = \frac{-2{,}86\,\text{kg}}{10{,}08\,\text{kg}} = -0{,}2837$</td></tr>
<tr><td colspan="3">Los tres valores tienen la misma unidad de medida (kg.)</td><td colspan="2">Al dividir la puntuación diferencial por la desviación típica, las unidades del numerador se pierden con las del denominador y el valor de z_i no tiene unidades, por lo que resulta ser la relación entre las dos.</td></tr>
<tr><td colspan="5">Valores de z_i *y su significado.*</td></tr>
<tr><td rowspan="3">$pd = (x_i - \overline{X})$</td><td rowspan="3">Puede ocurrir que</td><td>$x_i < \overline{X}$, $pd < 0$ ($z_i < 0$)</td><td colspan="2" rowspan="3">Al dividir por la *S,* la relación nos dice las veces que la *puntuación diferencial* contiene a la *S* o lo que es lo mismo a cuantas unidades de *S* está la unidad de observación respecto de la $\overline{X}$. También llamado unidades *z*. Si la *x* es menor que la media, *z* es negativa y si la *x* es mayor que la media, z es positiva.</td></tr>
<tr><td>$x_i = \overline{X}$, $pd = 0$ ($z_i = 0$)</td></tr>
<tr><td>$x_i > \overline{X}$, $pd > 0$ ($z_i > 0$)</td></tr>
</table>

Ejemplo y aclaración del significado de "*unidades de desviación típica*" o "*unidades z*". Un individuo realiza dos exámenes el *A* y el *B.* En el primero obtiene una nota de 8 y en el segundo de 20. A simple vista se puede decir que el 20 es un valor mayor que el 8, pero no se puede decir qué nota es mejor porque faltan referentes. Es necesario saber el rango de las dos calificaciones. Si el rango de la nota *A* es de 0 a 10 y el de la nota B de 0 a 100*,* entonces la mejor nota es el *8*. Pero no se sabe cuanto mejor es la calificación del examen *A* respecto del *B*. Para saber la magnitud de la diferencia y que proporción o porcentaje de individuos tienen mejor o peor nota, es necesario conocer la puntuación tipificada o *z*. Su cálculo requiere saber la media y desviación típica de los dos exámenes. El planteamiento y resolución es el de la Tabla 73.

<table>
<tr><td colspan="5">Tabla 73 Aplicación de *z*.</td></tr>
<tr><td></td><td>Examen A</td><td>Examen B</td><td>z_{iA}</td><td>z_{iB}</td></tr>
<tr><td>Caso</td><td>8</td><td>20</td><td rowspan="4">$z_{1A} = \frac{(8 - 5)}{2} = \frac{3}{2} = 1{,}50$</td><td rowspan="4">$z_{1B} = \frac{(20 - 50)}{20} = \frac{-30}{20} = \frac{-3}{2} = -1{,}50$</td></tr>
<tr><td>$\overline{X}$</td><td>5</td><td>50</td></tr>
<tr><td>S</td><td>2</td><td>20</td></tr>
<tr><td>Rango</td><td>0 ⇔ 10</td><td>0 ⇔ 100</td></tr>
</table>

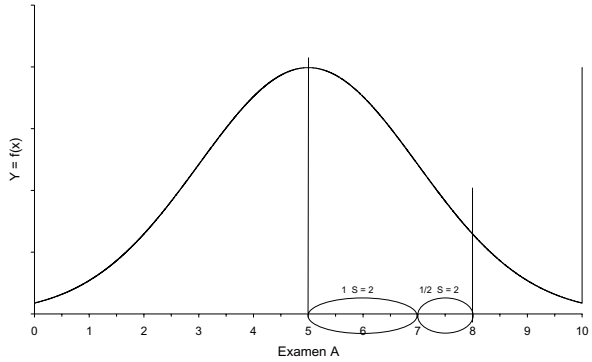
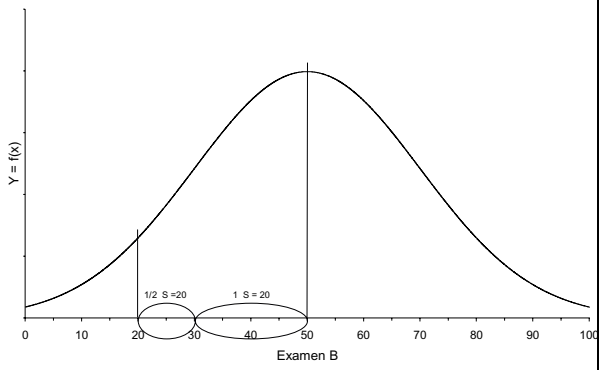
La nota de 8 en el examen *A* está a 1,5 veces la desviación típica (ó 1,5 *z's*) por encima de la media (por ser positiva), significa que la puntuación diferencial 3 contiene a la desviación típica 1,5 veces [3 = 2 + 1 (el 2 es 1 vez la *S* y el 1 es 0,5 veces la *S*)].

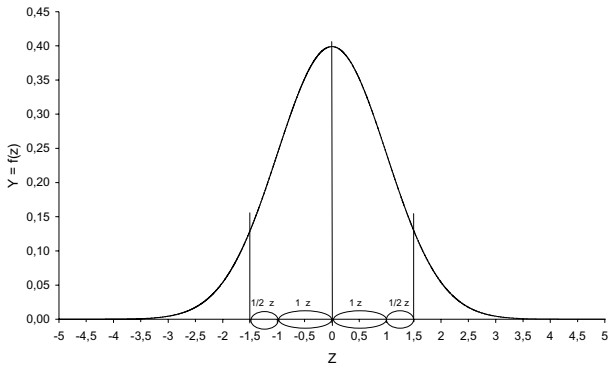
La nota de 20 en el examen *B* está a -1,5 veces la desviación típica (ó 1,5 *z's*) por debajo de la media (por ser negativa), significa que la puntuación diferencial 30 contiene a la desviación típica 1,5 veces [30 = 20 + 10 (el 20 es 1 vez la *S* y el 10 es 0,5 veces la *S*)].

El individuo está a la misma distancia de la media, pero en el examen *A* por encima y en el *B* por debajo de la media. Entonces el 8 es mejor nota que el 20, dentro del grupo.

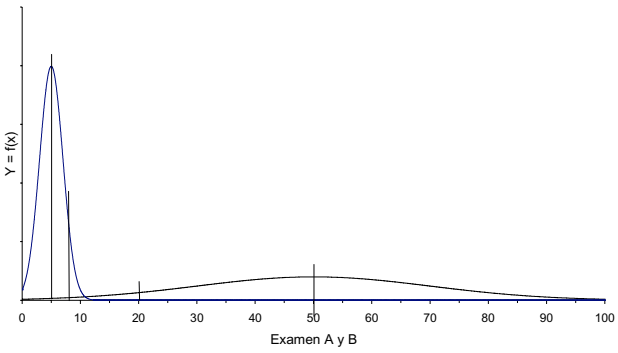
En la representación gráfica de las tres distribuciones (Tabla 74), que por el interés de la exposición se asumen normales, se muestra el concepto de *unidades de desviación típica* o *unidades z* y la transformación de la *puntuación directa* en *puntuación típica*.

Tabla 74 Representación gráfica de la distribución de las puntuaciones directas (x_i) y la típica (z_i).

Examen A	Examen B
	
La línea de referencia en el 8 indica la posición de la nota. La elipse mayor indica 1vez la *S* de 2, y la menor ½ vez la *S* de 2.	La línea de referencia en el 20 indica la posición de la nota. La elipse mayor indica 1vez la *S* de 20, y la menor ½ vez la *S* de 20.



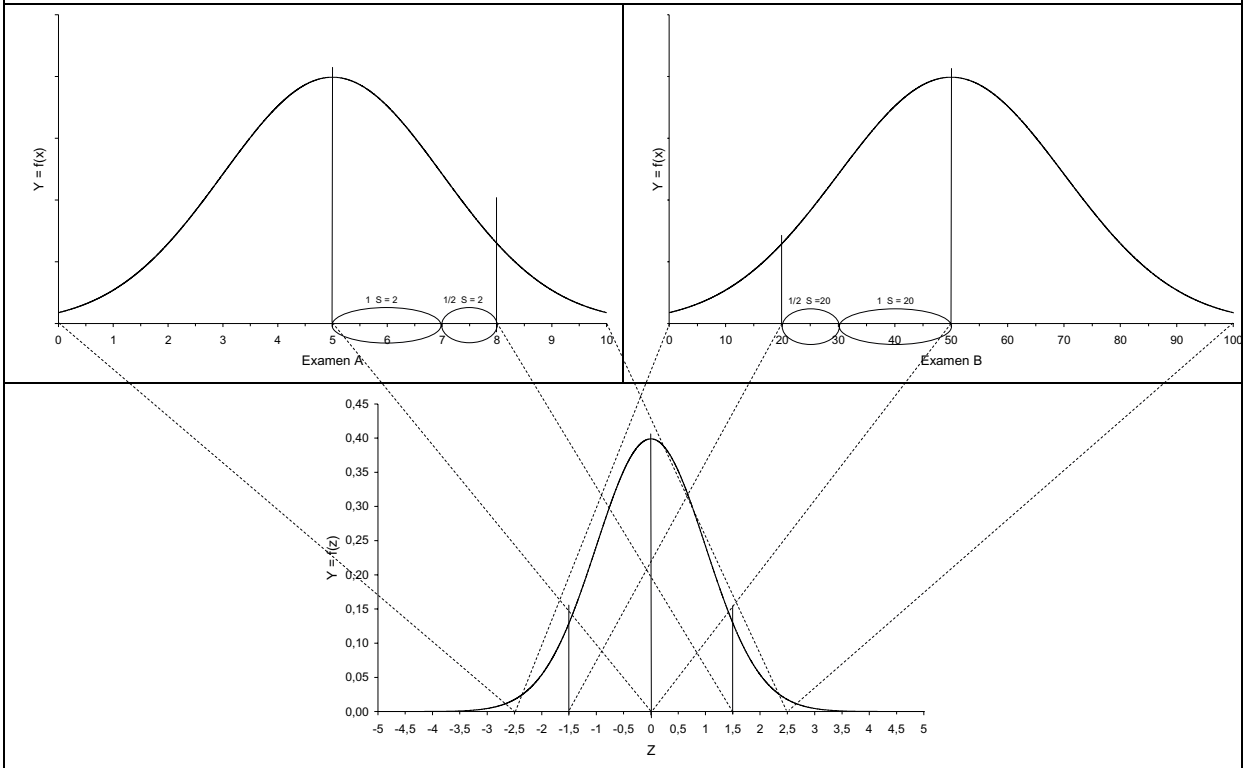
Al convertir las puntuaciones directas *x* en *z*, las unidades de desviación típica son *z's* y se representan en el mismo eje y en la misma escala.



Este gráfico muestra la superposición de la distribución de los dos exámenes en el mismo eje del examen B de escala 0 э 100.

El proceso gráfico de la transformación de la distribución de una variable a *z* es el de la Tabla 75,

Tabla 75 Gráfico de transformación de variables en unidades *z*.



Al convertir el valor x_i de cualquier variable a puntuación típica, pasa a un valor z_i en un mismo eje de unidades.

Puntuación examen A	Puntuación Z	Puntuación examen B	Puntuación Z
$X_i = 0$	$z_{iA} \mid \frac{(0\,4\,5)}{2} \mid \frac{45}{2} \mid 42{,}50$	$X_i = 0$	$z_{iB} \mid \frac{(0\,4\,50)}{20} \mid \frac{450}{20} \mid 42{,}50$
$X_i = 5$	$z_{iA} \mid \frac{(5\,4\,5)}{2} \mid \frac{0}{2} \mid 0{,}00$	$X_i = 20$	$z_{iB} \mid \frac{(20\,4\,50)}{20} \mid \frac{430}{20} \mid 41{,}50$
$X_i = 8$	$z_{iA} \mid \frac{(8\,4\,5)}{2} \mid \frac{3}{2} \mid 1{,}50$	$X_i = 50$	$z_{iB} \mid \frac{(50\,4\,50)}{20} \mid \frac{0}{20} \mid 0{,}00$
$X_i = 10$	$z_{iA} \mid \frac{(10\,4\,5)}{2} \mid \frac{5}{2} \mid 2{,}50$	$X_i = 100$	$z_{iB} \mid \frac{(100\,4\,50)}{20} \mid \frac{50}{20} \mid 2{,}50$

Para una puntuación x_i igual a la media de X la z_i siempre vale cero.

Aplicando el criterio de puntuación z_i a todos los casos, esto es, a todos los valores de una variable, se obtiene una nueva variable que se denomina *Z*. La tipificación de variables sólo se puede hacer con variables numéricas o consideradas numéricas. De la misma manera, se puede transformar cualquier variable a variable *Z's*. La Tabla 76 muestra las variables zp4*_1*, *zp4_2* y *zp4_3* de las variables *p4_1, p4_2* y *p4*_3 que son el peso, la estatura y la edad, respectivamente, de la matriz de datos de la Tabla 16.

Tabla 76 Matriz de datos con variables *Z*.

id	p4_1	p4_2	p4_3	zp4_1	zp4_2	zp4_3	Id	p4_1	p4_2	p4_3	zp4_1	zp4_2	zp4_3
1	63	1,63	21	-0,2841	-0,0087	0,0331	50	55	1,74	27	-1,0778	0,0022	0,6284
2	63	1,63	21	-0,2841	-0,0087	0,0331	51	67	1,7	20	0,1128	-0,0017	-0,0661
3	68	1,75	23	0,2120	0,0032	0,2315	52	77	1,87	19	1,1049	0,0151	-0,1654
4	80	1,75	19	1,4026	0,0032	-0,1654	53	77	1,87	19	1,1049	0,0151	-0,1654
5	73	1,82	24	0,7081	0,0102	0,3307	54	52	1,67	19	-1,3754	-0,0047	-0,1654
6	73	1,82	24	0,7081	0,0102	0,3307	55	78	1,85	21	1,2041	0,0132	0,0331
7	45	1,6	19	-2,0699	-0,0116	-0,1654	56	50	1,67	20	-1,5738	-0,0047	-0,0661
8	60	1,6	20	-0,5817	-0,0116	-0,0661	57	66	1,78	18	0,0136	0,0062	-0,2646
9	60	1,72	22	-0,5817	0,0003	0,1323	58	65	1,73	19	-0,0856	0,0013	-0,1654
10	55	1,63	18	-1,0778	-0,0087	-0,2646	59	58	1,63	21	-0,7801	-0,0087	0,0331
11	85	1,85	20	1,8986	0,0132	-0,0661	60	70	1,68	21	0,4104	-0,0037	0,0331
12	75	1,75	19	0,9065	0,0032	-0,1654	61	70	1,6	20	0,4104	-0,0116	-0,0661
13	75	1,75	19	0,9065	0,0032	-0,1654	62	65	1,77	18	-0,0856	0,0052	-0,2646
14	53	1,66	18	-1,2762	-0,0057	-0,2646	63	73	1,71	26	0,7081	-0,0007	0,5291
15	.	.	.	.	.	.	64	58	1,75	19	-0,7801	0,0032	-0,1654
16	52	1,66	17	-1,3754	-0,0057	-0,3638	65	75	1,58	18	0,9065	-0,0136	-0,2646
17	55	1,74	27	-1,0778	0,0022	0,6284	66	76	1,9	28	1,0057	0,0181	0,7276
18	67	1,7	20	0,1128	-0,0017	-0,0661	67	63	1,63	21	-0,2841	-0,0087	0,0331
19	77	1,87	19	1,1049	0,0151	-0,1654	68	52	1,63	25	-1,3754	-0,0087	0,4299
20	77	1,87	19	1,1049	0,0151	-0,1654	69	68	1,75	23	0,2120	0,0032	0,2315
21	52	1,67	19	-1,3754	-0,0047	-0,1654	70	80	1,75	19	1,4026	0,0032	-0,1654
22	78	1,85	21	1,2041	0,0132	0,0331	71	73	1,82	24	0,7081	0,0102	0,3307
23	78	1,85	21	1,2041	0,0132	0,0331	72	55	1,6	24	-1,0778	-0,0116	0,3307
24	66	1,78	18	0,0136	0,0062	-0,2646	73	45	1,6	19	-2,0699	-0,0116	-0,1654
25	65	1,73	19	-0,0856	0,0013	-0,1654	74	60	1,6	20	-0,5817	-0,0116	-0,0661
26	65	1,73	19	-0,0856	0,0013	-0,1654	75	60	1,72	22	-0,5817	0,0003	0,1323
27	70	1,68	21	0,4104	-0,0037	0,0331	76	55	1,63	18	-1,0778	-0,0087	-0,2646
28	70	1,6	20	0,4104	-0,0116	-0,0661	77	85	1,85	20	1,8986	0,0132	-0,0661
29	70	1,6	20	0,4104	-0,0116	-0,0661	78	75	1,75	19	0,9065	0,0032	-0,1654
30	73	1,71	26	0,7081	-0,0007	0,5291	79	58	1,63	19	-0,7801	-0,0087	-0,1654
31	58	1,75	19	-0,7801	0,0032	-0,1654	80	53	1,66	18	-1,2762	-0,0057	-0,2646
32	75	1,58	18	0,9065	-0,0136	-0,2646	81	.	.	.	.	.	.
33	76	1,9	28	1,0057	0,0181	0,7276	82	52	1,66	17	-1,3754	-0,0057	-0,3638
34	63	1,63	21	-0,2841	-0,0087	0,0331	83	55	1,74	27	-1,0778	0,0022	0,6284
35	63	1,63	21	-0,2841	-0,0087	0,0331	84	66	1,78	18	0,0136	0,0062	-0,2646
36	68	1,75	23	0,2120	0,0032	0,2315	85	77	1,87	19	1,1049	0,0151	-0,1654
37	80	1,75	19	1,4026	0,0032	-0,1654	86	.	1,65	20	.	-0,0067	-0,0661
38	73	1,82	24	0,7081	0,0102	0,3307	87	52	1,67	19	-1,3754	-0,0047	-0,1654
39	55	1,6	24	-1,0778	-0,0116	0,3307	88	78	1,85	21	1,2041	0,0132	0,0331
40	45	1,6	19	-2,0699	-0,0116	-0,1654	89	50	1,67	20	-1,5738	-0,0047	-0,0661
41	60	1,6	20	-0,5817	-0,0116	-0,0661	90	66	1,78	18	0,0136	0,0062	-0,2646
42	60	1,72	22	-0,5817	0,0003	0,1323	91	65	1,73	19	-0,0856	0,0013	-0,1654
43	55	1,63	18	-1,0778	-0,0087	-0,2646	92	70	1,6	20	0,4104	-0,0116	-0,0661
44	85	1,85	20	1,8986	0,0132	-0,0661	93	70	1,68	21	0,4104	-0,0037	0,0331
45	75	1,75	19	0,9065	0,0032	-0,1654	94	70	1,6	20	0,4104	-0,0116	-0,0661
46	75	1,75	19	0,9065	0,0032	-0,1654	95	76	1,9	28	1,0057	0,0181	0,7276
47	53	1,66	18	-1,2762	-0,0057	-0,2646	96	73	1,71	26	0,7081	-0,0007	0,5291
48	.	.	.	.	.	.	97	58	1,75	19	-0,7801	0,0032	-0,1654
49	52	1,66	17	-1,3754	-0,0057	-0,3638	98	75	1,58	18	0,9065	-0,0136	-0,2646
							99	76	1,9	28	1,0057	0,0181	0,7276

El caso 4 tiene un peso de 80 kg, 1,75 m de estatura y 19 años. Desde las puntuaciones *z* se puede saber que en cuanto a peso está por encima de la media de grupo 1,4026 veces la desviación típica. En la estatura está muy próximo a la media del grupo, ya que sólo está a 0,0032 unidades *z*. Y en cuanto a la edad, está por debajo de la media del grupo (*z* = -0,1654). Al estar todos los valores *z* en la misma unidad de medida, se han podido relacionar valores de variables de diferente unidad de medida. En cuanto a peso está muy por encima de la media, en estatura próximo a la media y en edad por debajo de la media.

10.1 Relación entre la distribución binomial y la normal

Si *n* es grande y *p* no es muy pequeña, la distribución binomial puede comportarse como una normal y se puede utilizar el criterio de transformación de *z* simbólicamente,

$$z_i \mid \frac{x_i \, 4 / n \Delta p 0}{\sqrt{n \Delta p \Delta q}}$$

En donde:

x_i:	Número de éxitos.
n:	Número de pruebas.
p:	Probabilidad *de éxito.*
q:	Probabilidad de fracaso.
n x *p:*	Media de una distribución binomial.
$\sqrt{n \Delta p \Delta q}$:	Desviación típica de una distribución binomial

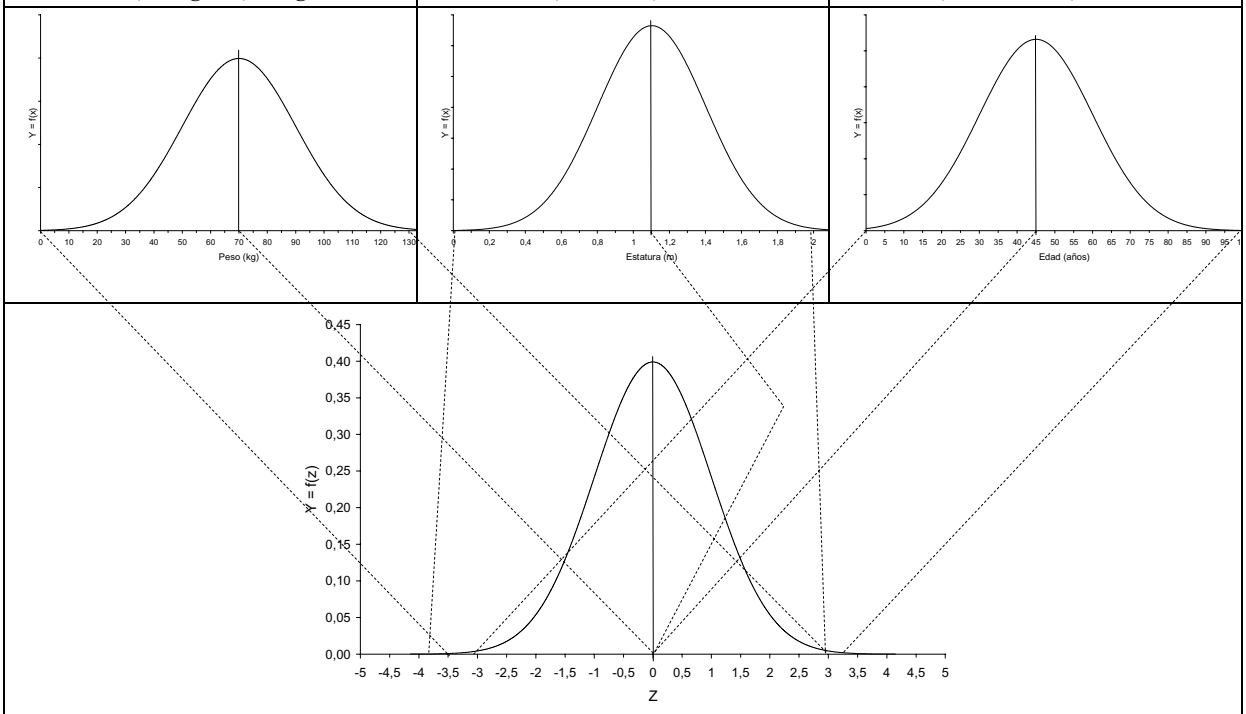
Fórmula 52

11 Concepto de probabilidad (variables continuas)

Según el epígrafe anterior, cualquier variable numérica puede ser transformada en puntuación *z* y por consiguiente en una variable *Z*. Tomando de una población o muestra grande tres variables, por ejemplo, el peso, la estatura y la edad, y asumiendo que las tres tiene una distribución normal por tener tamaños grandes y siguiendo el criterio de la Tabla 74, la representación gráfica de la transformación en *Z* se ve en la Tabla 77.

Tabla 77 Transformación de variables en puntuaciones z.

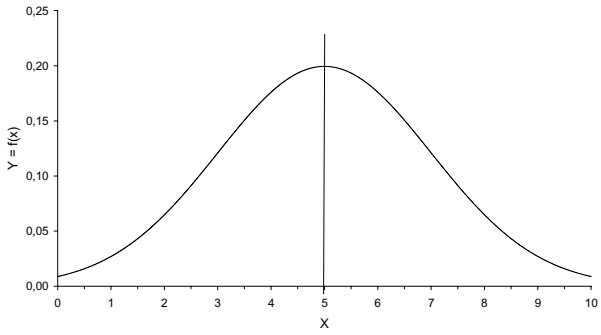
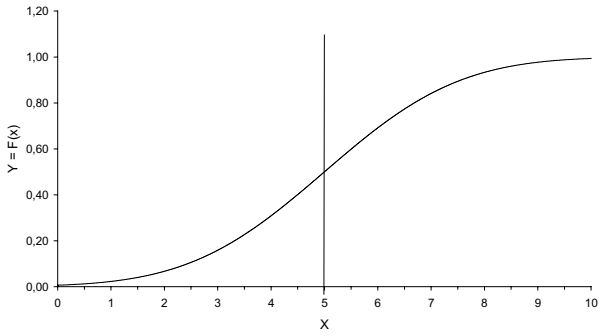
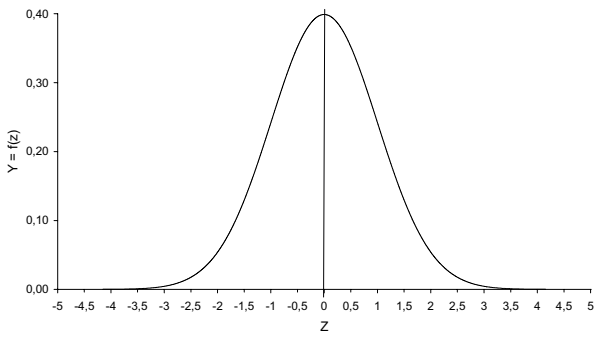
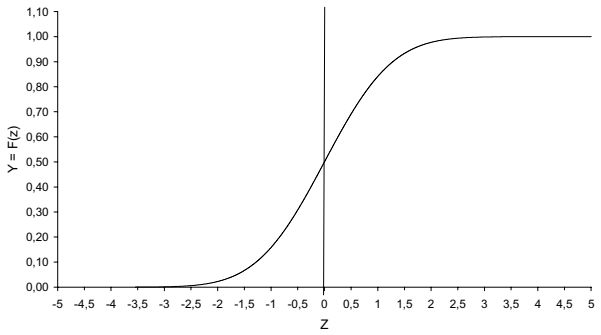
Gráfico de la variable peso	Gráfico de la variable estatura	Gráfico de la variable edad
$\overline{X} \mid 70kg , S \mid 20kg$	$\overline{X} \mid 1{,}1m , S \mid 0{,}3m$	$\overline{X} \mid 45años , S \mid 15años$



Igual que en la Tabla 74, con la puntuación de los exámenes *A* y *B,* al transformar las variables: *peso, estatura* y *edad* a puntuaciones *z,* las tres variables pasan a la misma unidad de medida, puntuaciones *z* o unidades de desviación típica. Y generalizando, cualquier variable que se transformen sus valores a puntuaciones *z*, se convierte en una variable *Z,* en unidades *z* o unidades de desviación típica.

Las cuatro variables tienen una distribución normal con su media y desviación típica ($N_{(\overline{X},S)}$), y estas distribuciones tienen su función *f(x)* definida, de tal manera que podemos decir que la *Y* está en función de *x,* simbólicamente, *Y = f(x),* y se puede generalizar a toda variable con distribución normal. Significa que para cualquier valor de la variable en el eje de abscisas, aplicando la función, obtenemos un valor en el eje de ordenadas o vertical (Tabla 78).

Tabla 78 Función y distribución de la normal.

Función de densidad de probabilidad (*fdp*) de la normal.	Gráfico
$y \mid f(x)$ $f(x) \mid \frac{1}{S\Delta\sqrt{2\Delta\phi}}\Delta e^{\left(-\frac{1}{2}\Delta\left(\frac{x-\overline{X}}{S}\right)^2\right)}$, $-\infty \leftarrow \{ x \{ \leftarrow +\infty$ En donde: ϕ = 3,141592654. *e* = 2,718281828. $\overline{X} \mid$ Media de la variable. *S* = Desviación típica de la variable. *x* Toma valores entre $-\infty \leftarrow e \leftarrow +\infty$. Gráfico: $N_{(5,2)}$ y $-\infty \leftarrow \{ x \{ \leftarrow +\infty$ La superficie bajo la curva; por encima del eje de abscisas, y entre $-\infty \leftarrow \{ x \{ \leftarrow +\infty$, vale la unidad.	
Función de distribución de la normal	Gráfico
$y \mid F(x)$ $F(x) \mid \frac{1}{S\Delta\sqrt{2\Delta\phi}}\Delta \int_{-\infty\leftarrow}^{x} e^{\left(-\frac{1}{2}\Delta\left(\frac{x-\overline{X}}{S}\right)^2\right)} \Delta dx$ En donde: ϕ = 3,141592654. *e* = 2,718281828. $\overline{X} \mid$ Media de la variable. *S* = Desviación típica de la variable. *x* Toma valores entre $-\infty \leftarrow e \leftarrow +\infty$. Gráfico: $N_{(5,2)}$ y $-\infty \leftarrow \{ x \{ \leftarrow +\infty$ La $F(x)$ cuando $x \Downarrow \leftarrow, -\infty$ es la unidad.	
fdp de la normal tipificada	Gráfico
$y \mid f(z)$ $f(z) \mid \frac{1}{\sqrt{2\Delta\phi}}\Delta e^{\left(-\frac{1}{2}\Delta z^2\right)}$, $-\infty \leftarrow \{ z \{ \leftarrow +\infty$ En donde: ϕ = 3,141592654. *e* = 2,718281828. $\overline{Z} \mid$ 0. *S* = 1. *z* Toma valores entre $-\infty \leftarrow e \leftarrow +\infty$. Gráfico: $N_{(0,1)}$ y $-\infty \leftarrow \{ z \{ \leftarrow +\infty$ La superficie bajo la curva; por encima del eje de abscisas, y entre $-\infty \leftarrow \{ z \{ \leftarrow +\infty$, vale la unidad.	
Función de distribución de la normal tipificada	Gráfico
$y \mid F(z)$ $F(z) \mid \frac{1}{\sqrt{2\Delta\phi}}\Delta \int_{-\infty\leftarrow}^{z} e^{\left(-\frac{1}{2}\Delta z^2\right)} dz$, $-\infty \leftarrow \{ z \{ \leftarrow +\infty$ En donde: ϕ = 3,141592654. *e* = 2,718281828. *z* Toma valores entre $-\infty \leftarrow e \leftarrow +\infty$. Gráfico: $N_{(0,1)}$ y $-\infty \leftarrow \{ z \{ \leftarrow +\infty$ La $F(z)$ cuando $z \Downarrow \leftarrow, -\infty$ es la unidad.	

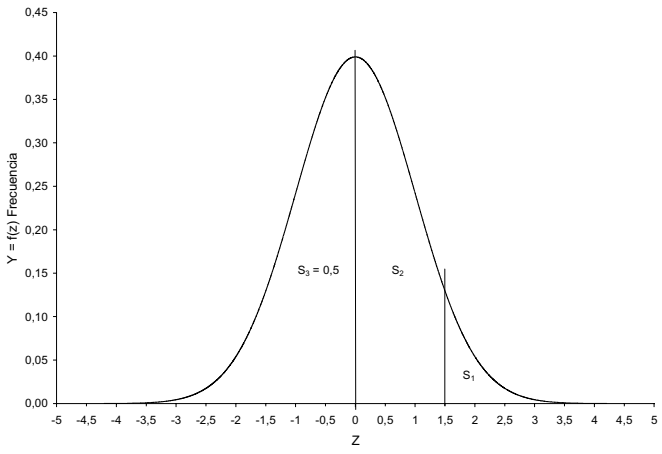
Las características del gráfico de la función de densidad de la normal tipificada o variable *Z*, son: su distribución es normal; la media vale cero; la desviación típica vale la unidad; la varianza vale también la unidad; la moda, la mediana y la media tienen el mismo valor; es simétrica por el eje que define la media, y la superficie contenida por debajo de la curva y por encima del eje de abscisas vale la unidad. Su distribución es normal de media 0 y desviación típica igual a la unidad, simbólicamente, $N_{(0,1)}$.

La simetría significa que las dos mitades que definen el eje que pasa por la media son iguales y tienen la misma forma. La superficie representa a todos los casos, y estos están representados por la unidad en términos de probabilidad o proporción y en porcentajes si multiplicamos la probabilidad por 100. Al ser simétrica, cada mitad vale 0,5 o 50 %.

Asociar *superficie* a *casos,* permite decir cual es la probabilidad de que un caso esté por debajo, por encima o entre cualesquiera dos valores de *z*. Al operar con el concepto de probabilidad objetiva frecuencista o "*a posteriori*", se puede asociar a un porcentaje y permite decir cual es el porcentaje de casos por debajo, por encima o entre cualesquiera dos valores de *z*. Para saber la probabilidad o porcentaje asociado a una superficie es necesario calcular el valor de la superficie.

Según este planteamiento, hallando la superficie se puede saber, en el caso del examen A (Tabla 73), cual es la probabilidad de que un individuo obtenga menos de 8 o más de 8, o que porcentaje de individuos del grupo han obtenido menos de 8 o más de 8.

Tabla 79 Planteamiento gráfico para el cálculo de superficies en la normal tipificada.

La línea de referencia trazada en z = 1,5, se corresponde con la transformación en puntuación típica de la nota 8 del examen A, y la superficie *S* que vale la unidad, queda dividida en dos partes. La superficie definida por S_1 y $S_2 + S_3$. La S_1 es la probabilidad de que un individuo obtenga más de un 8 en el examen A. La superficie por debajo del segmento en z = 1,5 ($S_2 + S_3$), es la probabilidad de que un individuo obtenga menos de 8. Si calculamos una de las dos superficies podemos obtener las restantes por suma o resta algebraica	

El concepto y cálculo de la superficie de un rectángulo ($S \mid b\Delta a$) no contempla dificultad debido a que es un polígono de lados paralelos. La dificultad de calcular la superficie bajo la curva de la normal y por encima del eje de abscisas, es que la altura es variable y la base está comprendida entre $4\leftarrow$ e $\leftarrow$. El problema es irresoluble, porque además de ser un polígono de altura variable, la base es infinita. Se debe proceder para simplificar el problema.

La simplificación empieza al dividir la superficie en dos partes por la línea de referencia en el punto z = 1,5. Ahora se debe proceder a calcular la superficie que queda por encima de la línea de referencia (S_1) o por debajo ($S_2 + S_3$) y la superficie restante se puede obtener por diferencia simple con 1, ya que $S_1 + S_2 + S_3$ es igual a la unidad. La superficie por debajo de la línea de referencia se puede descomponer en $S_2 + S_3$ y como S_3 vale 0,5 por ser la mitad de la curva, sólo falta obtener S_2.

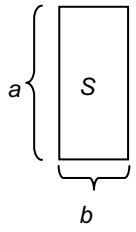
En este momento, la elección está entre hallar la superficie de S_2 o de S_1. La superficie de S_2 es la comprendida bajo la curva normal, el eje de abscisas y los valores de $z = 0$ y $z = 1{,}50$. Esta superficie plantea sólo un inconveniente, que la altura es variable, ya que la base es finita y conocida (en este caso, 1,5 – 0). La superficie de S_1 tiene dos inconvenientes, que la altura es variable y la base infinita, por lo tanto, optamos por calcular la superficie S_2.

El problema de calcular una superficie con lados curvos lo plantearon los griegos; aunque no dieron con la solución exacta sí hicieron aproximaciones. Fueron Newton y Leibniz quienes encontraron la solución a través del cálculo diferencial-integral. El procedimiento utilizado para el cálculo de S_2 es el atribuido a Leibniz (Tabla 80).[61]

Tabla 80 Cálculo de una superficie con lados curvos definida por la función de la normal tipificada.

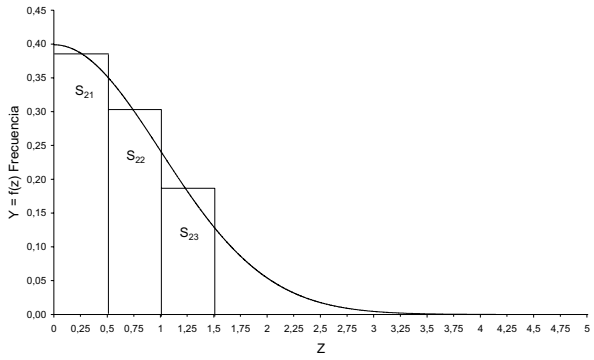
El cálculo de la superficie de un rectángulo es:

$$S = b\,\Delta a$$



La superficie S_2 tiene una base de $z = 1{,}5 = 1{,}5 - 0$, pero la altura es variable.
Se Procede a dividir la superficie S_2 en tres rectángulos: S_{21}, S_{22} y S_{23}. Por el mismo criterio que se expuso en el Gráfico 12, la superficie de S que se deja fuera, se asume equivalente a la que se incorpora, de fuera, en el rectángulo.
El procedimiento es calcular la superficie de cada uno de los rectángulos y la suma será la superficie que se busca,

$S_2 = S_{21} + S_{22} + S_{23}$.



Procedimiento de cálculo:

Datos de partida:	Base de los rectángulos:	Altura de los rectángulos:	Superficies:
$z_1 = 0{,}50$ $z_2 = 1{,}00$ $z_3 = 1{,}50$	$b_1 = 0{,}50 - 0{,}00 = 0{,}50$ $b_2 = 1{,}00 - 0{,}50 = 0{,}50$ $b_3 = 1{,}50 - 1{,}00 = 0{,}50$ La base es una diferencia constante y se llama diferencial de z (dz)	$a_1 = y_1 = f(z_1) = f(0{,}50)$ $a_2 = y_2 = f(z_2) = f(1{,}00)$ $a_3 = y_3 = f(z_3) = f(1{,}50)$ La altura de cada rectángulo está dada por el valor de $y = f(z)$ (Tabla 78).	$S_{21} = b_1 \times a_1 = d(z) \times f(z_1)$ $S_{22} = b_2 \times a_2 = d(z) \times f(z_2)$ $S_{23} = b_3 \times a_3 = d(z) \times f(z_3)$

Entonces:

$$S_2 = S_{21} + S_{22} + S_{23} = |b_1 \Delta a_1| + |b_2 \Delta a_2| + |b_3 \Delta a_3| \;\; dz \Delta f(z_1) + dz \Delta f(z_2) + dz \Delta f(z_3) = \sum_{i=1}^{3} f(z_i) \Delta dz$$

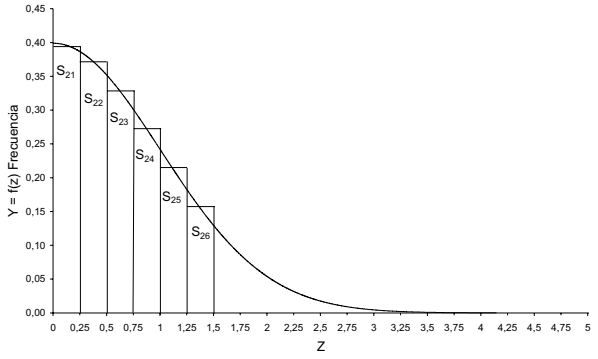
y de forma genérica,

$$S = \sum_{i=1}^{n} f(z_i) \Delta dz$$

[61] El método de Leibniz se considera el de la descomposición de la superficie en infinitos rectángulos y el método de Newton es de la descomposición de la superficie en infinitos triángulos para el incremento de la precisión del número ϕ.

Tabla 80 Cálculo de una superficie con lados curvos definida por la función de la normal tipificada.

Para tratar de corregir el error anterior, se puede reducir la base, aumentando el número de los rectángulos.



Entonces,

$$S_2 \mid \sum_{i\mid 1}^{n} f(z_i)\Delta dz \mid \sum_{i\mid 1}^{6} f(z_i)\Delta dz \mid f(z_1)\Delta dz + f(z_2)\Delta dz + \ldots + f(z_6)\Delta dz$$

Para seguir reduciendo el error que se produce al operar con rectángulos, se incrementa el número de rectángulos, hasta llegar a tener un rectángulo por cada punto de la curva *normal*. En este caso, la variable número de rectángulos tratada como discreta, es tan grande, que se ha convertido en continua y el sumatorio de variables discretas, se tiene que cambiar por el sumatorio de variables numéricas consideradas continuas y utilizar la integral definida,

Entonces, en vez de, $S_2 \mid \sum_{i\mid 1}^{\leftarrow} f(z_i)\Delta dz$, se utiliza la integral definida $\int_0^z f(z)\Delta dz$ y en este caso será, $S_2 \mid \int_0^{1,50} f(z)\Delta dz$

Resolviendo la integral definida se obtiene la superficie buscada. Pero para saber el resultado, no es necesario aplicar el cálculo integral, ya que este tipo de integrales están tabuladas y a través de su Tabla se puede resolver. El proceso seguido se considera necesario para tener el concepto de *integración* y de *superficie,* pero no es necesario saber cálculo diferencial-integral. Para resolver la integral se recurre a la tabla del Anexo I (Página 332) y se muestra el proceso en la Tabla 81, y en la Tabla 82 y Tabla 83 se muestra la lectura de la superficie correspondiente a la nota de los exámenes A y B, respectivamente.

Tabla 81 Resolución por tablas de la integral definida de la normal tipificada.

El valor de la superficie comprendido por debajo de la curva normal; por encima del eje de abscisa, y entre los valores de *z* = 0 y *z* = 1,50, se busca en la tabla del Anexo I. La tabla del Anexo 1 proporciona la superficie para cualquier integral de este tipo. El límite superior de la integral, que en este caso es 1,50, se descompone en dos partes "*unidades y décimas*" (1,5) y "*centésimas*" (0). El 1,5 es la entrada de filas en la Tabla, y el 0 la entrada de columnas y el punto donde se cruzan es el resultado de la integral, que es el de la superficie buscada, que en este caso es 0,4332.

$$S_2 \mid \int_0^z f(z)\Delta dz \mid \int_0^{1,50} f(z)\Delta dz \mid 0,4332$$

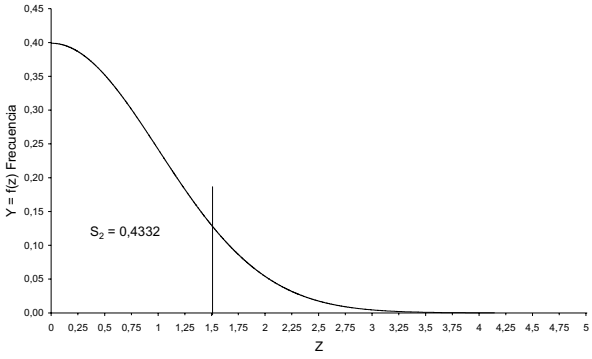


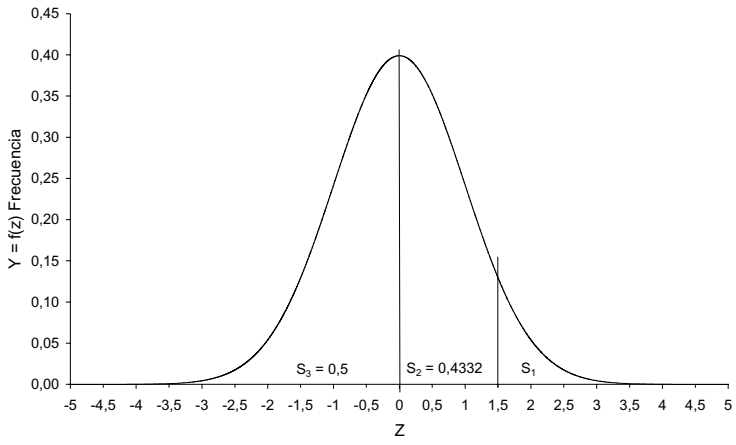
Tabla 81 Resolución por tablas de la integral definida de la normal tipificada.

Asignatura: Estadística Aplicada a las CC. SS. Profesor: Carlos de la Puente V.

Area bajo la curva normal estandarizada entre 0 y z

	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,03
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,07
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,11
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,15
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,18
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,22
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,25
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,28
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,31
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,33
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,36
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,38
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,40
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,41
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,43
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,44
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,45

Tabla 82 Lectura del valor de las superficies de la curva normal tipificada del examen A.

Comentario	Examen A
La probabilidad de obtener menos de 8 ($P_{(X < 8)}$ ó $P_{(x < 8)}$ ya que $P_{(X = 8)} = 0$, ver Epígrafe 11.1) en el examen *A*, es igual a la superficie que queda por debajo de $z = 1,50$, que es la superficie de S_2 más la superficie de S_3 (0,4332 + 0,5000 = 0,9332). Por lo tanto la probabilidad buscada es de 0,9332. Expresado en porcentaje sería que el 93,3% (0,9332 x 100) han obtenido menos de un 8 en el examen. La probabilidad de obtener más de 8 es 0,0668, que es el valor de la superficie S_1, que está dado por la diferencia entre la superficie total de la mitad derecha de la curva (0,5) menos la superficie S_2 (0,4332) (0,5 - 0,4332 = 0,0668). Expresado en porcentajes, el 6,7% han obtenido más de 8.	

Entonces,

$$F(x \mid 8) \mid P_{/X\Omega 80} \mid F/z \mid 1,50 \mid P_{/Z\Omega 1,50} \mid S_3 \, 2 \, S_2$$

$$S_3 \mid P_{/Z\Omega 00} \mid \left|_{4\leftarrow}^{0} f(z)\Delta dz \right| 0,5$$

$$S_2 \mid P_{/0\Omega Z\Omega 1,50} \mid \left|_{0}^{1,5} f(z)\Delta dz \right| 0,4332$$

$$P_{/Z\Omega 1,50} \mid \left|_{4\leftarrow}^{1,5} f(z)\Delta dz \right| \left|_{4\leftarrow}^{0} f(z)\Delta dz \, 2 \left|_{0}^{1,5} f(z)\Delta dz \right| 0,5 \, 2 \, 0,4332 \mid 0,9332$$

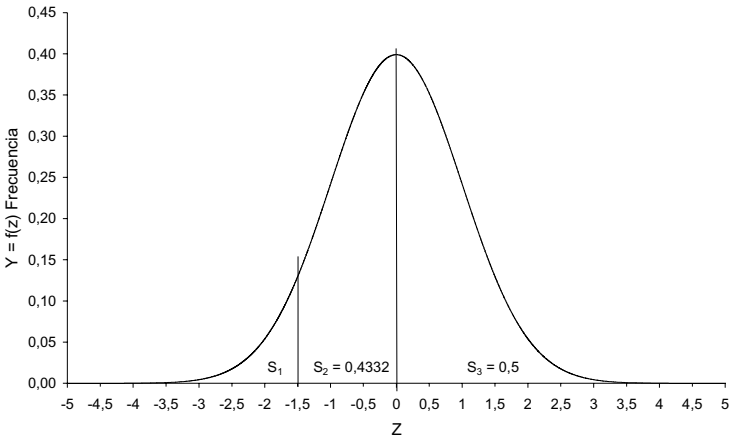
La probabilidad de obtener menos de un ocho en el examen A es:

$S_3 + S_2 = 0,5 + 0,4332 = 0,9332$ ó 93,3%

La probabilidad de obtener más de un ocho en el examen A es:

$S_1 = 1 - (S_3 + S_2) = 1 - 0,9332 = 0,0668$ ó 6,7%

Tabla 83 Lectura del valor de las superficies de la curva normal tipificada del examen B.

Comentario	Examen B
Como la distribución es simétrica y la *z* (-1,50) de la nota del examen B (20) es igual que la *z* (1,50) del examen A, pero con el signo cambiado, entonces la probabilidad o el porcentaje de obtener menos de 20 es la misma que la de obtener más de 8 (0,0668 ó 6,7%) y la de obtener más de 20 es la misma que la de obtener menos de 8 (0,9332 ó 93,3%)	

La probabilidad de obtener más de un 20 en el examen B es:

$S_3 + S_2 = 0,5 + 0,4332 = 0,9332$ ó 93,3%

La probabilidad de obtener menos de un 20 en el examen B es:

$S_1 = 1 - (S_3 + S_2) = 1 - 0,9332 = 0,0668$ ó 6,7%

El valor de la superficie bajo la curva normal tipificada o la probabilidad del suceso seguro, es igual a uno y está dada por la integral,

$$F_{|Z|\leftarrow 0}| \;\; P_{|Z\Omega\leftarrow 0}| \;\; \lfloor_{4\leftarrow}^{\leftarrow} f(z)dz \,| \;\; \lfloor_{4\leftarrow}^{0} f(z)dz \; 2 \; \lfloor_{0}^{\leftarrow} f(z)dz \,| \;\; 0{,}5 \; 2 \; 0{,}5 \,| \;\; 1$$

11.1 Relación entre probabilidad discreta y continua

La probabilidad de obtener un determinado valor ($P_{(X=x)}$), en el caso de variables discretas es igual o mayor que cero.

$$P_{|X|\;x0} \varnothing 0, \textit{ para} \text{ todo x}$$

En el caso de una variable continua la $P_{(X=x)}$ es siempre cero.

$$P_{|X|\;x0}| \;\; 0, \textit{ para} \text{ todo x}$$

La característica del cálculo de las probabilidades en el caso discreto es la relación entre los hechos favorables y los hechos posibles, por lo tanto siempre será igual o mayor a cero. En el caso de una variable continua al ser la probabilidad una superficie, siempre debe estar definida por dos valores o entre un valor e $\leftarrow$, por intuición (sin demostración matemática), la distancia o diferencia entre un valor consigo lo tomamos como cero y por lo tanto el cálculo de una superficie que tiene de base cero, también es cero y así mismo la probabilidad.

11.2 Aplicación de la probabilidad (variables continuas)

Cualquier variable que se le asuma que tiene una distribución normal o cualquier valor de una variable que se le asuma distribución normal ($N_{(\overline{X},S)}$), se le puede aplicar el criterio de transformación en puntuación típica o *z* ($N_{(0,1)}$) y calcular probabilidades o porcentajes. Además de calcular la superficie por debajo o por encima de cierto valor de la variable, otra posibilidad es la de calcular superficies entre dos valores que llamaremos intervalos.

Si una variable tiene la distribución normal según *f(x)* conocida (Tabla 78), se puede calcular la superficie para determinados valores y tomar la superficie como una probabilidad o porcentaje. Los resultados obtenidos a través de la función de la normal y la función de la normal tipificada son iguales. Se utiliza la *f(z)* por estar tabulada y su criterio de estandarización se puede aplicar a otras variables numéricas.

Para el cálculo de intervalos se plantea cuál es la probabilidad de que un caso esté en el intervalo de la media más/menos n-veces la desviación típica, como aplicación del Teorema de Tchebycheff, simbólicamente y el gráfico se muestran en la Tabla 84.

Tabla 84 Probabilidad por intervalos.

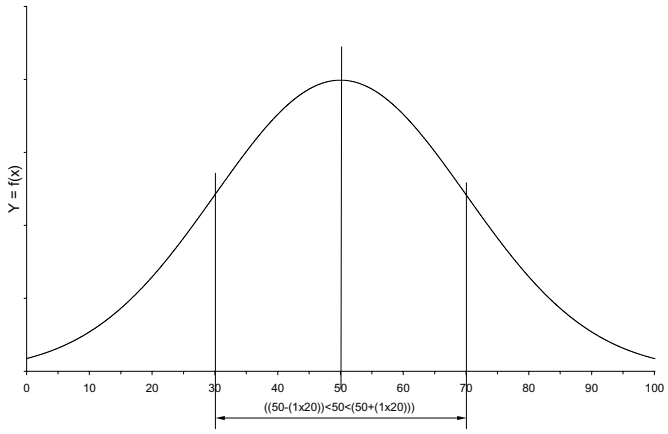
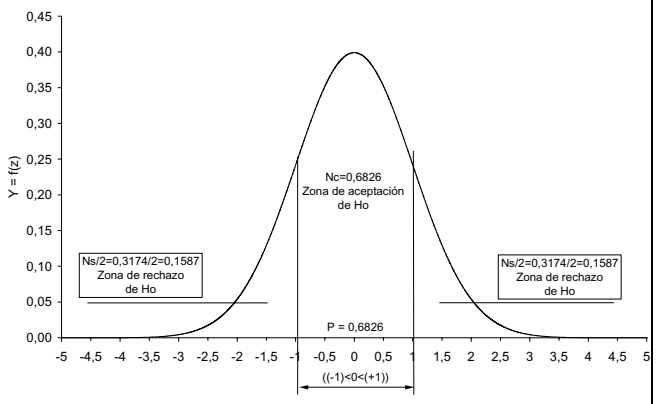
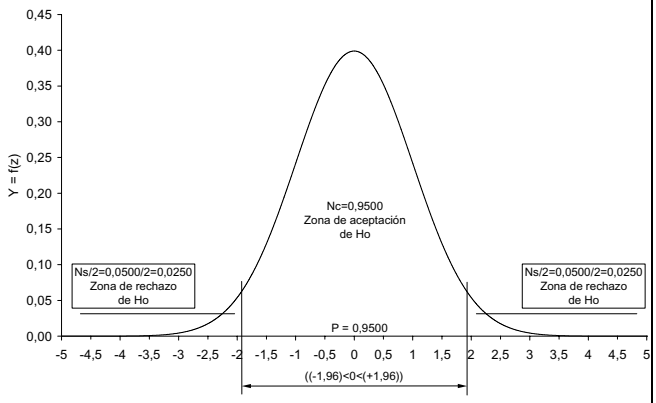
Intervalo	Gráfico
Genéricamente: $P_{\|\overline{X}4/n\Delta S\{\overline{X}\{\|\overline{X}2/n\Delta S}$ Aplicado a una variable $N_{(50,20)}$: El intervalo para *n* = 1 $\|\|504/1\Delta20\{50\{\|502/1\Delta20$	
El intervalo para *n* = 2: $\|\|504/2\Delta20\{50\{\|502/2\Delta20$	
Al transformar cualquier variable en puntuación típica se convierte en una variable *Z*, $N_{(0,1)}$. Al sustituir los valores de la media de 50 por la media de *z* = 0 y de la desviación típica de 20 por el de *z* = 1, los intervalos de probabilidad para diferentes valores de *n*, son:	

Tabla 84 Probabilidad por intervalos.

Intervalo	Gráfico

n = 1,
$//04/1\Delta 1000\{0\{/02/1\Delta 1000$ simplificando,
$//410\{0\{/2100$ y la probabilidad,
$P_{//410\{0\{/2100} \mid 0{,}6826$
Llamamos:
$(41) \ni (21)$: Intervalo de confianza.
Nc: Nivel de confianza.
Ns: Nivel de significación.
Nc + *Ns* = 1
Ns = 1 – *Nc* = 1 – 0,6826 = 0,3174
$P_{//410\{0\{/2100} \mid$
$\int_{41}^{21} f(z)\Delta dz \mid \int_{41}^{0} f(z)\Delta dz\ 2\ \int_{0}^{21} f(z)\Delta dz \mid$
$0{,}3413\ 2\ 0{,}3413 \mid 0{,}6826$
	

n = 1,96,
$//41{,}960\{0\{/21{,}9600$ y la probabilidad,
$P_{//41{,}960\{0\{/21{,}9600} \mid 0{,}9500$
$(41{,}96) \ni (21{,}96)$: Intervalo de confianza.
Nc: Nivel de confianza.
Ns: Nivel de significación.
Nc + *Ns* = 1
Ns = 1 – *Nc* = 1 – 0,9500 = 0,0500
$P_{//41{,}960\{0\{/21{,}9600} \mid$
$\int_{41{,}96}^{21{,}96} f(z)\Delta dz \mid \int_{41{,}96}^{0} f(z)\Delta dz\ 2\ \int_{0}^{21{,}96} f(z)\Delta dz \mid$
$0{,}4750\ 2\ 0{,}4750 \mid 0{,}9500$
	

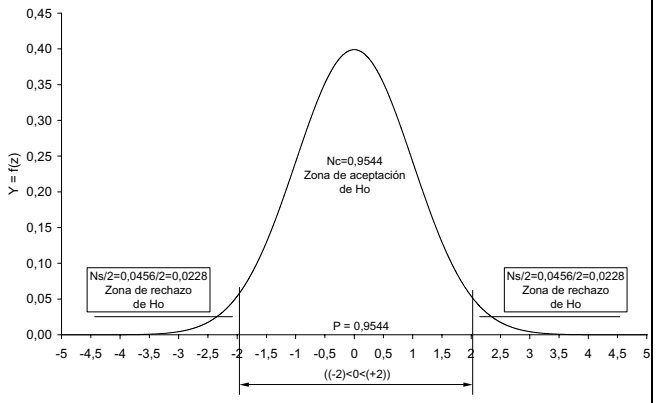
n = 2,
$//4\ 20\{0\{/2\ 200$ y la probabilidad,
$P_{//420\{0\{/2200} \mid 0{,}9544$
$(42) \ni (22)$: Intervalo de confianza.
Nc: Nivel de confianza.
Ns: Nivel de significación.
Nc + *Ns* = 1
Ns = 1 – *Nc* = 1 – 0,9544 = 0,0456
$P_{//420\{0\{/2200} \mid$
$\int_{42}^{22} f(z)\Delta dz \mid \int_{42}^{0} f(z)\Delta dz\ 2\ \int_{0}^{22} f(z)\Delta dz \mid$
$0{,}4772\ 2\ 0{,}4772 \mid 0{,}9544$
	

Tabla 84 Probabilidad por intervalos.

n = 3,

$//4\,3\{\ 0\ \{\ /2\,3$ y la probabilidad,

$P/43\{0\{/23 \mid\ 0{,}9974$

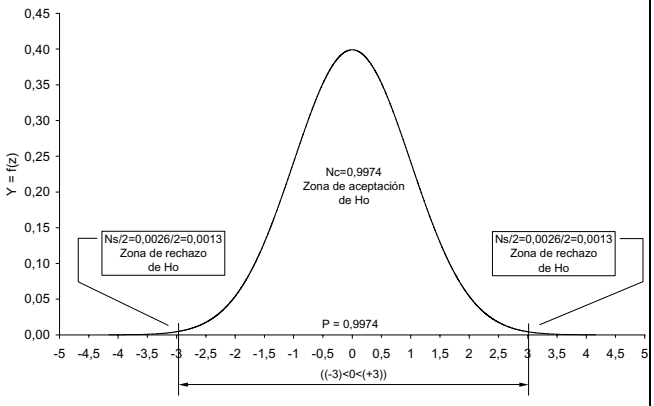
(43) э (23) : Intervalo de confianza.

Nc: Nivel de confianza.
Ns: Nivel de significación.
Nc + *Ns* = 1
Ns = 1 – *Nc* = 1 – 0,9974 = 0,0026

$P/43\{0\{/23 \mid$

$\int_{43}^{23} f(z)\Delta dz \mid \int_{43}^{0} f(z)\Delta dz\ 2\ \int_{0}^{23} f(z)\Delta dz \mid$

$0{,}4987\ 2\ 0{,}4987 \mid\ 0{,}9974$



La superficie contemplada dentro del intervalo de confianza se denomina *Nivel de Confianza* (simbólicamente *Nc*) y la superficie que queda por fuera a ambos lados del intervalo de confianza se denomina *Nivel de Significación* (*Ns*). La superficie total debajo de la curva normal, es la unidad, por lo tanto, *Nc + Ns = 1*, *Ns = 1 – Nc* y *Nc = 1 – Ns*. El *Ns* se distribuye por igual a ambos lados del intervalo de confianza. La superficie correspondiente al *Ns* también puede recibir el nombre de *p-valor* o ζ. La zona correspondiente al *Nc* es la de *aceptación de Ho* y la del *Ns rechazo de Ho*. Este aspecto se tratará detalladamente desde el Epígrafe 15. En la Tabla 85 se muestran algunos ejemplos de cálculo de probabilidades de una variable numérica continua.

Tabla 85 Otros ejemplos.

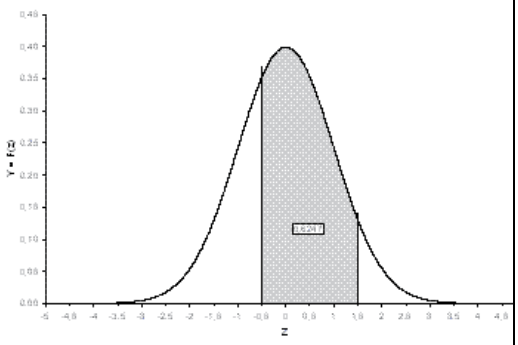
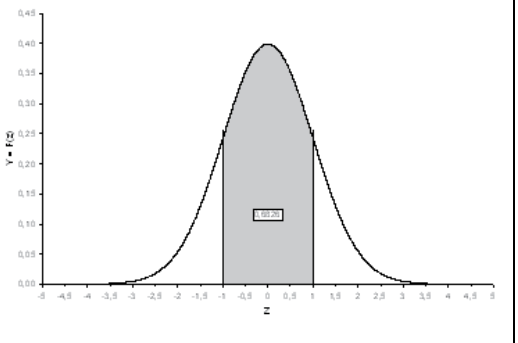
Ejemplo	Gráfico
Sea una variable *X* $N_{(50,20)}$, calcular la probabilidad de que un caso esté en el intervalo $x_1 = 40$ y $x_2 = 80$. $P/x_1 0\{x\{/x_2 0 \mid\ P/z_1 0\{z\{/z_2 0$ $P/40 0\{x\{/80 0 \mid\ P/\frac{40450}{20} 0\{z\{/\frac{80450}{20} 0, \ P/40 0\{x\{/80 0 \mid\ P/40{,}5 0\{z\{/1{,}5 0$ $P/40{,}5 0\{z\{/1{,}5 0 \mid\ \int_{40{,}5}^{1{,}5} f(z)dz$ $\int_{40{,}5}^{1{,}5} f(z)dz \mid\ \int_{40{,}5}^{0} f(z)dz\ 2\ \int_{0}^{1{,}5} f(z)dz \mid\ 0{,}1915\ 2\ 0{,}4332 \mid\ 0{,}6247$ La probabilidad de que un caso esté en el intervalo especificado de *X* es de 0,6247, o en el intervalo se encuentran el 62,5% de los casos (zona sombreada).	
Sea una variable *X* $N_{(10,2)}$, calcular la probabilidad de que un caso esté en el intervalo $x_1 = 8$ y $x_2 = 12$. $P/x_1 0\{x\{/x_2 0 \mid\ P/z_1 0\{z\{/z_2 0$ $P/8 0\{x\{/12 0 \mid\ P/\frac{8410}{2} 0\{z\{/\frac{12410}{2} 0, \ P/8 0\{x\{/12 0 \mid\ P/41 0\{z\{/1 0$ $P/41 0\{z\{/1 0 \mid\ \int_{41}^{1} f(z)dz$ $\int_{41}^{1} f(z)dz \mid\ \int_{41}^{0} f(z)dz\ 2\ \int_{0}^{1} f(z)dz \mid\ 0{,}3413\ 2\ 0{,}3413 \mid\ 0{,}6826$ La probabilidad de que un caso esté en el intervalo especificado de *X* es de 0,6826, o en el intervalo se encuentran el 68,3% de los casos (zona sombreada).	

Tabla 85 Otros ejemplos.

Sea una variable X $N_{(30,2)}$, calcular la probabilidad de que un caso esté por debajo de $x_1 = 25$.

$P(X < x_1) \quad P(Z < z_1), \quad P(X < 25) \quad P\left(Z < \frac{25-30}{2}\right)$

$P(X < 25) \quad P(z < -2,5)$

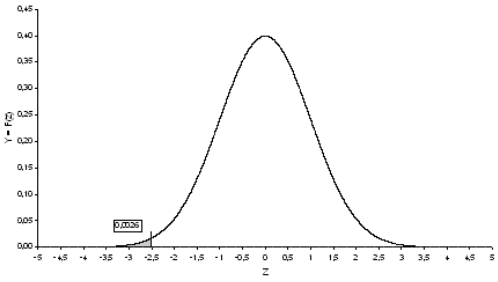
$P(z < -2,5) \quad \int_{-\infty}^{-2,5} f(z)dz \quad 0,0062$

Proceso de este cálculo:
La tabla de Z que se utiliza es entre $z = 0$ y z. La superficie pedida está por debajo de un valor negativo (-2,5). Al ser la curva simétrica, la superficie por debajo de $z = -2,5$, es la misma que la superficie por encima de $z = 2,5$. Para hallar la superficie por encima de $z = 2,5$, se ha de obtener la superficie que facilita la Tabla, que es entre $z = 0$ y $z = 2,5$, y proceder algebraicamente para obtener la superficie deseada.

$\int_{0}^{2,5} f(z)dz \quad \int_{-2,5}^{0} f(z)dz \quad 0,4938$

$\int_{-\infty}^{-2,5} f(z)dz \quad \int_{-\infty}^{-40} f(z)dz - \int_{-2,5}^{0} f(z)dz \quad 0,5 - 0,4938 \quad 0,0062$

La probabilidad de que un caso esté por debajo de 25 en la variable X es de 0,0062, o el 0,06% de los casos están por debajo de 25 (zona sombreada).



Sea una variable X $N_{(40,10)}$, calcular la probabilidad de que un caso esté por encima de $x_1 = 47$.

$P(x > x_1) \quad P(z > z_1), \quad P(x > 47) \quad P\left(z > \frac{47-40}{10}\right)$

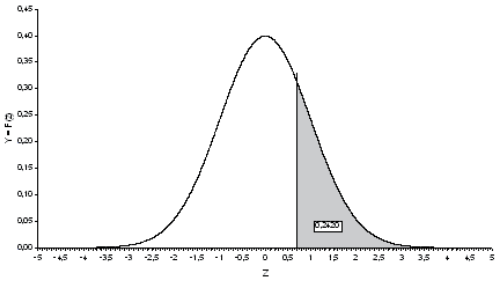
$P(x > 47) \quad P(z > 0,7)$

$P(z > 0,7) \quad \int_{0,7}^{\infty} f(z)dz \quad 0,2420$

Proceso:
Se busca en la Tabla la superficie comprendida entre $z = 0$ y $z = 0,7$ y se le resta a 0,5. La superficie buscada es 0,5 – 0,2580 = 0,2420

$P(z > 0,7) \quad \int_{0}^{\infty} f(z)dz - \int_{0}^{0,7} f(z)dz \quad 0,5 - 0,2580 \quad 0,2420$

La probabilidad de que un caso esté por encima de 47 en la variable X es de 0,2420, o el 24,2% de los casos están por encima de 47 (zona sombreada).



11.3 Otras funciones: θ^2, t y F (variables continuas).

Otras variables tipificadas que se utilizan en Sociología son: θ^2, t y F. La función que genera la curva es diferente a la Z y presentan la característica de tener grados de libertad (*gl*). Cada valor de grado de libertad genera una tabla de función de densidad de probabilidad distinta, pero los conceptos y aplicación de la probabilidad son iguales que los vistos para la función de densidad de probabilidad de la normal tipificada.

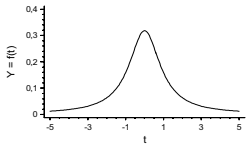
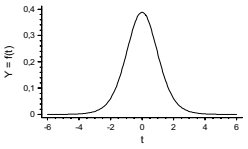
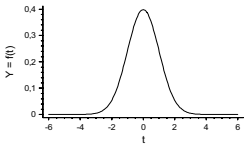
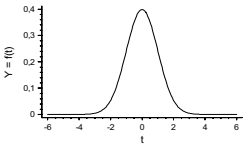
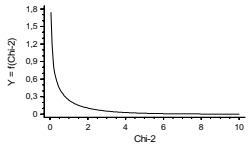
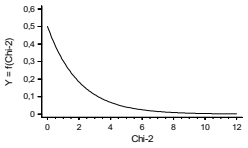
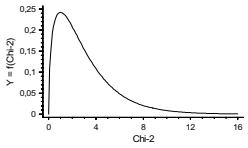
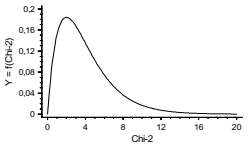
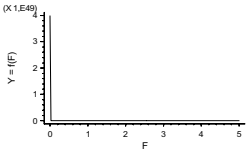
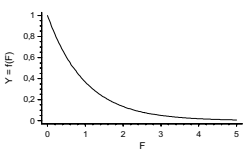
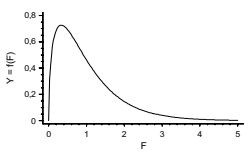
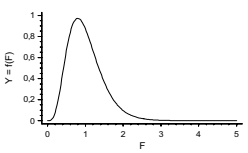
La representación de los gráficos de las variables mencionadas se realiza en un sistema de coordenadas cartesianas de dos dimensiones. En el eje de abscisas u horizontal se representa la variable y en el eje de ordenadas o vertical la *Y,* considerando que $y = f(x)$. Las variables Z y t, toman valores de $-\infty$ a ∞, y θ^2 y F sólo toman valores positivos. Las distribuciones tienden a normalizarse a medida que aumentan los *gl*. La superficie bajo la curva y por encima del eje de abscisas vale la unidad y representa al total de los casos, por lo

que se puede hablar en términos de probabilidad o de porcentajes. La forma de obtener los grados de libertad se verán en los Epígrafes correspondientes a los desarrollos de los estadísticos, en este apartado sólo se indicarán los valores (Tabla 86).

Tabla 86 Funciones de densidad de probabilidad: θ^2, t y F.

La distribución *t de Student* se hace más apuntada a medida que aumentan los *gl* y a partir de 10.000 se va aproximando a *Z*. En 30.000 *gl* se puede decir que es *Z*. Estos valores son orientativos para indicar la relación entre *t* y *Z*.

Los grados de libertad de la distribución *t* se obtienen restando una unidad al tamaño de la muestra (n-1). A medida que n aumenta, la diferencia entre la distribución *Z* y *t* desaparece.

	gl = 1	*gl* = 10	*gl* = 1.000	*gl* = 10.000
t				
	gl = 1	*gl* = 2	*gl* = 3	*gl* = 4
θ^2				
	gln = 1; *gld* = 1.000*	*gln* = 2; *gld* = 1.000	*gln* = 3; *gld* = 1.000	*gln* = 10; *gld* = 1.000
F				

Nota:
* *gln*: *gl* numerador y *gld*: *gl* denominador.

Tabla 87 Funciones de densidad de probabilidad: θ^2, t y F.

Se muestran las curvas superpuestas para ver la relación entre ellas

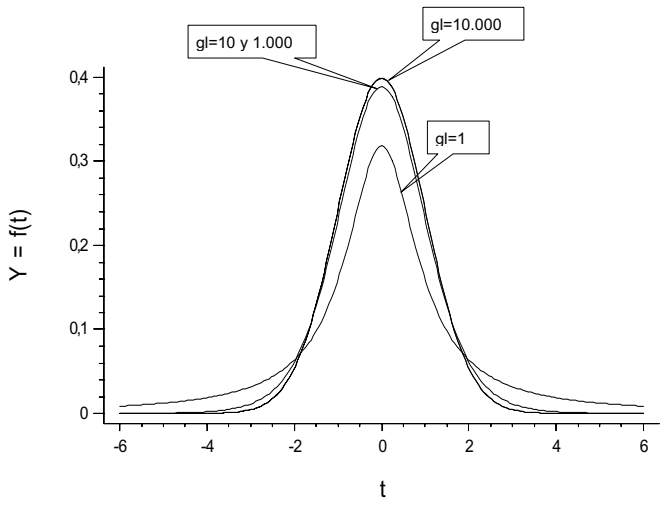
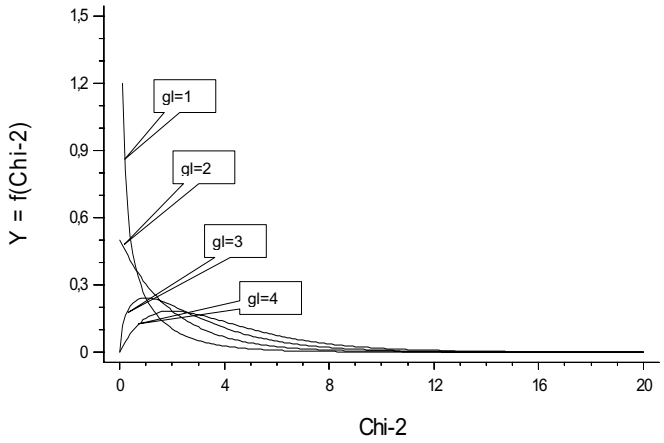
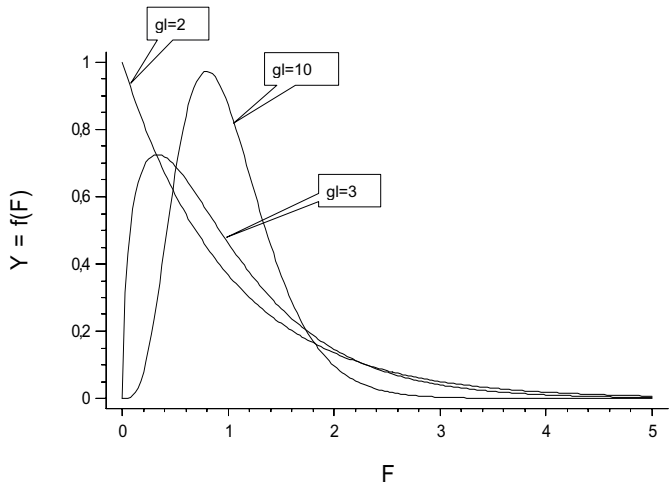
t	Las curvas de *gl* = 10 y *gl* = 1.000 están superpuestas. A medida que aumentan los grados de libertad aumenta la curtosis de la curva y las colas laterales se juntan al eje de abscisas. La curva tiende a normalizarse.	

Tabla 87 Funciones de densidad de probabilidad: θ^2, t y F.

θ^2	La curva tiende a normalizarse cuando aumentan los grados de libertad. Este comentario debe ser leído después de estudiar el capítulo de "Tablas de Contingencia": La distribución de θ^2, con *gl* = 1 no se puede utilizar en el contraste de Hipótesis porque su distribución no es normal y se aplica la corrección por la continuidad de θ^2 de Yates. Los programas estadísticos utilizados aplican automáticamente esta corrección. Debería ocurrir lo mismo con *gl* = 2 porque la distribución tampoco es normal, según el gráfico, pero en este caso es habitual que los manuales y programas estadísticos utilicen θ^2 sin corrección.	
F	Los grados de libertad son del numerador. En todos los casos los del denominador son 1.000. La curva de *gl* = 1 no se muestra porque distorsiona el gráfico. La curva tiende a normalizarse cuando aumentan los grados de libertad del numerador. También se puede utilizar el símbolo F_S. La "*S*" es de George Waddel Snedecor que desarrolló la distribución de la variable aleatoria continua y le puso el símbolo de "*F*" en memoria de Ronald Aylmer Fisher. Este comentario debe ser leído después de estudiar el capítulo de "Análisis de Varianza". Un grado de libertad asociados al numerador le corresponde a 2 grupos, pero como la distribución de la variable *F* no es normal, no sería posible su aplicación, pero al ser dos grupos se puede utilizar la diferencia de medias (t-Student) y se llega a las mismas conclusiones en el contraste de Hipótesis que con la F. Cuando son dos grupos $F = t^2$. Con *gl* = 2 le corresponde a 3 grupos y se aplica F.	

12 Asociación de tablas de contingencia

El *análisis de asociación* de este capítulo es frecuencista o de frecuencias. Es el referente al análisis de asociación entre variables categóricas. Trata de detectar la existencia de asociación o dependencia entre las categorías de las variables categóricas de la tabla de contingencia a través del análisis de las frecuencias absolutas de las celdas. El otro análisis de asociación es el lineal del coeficiente de correlación de Pearson y la ecuación de la línea recta (Ver Epígrafe 16).

El tratamiento estadístico de las tablas de doble entrada se divide en dos partes, *descriptivo* y *analítico (análisis)*. El primero comprende la creación de la tabla, hacer el recuento para expresar las *frecuencias absolutas* y el cálculo de los porcentajes o proporciones para expresar las *frecuencias relativas*.

El Análisis consta de tres partes: detección de la Asociación (θ^2); fuerza y dirección de la asociación, y a qué celdas es debida la asociación. El proceso se resume en la Tabla 88.

Tabla 88 Estadísticos de la tabla de contingencia.

Secuencia lógica	Proceso de lectura*	Estadística	Partes	Estadístico	
Lectura, haya o no asociación	2º	Descriptiva	Frecuencias absolutas	Recuento	
Lectura, haya o no asociación	2º	Descriptiva	Frecuencias relativas	%TF %TC %TT Regla de Zeisel	
Determina la existencia de asociación	1º	Análisis	Asociación entre las variables	θ^2	
Sólo si se detecta asociación	3º	Análisis	Fuerza de la asociación	Coeficiente de cont. (cc) V de Cramer λ (fi ó phi)	Basados en θ^2
				Lambda (ς)	RPE**
Sólo si se detecta asociación	4º	Análisis	Dirección de la asociación	Gamma *d* de Somers tau-b de kendall tau-c de kendall	
Sólo si se detecta asociación	5º	Análisis	Asociación por celdas	Residuo Residuo tipificado (Zresiduo)	

Nota:

* El proceso de lectura es orientativo y secuencial, porque la lectura de la tabla es holístico, global. También puede ocurrir que aunque no haya asociación se quiera ver la asociación por celdas, porque como ya se tratará, puede ocurrir que la tabla presente asociación por el carácter acumulativo de la estandarización de los residuos, pero ninguna celda presente una asociación significativa y viceversa, puede que la tabla no presente asociación, pero que alguna celda tenga asociación significativa y el resto de las celdas la neutralizan en el global.

** Estadísticos RPE (estadísticos basados en la Reducción Proporcional del Error).

12.1 Cálculo de la asociación y contraste de hipótesis.

Este epígrafe trata el proceso de cálculo e interpretación estadística, dejando las cuestiones teóricas e interpretativas en un segundo plano. La tabla de doble entrada que se va a utilizar para seguir el procedimiento de cálculo estadístico, para ver las contingencias, es la Tabla 51 y se genera la Tabla 89.

Tabla 89 Contraste de hipótesis en una tabla de contingencia.

Estado civil según el sexo

			Sexo Varón	Mujer	Total fila
Estado civil	*Soltero/a*	*fo*	36	41	77
		fe	38,1	38,9	77,0
		%TC	73,5%	82,0%	77,8%
		Residuo	-2,1	2,1	
		ZResiduo	-0,3	0,3	
	Casado/a	*fo*	6	3	9
		fe	4,5	4,5	9,0
		%TC	12,2%	6,0%	9,1%
		Residuo	1,5	-1,5	
		Zresiduo	0,7	-0,7	
	Pareja	*fo*	7	6	13
		fe	6,4	6,6	13,0
		%TC	14,3%	12,0%	13,1%
		Residuo	0,6	-0,6	
		Zresiduo	0,2	-0,2	
	Total columna	*fo*	49	50	99
		fe	49,0	50,0	99,0
		%TF	49,5%	50,5%	100,0%
		%TC	100,0%	100,0%	100,0%

En donde:
fo: Frecuencia observada.
fe: frecuencia esperada.
%TC: % total de columna.

La asociación entre variables categóricas pretende ver si existe relación entre la distribución de las frecuencias absolutas obtenidas por el cruce de las categorías de la variable de filas o considerada dependiente y la variable de columnas o considerada independiente. Esto es, si el hecho de pertenecer a una de las categorías de la variable dependiente, está relacionado con el hecho de pertenecer a una de las categorías de la variable independiente, determinando si la relación tiene alguna significación estadística, lo que no es garantía de que esa relación se encuentre en la realidad (porque la relación estadística sea espuria (falsa, engañosa)). Y si la asociación se da en la realidad, no quiere decir que sea única, ya que puede haber otras variables que también presenten asociación con la variable considerada independiente, con la considerada dependiente, y que haya otras variables intervinientes.

Cuando se realiza un análisis de estas características, se está aplicando el criterio "*céteris páribus"* que significa considerando todo lo demás constante, lo que difícilmente se puede asumir como cierto, ya que la complejidad de la realidad está influida por infinitas variables, aunque no todas tienen el mismo peso o la misma importancia. Con este escenario, el objetivo es sencillo, aplicar el estadístico y observar y utilizar la información que nos facilita, teniendo en cuenta la complejidad de la realidad humana que es la que normalmente se analiza y describe en sociología.

Para ver la existencia de asociación estadística entre dos variables categóricas, a partir de la Hipótesis científica o la Hipótesis de la investigación, se proponen las hipótesis estadísticas: la *Hipótesis Alternativa* y la *Hipótesis Nula*, simbólicamente representadas por H_1 y H_0, respectivamente.

La H_1 propone que existe relación de dependencia o asociación entre las variables, y H_0 que las variables son independientes, de tal manera que las dos hipótesis son mutuamente excluyentes. El proceso consiste en proponer la H_1 y contrastar la H_0, su aceptación supone rechazar la H_1, y su rechazo la aceptación de la H_1. Simbólicamente,

Tabla 90 Relación H_1 y H_0.

H_1	Aceptar
	Rechazar
H_0	Aceptar
	Rechazar

12.2 Protocolo de contraste de Hipótesis

El protocolo de contraste de hipótesis propuesto es:

1. En la H_1 se propone la relación de asociación o dependencia entre las variables.
2. Expresión del nivel de medida de las variables.
3. Expresión de la relación entre las variables
4. En la H_0 se propone la negación de la H_1 que es la no asociación o independencia entre las variables.
5. La decisión del estadístico para realizar el contraste de la H_0 está determinada por el nivel de medida de las variables y la estructura de la matriz de datos.
6. Criterio para aceptar o rechazar la H_0. Normalmente $Ns = 0{,}05$ ó $Ns = 0{,}01$.

El contraste de hipótesis es un proceso lógico-matemático-estadístico. Comienza desde un planteamiento en formato teórico-texto, después se procede desde el nivel estadístico-matemático para resolver la aceptación o rechazo de la H_0. Al ser mutuamente excluyentes, la aceptación de H_0 supone el rechazo de H_1 y el rechazo de H_0 supone la aceptación de H_1. El proceso se completa con el retorno al formato teórico-texto pero aceptando o rechazando la H_1.

En el caso de la Tabla 89, el protocolo es:

1. H_1: "Existe asociación, relación o dependencia entre el sexo de los individuos y el estado civil de los mismos" o de forma abreviada "El sexo influye en el estado civil de los individuos". Este formato puede confundir con una relación de causa-efecto. Se plantea que los sucesos son dependientes.
2. *Sexo*: variable categórica nominal. *Estado civil*: variable categórica nominal.
3. Variable considerada como independiente: *sexo*. Variable considerada como dependiente: *estado civil*.
4. H_0: "No existe asociación o dependencia entre el sexo de los individuos y el estado civil de los mismos" o de forma abreviada "El sexo no influye en el estado civil de los individuos". Se plantea que los sucesos son independientes.
5. Estadístico: θ^2, por ser las dos variables categóricas:
6. Criterio de aceptación/rechazo de H_0, $Ns = 0{,}05$.

12.3 Proceso de contraste de Hipótesis

Según el Epígrafe de las probabilidades (pág. 117) la probabilidad de que ocurra un suceso elemental es igual a los hechos favorables dividido por los hechos posibles. Simbólicamente:

$P_{(S_1)} \mid \frac{hechos\_favorables}{Hechos\_posibles}$	Fórmula 53

Entonces la probabilidad de ser *varón* o la probabilidad de estar *soltero/a*, sería:

$$P_{(S_v)} \mid \frac{49}{99} \mid 0{,}495\,,\ P_{(S_s)} \mid \frac{77}{99} \mid 0{,}778$$

Que multiplicado por 100 quedaría expresado en porcentaje del total de columna o del total de fila. Pero el interés no es la probabilidad de ocurrencia de los sucesos elementales, sino la probabilidad de la intersección de los sucesos elementales. Esto es, la probabilidad de ocurrencia de la intersección de los sucesos elementales de ser *varón* y *soltero*. Simbólicamente:

$P_{(S_v \sim S_s)}$	Fórmula 54

Al contrastar la H_0, asumimos la independencia de las variables y la independencia de los sucesos elementales. Según el capítulo de probabilidades (pág. 117) si dos sucesos elementales son independientes pero mutuamente no excluyentes, la probabilidad de la intersección de dos sucesos es igual al producto de sus probabilidades. Simbólicamente:

$$P_{(S_v \sim S_s)} \mid P_{(S_v)} \Delta P_{(S_s)} \mid 0{,}495 \Delta 0{,}778 \mid 0{,}3851$$

Entonces, si la probabilidad de ser *varón* y *estar soltero* es 0,3851, asumiendo que los sucesos son independientes, mutuamente no excluyentes, entonces el número de *varones solteros* se espera que sea de 38,1. Simbólicamente:

$$P_{(S_v \sim S_s)} \mid \frac{Hechos\_favorables}{Hechos\_posibles} \mid \frac{Hf}{99}, \text{ entonces, } 0{,}3851 \mid \frac{Hf}{99} \text{ y } Hf \mid 0{,}3851 \Delta 99 \mid 38{,}1$$

Entonces los *Hechos favorables*, que son el número de *varones solteros esperados*, son 38,1 y se llama la *frecuencia esperada* (*fe*), y son los casos que debería haber si los sucesos fuesen independientes mutuamente no excluyentes (la repetición de esta característica es debido a que es la clave del proceso). Este valor se llama *valor teórico* o *modelo teórico* porque es resultado de un modelo probabilístico. Los 36 *varones solteros* es la *frecuencia observada* (*fo*) o el modelo empírico, aquello que ha ocurrido en la realidad, lo que se ha observado, y los 38,1 los que deberían haber sido si los sucesos fuesen independientes mutuamente no excluyentes. Entonces lo que se ha conseguido es un modelo probabilístico de referencia con el que comparar el modelo empírico. La diferencia entre el valor empírico y el teórico se llama *residuo* o *residual*. Simbólicamente,

$Res \mid\ fo\ 4\ fe$	Fórmula 55

El proceso se debe aplicar a todas las celdas de la tabla.

Tabla 91 Cálculo de las frecuencias esperadas y los residuos.

Intersección	Probabilidad de la intersección de los sucesos elementales	Frecuencias esperadas	Residuos
Varón – soltero	$P_{(S_v \sim S_s)} \mid\ P_{(S_v)} \Delta P_{(S_s)} \mid\ 0{,}495 \Delta 0{,}778 \mid\ 0{,}3851$	$fe \mid\ 0{,}3851 \Delta 99 \mid\ 38{,}1$	$Res_{11} \mid\ fo_{11}\ 4\ fe_{11} \mid\ 36\ 4\ 38{,}1 \mid\ 42{,}1$
Mujer – soltera	$P_{(S_m \sim S_s)} \mid\ P_{(S_m)} \Delta P_{(S_s)} \mid\ 0{,}505 \Delta 0{,}778 \mid\ 0{,}3929$	$fe \mid\ 0{,}3929 \Delta 99 \mid\ 38{,}9$	$Res_{12} \mid\ fo_{12}\ 4\ fe_{12} \mid\ 41\ 4\ 38{,}9 \mid\ 2{,}1$
Varón – casado	$P_{(S_v \sim S_c)} \mid\ P_{(S_v)} \Delta P_{(S_c)} \mid\ 0{,}495 \Delta 0{,}091 \mid\ 0{,}0450$	$fe \mid\ 0{,}0450 \Delta 99 \mid\ 4{,}5$	$Res_{21} \mid\ fo_{21}\ 4\ fe_{21} \mid\ 6\ 4\ 4{,}5 \mid\ 1{,}5$
Mujer – casada	$P_{(S_m \sim S_c)} \mid\ P_{(S_m)} \Delta P_{(S_c)} \mid\ 0{,}505 \Delta 0{,}091 \mid\ 0{,}0460$	$fe \mid\ 0{,}0460 \Delta 99 \mid\ 4{,}5$	$Res_{22} \mid\ fo_{22}\ 4\ fe_{22} \mid\ 3\ 4\ 4{,}5 \mid\ 41{,}5$
Varón – pareja	$P_{(S_v \sim S_p)} \mid\ P_{(S_v)} \Delta P_{(S_p)} \mid\ 0{,}495 \Delta 0{,}131 \mid\ 0{,}0648$	$fe \mid\ 0{,}0648 \Delta 99 \mid\ 6{,}4$	$Res_{31} \mid\ fo_{31}\ 4\ fe_{31} \mid\ 7\ 4\ 6{,}4 \mid\ 0{,}6$
Mujer – pareja	$P_{(S_m \sim S_p)} \mid\ P_{(S_m)} \Delta P_{(S_p)} \mid\ 0{,}505 \Delta 0{,}131 \mid\ 0{,}0662$	$fe \mid\ 0{,}0662 \Delta 99 \mid\ 6{,}6$	$Res_{32} \mid\ fo_{32}\ 4\ fe_{21} \mid\ 6\ 4\ 6{,}6 \mid\ 40{,}6$

Las frecuencias observadas son el modelo empírico del cruce de las variables *sexo* y *estado civil*, del que desconocemos si su relación es dependencia o independencia. Las frecuencias esperadas son el modelo probabilístico teórico, del que sabemos que son las frecuencias que deberían ser si los sucesos fuesen independientes, mutuamente no excluyentes. La comparación del modelo empírico con el teórico es mediante resta simple de las frecuencias observadas menos las frecuencias esperadas de cada celda y que se han llamado *residuos*.

Si todos los residuos fuesen cero, entonces es que los dos modelos son iguales y lo que era desconocido se hace conocido, las frecuencias observadas son como las esperadas, independientes. Como este es el planteamiento de la H_0, entonces nos llevaría a aceptarla. Significaría que no tiene ninguna relación el género en cuanto a *sexo* con el *estado civil* de las personas. Significa que hay *solteros, casados* y *en pareja* tanto entre los varones como entre las mujeres.

Pero si los residuos son distintos de cero, muy distintos, con una diferencia enorme, entonces el modelo empírico es distinto del teórico y si no son independientes, entonces son dependientes. Este proceso nos lleva a rechazar la H_0 y por lo tanto a aceptar la H_1. La conclusión sería que el *estado civil* de las personas está relacionado con el *sexo*.

Pero entre residuos igual a cero y los residuos enormes, hay una escala dentro de la cual hay que determinar hasta que valor se acepta H_0, o lo que es lo mismo a partir de que valor se consideran grandes los residuos para rechazar la H_0 y aceptar la H_1. También se presenta el inconveniente de que una cosa es grande o pequeña dependiendo de con qué se compare. Por ejemplo, una diferencia de 5, entre 20 y 25, puede ser grande. Pero puede resultar pequeña si es entre 995 y 1.000. Incluso puede llegar a ser insignificante si las cantidades son mayores.

Hay que resolver dos problemas, estandarizar los valores y establecer un criterio para determinar cuando los residuos se pueden considerar grandes o pequeños o de forma más precisa cuándo son significativamente grandes o significativamente pequeños. La estandarización de los residuos se consigue aplicando el estadístico θ^2 de Pearson y el criterio para determinar cuándo son grandes o pequeños es aplicar los conceptos de probabilidad de la distribución de la variable estandarizada θ^2. La estandarización de los residuos es:

$\theta_e^2 = \sum_{i=1}^{F}\sum_{j=1}^{C} \frac{(fo_{ij} - fe_{ij})^2}{fe_{ij}}$	Fórmula 56

Que es el sumatorio para todas las celdas de la diferencia entre las frecuencias observadas menos las frecuencias esperadas, elevado al cuadrado, dividido (relativizado o estandarizado) por las frecuencias esperadas. Y es el estadístico chi-cuadrado estimado, y en el caso de la Tabla 89 es:

$$\theta_e^2 = \sum_{i=1}^{F}\sum_{j=1}^{C} \frac{(fo_{ij} - fe_{ij})^2}{fe_{ij}} = \sum_{i=1}^{F=3}\sum_{j=1}^{C=2} \frac{(fo_{ij} - fe_{ij})^2}{fe_{ij}} =$$

$$\frac{(fo_{11} - fe_{11})^2}{fe_{11}} + \frac{(fo_{12} - fe_{12})^2}{fe_{12}} + \frac{(fo_{21} - fe_{21})^2}{fe_{21}} + \frac{(fo_{22} - fe_{22})^2}{fe_{22}} + \frac{(fo_{31} - fe_{31})^2}{fe_{31}} + \frac{(fo_{32} - fe_{32})^2}{fe_{32}} =$$

$$\frac{(36 - 38{,}1)^2}{38{,}1} + \frac{(41 - 38{,}9)^2}{38{,}9} + \frac{(6 - 4{,}5)^2}{4{,}5} + \frac{(3 - 4{,}5)^2}{4{,}5} + \frac{(7 - 6{,}4)^2}{6{,}4} + \frac{(6 - 6{,}6)^2}{6{,}6} = 1{,}39$$

Para saber si el valor de θ_e^2 es grande o no, se comprueba con la distribución que sigue el valor de este estadístico, que es la distribución θ^2, con *n* grados de libertad. Los grados de libertad de la distribución θ^2 para tablas de contingencia se calculan,

$gl = (f - 1)\Delta(c - 1)$	En donde: *gl*: Grados de libertad. *f*: Número de filas de la tabla. *c*: Número de columnas de la tabla.	Fórmula 57

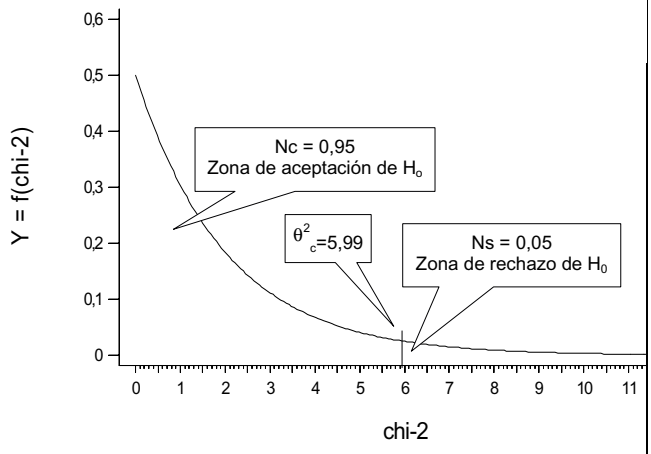
En el caso de la Tabla 89 es $gl = (f-1)\Delta(c-1) = (3-1)\Delta(2-1) = 2\Delta1 = 2$ y la distribución de θ^2 que se utiliza es la de la Tabla 92.

Tabla 92 Cálculo de chi-cuadrado crítico.

Cálculo de θ_c^2:

θ_c^2	*gl* = 2 *Ns* = 0,05	= 5,9915

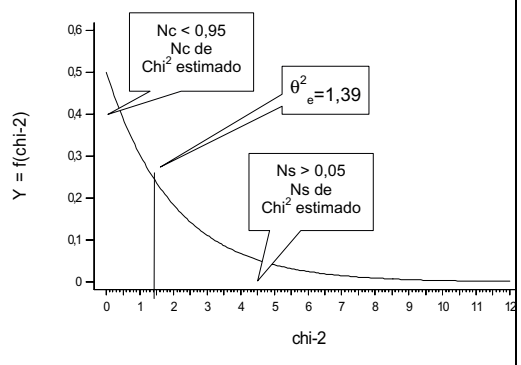
El valor del θ_c^2 (Chi-cuadrado crítico) es el que deja una superficie a su derecha, por debajo de la curva de θ^2 y por encima del eje de abscisas, igual al valor del *Ns*, que en este caso es 0,05.
Este valor se obtiene en la Tabla de θ^2 del Anexo 2 (pág. 333). Los *gl* = 2 son la entrada de filas y el *Ns* = 0,05 es la entrada de columnas. La celda en la que se cruzan es el valor de θ_c^2 que cumple estas condiciones.

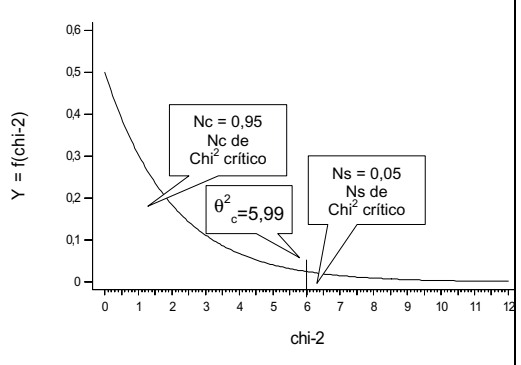


Si el valor de los residuos estandarizados, o sea, θ_e^2 (Chi-cuadrado estimado) fuese cero, entonces las frecuencias observadas coincidirían con las esperadas, y el valor de θ_e^2 estaría en la zona de *aceptación* y se aceptaría H_o. Suponiendo que θ_e^2 no es cero, que está en la zona de $\theta^2 = 2$, entonces no siendo cero, se asume que las diferencias entre las frecuencias observadas y las esperadas son tan pequeñas que son debidas al azar, sigue en la zona de *aceptación*, y por lo tanto no se puede asumir rechazar la H_o, no siendo cero la diferencia, no es significativamente distinta de cero. Este proceso se podría repetir hasta que la pregunta fuese ¿Cuándo las diferencias son lo suficientemente grandes como para asumir que podemos rechazar la H_o y aceptar la H_1? Cuando θ_e^2 sea igual o mayor al valor del θ_c^2 (Chi-cuadrado crítico), entonces estaría en la zona de *rechazo* y se rechazaría H_o, aceptando la H_1.

El θ_e^2 de la Tabla 89 es 1,39, para saber si se acepta o rechaza la H_o, tenemos que comparar el θ_e^2 con el θ_c^2. Mediante la representación gráfica se ilustra el contraste (Gráfico 16).

Gráfico 16 Comparación de θ_e^2 con θ_c^2.





Ahora se puede establecer el siguiente esquema para determinar la aceptación o rechazo de H_0.

Se acepta H_o si: $\theta_e^2 \{ \theta_c^2 \sum Nc_e \{ Nc_c \sum Ns_e \} Ns_c$

Se rechaza H_o si: $\theta_e^2 \varnothing \theta_c^2 \sum Nc_e \varnothing Nc_c \sum Ns_e \Omega Ns_c$

En este caso, se acepta H_o porque el $\theta_e^2 \{ \theta_c^2$, que equivale a decir que $Nc_e \{ Nc_c$ y equivalente también a que $Ns_e \} Ns_c$. Por lo tanto se puede concluir que al *Ns* de 0,05 no existe asociación entre las variables *sexo* y *estado civil*, o que son independientes.

Como no hay asociación entre las dos variables, no tiene sentido mirar la fuerza, la dirección y a que celdas es debida la asociación. Sí es de interés la estadística descriptiva bivariable o la lectura de las frecuencias relativas (ver Tabla 51 y Tabla 52).

12.4 Contraste de hipótesis de una tabla de contingencia que presenta asociación

El proceso completo (Tabla 88) de una tabla que presenta asociación se realiza con dos variables extraídas de la Encuesta Social Europea.[62] Se ha utilizado el criterio de ponderación y se ha aplicado a la muestra del Reino Unido por el interés de los resultados (Tabla 93).

[62] R. Jowell and the Central Co-ordinating Team, European Social Survey 2006/2007: Technical Report, London: Centre for Comparative Social Surveys, City University (2007). El servicio de Datos de las Ciencias Sociales Noruego (NDS) ha realizado la distribución de los datos. Disponible en: http://www.europeansocialsurvey.org/.

Tabla 93 Contraste de hipótesis de una tabla con asociación.

Pregunta B4:	Pregunta F4:
Por favor dígame en una escala de 0-10 cuál es la confianza que usted tiene en el Parlamento de su país. El 0 significa que usted no confía en absoluto, y el 10 significa que usted tiene plena confianza. Esta pregunta genera la variable *confianza en el parlamento de su país*. Para reducir el número de categorías se ha recodificado siguiendo el criterio tradicional de nota semántica: 1-4: *Suspenso*. 5-6: *Aprobado*. 7-8: *Notable*. 9-10: *Sobresaliente*.	Género de la persona entrevistada (seleccionada la persona por procedimientos aleatorios). Esta pregunta genera la variable: *sexo del entrevistado*.

Confianza en el Parlamento del país según el sexo, en Reino Unido.

			Sexo		
			Varón	*Mujer*	*Total fila*
Confianza en el Parlamento del país.	*Suspenso*	*fo*	569	630	1.199
		fe	574,5	624,5	1.199,0
		%TC	50,6	51,6	51,1
		%TT	24,3	26,9	
		Residuo	-5,5	5,5	
		ZResiduo	-0,23	0,22	
	Aprobado	*fo*	313	421	734
		fe	351,7	382,3	734
		%TC	27,85	34,45	31,3
		%TT	13,3	17,9	
		Residuo	-38,7	38,7	
		Zresiduo	-2,06	1,98	
	Notable	*fo*	213	140	353
		fe	169,1	183,9	353
		%TC	19,0	11,5	15,0
		%TT	9,1	6,0	
		Residuo	43,9	-43,9	
		Zresiduo	3,37	-3,24	
	Sobresaliente	*fo*	29	31	60
		fe	28,7	31,3	60
		%TC	2,6	2,5	2,6
		%TT	1,2	1,3	
		Residuo	0,3	-0,3	
		ZResiduo	0,05	-0,05	
	Total columna	*fo*	1.124	1.222	2.346
		fe	1.124,0	1.222,0	2.346,0
		%TF	100,0	100,0	100,0
		%TC	47,9	52,1	100,0

Proceso de contraste de hipótesis:

1. H_1: "Existe asociación o dependencia entre el sexo de los individuos y la confianza que tienen en el Parlamento" o de forma abreviada "El sexo influye en la confianza en el Parlamento". Se plantea que los sucesos son dependientes.
2. *Sexo*: variable categórica nominal. *Confianza en el Parlamento*: variable categórica ordinal.
3. Variable considerada como independiente: *sexo*. Variable considerada como dependiente: *Confianza en el Parlamento*.
4. H_0: "No existe asociación o dependencia entre el sexo de los individuos y la confianza que tienen en el Parlamento" o de forma abreviada "El sexo no influye en la confianza en el Parlamento". Se plantea que los sucesos son independientes.
5. Estadístico: θ^2, por ser las dos variables categóricas:
6. Criterio de aceptación/rechazo de H_0, $Ns = 0{,}05$.

Tabla 94 Cálculo de las frecuencias esperadas y los residuos.			
Intersección	Probabilidad de la intersección de los sucesos elementales	Frecuencias esperadas	Residuos
Varón – suspenso	$P_{(S_v \cap S_s)} = P_{(S_v)} \cdot P_{(S_s)} = 0{,}479 \cdot 0{,}511 = 0{,}2449$	$fe = 0{,}2449 \cdot 2.346 = 574{,}5$	$\mathrm{Re}\, s_{11} = fo_{11} - fe_{11} = 569 - 574{,}5 = -45{,}5$
Mujer – suspenso	$P_{(S_m \cap S_s)} = P_{(S_m)} \cdot P_{(S_s)} = 0{,}521 \cdot 0{,}511 = 0{,}2662$	$fe = 0{,}2662 \cdot 2.346 = 624{,}5$	$\mathrm{Re}\, s_{12} = fo_{12} - fe_{12} = 630 - 624{,}5 = 5{,}5$
Varón – aprobado	$P_{(S_v \cap S_a)} = P_{(S_v)} \cdot P_{(S_a)} = 0{,}479 \cdot 0{,}313 = 0{,}1499$	$fe = 0{,}1499 \cdot 2.346 = 351{,}7$	$\mathrm{Re}\, s_{21} = fo_{21} - fe_{21} = 313 - 351{,}7 = -438{,}7$
Mujer – aprobado	$P_{(S_m \cap S_a)} = P_{(S_m)} \cdot P_{(S_a)} = 0{,}521 \cdot 0{,}313 = 0{,}1630$	$fe = 0{,}1630 \cdot 2.346 = 382{,}3$	$\mathrm{Re}\, s_{22} = fo_{22} - fe_{22} = 421 - 382{,}3 = 38{,}7$
Varón – notable	$P_{(S_v \cap S_n)} = P_{(S_v)} \cdot P_{(S_n)} = 0{,}479 \cdot 0{,}150 = 0{,}0721$	$fe = 0{,}0721 \cdot 2.346 = 169{,}1$	$\mathrm{Re}\, s_{31} = fo_{31} - fe_{31} = 213 - 169{,}1 = 43{,}9$
Mujer – notable	$P_{(S_m \cap S_n)} = P_{(S_m)} \cdot P_{(S_n)} = 0{,}521 \cdot 0{,}150 = 0{,}0784$	$fe = 0{,}0784 \cdot 2.346 = 183{,}9$	$\mathrm{Re}\, s_{32} = fo_{32} - fe_{21} = 140 - 183{,}9 = -443{,}9$
Varón – sobresaliente	$P_{(S_v \cap S_s)} = P_{(S_v)} \cdot P_{(S_s)} = 0{,}479 \cdot 0{,}026 = 0{,}0123$	$fe = 0{,}0123 \cdot 2.346 = 28{,}7$	$\mathrm{Re}\, s_{41} = fo_{41} - fe_{41} = 29 - 28{,}7 = 0{,}3$
Mujer – sobresaliente	$P_{(S_m \cap S_s)} = P_{(S_m)} \cdot P_{(S_s)} = 0{,}521 \cdot 0{,}026 = 0{,}0133$	$fe = 0{,}0133 \cdot 2.346 = 31{,}3$	$\mathrm{Re}\, s_{42} = fo_{42} - fe_{41} = 31 - 31{,}3 = -40{,}3$

Para contrastar la hipótesis nula (H_0) hay que calcular el chi-cuadrado estimado, y en el caso de la Tabla 93 es:

$$\theta_e^2 = \sum_{i=1}^{F} \sum_{j=1}^{C} \frac{(fo_{ij} - fe_{ij})^2}{fe_{ij}} = \sum_{i=1}^{F=4} \sum_{j=1}^{C=2} \frac{(fo_{ij} - fe_{ij})^2}{fe_{ij}}$$

$$\frac{(fo_{11} - fe_{11})^2}{fe_{11}} + \frac{(fo_{12} - fe_{12})^2}{fe_{12}} + \frac{(fo_{21} - fe_{21})^2}{fe_{21}} + \frac{(fo_{22} - fe_{22})^2}{fe_{22}} + \frac{(fo_{31} - fe_{31})^2}{fe_{31}} + \left[\frac{(fo_{32} - fe_{32})^2}{fe_{32}} + \frac{(fo_{41} - fe_{41})^2}{fe_{41}} + \frac{(fo_{42} - fe_{42})^2}{fe_{42}} \right] =$$

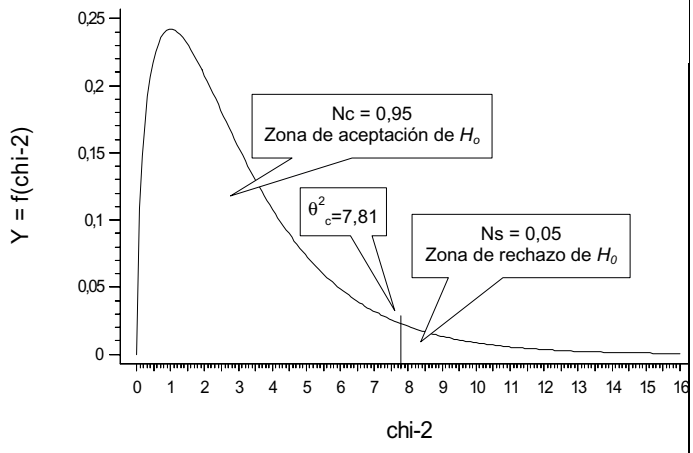
$$\frac{(569 - 574{,}5)^2}{574{,}5} + \frac{(630 - 624{,}5)^2}{624{,}5} + \frac{(313 - 351{,}7)^2}{351{,}7} + \frac{(421 - 382{,}3)^2}{382{,}3} + \frac{(213 - 169{,}1)^2}{169{,}1} + \left[\frac{(140 - 183{,}9)^2}{183{,}9} + \frac{(29 - 28{,}7)^2}{28{,}7} + \frac{(31 - 31{,}3)^2}{31{,}3} \right] = 30{,}12$$

Tabla 95 Cálculo de chi-cuadrado crítico.

Cálculo de θ_c^2:

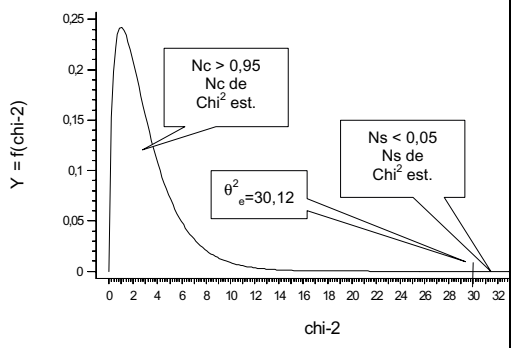
θ_c^2	*gl* = 3 *Ns* = 0,05	= 7,8147

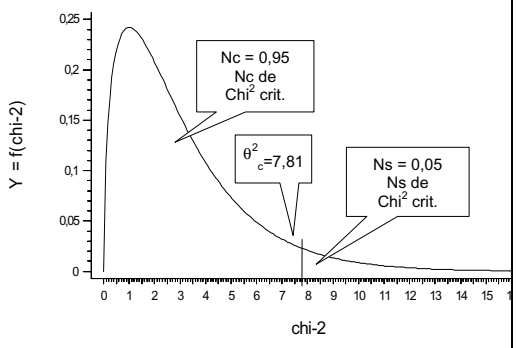
El valor de θ_c^2 | 7,8147 es el que deja una superficie a su derecha, por debajo de la curva de θ^2 y por encima del eje de abscisas, igual a 0,05, que se llama *Ns* y para *gl* = 3.
Este valor se obtiene en la Tabla de θ^2 del Anexo 2 (pág. 333). Los *gl* = 3 son la entrada de filas y el *Ns* = 0,05 es la entrada de columnas. La celda en la que se cruzan es el valor de θ_c^2 que cumple estas condiciones.



El θ_e^2 de la Tabla 93 es 30,12, para saber si se acepta o rechaza la H_o, se compara el θ_e^2 con el θ_c^2. Mediante la representación gráfica se ilustra el contraste (Gráfico 17).

Gráfico 17 Comparación de θ_e^2 con θ_c^2.





El siguiente esquema sirve de ayuda para la aceptación o rechazo de H_0.

Se acepta H_o si: $\theta_e^2 \{ \theta_c^2 \sum Nc_e \{ Nc_c \sum Ns_e \} Ns_c$

Se rechaza H_o si: $\theta_e^2 \varnothing \theta_c^2 \sum Nc_e \varnothing Nc_c \sum Ns_e \Omega Ns_c$

En este caso se rechaza H_o porque el $\theta_e^2 \varnothing \theta_c^2$, que equivale a decir que $Nc_e \varnothing Nc_c$ y equivalente también a que $Ns_e \Omega Ns_c$. Por lo tanto se puede concluir que al *Ns* de 0,05

(incluso al *Ns* = 0,01 ya que con *gl* = 3, $\theta_c^2 \mid 11{,}35$) existe asociación entre las variables *sexo* y *confianza en el Parlamento del país*.

Al existir asociación entre las variables, hay que ver: fuerza de la asociación, dirección de la asociación (en esta ocasión no se trata porque no son de nivel de medida ordinal las dos variables) y a qué celdas es debida la asociación. La lectura de porcentajes ya se ha trabajado y salvo excepción, en esta tabla no se tratará.

Los estadísticos que indican la fuerza de la asociación son: Coeficiente de contingencia (*cc*), *V* de Cramer, λ (fi ó phi) y Lambda (ς) de Goodman y Kruskal.

$cc \mid \sqrt{\frac{\theta_e^{\,2}}{\theta_e^{\,2} \; 2\, n}}$	*n*: número de casos de la tabla.	Fórmula 58
$cc_m \mid \sqrt{\frac{k \; 4 \; 1}{k}}$	El coeficiente de contingencia máximo se puede calcular con tablas cuadradas (*k* = *f* ó *c*).	Fórmula 59
$V \mid \sqrt{\frac{\theta_e^{\,2}}{n \Delta (k \; 4 \; 1)}}$	Siendo *k* el menor del número de filas o columnas, Si *k* = 2 y $\theta_e^{\,2} \} \, n$, V vale más que la unidad. Si *k* = 2, *V* es igual a λ	Fórmula 60
$\lambda \mid \sqrt{\frac{\theta_e^{\,2}}{n}}$	Si $\theta_e^{\,2} \} \, n$, λ vale más que la unidad.	Fórmula 61
$\varsigma_{yx} \mid \frac{P_{1y} \; 4 \; P_{2y}}{P_{1y}}$	RPE al considerar la variable de filas como dependiente (estadístico asimétrico). P_{1y}: Probabilidad de error en *y*. Es uno menos la probabilidad modal de *y*. P_{2y}: Sumatorio de las probabilidades de las celdas que no son moda de columnas (*x*).	Fórmula 62
$\varsigma_{xy} \mid \frac{P_{1x} \; 4 \; P_{2x}}{P_{1x}}$	RPE al considerar la variable de columnas como dependiente (estadístico asimétrico). P_{1x}: Probabilidad de error en *x*. Es uno menos la probabilidad modal de *x*. P_{2x}: Sumatorio de las probabilidades de las celdas que no son moda de filas (*y*).	Fórmula 63

La asociación de cada una de las celdas se analiza con los residuos tipificados. Estos residuos tienen la distribución de una variable $N_{(0,1)}$, la puntuación tipificada ó *z*, que permitirá saber a través de los residuos que frecuencias observadas y esperadas son significativamente distintas. Se puede repasar el Epígrafe 11.

$zres = \frac{fo_{ij} - fe_{ij}}{\sqrt{fe_{ij}}}$	Fórmula 64

Los valores de los estadísticos *cc, V,* λ y ς, de la Tabla 93 son:

$$cc = \sqrt{\frac{\theta_e^2}{\theta_e^2 + n}} = \sqrt{\frac{30{,}12}{30{,}12 + 2.346}} = 0{,}11$$

$$V = \sqrt{\frac{\theta_e^2}{n\Delta(k-1)}} = \sqrt{\frac{30{,}11}{2.346\Delta(2-1)}} = 0{,}11$$

$$\lambda = \sqrt{\frac{\theta_e^2}{n}} = \sqrt{\frac{30{,}11}{2.346}} = 0{,}11$$

$$\varsigma_{yx} = \frac{P_{1y} - P_{2y}}{P_{1y}} = \frac{/1 - 0{,}511\emptyset - /0{,}133 + 0{,}091 + 0{,}012 + 0{,}179 + 0{,}060 + 0{,}013\emptyset}{/1 - 0{,}511\emptyset} = \frac{0{,}489 - 0{,}488}{0{,}489} = 0{,}00$$

$$\varsigma_{xy} = \frac{P_{1x} - P_{2x}}{P_{1x}} = \frac{/1 - 0{,}521\emptyset - /0{,}243 + 0{,}133 + 0{,}060 + 0{,}012\emptyset}{/1 - 0{,}521\emptyset} = \frac{0{,}479 - 0{,}448}{0{,}479} = \frac{0{,}031}{0{,}479} = 0{,}06$$

Para una interpretación de los valores de *cc*, *V* y λ, hay que considerar dos cosas, la escala o rango de valores en los que se mueven y la característica de las variables que se están comparando o relacionando. El rango de valores que puede tomar *cc* es de 0 ≤ <1, no alcanza el valor de 1 y además el valor máximo es desconocido (excepto en las tablas cuadradas, ver Fórmula 59), esta característica hace que la interpretación del valor resulte más difusa. Los otros dos estadísticos toman valores de 0 ≤ 1 (Tabla 96). En los tres casos cuanto más cerca está el valor de cero, menor es la fuerza de la asociación y según se aproxima al valor máximo mayor es la fuerza de la asociación. Pero utilizando una metáfora, los valores pocas veces van a ser "*blanco*", ni "*negro*", sino que se van a mover en una escala de "*grises*" que es la que hay que interpretar. Los estadísticos *V* (para *k* = 2) y λ pueden llegar a valer más de la unidad si $\theta_e^2 > n$.

Tabla 96 Escala de cc, V y λ.

cc	0 0,1 0,2 0,3 0,4 0,5 0,6 0,7 0,8 ¿?
V	0 0,1 0,2 0,3 0,4 0,5 0,6 0,7 0,8 0,9 1
λ	0 0,1 0,2 0,3 0,4 0,5 0,6 0,7 0,8 0,9 1

En una escala que tiene los límites en el 0 y en el 1, la mitad es 0,5, que puede servir de orientación, pero sin decir mucho sobre su interpretación. La escala del coeficiente de contingencia, excepto cuando la tabla es cuadrada, no permite determinar el punto medio al ser desconocido el punto máximo.

Las características de las variables que se relacionan también influyen en la interpretación del estadístico. Si su relación es funcional (está sujeta a alguna fórmula o ecuación conocida) o si no se corresponde con una función o fórmula conocida. En el primer caso la asociación debe ser muy alta o próxima a 1, porque pequeñas variaciones sería indicativo de que algo ha ocurrido. En el segundo caso, si dos variables no tienen relación aparente o no tienen porqué estar relacionadas, una asociación, aunque esta pueda parecer baja, es indicativo de que hay asociación cuando no la debiera haber.

Un ejemplo del primer caso puede ser la *velocidad*, el *espacio* y el *tiempo*. Su relación es funcional, por lo tanto cualquier relación entre esas variables debe ser alta y una pequeña variación sería indicativo de que algo ha pasado. En el segundo caso, la asociación entre dos características o atributos humanos que no tengan relación funcional conocida como puede ser la *categoría profesional* y el *sexo,* entendiendo que las demás características no influyen, un valor de asociación que puede parecer bajo, puede ser indicativo de que "algo hay" cuando "no debería haber nada".

La interpretación de los valores de los estadísticos y más cuando se trata de comportamientos humanos es compleja, por eso es más cómodo la comparación que la interpretación. Si el análisis que se hace es comparado con un estudio anterior, es más cómodo y sencillo y no está sujeto a interpretación decir si el resultado es más bajo, igual o más alto que el resultado anterior. Para poder comparar se debe utilizar el mismo estadístico.

En este caso que los tres estadísticos toman el valor de 0,11 se puede decir que aunque hay asociación, esta es débil. V de Cramer y fí siempre tendrán el mismo valor cuando el menor del número de filas o columnas sea 2.

El estadístico ς, que es asimétrico, es una herramienta que puede orientar para determinar que variable puede ser considerada como la independiente y cual la dependiente. En el caso de ς_{yx} = 0,00, la reducción proporcional de error al predecir la variable *confianza en el Parlamento del país* a partir de la variable sexo es de 0,0% y la reducción proporcional de error al predecir la variable *sexo* a partir de la variable *confianza en el Parlamento del país* es de 0,06%. No se puede decir que las variables puedan servir para la reducción proporcional del error. El estadístico ς no es específicamente de asociación, cada estadístico busca la relación de una forma determinada.

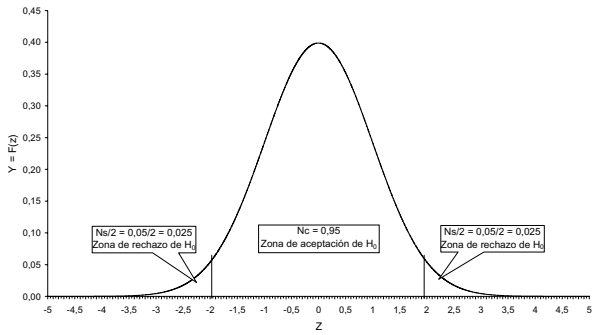
Lambda toma valores en el rango de 0 a 1. Un valor de 0 significa que la variable considerada como independiente no ayuda en la predicción de la variable dependiente. Un valor de 1 significa que la variable considerada como independiente predice perfectamente las categorías de la variable dependiente. Esta perfección sólo se consigue cuando cada una de las columnas tiene por lo menos una celda con valor no-cero.

Un lambda de cero no implica que no haya asociación estadística. Como con todas las medidas de asociación, lambda se construye para medir el grado de asociación en una forma muy específica. Concretamente, lambda refleja la reducción del error cuando los valores de una variable se usan para predecir los valores de la otra. Si no existe esta particular asociación entre las variables, lambda es cero, pero otros índices pueden encontrar asociación de una forma diferente, aunque lambda sea cero.

La interpretación de los residuos tipificados está sujeta a contraste de hipótesis y se hace aplicando el criterio de *Z* por seguir la distribución de la variable *Z*, tipificada o estandarizada. El estadístico θ_e^2, es un estadístico holístico, global, para toda la tabla, dice si hay asociación o no, pero en toda la tabla. Con los residuos tipificados se determina en que celda o celdas se produce o a qué celdas es debida esa asociación. Los residuos se calculan por diferencia entre frecuencias observadas y las esperadas. Los resultados posibles son: cero

si son iguales, negativo si la *fo* es menor que la *fe* y positivo si la *fo* es mayor que la *fe*. Determinar si las diferencias son pequeñas o grandes o en términos probabilísticos, si la diferencia es lo suficientemente grande como para asumir que las diferencias son significativas, se ve por medio de la estandarización de los residuos y aplicando un contraste de hipótesis. El proceso del contraste de hipótesis de los residuos tipificados se muestra en la Tabla 97. La estandarización de los residuos es según la Fórmula 64.*

Tabla 97 Contraste de hipótesis de los residuos tipificados.

1º H_1: *fo* ∏ *fe*.	
2º H_0: *fo* = *fe*.	
3º Estadístico: *Z*.	
4º Criterio *Ns* = 0,05. Z_c = 1,96; *Nc* = 0,95	
Se acepta H_o si: $\lvert zres_e \rvert \{ \lvert zres_c \rvert \sum Nc_e \{ Nc_c \sum Ns_e \} Ns_c$	
Se rechaza H_0 si: $\lvert zres_e \rvert \varnothing \lvert zres_c \rvert \sum Nc_e \varnothing Nc_c \sum Ns_e \Omega Ns_c$	

Cálculo de los residuos tipificados de la Tabla 93

Varones	Mujeres
$zres_{11} \mid \frac{fo_{11} 4 fe_{11}}{\sqrt{fe_{11}}} \mid \frac{569 4 574,5}{\sqrt{574,5}} \mid 40,23$	$zres_{12} \mid \frac{fo_{12} 4 fe_{12}}{\sqrt{fe_{12}}} \mid \frac{630 4 624,5}{\sqrt{624,5}} \mid 0,22$
$zres_{21} \mid \frac{fo_{21} 4 fe_{21}}{\sqrt{fe_{21}}} \mid \frac{313 4 351,7}{\sqrt{351,7}} \mid 42,06$	$zres_{22} \mid \frac{fo_{22} 4 fe_{22}}{\sqrt{fe_{22}}} \mid \frac{421 4 382,3}{\sqrt{382,3}} \mid 1,98$
$zres_{31} \mid \frac{fo_{31} 4 fe_{31}}{\sqrt{fe_{31}}} \mid \frac{213 4 169,1}{\sqrt{169,1}} \mid 3,37$	$zres_{32} \mid \frac{fo_{32} 4 fe_{32}}{\sqrt{fe_{32}}} \mid \frac{140 4 183,9}{\sqrt{183,9}} \mid 43,24$
$zres_{41} \mid \frac{fo_{41} 4 fe_{41}}{\sqrt{fe_{41}}} \mid \frac{29 4 28,7}{\sqrt{28,7}} \mid 0,05$	$zres_{42} \mid \frac{fo_{42} 4 fe_{42}}{\sqrt{fe_{42}}} \mid \frac{31 4 31,3}{\sqrt{31,3}} \mid 40,05$

Para todos los residuos tipificados que su valor está comprendido en el intervalo –1,96 ϶ 1,96, se acepta la H_0, y los *zres* ∅1,96 o *zres* Ω-1,96 o |*zres*/ ∅|1,96| produce el rechazo de la H_0 y consecuentemente la aceptación de la H_1. Entonces las diferencias significativas entre las *fo* y las *fe* se dan en las celdas de *aprobado* y *notable* con *varón* y *mujer*. No hay asociación en las celdas *suspenso* y *sobresaliente* con *varón* y *mujer*.

En las celdas que la diferencia entre las *fo* y las *fe* no es significativa, quiere decir que lo observado es lo esperado y por lo tanto son sucesos independientes mutuamente no excluyentes. En las celdas en las que la diferencia entre las *fo* y las *fe* es significativa, quiere decir que lo observado es significativamente distinto a lo esperado, en este caso al nivel de significación de 0,05, y por lo tanto lo observado no es lo esperado. Cuando la diferencia es negativa, lo observado es significativamente menor que lo esperado y cuando es positiva, lo observado es significativamente mayor que lo esperado.

La interpretación es: en la valoración del Parlamento del Reino Unido por los ciudadanos, no hay diferencias en la valoración de *suspenso* y *sobresaliente* entre las *fo* y las *fe* ni en los varones ni en las mujeres. Pero en la valoración de *aprobado* hay menos varones de los que debería haber y más mujeres de las que debería haber si los sucesos fuesen independientes, entonces se puede decir que hay significativamente menos varones de los que debería haber que aprueban al Parlamento y más mujeres de las que debería haber que aprueban al Parlamento.

En la valoración de *notable* son más los varones que hay y menos las mujeres que hay de los que debería haber si los sucesos fuesen independientes, mutuamente no excluyentes, entonces se puede decir que hay significativamente más varones de los que debería haber que dan notable al Parlamento mientras que hay menos mujeres de las que debería haber que dan notable al Parlamento.

Es un análisis frecuencista, se ve si hay más o menos casos. La lectura que no se puede hacer es que las mujeres "aprueban" más la actuación del Parlamento que los varones, pero lo valoran peor en "notable". Esta lectura es falsa, el hecho de que sean más o menos individuos, no varía la calificación, sino el número de personas que dan esa calificación, por lo tanto la calificación es la misma. La diferencia es el número de casos.

12.5 Contraste de hipótesis de una tabla de contingencia con variables ordinales

El siguiente ejemplo de análisis se realiza con dos variables categóricas de nivel de medida ordinal para completarlo con los estadísticos que miden este tipo de relación. Se van a utilizar preguntas de un estudio de CIRES[63] (Centro de Investigaciones de la Realidad Social) (Tabla 98). Se ha ponderado la muestra según el criterio de CIRES.

Tabla 98 Contraste de hipótesis con variables ordinales.

Pregunta P.15: En general, ¿tiene la sensación de que le falta tiempo o de que tiene tiempo de sobra?		Pregunta P.16: En general, ¿hace las cosas con prisa o tranquilamente?
Categorías originales 1. Falta. 2. Sobra. 3. No falta ni sobra. 9. Ns/Nc	Recodificación para mantener la ordinalidad. 1. Falta. 2. No falta ni sobra. 3. Sobra. 9. Ns/Nc	1. Con prisa. 2. Con tranquilidad. 9. Ns/Nc
Esta pregunta genera la variable B1.		Esta pregunta genera la variable B2.

Forma de hacer las cosas y sensación o sentimiento de falta o sobra del tiempo.

			Sensación de falta o sobra tiempo			
			Falta	*No falta ni sobra*	*Sobra*	*Total fila*
Forma de hacer las cosas.	*Con prisa*	*fo*	336	104	65	505
		fe	219,0	128,4	157,6	505,0
		%TC	65,9	34,8	17,7	42,9
		%TT	28,6	8,8	5,5	42,9
		Residuo	117,0	-24,4	-92,6	
		ZResiduo	7,91	-2,15	-7,38	
	Con tranquilidad	*fo*	174	195	302	671
		fe	291,0	170,6	209,4	671,0
		%TC	34,1	65,2	82,3	57,1
		%TT	14,8	16,6	25,7	57,1
		Residuo	-117,0	24,4	92,6	
		Zresiduo	-6,86	1,87	6,40	
	Total columna	*fo*	510	299	367	1.176
		fe	510,0	299,0	367,0	1.176,0
		%TF	100,0	100,0	100,0	100,0
		%TC	43,4	25,4	31,2	100,0

Proceso de contraste de hipótesis:

1. H_1: "Existe asociación o dependencia entre la sensación o sentimiento de falta o sobra

[63] Realizado en enero de 1996 sobre "Usos del tiempo". Este Centro, dirigido por Juan Díez Nicolas, estaba patrocinado por la Fundación BBVA, Caja Madrid y BILBAO-BIZKAIA-KUTXA.

tiempo y la forma de hacer las cosas" o de forma abreviada "El sentimiento o sensación de falta o sobra tiempo influye en la forma de hacer las cosas". Se plantea que los sucesos son dependientes.

2. *Sensación de falta o sobra tiempo*: variable categórica ordinal. *Forma de hacer las cosas*: variable categórica ordinal.
3. Variable considerada como independiente: *sensación de falta o sobra tiempo*. Variable considerada como dependiente: *forma de hacer las cosas*.
4. H_0: "Existe asociación o dependencia entre la sensación o sentimiento de falta o sobra tiempo y la forma de hacer las cosas" o de forma abreviada "El sentimiento o sensación de falta o sobra tiempo no influye en la forma de hacer las cosas". Se plantea que los sucesos son independientes.
5. Estadístico: θ^2, por ser las dos variables categóricas:
6. Criterio de aceptación/rechazo de H_0, $Ns = 0{,}05$.

Tabla 99 Cálculo de las frecuencias esperadas y los residuos.

Intersección	Probabilidad de la intersección de los sucesos elementales	Frecuencias esperadas	Residuos
Falta – con prisa	$P_{(S_f - S_p)} = P_{(S_f)} \Delta P_{(S_p)} = 0{,}434 \Delta 0{,}429 = 0{,}1862$	$fe = 0{,}1862 \Delta 1.176 = 219{,}0$	$\mathrm{Re}s_{11} = fo_{11} - fe_{11} = 336 - 219{,}0 = 117{,}0$
No f/s– con prisa	$P_{(S_n - S_p)} = P_{(S_n)} \Delta P_{(S_p)} = 0{,}254 \Delta 0{,}429 = 0{,}1092$	$fe = 0{,}1092 \Delta 1.176 = 128{,}4$	$\mathrm{Re}s_{12} = fo_{12} - fe_{12} = 104 - 128{,}4 = 424{,}4$
Sobra – con prisa	$P_{(S_s - S_p)} = P_{(S_s)} \Delta P_{(S_p)} = 0{,}312 \Delta 0{,}429 = 0{,}1340$	$fe = 0{,}1340 \Delta 1.176 = 157{,}6$	$\mathrm{Re}s_{13} = fo_{13} - fe_{13} = 65 - 157{,}6 = 492{,}6$
Falta – con tranquilidad	$P_{(S_f - S_t)} = P_{(S_f)} \Delta P_{(S_t)} = 0{,}434 \Delta 0{,}571 = 0{,}2474$	$fe = 0{,}2474 \Delta 1.176 = 291{,}0$	$\mathrm{Re}s_{21} = fo_{21} - fe_{21} = 174 - 291{,}0 = 4117{,}0$
No f/s– con tranquilidad	$P_{(S_n - S_t)} = P_{(S_n)} \Delta P_{(S_t)} = 0{,}254 \Delta 0{,}571 = 0{,}1451$	$fe = 0{,}1451 \Delta 1.176 = 170{,}6$	$\mathrm{Re}s_{22} = fo_{22} - fe_{22} = 195 - 170{,}6 = 24{,}4$
Sobra – con tranquilidad	$P_{(S_s - S_t)} = P_{(S_s)} \Delta P_{(S_t)} = 0{,}312 \Delta 0{,}571 = 0{,}1781$	$fe = 0{,}1781 \Delta 1.176 = 209{,}4$	$\mathrm{Re}s_{23} = fo_{23} - fe_{23} = 302 - 209{,}4 = 92{,}6$

Para contrastar la hipótesis nula (H_0) hay que calcular el chi-cuadrado estimado, y en el caso de la Tabla 98 es:

$$\theta_c^2 = \sum_{i=1}^{F} \sum_{j=1}^{C} \frac{(fo_{ij} - fe_{ij})^2}{fe_{ij}} = \sum_{i=1}^{F=2} \sum_{j=1}^{C=3} \frac{(fo_{ij} - fe_{ij})^2}{fe_{ij}}$$

$$\frac{(fo_{11} - fe_{11})^2}{fe_{11}} + \frac{(fo_{12} - fe_{12})^2}{fe_{12}} + \frac{(fo_{13} - fe_{13})^2}{fe_{13}} + \frac{(fo_{21} - fe_{21})^2}{fe_{21}} + \frac{(fo_{22} - fe_{22})^2}{fe_{22}} +$$

$$+ \frac{(fo_{23} - fe_{23})^2}{fe_{23}} = \frac{(316 - 219{,}0)^2}{219{,}0} + \frac{(104 - 128{,}4)^2}{128{,}4} + \frac{(65 - 157{,}6)^2}{157{,}6} + \frac{(174 - 291{,}0)^2}{291{,}0} +$$

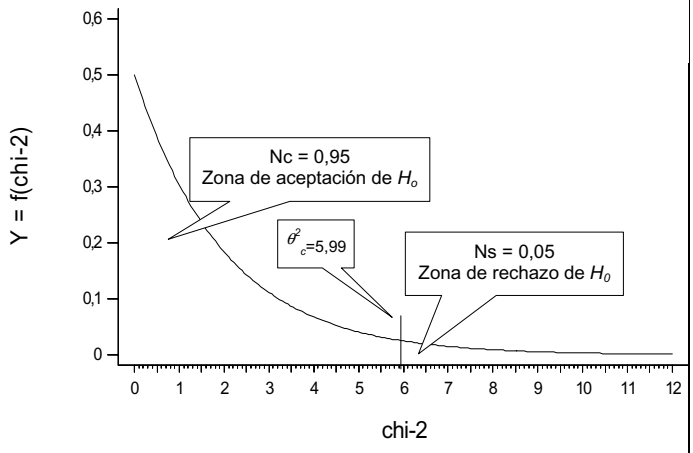
$$+ \frac{(195 - 170{,}6)^2}{170{,}6} + \frac{(302 - 209{,}4)^2}{209{,}4} = 213{,}02$$

Tabla 100 Cálculo de chi-cuadrado crítico.

Cálculo de θ_c^2:

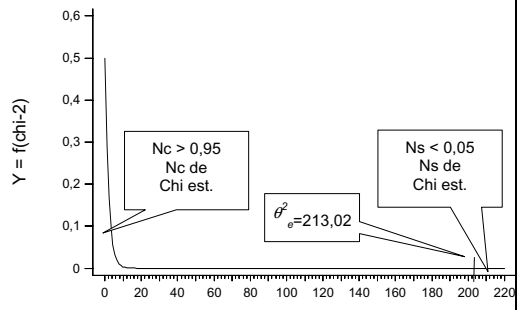
θ_c^2	*gl* = 2 *Ns* = 0,05	= 5,9915

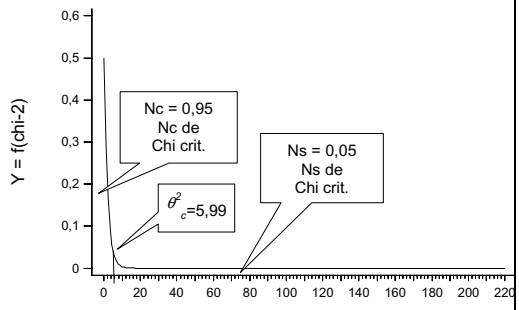
El valor de $\theta_c^2 \mid 5{,}9915$ es el que deja una superficie a su derecha, por debajo de la curva de θ^2 y por encima del eje de abscisas, igual a 0,05, que se llama *Ns* y para *gl* = 2.
Este valor se obtiene en la Tabla de θ^2 del Anexo 2 (pág. 333). Los *gl* = 2 son la entrada de filas y el *Ns* = 0,05 es la entrada de columnas. La celda en la que se cruzan es el valor de θ_c^2 que cumple estas condiciones.



El θ_e^2 de la Tabla 98 es 213,02, para saber si se acepta o rechaza la H_o, tenemos que comparar el θ_e^2 con el θ_c^2. Mediante la representación gráfica se ilustra el contraste (Gráfico 17).

Gráfico 18 Comparación de θ_e^2 con θ_c^2.





Ahora se puede establecer el siguiente esquema para determinar la aceptación o rechazo de H_0.

Se acepta H_o si: $\theta_e^2 \{ \theta_c^2 \sum Nc_e \{ Nc_c \sum Ns_e \} Ns_c$

Se rechaza H_o si: $\theta_e^2 \varnothing \theta_c^2 \sum Nc_e \varnothing Nc_c \sum Ns_e \Omega Ns_c$

En este caso, se rechaza H_o porque el $\theta_e^2 \varnothing \theta_c^2$, que equivale a decir que $Nc_e \varnothing Nc_c$ y equivalente también a que $Ns_e \Omega Ns_c$. Por lo tanto se puede concluir que al *Ns* de 0,05 (incluso al *Ns* = 0,01 ya que con *gl* = 2, $\theta_c^2 \mid 9{,}2104$) existe asociación entre las variables *sensación de falta o sobra el tiempo y forma de hacer las cosas*.

Al existir asociación entre las variables, hay que ver: fuerza de la asociación, dirección de la asociación (en esta ocasión sí, porque las dos variables se consideran de nivel de medida ordinal) y a qué celdas es debida la asociación. La lectura de porcentajes ya se ha trabajado y salvo excepción, en esta tabla no se tratará.

Los estadísticos que indican la fuerza de la asociación son: Coeficiente de contingencia (*cc*), *V* de Cramer, λ (fi ó phi) y Lambda (ς).

La asociación de cada una de las celdas se analiza con los residuos tipificados, que tienen la distribución de una variable $N_{(0,1)}$, la puntuación tipificada ó *z*, que permitirá saber a través de los residuos que frecuencias observadas y esperadas son significativamente distintas. Se puede repasar el Epígrafe 11.

Los valores de los estadísticos *cc, V,* λ y ς, de la Tabla 98 son:

$$cc \mid \sqrt{\frac{\theta_e^{\,2}}{\theta_e^{\,2} 2\, n}} \mid \sqrt{\frac{213{,}02}{213{,}02\, 2\, 1.176}} \mid 0{,}39$$

$$V \mid \sqrt{\frac{\theta_e^{\,2}}{n\Delta(k\, 4\, 1)}} \mid \sqrt{\frac{213{,}02}{1.176\Delta(2\, 4\, 1)}} \mid 0{,}43$$

$$\lambda \mid \sqrt{\frac{\theta_e^{\,2}}{n}} \mid \sqrt{\frac{213{,}02}{1.176}} \mid 0{,}43$$

$$\varsigma_{yx} \mid \frac{P_{1y}\, 4\, P_{2y}}{P_{1y}} \mid \frac{/1\, 4\, 0{,}5710 4\, /0{,}148\, 2\, 0{,}088\, 2\, 0{,}0550}{/1\, 4\, 0{,}5710} \mid \frac{0{,}429\, 4\, 0{,}291}{0{,}429} \mid 0{,}32$$

$$\varsigma_{xy} \mid \frac{P_{1x}\, 4\, P_{2x}}{P_{1x}} \mid \frac{/1\, 4\, 0{,}4340 4\, /0{,}088\, 2\, 0{,}055\, 2\, 0{,}148\, 2\, 0{,}1660}{/1\, 4\, 0{,}4340} \mid \frac{0{,}566\, 4\, 0{,}457}{0{,}566} \mid \frac{0{,}109}{0{,}566} \mid 0{,}19$$

Los valores de *cc*, *V* y λ, (0,39 y 0,43 respectivamente) indican cierta asociación, considerando que las variables son sentimientos o sensaciones. En el caso de ς, la variable que mejor ayuda a predecir por tener una reducción proporcional de error mayor es cuando se considera *forma de hacer las cosas* como variable dependiente y *sensación de falta o sobra tiempo* como variable independiente (ς = 0,32).

Entonces las categorías de la variable *sensación de falta o sobra tiempo* ayuda a predecir mejor las categorías de la variable *forma de hacer las cosas*. En ocasiones, como es el caso de estas dos variables, puede haber dudas para decidir qué variable considerar la dependiente y cual la independiente y ς puede ser una ayuda.

12.5.1 Estadísticos de dirección de la asociación con variables ordinales

Los estadísticos que miden la dirección de la asociación, están basados en la

comparación de los valores de las dos variables para todos los casos posibles de la tabla tomados por pares. Se consideran pares concordantes, discordantes o empatados. Un par de casos se consideran *concordantes* cuando uno de ellos tiene los valores, en ambas variables de la tabla, altos (o bajos) coincidiendo con los correspondientes valores del otro caso en ambas variables.

El par es *discordante* si el valor de un caso en una de las variables es mayor que el correspondiente valor del otro caso y la dirección en la segunda variable es inversa. Se dice pares *empatados* cuando los individuos tienen idénticos valores en una o las dos variables.

Así, para cualquier par de casos con valores en las variables X e Y pueden ser concordantes, discordantes o estar empatados en X pero no en Y, empatados en Y pero no en X o estar empatados en las dos.

Si hay mayor número de pares concordantes, la asociación es positiva: cuando los casos se incrementan (o decrementan) con los valores de la variable X, lo mismo ocurre en la variable Y. Si la mayoría de los pares son discordantes, la asociación es negativa: cuando los casos se incrementan con los valores de una variable, tienden a decrementar en la otra. Si los pares concordantes y discordantes son igualmente probables, entonces no se puede decir que haya asociación.

Se debe tener precaución al considerar nivel superior e inferior en una variable categórica ordinal porque la codificación, como ya se vio, es arbitraria y aleatoria, puede tener codificación inversa al orden de las categorías de la variable. Un ejemplo simple de una variable ordinal con los sucesos elementales: *bajo, medio* y alto, puede estar codificada como: *1, 2* y *3* o como *3, 2* y 1. En el primer caso la codificación se considera directa (*bajo* (1); *medio* (2), y *alto* (3)) y en el segundo caso se considera inversa (*bajo* (3); *medio* (2), y *alto* (1)). Cuando se habla de concordancia se considera la codificación directa. La expresión simbólica del cálculo de la concordancia es,

$P_c \mid \sum_{i\mid 1}^{f-1} \sum_{j\mid 1}^{c-1} n_{ij} \Delta \sum_{k\mid i21}^{f} \sum_{l\mid j21}^{c} n_{kl}$	P_c: Pares concordantes *f:* Filas *c:* Columnas	Fórmula 65
$P_d \mid \sum_{i\mid 1}^{f-1} \sum_{j\mid c}^{1} n_{ij} \Delta \sum_{k\mid i21}^{f} \sum_{l\mid c41}^{1} n_{kl}$	P_d: Pares discordantes *f:* Filas *c:* Columnas	Fórmula 66
$T_X \mid \sum_{i\mid 1}^{f} n_{i.} \Delta \sum_{k\mid i21}^{f} n_{k.}$	T_X: Empates en la variable independiente (por filas). *f:* Filas. *c:* Columnas.	Fórmula 67
$T_Y \mid \sum_{j\mid 1}^{c} n_{.j} \Delta \sum_{l\mid j21}^{c} n_{.l}$	T_X: Empates en la variable dependiente (por columnas). *f:* Filas. *c:* Columnas.	Fórmula 68
$T_{XY} \mid \sum_{i\mid 1}^{f} \sum_{j\mid 1}^{c} \frac{n_{ij} \Delta n_{ij} 41}{2}$	T_{XY}: Empates simultáneos en filas y columnas.	Fórmula 69

$T = \frac{N(N-1)}{2}$	*T*: Pares totales de la tabla. *N*: Número de casos.	Fórmula 70

Es recomendable ver el desarrollo gráfico del cálculo de la concordancia en M. García Ferrando (1982: 239-243.).

Los estadísticos ordinales son: *Tau-a*, *Tau-b*, *Tau-c*, *Gamma* (ν) y *d* de Somers (asimétrico). Al indicar dirección estos estadísticos, el rango está comprendido entre -1 ≤ +1, aunque en los límites hay algunas excepciones. Simbólicamente,

$t_a = \frac{P_c - P_d}{T}$	t_a: Tau-a. P_c: Pares concordantes. P_d: Pares discordantes. *T*: Total de pares.	Fórmula 71

El coeficiente Tau-a resulta de dividir los pares concordantes (P_c) menos los pares discordantes (P_d) entre el número total de pares, Si no hay pares coincidentes, este coeficiente está comprendido en el rango -1 ≤ +1. Si hubiese pares coincidentes, el rango de valores posibles es más limitado; dependiendo del número de pares coincidentes.

Un coeficiente que intenta normalizar (P_c - P_d) considerando los pares coincidentes para cada una de las variables en el par de casos separadamente, pero no los coincidentes en ambas variables para el par de casos, es Tau-b.

$t_b = \frac{P_c - P_d}{\sqrt{(P_c + P_d + T_X)(P_c + P_d + T_Y)}}$	t_b: Tau-b. P_c: Pares concordantes. P_d: Pares discordantes. T_X: Pares coincidentes en *X* pero no en *Y*. T_Y: Pares coincidentes en *Y* pero no en *X*.	Fórmula 72

Si no hay frecuencias marginales de cero, los valores de Tau-b están en el rango -1 ≤ +1, sólo para tablas cuadradas (igual número de filas que de columnas).

$t_c = \frac{2k(P_c - P_d)}{N^2(k-1)}$	t_c: Tau-c. P_c: Pares concordantes. P_d: Pares discordantes. *N*: Número de casos. *k*. El número menor de filas o columnas.	Fórmula 73

Un coeficiente que puede obtener valores en el rango -1 ≤ +1 para cualquier tabla, es Tau-c. Los coeficientes de Tau-b y de Tau-c, no difieren mucho en sus resultados si cada marginal contiene, aproximadamente, las mismas frecuencias.

$\nu = \frac{P_c - P_d}{P_c + P_d}$	ν: Gamma de Goodman y kruskal. P_c: Pares concordantes. P_d: Pares discordantes.	Fórmula 74

El valor de *gamma* está en el rango de -1 a +1, *gamma* es 1 si todas las observaciones están concentradas en una diagonal; el signo será positivo si las observaciones se concentran en la diagonal positiva o concordante (desde el extremo *XY* altos al extremo *XY* bajos), y el signo será negativo si las observaciones se concentran en la diagonal negativa o discordante (desde el extremo *X* alto, *Y* bajo al extremo *X* bajo, *Y* alto). En el caso de independencia entre las variables, gamma será igual a 0. El resumen es que *gamma* indica en qué diagonal tienden a concentrarse los casos. Se puede considerar también en sentido predictivo, cuanto más se acerque a la unidad se puede considerar con mayor fuerza predictiva.[64] El estadístico *gamma* de Goodman y Kruskal es similar a los *Tau*.

$d_X \mid \frac{P_c 4 P_d}{P_c 2 P_d 2 T_X}$	d_X: *d* de Somers. Variable de filas como dependiente. P_c: Pares concordantes. P_d: Pares discordantes. T_X: Pares coincidentes en *X* pero no en *Y*.	Fórmula 75

$d_Y \mid \frac{P_c 4 P_d}{P_c 2 P_d 2 T_Y}$	*dy*: *d de Somers.* Variable de columnas como dependiente. P_c: Pares concordantes. P_d: Pares discordantes. T_Y: Pares coincidentes en *Y* pero no en *X*.	Fórmula 76

$d \mid \frac{P_c 4 P_d}{\frac{/P_c 2 P_d 2 T_X 0 2 /P_c 2 P_d 2 T_Y 0}{2}}$	d: *d* de Somers. Simétrico. P_c: Pares concordantes. P_d: Pares discordantes. T_X: Pares coincidentes en *X* pero no en *Y*. T_Y: Pares coincidentes en *Y* pero no en *X*.	Fórmula 77

Somers (1962) propone una extensión asimétrica de *gamma* que difiere sólo en la inclusión en el denominador del número de pares no coincidentes sobre la variable *X* pero que coinciden en la variable *Y* (T_Y) o incluyendo el número de pares no coincidentes sobre la variable *Y* pero que coinciden en la variable *X* (T_X). El coeficiente d_Y indica la proporción en exceso de pares concordantes sobre los pares discordantes a través de los pares no coincidentes en la variable independiente. El coeficiente d_X indica la proporción en exceso de pares concordantes sobre los pares discordantes a través de los pares no coincidentes en la variable dependiente. En la variante simétrica de la *d* de Somers, en el denominador se usa el valor medio de los denominadores de los dos coeficientes asimétricos. Ver Manuel García Ferrando (1982: 246). El cálculo de los estadísticos de la Tabla 98 se muestra a continuación en la Tabla 101

[64] Los criterios predictivos, en el sentido causa-efecto, se recomienda que cada lector siga la tradicción del marco teórico en el que esté inmerso. En nuestro caso consideramos que la predicción sería *a posteriori,* esto es, "decir lo que ha pasado cuando ya ha pasado" no consideramos la predicción *a priori*, "predecir lo que va a pasar". Este último caso se aceptaría cuando incluyese un cierto nivel de probabilidad y considerando que podría "no pasar".

Tabla 101 Estadísticos de dirección.

$$P_c = \sum_{i=1}^{f}\sum_{j=1}^{c} n_{ij} \cdot \sum_{k=i+1}^{f}\sum_{l=j+1}^{c} n_{kl} = n_{11}\cdot(n_{22}+n_{23})+n_{12}\cdot(n_{23})+n_{13}\cdot(n_{21}+n_{22}) = 336\cdot(195+302)+104\cdot(302)+65\cdot(104+174) = 198.400$$

$$P_d = \sum_{i=1}^{f}\sum_{j=c}^{1} n_{ij} \cdot \sum_{k=i+1}^{f}\sum_{l=c-1}^{1} n_{kl} = n_{13}\cdot(n_{22}+n_{21})+n_{12}\cdot(n_{21})+n_{11}\cdot(n_{22}+n_{23}) = 65\cdot(195+174)+104\cdot(174)+336\cdot(195+302) = 42.081$$

$$T_X = \sum_{i=1}^{f} n_{i.} \cdot \sum_{k=i+1}^{f} n_{k.} = n_{11}\cdot n_{21}+n_{12}\cdot n_{22}+n_{13}\cdot n_{23} = 336\cdot 174+104\cdot 195+65\cdot 302 = 98.374$$

$$T_Y = \sum_{j=1}^{c} n_{.j} \cdot \sum_{l=j+1}^{c} n_{.l} = n_{11}\cdot(n_{12}+n_{13})+n_{12}\cdot n_{13}+n_{21}\cdot(n_{22}+n_{23})+n_{22}\cdot n_{23} =$$
$$336\cdot(104+65)+104\cdot 65+174\cdot(195+302)+195\cdot 302 = 208.912$$

$$T_{XY} = \sum_{i=1}^{f}\sum_{j=1}^{c} \frac{n_{ij}\cdot(n_{ij}-1)}{2} = \sum_{i=1}^{2}\sum_{j=1}^{3} \frac{n_{ij}\cdot(n_{ij}-1)}{2} = \frac{n_{11}\cdot(n_{11}-1)}{2}+\frac{n_{12}\cdot(n_{12}-1)}{2}+\frac{n_{13}\cdot(n_{13}-1)}{2}+$$
$$+\frac{n_{21}\cdot(n_{21}-1)}{2}+\frac{n_{22}\cdot(n_{22}-1)}{2}+\frac{n_{23}\cdot(n_{23}-1)}{2} = \frac{336\cdot(336-1)}{2}+\frac{104\cdot(104-1)}{2}+\frac{65\cdot(65-1)}{2}+$$
$$+\frac{174\cdot(174-1)}{2}+\frac{195\cdot(195-1)}{2}+\frac{302\cdot(302-1)}{2} =$$
$$56.280+5.356+2.080+15.051+18.915+45.451 = 143.133$$

$$T = \frac{N\cdot(N-1)}{2} = \frac{1.176\cdot(1.176-1)}{2} = 690.900$$
$$T = P_c+P_d+T_X+T_Y+T_{XY} = 198.400+42.081+98.374+208.912+143.133 = 690.900$$

$$t_a = \frac{P_c-P_d}{T} = \frac{198.400-42.081}{690.900} = \frac{156.319}{690.900} = 0{,}23$$

$$t_b = \frac{P_c-P_d}{\sqrt{(P_c+P_d+T_X)\cdot(P_c+P_d+T_Y)}} = \frac{198.400-42.081}{\sqrt{(198.400+42.081+98.374)\cdot(198.400+42.081+208.912)}} =$$
$$\frac{156.319}{\sqrt{(338.855)\cdot(449.393)}} = \frac{156.319}{390.230} = 0{,}40$$

$$t_c = \frac{2\cdot k\cdot(P_c-P_d)}{N^2\cdot(k-1)} = \frac{2\cdot 2\cdot(198.400-42.081)}{1.176^2\cdot(2-1)} = \frac{625.276}{1.382.976} = 0{,}45$$

$$\nu = \frac{P_c-P_d}{P_c+P_d} = \frac{198.400-42.081}{198.400+42.081} = \frac{156.319}{240.481} = 0{,}65$$

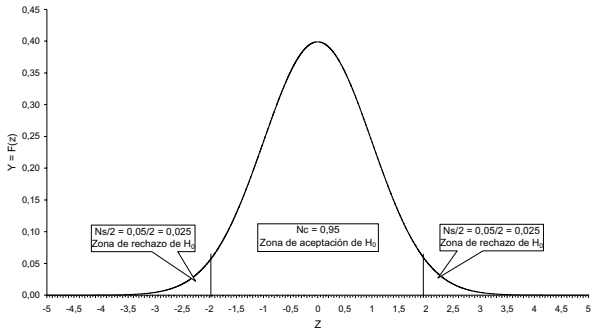
$$d_X = \frac{P_c-P_d}{P_c+P_d+T_X} = \frac{198.400-42.081}{198.400+42.081+98.374} = \frac{156.319}{338.855} = 0{,}46$$

$$d_Y = \frac{P_c-P_d}{P_c+P_d+T_Y} = \frac{198.400-42.081}{198.400+42.081+208.912} = \frac{156.319}{449.393} = 0{,}35$$

$$d = \frac{P_c-P_d}{\frac{(P_c+P_d+T_X)+(P_c+P_d+T_Y)}{2}} = \frac{198.400-42.081}{\frac{(198.400+42.081+98.374)+(198.400+42.081+208.912)}{2}} = \frac{156.319}{\frac{788.248}{2}} = \frac{156.319}{394.124} = 0{,}40$$

El proceso del contraste de hipótesis de los residuos tipificados se muestra en la Tabla 102.

Tabla 102 Contraste de hipótesis de los residuos tipificados.

1º H_1: *fo* ∏ *fe*.	
2º H_0: *fo* = *fe*.	
3º Estadístico o distribución de tipo: *Z*	
4º Criterio *Ns* = 0,05. Z_c = 1,96; *Nc* = 0,95	
Se acepta H_o si: $\lvert zres_e \rvert \{ \lvert zres_c \rvert \sum Nc_e \{ Nc_c \sum Ns_e \} Ns_c$ Se rechaza H_0 si: $\lvert zres_e \rvert \varnothing \lvert zres_c \rvert \sum Nc_e \varnothing Nc_c \sum Ns_e \Omega Ns_c$	

Cálculo de los residuos tipificados de la Tabla 98

Hace las cosas con prisa:	Hace las cosas tranquilamente:
$zres_{11} \lvert \frac{fo_{11} 4 fe_{11}}{\sqrt{fe_{11}}} \rvert \frac{336 4 219,0}{\sqrt{219,0}} \rvert$ 7,91	$zres_{21} \lvert \frac{fo_{21} 4 fe_{21}}{\sqrt{fe_{21}}} \rvert \frac{174 4 291,0}{\sqrt{291,0}} \rvert$ 46,86
$zres_{12} \lvert \frac{fo_{12} 4 fe_{12}}{\sqrt{fe_{12}}} \rvert \frac{104 4 128,4}{\sqrt{128,4}} \rvert$ 42,15	$zres_{22} \lvert \frac{fo_{22} 4 fe_{22}}{\sqrt{fe_{22}}} \rvert \frac{195 4 170,6}{\sqrt{170,6}} \rvert$ 1,87
$zres_{13} \lvert \frac{fo_{13} 4 fe_{13}}{\sqrt{fe_{13}}} \rvert \frac{65 4 157,6}{\sqrt{157,6}} \rvert$ 47,38	$zres_{23} \lvert \frac{fo_{23} 4 fe_{23}}{\sqrt{fe_{23}}} \rvert \frac{302 4 209,4}{\sqrt{209,4}} \rvert$ 6,40

Para todos los residuos tipificados que su valor está comprendido en el intervalo de –1,96 э 1,96, se acepta la H_0, y los *zres* ∅ 1,96 o *zres* Ω -1,96 o |*zres*/ ∅ |1,96| produce el rechazo de la H_0 y consecuentemente la aceptación de la H_1. Entonces las diferencias significativas entre las *fo* y las *fe* se da en todas las celdas, excepto en el cruce "hace las cosas con tranquilidad" y "ni le sobra ni falta tiempo".

En las celdas que la diferencia entre las *fo* y las *fe* no es significativa, quiere decir que lo observado es lo esperado y por lo tanto son sucesos independientes mutuamente no excluyentes. En las celdas en las que la diferencia entre las *fo* y las *fe* es significativa, significa que lo observado es significativamente distinto a lo esperado, en este caso al nivel de significación de 0,05, y por lo tanto lo observado no es lo esperado. Cuando la diferencia es negativa, lo observado es menor que lo esperado y cuando es positiva, lo observado es mayor que lo esperado.

En las celdas "hace las cosas con prisa, sensación de falta de tiempo" y "hace las cosas con tranquilidad, sensación de que le sobra el tiempo", la *fo* es significativamente mayor que la *fe*. Entonces tiende a haber más individuos cuando la sensación es de que les falta el tiempo y hacen las cosas con prisa y cuando la sensación es de que les sobra el tiempo y hacen las cosas con tranquilidad.

Se produce el efecto contrario, esto es la *fo* es significativamente menor que la *fe*, cuando las cosas "se hacen con tranquilidad, sensación de falta el tiempo" y "se hacen con prisa, sobra el tiempo". Entonces tiende a haber menos individuos cuando hacen las cosas tranquilamente y la sensación es de que les falta el tiempo, y hacen las cosas con prisa y la sensación es de que les sobra el tiempo.

La interpretación puede ser que cuando los individuos hacen las cosas con prisa, tiende a haber más individuos que tienen sensación de que les falta el tiempo y menos que

sienten que les sobre e inversamente, cuando las cosas se hacen con tranquilidad hay más individuos que sienten que les sobra el tiempo y menos que sienten que les falte el tiempo. En este sentido, los estadísticos de dirección indican que hay concentración de casos hacia la diagonal positiva ($t_a = 0{,}23$; $t_b = 0{,}40$; $t_c = 0{,}45$; $v = 0{,}65$).

Se recomienda a los lectores que como ejercicio trabajen con una tabla de contingencia que la variable independiente y la dependiente sea la misma, porque en este caso puede resultar clarificador el significado de los estadísticos para facilitar su lectura e interpretación (Tabla 103). La pregunta seleccionada es de un estudio de CIRES (Centro de Investigaciones de la Realidad Social) (Ver nota 63). Se ha ponderado la muestra según el criterio de CIRES.

Tabla 103 Análisis del cruce de una variable ordinal por sí misma.

Pregunta P.15: En general, ¿tiene la sensación de que le falta tiempo o de que tiene tiempo de sobra?		Pregunta P.15: En general, ¿tiene la sensación de que le falta tiempo o de que tiene tiempo de sobra?	
Categorías originales	Recodificación para mantener ordinalidad.	Categorías originales	Recodificación para mantener ordinalidad.
1. Falta.	1. Falta.	1. Falta.	1. Falta.
2. Sobra.	2. No falta ni sobra.	2. Sobra.	2. No falta ni sobra.
3. No falta ni sobra.	3. Sobra.	3. No falta ni sobra.	3. Sobra.
9. Ns/Nc	9. Ns/Nc	9. Ns/Nc	9. Ns/Nc
Esta pregunta genera la variable B1.		Esta pregunta genera la variable B1.	

Sensación o sentimiento de falta o sobra del tiempo y sensación o sentimiento de falta o sobra del tiempo.

			Falta o sobra tiempo			
			Falta	*No falta ni sobra*	*Sobra*	*Total fila*
Falta o sobra tiempo	*Falta*	*fo*	518	0	0	518
		fe	223,8	133,1	161,1	518,0
		%TC	100,0	0,0	0,0	43,2
		%TT	43,2	0,0	0,0	43,2
		Residuo	294,2	-133,1	-161,1	
		ZResiduo	19,67	-11,54	-12,69	
	No falta ni sobra	*fo*	0	308	0	308
		fe	133,1	79,1	95,8	308,0
		%TC	0,0	100,0	0,0	25,7
		%TT	0,0	25,7	0,0	25,7
		Residuo	-133,1	228,9	-95,8	
		Zresiduo	-11,54	25,73	-9,79	
	Sobra	*fo*	0	0	373	373
		fe	161,1	95,8	116,0	373,0
		%TC	0,0	0,0	100,0	31,1
		%TT	0,0	0,0	31,1	31,1
		Residuo	-161,1	-95,8	257,0	
		Zresiduo	-12,69	-9,79	23,85	
	Total columna	*fo*	518	308	373	1.199
		fe	518,0	308,0	373,0	1.199,0
		%TF	100,0	100,0	100,0	100,0
		%TC	43,2	25,7	31,1	100,0

Proceso de contraste de hipótesis:

1. H_1: "Existe asociación o dependencia entre la sensación o sentimiento de falta o sobra tiempo y la sensación o sentimiento de falta o sobra tiempo" o de forma abreviada "El sentimiento o sensación de falta o sobra tiempo influye en la sensación o sentimiento de falta o sobra tiempo". Se plantea que los sucesos son dependientes.
2. *Falta o sobra tiempo*: variable categórica ordinal.

3. Variable considerada como independiente: *falta o sobra tiempo*. Variable considerada como dependiente: la misma variable.
4. H_0: "Existe asociación o dependencia entre la sensación o sentimiento de falta o sobra tiempo y la sensación o sentimiento de falta o sobra tiempo" o de forma abreviada "El sentimiento o sensación de falta o sobra tiempo no influye en la sensación o sentimiento de falta o sobra tiempo". Se plantea que los sucesos son independientes.
5. Estadístico: θ^2, por ser las dos variables categóricas:
6. Criterio de aceptación/rechazo de H_0, Ns = 0,05.

Para contrastar la hipótesis nula (H_0) hay que calcular el θ_e^2, y en el caso de la Tabla 103 es,

$$\theta_e^2 = \sum_{i=1}^{F}\sum_{j=1}^{C}\frac{(fo_{ij} - fe_{ij})^2}{fe_{ij}} = \sum_{i=1}^{F=2}\sum_{j=1}^{C=3}\frac{(fo_{ij} - fe_{ij})^2}{fe_{ij}} = 2.398$$

En esta tabla se corresponde con el $\theta^2_{máximo}$ que se obtiene también por $\theta^2_{máximo} = N(K-1)$, siendo N el total de casos de la tabla y k el menor de las filas o las columnas. Se rechaza H_0 y se acepta H_1, al Ns = 0,05 porque $\theta_e^2 > \theta_c^2$ (gl = 4; Ns = 0,05; $\theta_c^2 = 9{,}4877$) e incluso al Ns = 0,01 (gl = 4; Ns = 0,01; $\theta_c^2 = 13{,}2767$). Según el planteamiento de contraste de hipótesis, equivale a decir que $Nc_e > Nc_c$ y equivalente también a que $Ns_e < Ns_c$. Aunque es trivial decirlo, la asociación entre una variable y ella misma es máxima.

El coeficiente c de contingencia es máximo (conocido al ser una tabla cuadrada) y toma el valor de 0,8165. La V de Cramer es igual a 1, y λ se hace mayor que la unidad (1,41) al ser θ_e^2 mayor que N. El estadístico ς es igual a uno, porque la reducción proporcional del error al predecir las categorías de una variable a partir de la otra, cuando son la misma, es del 100,0%. La conclusión es que la fuerza de la asociación de una variable con ella misma es máxima.

La dirección de la asociación es directa y máxima, ya que a partir de los valores de una variable podemos predecir a ella misma. Las personas que tienen sensación de que les falta el tiempo "son personas que tienen sensación de que les falta el tiempo" y todos los casos están situados en la diagonal principal. Así, los valores que toman *tau-b* (1,00), *tau-c* (0,98), *gamma* (1,00) y *d* de Somers (1,00), son máximos también.

12.6 Restricciones de chi-cuadrado

1. En las tablas de contingencia puede ocurrir que haya frecuencias esperadas menores de 5. Hay dos líneas a seguir en esta circunstancia. Que no se aplique θ^2, o que se aplique si son pocas las celdas con frecuencias esperadas menores de 5. Entonces hay que decidir cuando son "pocas". Esta característica suele ocurrir cuando la tabla tiene muchas filas y/o muchas columnas. Una posible solución es reducir las filas y/o las columnas por recodificación.

2. En las tablas de 2x2, no se recomienda usar chi-cuadrado porque la distribución de la curva se aleja de la normal y no se pueden utilizar criterios de probabilidad. En su lugar se puede utilizar θ^2 con la corrección por la continuidad de Yates, simbólicamente,

$\theta_e^2 = \sum_{i=1}^{F} \sum_{j=1}^{C} \frac{(\lvert fo_{ij} - fe_{ij} \rvert - 0{,}5)^2}{fe_{ij}}$		Fórmula 78

3. Si además tiene menos de 20 casos, se aplica el test exacto de Fisher. En los dos casos, la corrección de Yates o el exacto de Fisher, el contraste de hipótesis se realiza con los mismos criterios vistos para θ^2.
4. El aumento de las frecuencias observadas de las celdas puede hacer que una tabla que no tenía asociación, se vuelva significativa. Se dice que θ^2 es sensible al número de casos.
5. Para poder aplicar chi-cuadrado, la muestra tiene que estar obtenida aleatoriamente desde una distribución multinomial.

13 Tabla de medias

Otro grupo de la Estadística Descriptiva Bivariable es el que cruza una variable numérica con otra u otras categóricas y se denomina *tabla de medias*. Este tipo de tablas se obtienen por el cruce de una variable numérica que puede ser considerada como dependiente y que es la agrupada por otra variable categórica que puede ser considerada como la independiente y que es la de agrupamiento.

En el cruce se producen tantos grupos en o de la variable numérica como categorías tiene la variable de agrupamiento. Este esquema se denomina *muestras independientes*. El conjunto de valores de la variable numérica por cada categoría de la variable de agrupamiento, constituyen un subgrupo o submuestra y sobre ellos se pueden calcular todos los estadísticos de la Estadística Descriptiva Univariable: *moda, mediana, media, rango, varianza, desviación típica, coeficiente de variación, asimetría, apuntamiento, percentiles, gráficos, etc.*

La variable numérica puede ser agrupada por más de una variable categórica o de agrupamiento. Entonces se produce un grupo, subgrupo o submuestra en la variable numérica por cada combinación de categorías de las variables de agrupamiento o categóricas.

Sea una variable numérica Y y una variable categórica X con dos categorías x_1 y x_2, el cruce de la variable numérica por la variable categórica produce el esquema de la Tabla 104.

Tabla 104 Cruce de variable numérica por variable categórica de dos categorías.

Muestra total	Submuestras	Y	X
$\overline{Y}_t, S_t^2, S_t, n_t$	Submuestra x_1 $\overline{Y}_1, S_1^2, S_1, n_1$	y_1	x_1
		Y_2	x_1
		Y_3	x_1
		Y_4	x_1
		Y_5	x_1
		Y_6	x_1
		Y_7	x_1
		Y_8	x_1
		Y_9	x_1
		y_{10}	x_1
	Submuestra x_2 $\overline{Y}_2, S_2^2, S_2, n_2$	y_{11}	x_2
		y_{12}	x_2
		y_{13}	x_2
		y_{14}	x_2
		y_{15}	x_2
		y_{16}	x_2
		y_{17}	x_2
		y_{18}	x_2
		y_{19}	x_2
		y_{20}	x_2

Sea una variable numérica Y y dos variables categóricas X y Z con dos categorías cada una x_1 y x_2, y z_1 y z_2 el cruce de la variable numérica por la dos variables categóricas produce el esquema de la Tabla 105.

Tabla 105 Cruce de variable numérica por dos variables categóricas de dos categorías cada una.

Muestra total	Submuestras	*Y*	*X*	*Z*
$\overline{Y}_t, S_t^2, S_t, n_t$	Submuestra x_1 z_1 $\overline{Y}_1, S_1^2, S_1, n_1$	y_1	x_1	z_1
		y_2	x_1	z_1
		y_3	x_1	z_1
		y_4	x_1	z_1
		y_5	x_1	z_1
	Submuestra x_1 z_2 $\overline{Y}_2, S_2^2, S_2, n_2$	y_6	x_1	z_2
		y_7	x_1	z_2
		y_8	x_1	z_2
		y_9	x_1	z_2
		y_{10}	x_1	z_2
	Submuestra x_2 z_1 $\overline{Y}_3, S_3^2, S_3, n_3$	y_{11}	x_2	z_1
		y_{12}	x_2	z_1
		y_{13}	x_2	z_1
		y_{14}	x_2	z_1
		y_{15}	x_2	z_1
	Submuestra x_2 z_2 $\overline{Y}_4, S_4^2, S_4, n_4$	y_{16}	x_2	z_2
		y_{17}	x_2	z_2
		y_{18}	x_2	z_2
		y_{19}	x_2	z_2
		y_{20}	x_2	z_2

14 Muestreo. Probabilístico y no probabilístico[65]

Antes de comenzar con la Estadística Inferencial Paramétrica, se considera conveniente ver el capítulo de Muestreo para introducir los conceptos adecuados a los contrastes de hipótesis con subgrupos.

La Teoría y la Técnica del diseño de muestras, igual que la Estadística, se puede considerar que es un descubrimiento y no un invento, la actividad de todos los seres vivos, probablemente, implica operaciones de muestreo. Por cuestiones de comunicación, los ejemplos de referencia serán con los humanos y porque se pueden hacer autocomprobaciones empíricas.

Como definiciones, *muestreo* es "Un método para recoger información y hacer las inferencias sobre una población más grande o universo, a partir del análisis de sólo una parte, la muestra".[66] Y *muestra* "una parte pequeña que tiene la intención de mostrar lo que es el todo".[67]

En la vida cotidiana se hacen muchas operaciones que son *muestreo* o prueba de la realidad que nos rodea para, en base a ellas, hacer inferencias de cómo es la generalidad. Hechos sencillos que lo demuestran son el plato de sopa que previamente se toma una cucharada de algún lado del plato y se aproxima de forma prudente a los labios para comprobar su temperatura. Cuando se compra un jamón y se le pide al dependiente un trocito de algún lado para probar si el jamón está salado. Cuando se comprueba extendiendo la mano fuera de la ventana para ver cuánto llueve. Hay referencias bíblicas de muestreo como la paloma que Noé envía para ver si hay tierra seca y al traer una ramita entiende que ya pueden salir del Arca.

Procedimientos científicos de muestreo son el análisis de sangre para ver cuál es el estado de salud general, que extraen sangre de una cierta parte del cuerpo y sólo un poquito (muestra). No necesitan extraer toda la sangre del cuerpo para hacer las comprobaciones. No es necesario comerse el jamón entero para ver si está salado, ni hay que comerse toda la sopa para comprobar si está caliente. Tampoco hay que salir todo el día a la calle ni recorrer grandes distancias para comprobar que está lloviendo y salir a la calle, dentro de un cierto perímetro, va a suponer el mojarse.

Hay un factor común en todos estos casos y es que son hechos homogéneos. Toda la sopa tiene la misma temperatura, todo el jamón tiene el mismo sabor, si llueve lo hará de forma semejante en un radio a la redonda tan amplio que andando no saldríamos de él. Y toda la sangre que quieren analizar es muy homogénea.

Pero en Sociología interesa analizar los aspectos sociales, políticos, económicos, demográficos, etc. de las poblaciones de personas y la característica principal es la heterogeneidad. Esta característica hace que una sola persona no sea reflejo (representativa) de toda una población. Para poder hablar (inferir) cosas de una población, necesitamos un grupo (muestra) que sea representativa de toda la población. Para que una muestra sea

---

[65] Este capítulo de muestreo no es exhaustivo, ni pretende ser completo. La pretensión es introducir el *muestreo* a nivel conceptual, trata de mostrar a los lectores la "idea" y poder diseñar muestras sencillas. Posteriormente, el lector tiene que acceder a los libros específicos de la materia (Cochram, 1974; Mirás, 1985; Sánchez-Crespo, 1986; Fernández García, 1995).

[66] "sampling" A Dictionary of Sociology. John Scott and Gordon Marshall. Oxford University Press 2005. Oxford Reference Online. Oxford University Press. Universidad Complutense de Madrid. 17 September 2008 http://www.oxfordreference.com/views/ENTRY.html?subview=Main&entry=t88.e1995. Traducción propia.

[67] "sample noun" The Oxford Dictionary of English (revised edition). Ed. Catherine Soanes and Angus Stevenson. Oxford University Press, 2005. Oxford Reference Online. Oxford University Press. Universidad Complutense de Madrid. 17 September 2008 http://www.oxfordreference.com/views/ENTRY.html?subview=Main&entry=t140.e68058. Traducción propia.

representativa de una población tenemos que aplicar la Teoría y las Técnicas de muestreo.

Para conseguir que una muestra sea representativa de una población hay que aplicar Técnicas de Muestreo y Técnicas de Cálculo de Tamaño de Muestra. Con las Técnicas de Cálculo de Tamaño de Muestra sabemos a cuántas personas hay que seleccionar y con las Técnicas de Muestreo, a cuáles y cómo seleccionarlas o buscarlas.

La Teoría y Técnicas de muestreo se aplican porque no se dispone de los recursos económicos y materiales suficientes para trabajar con toda la población o censo.[68] También se producen menos errores porque el trabajo se controla mejor y se emplea a personal más especializado cuando se trabaja con un número pequeño de observaciones (muestra) que cuando se trabaja con un número grande (población). En realidad la muestra puede producir datos más exactos que trabajar con la población o el censo (Cochram, 1974: 19-31).

Las Técnicas de Muestreo pueden ser probabilísticas y no probabilísticas. En las primeras la probabilidad de selección de una de las múltiples muestras que pueden ser extraídas de la población puede ser distinta y entonces debe ser conocida esta probabilidad o consideramos que es igual para todas las muestras. Por comodidad para los procesos de cálculo se asume que es igual para todas las muestras. De la misma manera, la probabilidad de selección de los individuos que componen cada muestra puede ser distinta y entonces debe ser conocida esta probabilidad o consideramos que esta probabilidad es igual para todos los individuos. Por comodidad para los procesos de cálculo se asume que es igual para todos los individuos. En las no probabilísticas esta probabilidad es desconocida. Esta característica hace que con las primeras Técnicas de Muestreo obtengamos muestras representativas, numéricamente hablando, y las segundas no producen muestras representativas o su representatividad es estructural, y la información que facilitan asumimos que puede ser generalizable.[69] Las primeras técnicas de muestreo se usan en las Técnicas de Investigación del Paradigma Técnico Cuantitativo y las segundas en las Técnicas de Investigación del Paradigma Técnico Cualitativo. La Tabla 106 muestra las Técnicas de Muestreo, y la Tabla 107 algunas definiciones previas.

Tabla 106 Técnicas de Muestreo.

Técnicas de Muestreo	Probabilísticas	Muestreo Aleatorio Simple Muestreo Aleatorio Sistemático Muestreo Aleatorio Estratificado Muestreo por Conglomerados
	No probabilísticas	Muestreo Intencional Muestreo Accidental Muestreo de Bola de Nieve Muestreo por Cuotas

[68] Se consideran distintos *censo* y *población*. La *población* puede ser un conjunto de indiviuos del *censo* que cumplen ciertos requisitos y sobre los cuales se quiere obtener una muestra. El *censo* es el conjunto total de los individuos de una zona geográfica delimitada administrativamente: municipio, provincia, Comunidad, Nación, etc. La *población* siempre será menor o igual al *censo*.

[69] Por ejemplo, si se pregunta a un grupo de personas cuántos vasos de agua han bebido la último semana, van a dar una información numérica, el número de vasos que han bebido y nos permite calcular la media de vasos de agua bebidos la última semana. La cuestión es ¿Este estadístico es representativo de alguna población? Si la muestra es representativa, entonces la media se puede inferir a esa población. Pero tenemos otra información, y es que salvo muy raras excepciones, es que todos habrán bebido agua. Por lo tanto podemos generalizar que todos los humanos beben agua. Aunque para hacer esta generalización necesitamos otra información de tipo fisiológica, y es que toda forma de vida conocida, necesita agua para sobrevivir y por lo tanto se puede aplicar también a los humanos.

Tabla 107 Glosario de términos.	
Censo	Relación completa de los elementos de una población.
Elemento, objeto o unidad de observación	Cada una de las unidades de la población sobre las que interesa obtener información.
Error exacto	La diferencia entre el parámetro y el estimador o estadístico. Por ejemplo la diferencia entre la media de la población y la media de la muestra es lo que se entiende en este manual como error exacto, pero normalmente no es conocido porque los parámetros de la población (media de la población en este caso) son desconocidos.
Error de muestreo o muestral	Es el error obtenido a partir de una muestra obtenida por los procedimientos de técnicas de muestreo probabilísticas y permite definir el intervalo de confianza dentro del cual estará el parámetro desconocido de la población. El error muestral o error absoluto es el error típico multiplicado por el valor de *Z* que define un determinado *Nc*. La muestra permite obtener el error típico y a partir de este obtener el error absoluto o error muestral.
Error no muestral	Es el que se produce en toda la investigación como consecuencia de definiciones conceptuales incorrectas, de fallos en los instrumentos de medida, fallos de los entrevistadores, fallos de los entrevistados, fallos en el desarrollo del trabajo de campo (Cochram, 1974: 443-496; Cea, 2004: 40-45).
Estadístico	Función aplicada sobre una característica medida en una muestra. Ejemplo: media, varianza, etc.
Estimador	Es el valor muestral utilizado para inferir un valor poblacional. Un estimador *insesgado* es un estimador cuya esperanza matemática es el parámetro poblacional que estima (la media es una esperanza matemática). Se dice que un estimador es *consistente* si al sustituir el tamaño de la muestra por el del total de la población la estimación coincide con el parámetro poblacional.
Inferencia estadística	Proceso de estimación de los parámetros de una población a partir de los estadísticos obtenidos de una muestra de esa población.
Intervalo de confianza	Intervalo con una determinada probabilidad (*Nc*) de contener un parámetro. Normalmente la media o la proporción. Se calcula a partir de los estadísticos y el error muestral.
Ley de los Grandes Números	Si se aumenta *n* hasta llegar a ser igual que *N*, entonces la muestra se convierte en la población y por lo tanto los estadísticos de la muestra son los parámetros de la población. El error exacto tenderá a ser cero.
Marco muestral	Listado o delimitación que identifica a los elementos de la población objetivo desde la que se va a extraer la muestra.
Técnica de muestreo	Procedimiento utilizado para seleccionar las unidades muestrales y que se puedan considerar representativas de la población.
Muestra	Subconjunto de elementos de la población elegidos para estudiar y así tratar de inferir características de la población. Tiene la misma delimitación geográfica que la población.
Muestreo	Conjunto de operaciones encaminadas a determinar una muestra, su tamaño y demás características necesarios para identificar a los elementos que la forman.
Nivel de confianza	Probabilidad de que un parámetro esté dentro del intervalo de confianza o si obtenemos 100 muestras, es la proporción o porcentaje de muestras que contendrían el parámetro desconocido de la población.
Parámetro	Función aplicada sobre una característica medida en una población. Ejemplo: media, varianza, etc.
Población	Conjunto formado por la totalidad de elementos con arreglo a unas características concretas y con una delimitación geográfica. La población puede ser unidimensional si sólo consideramos una variable. Por ejemplo el peso. Es pluridimensional si se consideran muchas variables. En sociología las poblaciones se consideran pluridimensionales porque se estudian muchas variables (Sánchez-Crespo, 1986: 17).
Sesgo	Error específico de la muestra por falta de representatividad.
Teorema del Límite Central	Si el tamaño de cada muestra es lo suficientemente grande (a partir de 30) y si se extraen muchas muestras (más de 30) aleatoriamente, este teorema nos dice que la distribución de las medias muestrales tiene una distribución normal con media igual a la media de la población y con una varianza igual a la varianza de la población dividida por el tamaño de la muestra.
Unidad muestral	Conjunto de elementos de la población que contiene varias unidades u objetos de observación de la población y es el conglomerado en el Muestreo por Conglomerados.

El concepto de *generalizable* en el paradigma técnico cualitativo está basado sobre la idea de la representatividad social, que va más allá de los límites de la representatividad estadística. La finalidad es observar las relaciones entre variables, en vez de evaluar el número de personas que poseen una característica (Gobo, 2004: 453).[70]

Pero la representatividad es un concepto amplio y complejo. En el Paradigma Técnico Cuantitativo, se pretende la representatividad estadística y numérica. En el Paradigma Técnico Cualitativo "Se pretende, a través de la elaboración de ejes o tipologías discursivas, la

[70] Son ejemplos de estudios cualitativos que se pueden considerar generalizables: E. Goffman (1961); W. F. Whyte (1943); A. G. Gouldner (1954); A. V. Cicourel (1968); T. A. van Dijk (1983); D. A. Norman (1988). Citados en G. Gobo (2004: 453).

representación socio-estructural de los sentidos circulantes en un determinado universo y con relación al tema a investigar" (Serbia, 2007: 133).

Hasta ahora se ha hablado de representatividad de comportamientos cuantificables, de valores, actitudes, creencias. Todos son comportamientos sociales correspondientes a la Cultura desarrollada por el ser humano. Pero hay otros comportamientos adquiridos a través del proceso *filogenético* como subespecie. Cuando se trata de instintos y emociones el comportamiento es homogéneo entre todos los seres vivos en general y del ser humano en particular, considerados normales.[71] Sólo hay que privar a un humano de agua un número de días para saber que es lo que le pasaría a cualquier otro humano. A través del estudio de cualquier humano se puede ver que el instinto de supervivencia es muy fuerte y que la tendencia es a permanecer con vida. Un solo individuo es una muestra suficiente para saber que a cualquiera de ellos le resulta imposible, a través del acto-reflejo, introducir la mano en un puchero con agua hirviendo o en el fuego. Que si a cualquier individuo recién nacido se le priva de contacto con toda forma de vida, además de por inanición, terminará empobrecido por falta de desarrollo mental y finalmente muriendo. O expresado de otra manera, si se aísla a un solo recién nacido de todo contacto humano, no sólo no aprendería a hablar, sino que después de pasado cierto nivel de desarrollo nunca lo conseguiría y serviría para saber que a cualquiera otro humano le ocurriría lo mismo. Los lectores pueden aportar infinidad de ejemplos que pueden ver en la prensa o los informativos.

En alguno de los casos planteados anteriormente no hay ninguna excepción, porque si los individuos de alguna especie no cumplen el mandato de los instintos, las emociones o los acto-reflejo, probablemente, entrarían en proceso de extinción. Entonces para hablar de representatividad es necesario saber que es lo que se quiere representar para saber como hay que estudiarlo, aunque no sea una tarea fácil. El comportamiento humano es una amalgama de instintos, emociones y comportamiento social y no es fácil descomponerlos y separarlos para su estudio y análisis.

[71] Si se asumen las Teorías e Hipótesis de la Evolución, se puede considerar que los instintos (asociados al Mesencéfalo o cerebro medio) se originan en el Período Carbonífero (hace 360 a 290 Millones de años). Las emociones (asociadas al Sistema Límbico o Diencéfalo) en el Triásico (hace 250 a 199 Ma. años). Lo social de "manada" se puede atribuir a la corticalización de los neomamíferos, hace 65 a 55 Ma. (Época Paleoceno). Y la Cultura, si se asocia con el área de neocorticalización fuerte, su origen tiene una antigüedad de 3,5 y 2 Ma. (Período Terciario), por lo que el comportamiento asociado a esta parte del cerebro tiene un período de maduración menor y al estar menos consolidado, su estudio y predicción debe resultar más complejo (Confrontar De la Puente 2007 a). No obstante, aunque este esquema se muestra simple y sencillo, la Evolución del cerebro es compleja y llena de lagunas (Para ampliar ver Aboitiz *et al.,* 2007).

14.1 Conceptos previos

Antes de desarrollar las Técnicas de Muestreo se van a establecer unos conceptos previos.

PARÁMETROS Y ESTADÍSTICOS

La población y la muestra están representadas por valores. En la primera se denominan *parámetros* y en la segunda *estadísticos*.

Tabla 108 Valores de la población y de la muestra

	Muestra	Población
	Estadísticos	Parámetros
Número de casos	n	N
Media	$\overline{X}$	σ
Varianza	S^2	ω^2
Desviación típica	S	ω
Porcentaje o proporción.	p	P
Delimitación geográfica	Es la misma	

A través de una muestra representativa se trata de inferir desde los *estadísticos* conocidos de una muestra los *parámetros* desconocidos de una población.

RELACIÓN ENTRE LA POBLACIÓN Y LA MUESTRA

Entre la *población* y la *muestra* existe una relación *cualitativa* y otra *cuantitativa*. La primera significa que la muestra debe ser heterogénea como la población, esto es, debe tener las mismas características que la población. Si la población tiene varones y mujeres, la muestra debe tener varones y mujeres. Si la primera tiene individuos de todas las edades la muestra también, si la población tiene individuos de diferentes niveles de instrucción, la segunda también y así sucesivamente de tal manera que la muestra se considere que es heterogénea como la población.

La relación *cuantitativa* se concreta en dos ratios, el *coeficiente de elevación (ce)* y la *fracción de muestreo (fm)*. Simbólicamente,

$$ce \mid \frac{N}{n}$$ Fórmula 79

$$fm \mid \frac{n}{N} \text{ ó } fm \mid \frac{n}{N}\Delta 100$$ Fórmula 80

El *ce* es el número de veces que la muestra está contenida en la población o el valor por el que hay que multiplicar *n* para obtener *N*. La *fm* es la proporción de *n sobre N*, y es el inverso del *ce*.

En el muestreo aleatorio simple, la *fm* también se puede considerar como la probabilidad de que un individuo de la población sea seleccionado. Según Fórmula 53, *n* son los *hechos favorables* y *N* los *hechos posibles*. Simbólicamente,

$$P_{(S_1)} = \frac{hechos\_favorables}{hechos\_posibles} = \frac{n}{N}$$

Igualmente, en las condiciones de muestreo aleatorio simple, la probabilidad de obtener una de esas muestras estará dado por la inversa de las combinaciones sin repetición de *N* elementos tomados de *n* en *n*, que es uno (una muestra) dividido por el total de las muestras que se pueden extraer. Entonces se considera que la probabilidad de obtener una de las muestras es, simbólicamente,

$$P_S = \frac{hechos\_favorables}{Hechos\_posibles} = \frac{1\,\text{muestra}}{\text{total de muestras}} = \frac{1}{C_{N,n}} = \frac{1}{\binom{N}{n}} = \frac{1}{\frac{N!}{n!(N-n)!}} = \frac{n!(N-n)!}{N!}$$

$P_S = \frac{1}{C_{N,n}} = \frac{1}{\binom{N}{n}} = \frac{1}{\frac{N!}{n!(N-n)!}} = \frac{n!(N-n)!}{N!}$	Fórmula 81

LEY DE LOS GRANDES NÚMEROS

La *ley de los grandes números* establece que cuando *n* tiende a *N*, así mismo ocurre con los *estadísticos* de la muestra que tienden a los *parámetros* de la población. Cuando *n* se hace *N* entonces los *estadísticos* son los *parámetros*.

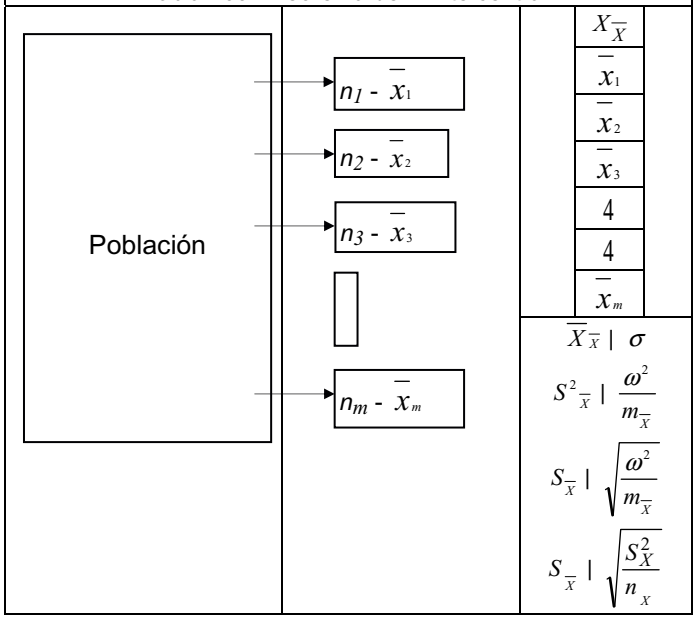
TEOREMA DEL LÍMITE CENTRAL

El *teorema del límite central* establece que si extraemos *m* muestras de una población de tamaño *n,* siendo *n* en todos los casos mayor que 30, si calculamos las medias muestrales de las *m* muestras, obtenemos *m* medias muestrales. Si creamos una variable ($X_{\overline{X}}$) con las *m* medias muestrales, esta variable tiene una distribución *normal*[72] ($N(\sigma, S_{\overline{X}})$)[73]. Entonces, la media de las medias muestrales es igual a la media de la población y la varianza es igual a la varianza de la población partido por la *m* de las medias muestrales (Tabla 109).

[72] La distribución de las medias muestrales es normal aunque la distribución de la población sea uniforme (Norusis, 1986: B-119).

[73] Una distribución normal (*N*) con media igual a la media de la población (σ) y desviación típica igual a la desviación típica de la variable de las medias muestrales ($S_{\overline{X}}$).

Tabla 109 Teorema del límite central.



La media de las medias muestrales es igual a la media de la población, simbólicamente,

$\overline{X}_{\bar{X}} \mid \sigma$	Fórmula 82

La varianza de las medias muestrales es igual a la varianza de la población partido por la m de las medias muestrales, simbólicamente.

$S^2_{\bar{X}} \mid \frac{\omega^2}{m_{\bar{X}}}$	Fórmula 83

Y por lo tanto

$S_{\bar{X}} \mid \sqrt{\frac{\omega^2}{m_{\bar{X}}}}$	Fórmula 84

Y se denomina *error típico* o *desviación típica de las medias muestrales*.

Como normalmente la varianza de la población es desconocida y no existe una distribución de medias muestrales, entonces se acepta como *error típico* o *desviación típica de las medias muestrales*, la raíz cuadrada de la varianza de la variable dividido por la *n* de la variable. Indica la dispersión de la media de la muestra obtenida respecto de la media desconocida de la población y está relacionado inversamente con el tamaño de la muestra. Cuanto mayor sea la muestra menor será el error y viceversa cuanto menor sea la muestra mayor será el error. Simbólicamente,

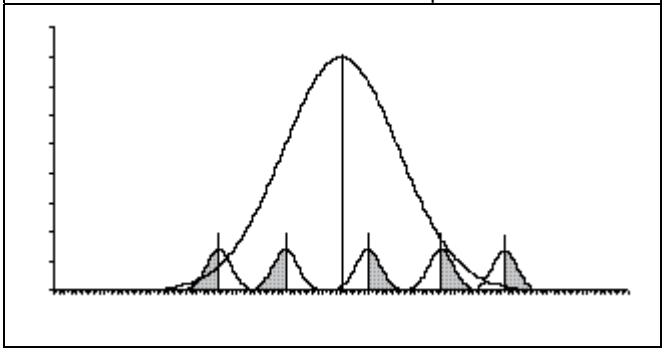
$S_{\overline{X}} \mid \sqrt{\frac{S_X^2}{n_X}}$	Fórmula 85

Si se considera la población finita, entonces $N < 100.000$ y se aplica el *corrector por población finita* (*cpf*),

$S_{\overline{X}} \mid \sqrt{\frac{S_X^2}{n_X}} \Delta \sqrt{1 4 fm}$	$cpf \mid \sqrt{\frac{N 4 n}{N}} \mid \sqrt{\frac{N}{N} 4 \frac{n}{N}} \mid \sqrt{1 4 fm}$ Se denomina corrector por población finita *(cpf)*	Fórmula 86

La demostración intuitiva (no matemática) de que la media de las medias es la media de la población se deriva de que al obtener muestras de una distribución, la media tiende a la media de la población, por tender a haber más de lo que más hay y menos de lo que menos hay, asumiendo que la distribución es normal (Ver nota 72). La varianza, al ser la dispersión de una variable que sus valores son las medias de las muestras obtenidas, se pierde dispersión al eliminar los valores de los casos que quedan en la parte externa (zona sombreada en el Gráfico 19) de las distribuciones de las muestras.

Gráfico 19 Pérdida de dispersión.



ERROR EXACTO

El *error exacto* (*ee*) es la diferencia entre el parámetro de la población y el estadístico de la muestra, simbólicamente en el caso de la media y de la proporción,

$ee_{\overline{X}} \mid \sigma 4 \overline{X}$	Fórmula 87

$ee_P \mid P 4 p$	Fórmula 88

Podemos conocer el *error* por diferencia entre el parámetro y el estadístico, pero como normalmente no se conoce el parámetro, no se puede conocer el error exacto cometido.

Entonces la estimación del *parámetro* desconocido se puede hacer por *estimación puntual* o *estimación por intervalo*. La *estimación puntual* es asignar al *parámetro* de la población el valor del *estadístico* de la muestra, simbólicamente,

$\sigma \mid \overline{X}$	Fórmula 89

$P \mid p$	Fórmula 90

ERROR MUESTRAL

Según se ha visto en el *teorema del límite central*, excepto por azar, los valores de los *estadísticos* no coinciden con los valores de los *parámetros*, por lo que hacer asignaciones directas lleva tener errores y además no conocer la magnitud del error.

Entonces, la estimación de parámetros se hace mediante la *estimación de intervalos* a partir del estadístico de la muestra. La estimación del intervalo se hace a partir del *error muestral*. El error muestral es el *error típico* multiplicado por *Z*, estando esta definida por el *Nc* que define el intervalo de confianza. El error típico, que es la desviación típica de las medias muestrales, según el *Teorema del Límite Central*, y según la Fórmula 85, para las medias es,

$S_{\bar{x}} \mid \sqrt{\frac{S_X^2}{n_X}}$	Fórmula 91

Y para proporciones (Ver Epígrafe 15.1),

$S_p \mid \sqrt{\frac{p \Delta q}{n_X}}$	Fórmula 92

El error absoluto o error muestral para medias es, simbólicamente,

$e_{\bar{x}} \mid z \Delta \sqrt{\frac{S_X^2}{n_X}}$	*Nc* = 95,00, entonces *z* = 1,96 *Nc* = 95,44, entonces *z* = 2,00 *Nc* = 99,74, entonces *z* = 3,00	Fórmula 93

Y para proporciones,

$em_p \mid z \Delta \sqrt{\frac{p \Delta q}{n_X}}$	*Nc* = 95,00, entonces *z* = 1,96 *Nc* = 95,44, entonces *z* = 2,00 *Nc* = 99,74, entonces *z* = 3,00	Fórmula 94

La *estimación por intervalo* tampoco permite saber el valor exacto del *parámetro* desconocido pero define un *intervalo de confianza* dentro del cual se encuentra y a un cierto *nivel de confianza* (*Nc*) o lo que es lo mismo cual es la probabilidad de que el *parámetro* se encuentre dentro de ese intervalo o que de 100 muestras cuántas contendrían en su *intervalo de confianza* el *parámetro* desconocido.

La *estimación por intervalo*, supone conocer los límites dentro de los cuales se encuentra el *parámetro* desconocido de la población, pero siempre existirá la probabilidad de que el intervalo no contenga este *parámetro*. Las probabilidades supone asumir este nivel de incertidumbre.

Para la *estimación por intervalo* se aplican los conceptos de *intervalo de confianza, teorema de Tchebycheff* y *probabilidades* (ver epígrafe 11.2). Conocida la media, el número de casos y la desviación típica de una variable y asumiendo que su distribución es normal, marcadamente normal o supuestamente normal, se puede calcular la probabilidad y el intervalo dentro del cual estará un caso. La probabilidad se llama *nivel de confianza (Nc)* y el intervalo, *intervalo de confianza*, simbólicamente,

$P_{\left\|\overline{X}-n\Delta S\right.\left\{\overline{X}\right\|\overline{X}+n\Delta S\left.\right\|} = Nc$	Según fórmula de Tabla 84	Fórmula 95

Y se puede leer como: la probabilidad (es una superficie) definida por el intervalo (es un segmento) comprendido entre la media menos *n* veces la desviación típica que es menor que la media y menor que la media más *n* veces la desviación típica y esta probabilidad se llama *nivel de confianza* (*Nc*).

Y en el caso de la distribución de las medias muestrales, tenemos que,

$P_{\left\|\overline{X}-n\Delta S_{\overline{X}}\right.\left\{\sigma\right\|\overline{X}+n\Delta S_{\overline{X}}\left.\right\|} = Nc$	Según fórmula de Tabla 84	Fórmula 96

En donde $\overline{X}$ es la media de la variable; $S_{\overline{X}}$ es la desviación típica de las medias muestrales o *error típico de la media*; $n\Delta S_{\overline{X}}$ es el *error absoluto* o *error muestral*; $\overline{X}-n\Delta S_{\overline{X}}$ es el límite inferior del *intervalo de confianza;* $\overline{X}+n\Delta S_{\overline{X}}$ es el límite superior del *intervalo de confianza*, y la media de la población es σ. El valor de *n* está definido por el *Nc*. Si $Nc = 95{,}00\%$ ó $0{,}9500 \rightarrow n = 1{,}96$, si $Nc = 95{,}44\%$ ó $0{,}9544 \rightarrow n = 2$ [74] y si $Nc = 99{,}74\%$ ó $0{,}9974 \rightarrow n = 3$, por lo que *n* es el valor de la *Z.*

14.2 Intervalo de confianza para la media

Tabla 110 Estimación por intervalo de la media de la población.

$P_{\left\|\overline{X}-1{,}96\Delta S_{\overline{X}}\right.\left\{\sigma\right\|\overline{X}+1{,}96\Delta S_{\overline{X}}\left.\right\|} = 0{,}9500$	$\overline{X}-1{,}96\Delta S_{\overline{X}}$ Límite inferior del intervalo de confianza.
	$Nc = 95{,}00\%$; *error típico* = $S_{\overline{X}}$; *error absoluto* = $1{,}96\Delta S_{\overline{X}}$
	$\overline{X}+1{,}96\Delta S_{\overline{X}}$ Límite superior del intervalo de confianza.
$P_{\left\|\overline{X}-2{,}00\Delta S_{\overline{X}}\right.\left\{\sigma\right\|\overline{X}+2{,}00\Delta S_{\overline{X}}\left.\right\|} = 0{,}9544$	$\overline{X}-2{,}00\Delta S_{\overline{X}}$ Límite inferior del intervalo de confianza.
	$Nc = 95{,}44\%$; *error típico* = $S_{\overline{X}}$; *error absoluto* = $2{,}00\Delta S_{\overline{X}}$
	$\overline{X}+2{,}00\Delta S_{\overline{X}}$ Límite superior del intervalo de confianza.

[74] En la tabla de probabilidad de *Z* obtenida por el autor, así como otras tablas, el valor del *Nc* para *Z* = 2, es 95,44%, aunque habitualmente en algunos manuales y estudios aparece 99,45% y se redondea a 95,5%.

Tabla 110 Estimación por intervalo de la media de la población.	
$P_{\|\overline{X}-3{,}00\Delta S_{\overline{X}} \le \mu \le \overline{X}+3{,}00\Delta S_{\overline{X}}\|} = 0{,}9974$	$\overline{X}-3{,}00\Delta S_{\overline{X}}$ Límite inferior del intervalo de confianza.
	Nc = 99,74% ; *error típico* = $S_{\overline{X}}$; *error absoluto* = $3{,}00\Delta S_{\overline{X}}$
	$\overline{X}-3{,}00\Delta S_{\overline{X}}$ Límite superior del intervalo de confianza.

Aplicándolo al estudio de CIRES de enero de 1996, *Usos del tiempo*, para estimar la media de edad de la población española de ambos sexos y de 18 años o más en enero de 1996, se calcula la media y la desviación típica de la variable edad y se procede,

Tabla 111 Estadísticos de la variable edad.	
Encuesta CIRES: *Usos del tiempo.* Enero de 1996. Ámbito nacional. Población española de ambos sexos de 18 años o más	
$\overline{X}_X$	44,95 años
S_X	18,33 años
S_X^2	335,97
n_X	1.200
$S_{\overline{X}}$	$\sqrt{\frac{S_X^2}{n_X}}$ años
$S_{\overline{X}}$ $(cpf)^{75}$	$\sqrt{\frac{S_X^2}{n_X}}\Delta\sqrt{1-fm}$ años
$S_{\overline{X}}$	$\sqrt{\frac{335{,}97}{1.200}} = 0{,}53$ años

Aplicándolo a la encuesta de CIRES,

Tabla 112 Estimación por intervalo de la media de edad de la población española para distintos *Nc*.		
A	$P_{\|44{,}954-1{,}96\Delta 0{,}53 \le \mu \le 44{,}952+1{,}96\Delta 0{,}53\|} = 0{,}9500$ $P_{\|43{,}91 \le \mu \le 45{,}99\|} = 0{,}9500$	$\|44{,}95-1{,}96\Delta 0{,}53\| = \|44{,}95-1{,}04\| = 43{,}91$ Limite inferior
		Nc = 95,00% ; *error típico* = 0,53; *error absoluto* = 1,04
		$\|44{,}95+1{,}96\Delta 0{,}53\| = \|44{,}95+1{,}04\| = 45{,}99$ Limite superior
B	$P_{\|44{,}954-2{,}00\Delta 0{,}53 \le \mu \le 44{,}952+2{,}00\Delta 0{,}53\|} = 0{,}9544$ $P_{\|43{,}89 \le \mu \le 46{,}01\|} = 0{,}9544$	$\|44{,}95-2{,}00\Delta 0{,}53\| = \|44{,}95-1{,}06\| = 43{,}89$ Limite inferior
		Nc = 95,44% ; *error típico* = 0,53; *error absoluto* = 1,06
		$\|44{,}95+2{,}00\Delta 0{,}53\| = \|44{,}95+1{,}06\| = 46{,}01$ Limite superior
C	$P_{\|44{,}954-3{,}00\Delta 0{,}53 \le \mu \le 44{,}952+3{,}00\Delta 0{,}53\|} = 0{,}9974$ $P_{\|43{,}36 \le \mu \le 46{,}54\|} = 0{,}9974$	$\|44{,}95-3{,}00\Delta 0{,}53\| = \|44{,}95-1{,}59\| = 43{,}36$ Limite inferior
		Nc = 99,74% ; *error típico* = 0,53; *error absoluto* = 1,59
		$\|44{,}95+3{,}00\Delta 0{,}53\| = \|44{,}95+1{,}59\| = 46{,}54$ Limite superior

[75] El *cpf* (corrector por poblaciones finitas) se aplica cuando la población se considera finita (N < 100.000).

En la Tabla 112 el caso *A* significa que con la probabilidad de 0,95 (*Nc* = 0,95 ó 95,0 %), la media de edad de la población española en enero de 1996 de 18 años o más, está comprendida en el intervalo (de confianza) de 43,91 años y 45,99 años, o lo que es lo mismo, que si se extraen 100 muestras de esa población, 95 tendrían en su intervalo el parámetro desconocido de la población. Como se comentó anteriormente, al existir la probabilidad de 0,05 o 5,0 % de que la muestra no contenga el parámetro desconocido de la población, puede ser que la muestra extraída sea una de esas cinco. Este comentario es obligado decirlo, pero en la investigación se asume que la muestra lo contiene.

En el caso B, con la probabilidad de 0,9544 (*Nc* = 0,9544 ó 95,44%) la media de edad de la población española está comprendida en el intervalo (de confianza) de 43,89 años y 46,01 años, o lo que es lo mismo, que si se extraen 100 muestras de esa población, 95,44 tendrían en su intervalo el parámetro desconocido de la población.

Y en el caso C, para un *Nc* = 0,9974 el intervalo de confianza es entre 43,36 años y 46,54 años.

14.3 Intervalo de confianza para proporciones

Tabla 113 Estimación por intervalo de la proporción de la población.	
$P(p \pm 1{,}96\,\Delta S_p) \{ P(p \pm 1{,}96\,\Delta S_p) = 0{,}9500$	$p - 1{,}96\,\Delta S_p$ Límite inferior del intervalo de confianza.
	Nc = 95,00% ; *error típico* = S_p ; *error absoluto* = $1{,}96\,\Delta S_p$
	$p + 1{,}96\,\Delta S_p$ Límite superior del intervalo de confianza.
$P(p \pm 2{,}00\,\Delta S_p) \{ P(p \pm 2{,}00\,\Delta S_p) = 0{,}9544$	$p - 2{,}00\,\Delta S_p$ Límite inferior del intervalo de confianza.
	Nc = 95,44% ; *error típico* = S_p ; *error absoluto* = $2{,}00\,\Delta S_p$
	$p + 2{,}00\,\Delta S_p$ Límite superior del intervalo de confianza.
$P(p \pm 3{,}00\,\Delta S_p) \{ P(p \pm 3{,}00\,\Delta S_p) = 0{,}9974$	$p - 3{,}00\,\Delta S_p$ Límite inferior del intervalo de confianza.
	Nc = 99,74% ; *error típico* = S_p ; *error absoluto* = $3{,}00\,\Delta S_p$
	$p + 3{,}00\,\Delta S_p$ Límite superior del intervalo de confianza.

Aplicándolo al estudio de CIRES de enero de 1996, *Usos del tiempo*, para estimar la proporción o porcentaje de varones de la población española de ambos sexos y de 18 años o más en enero de 1996, se calcula la proporción de varones de la variable sexo y se procede,

Tabla 114 Estadísticos de la variable sexo.	
Encuesta CIRES: *Usos del tiempo.* Enero de 1996. Ámbito nacional. Población española de ambos sexos de 18 años o más	
p	48,2% (varones)
n_X	1.200
S_p	$\sqrt{\frac{p \cdot q}{n_X}}$
S_p (cpf)[76]	$\sqrt{\frac{p \cdot q}{n_X}} \cdot \sqrt{1 - fm}$ años
S_p	$\sqrt{\frac{48{,}2 \cdot 51{,}8}{1.200}} = 1{,}44$ años

Aplicándolo a este caso,

Tabla 115 Estimación por intervalo de la proporción de varones de la población española para distintos *Nc*.		
A	$P(0{,}48{,}24 - 1{,}96 \cdot 1{,}44 < P < 0{,}48{,}22 + 1{,}96 \cdot 1{,}44) = 0{,}9500$ $P(45{,}38 < P < 51{,}02) = 0{,}9500$	$48{,}2 - 1{,}96 \cdot 1{,}44 = 48{,}2 - 2{,}82 = 45{,}38$ Limite inferior
		Nc = 95,00% ; *error típico* = 1,44; *error absoluto* = 2,82
		$48{,}2 + 1{,}96 \cdot 1{,}44 = 48{,}2 + 2{,}82 = 51{,}02$ Limite superior
B	$P(0{,}48{,}24 - 2{,}00 \cdot 1{,}44 < P < 0{,}48{,}22 + 2{,}00 \cdot 1{,}44) = 0{,}9544$ $P(45{,}32 < P < 51{,}08) = 0{,}9544$	$48{,}2 - 2{,}00 \cdot 1{,}44 = 48{,}2 - 2{,}88 = 45{,}32$ Limite inferior
		Nc = 95,44% ; *error típico* = 1,44; *error absoluto* = 2,88
		$48{,}2 + 2{,}00 \cdot 1{,}44 = 48{,}2 + 2{,}88 = 51{,}08$ Limite superior
C	$P(0{,}48{,}24 - 3{,}00 \cdot 1{,}44 < P < 0{,}48{,}22 + 3{,}00 \cdot 1{,}44) = 0{,}9974$ $P(43{,}88 < P < 52{,}52) = 0{,}9974$	$48{,}2 - 3{,}00 \cdot 1{,}44 = 48{,}2 - 4{,}32 = 43{,}88$ Limite inferior
		Nc = 99,74% ; *error típico* = 1,44; *error absoluto* = 4,32
		$48{,}2 + 3{,}00 \cdot 1{,}44 = 48{,}2 + 4{,}32 = 52{,}52$ Limite superior

En la Tabla 112 el caso *A* significa que con la probabilidad de 0,95 (*Nc* = 0,95 ó 95,0 %), el porcentaje de varones en la población española está comprendido en el intervalo (de confianza) de 45,4% y 51,0%, o lo que es lo mismo, que si se extraen 100 muestras de esa población, 95 tendrían en su intervalo el parámetro desconocido de la población. Como se comentó anteriormente, al existir la probabilidad de 0,05 o 5,0 % de que la muestra no contenga el parámetro desconocido de la población, puede ser que la muestra extraída sea una de esas cinco. Este comentario es obligado decirlo, pero en la investigación se asume que la muestra lo contiene.

En el caso B, con la probabilidad de 0,9544 (*Nc* = 0,9544 ó 95,44% ó 95,5 %), el porcentaje de varones en la población española está comprendida en el intervalo (de confianza) de 45,3% y 51,1%, o lo que es lo mismo, que si se extraen 100 muestras de esa población, 95,44 tendrían en su intervalo el parámetro desconocido de la población.

Y en el caso C, para un *Nc* = 0,9974 el intervalo de confianza es entre 43,9% y 52,5%.

[76] El *cpf* (corrector por poblaciones finitas) se aplica cuando la población se considera finita (N < 100.000).

14.4 Técnicas de muestreo no probabilísticas

Las técnicas de muestreo *no probabilísticas* (Tabla 106) son: *intencional (muestreo útil y de casos típicos), accidental, bola de nieve* y *por cuotas.*

Estas son las técnicas de muestreo que se utilizan en el paradigma técnico cualitativo, y *por cuotas* también se utiliza como una de las etapas del muestreo probabilístico polietápico.

Estas muestras no se consideran representativas estadísticamente o numéricamente. Su representatividad es social, estructural o de características "La representatividad de estas muestras no radica en la cantidad de las mismas, sino en las posibles configuraciones subjetivas (valores-creencias-motivaciones) de los sujetos con respecto a un objeto o fenómeno determinado" (Serbia, 2007: 133).

En el muestreo *intencional* el investigador selecciona las unidades de observación en base a algún criterio como puede ser el *muestreo útil* (*purposive sampling*) (Gobo, 2004: 448), que consiste en seleccionar casos en situaciones extremas o dentro de un rango amplio de situaciones para maximizar la variación.[77] En el muestreo de *casos típicos* (Gobo, 2004: 449) se pueden seleccionar unidades teniendo en consideración tres características: que sea un caso considerado medio, que sea un caso destacado o un fenómeno emergente.[78]

En el muestreo *accidental*, las unidades de observación son seleccionadas sin atender a criterios, como puede ser el hecho de personas que circulan por un determinado lugar en un cierto momento. Ejemplos de este tipo son los estudios de mercado o de opinión que entrevistan a personas para recoger la opinión de cierto producto, líder político, publicación o acontecimiento.

El muestreo de *bola de nieve*, es útil para contactar con personas que por sus características son de difícil acceso como puede ser: inmigrantes ilegales, grupos considerados marginales, etc.[79]

En el muestreo por *cuotas,* la población se divide en subgrupos en base a algún criterio de interés para el estudio y se establece la proporción de los individuos de la población que hay en cada subgrupo. La pretensión es que en la muestra existan estos mismos subgrupos y en la misma proporción a los grupos de la población. Es un concepto similar a los *estratos*, que se verá posteriormente y a la *afijación proporcional* o *reparto proporcional.* Se puede utilizar uno o más criterios para establecer las *cuotas*. Algunos ejemplos son: establecer cuotas en base a la edad y el sexo; edad, sexo y status socioeconómico; edad, sexo, status socioeconómico y estudios; etc. En el *muestreo aleatorio estratificado* se verá un ejemplo. Puede acontecer que se usen diversas técnicas de muestreo de forma conjunta.

14.5 Técnicas de muestreo probabilísticas

Las técnicas de muestreo *probabilísticas* (Tabla 106) son: *muestreo aleatorio simple*; *muestreo aleatorio sistemático*; *muestreo aleatorio estratificado*, y *muestreo por conglomerados.*

---

[77] Para ver ejemplos de este tipo de muestreo confrontar: A. Davis (1941); St. C. Drake (1945); L. W. Warner (1949). Citados en G. Gobo (2004: 449). Un ejemplos más reciente es J. W. Harris (2001).

[78] Ejemplos de este tipo de muestreo son: R. S. Lynd and H. M. Lynd (1937); A. G. Gouldner (1954); M. Dalton (1959); R. M. Kanter (1977). Citados en G. Gobo (2004: 449).

[79] W. F. Whyte (1943); W. D. TenHouten (1971). Pueden ser ejemplos. Citados en G. Gobo (2004: 449).

MUESTREO ALEATORIO SIMPLE

El *muestreo aleatorio simple* consiste en extraer un conjunto de *n* individuos que llamamos muestra a partir de un conjunto más grande *N* de individuos, que es la población. Para que la muestra se pueda considerar representativa y poder inferir los resultados a la población, los individuos o unidades deben ser extraídos por cualquier procedimiento que suponga aleatoriedad y además hacerlo de tal manera que se pueda considerar que todos han tenido la misma probabilidad de ser seleccionados. Todos los procesos de muestreo y cálculos se realizan en base a esta consideración. Los procedimientos de extracción pueden ser: tablas de números aleatorios, hojas de cálculo o programas estadísticos.

Para utilizar este procedimiento de muestreo es necesario tener el listado de toda la población. Si se trata de personas, no es posible disponer de él por la Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal. Este tipo de muestreo es adecuado y útil y por lo tanto se aplicará para seleccionar: empleados en una organización, productos, Secciones Censales, calles, portales, plantas, puertas, organizaciones, municipios, etc. En general elementos que no estén afectados por la mencionada Ley Orgánica o que la organización que solicita la muestra tenga acceso al listado de empleados. El anonimato de la población se puede mantener disponiendo únicamente de un código asignado a cada unidad de observación que posteriormente permita acceder a la persona, por medio del propietario de los datos, sin conocer la identidad por parte del investigador o investigadora.

Puede ser el caso de una Administración que dispone de los datos de todos los ciudadanos. Una empresa privada puede tener acceso a un código asignado a cada ciudadano y la Administración disponer del enlace que relaciona el código con las personas. Se puede considerar como acceso a las personas *por direccionamiento* sin quebrantar la Ley Orgánica.

En la Tabla 116 se muestra un ejemplo de extracción de unidades de observación de un censo por muestreo aleatorio simple.

Tabla 116 Ejemplo de Muestreo Aleatorio Simple.

Suponiendo una población de 100 individuos numerados del 1 al 100, consiste en extraer una muestra de 20, mediante un procedimiento aleatorio.

Utilizamos la Tabla de Números Aleatorios uniformemente distribuidos del Anexo 5.

Caso 1:

Como la Tabla de Números Aleatorios son números de dos cifras y la población llega hasta las tres cifras (100), necesitamos coger números de tres cifras, por lo que ignoramos las columnas existentes y las definimos de tres en tres cifras. Por ejemplo, el número de la primera celda arriba a la izquierda no sería el 43 sino el 436.

Este procedimiento lleva a ignorar todos los números que están fuera del rango de la población, como por ejemplo el 436.

Caso 2:

Para evitar este inconveniente, en vez de definir la población del 1 al 100, se realiza un desplazamiento hacia atrás y se numera del 0 al 99, operando con las columnas de números de dos cifras. Pero si extraemos el cero, no está en la población y el 100 nunca entraría en la muestra por lo que se altera la regla de *igual probabilidad*. Entonces una vez extraídos los números de la Tabla de Números Aleatorios, se le suma una unidad y se elimina el retraso anterior.

Utilizando el Caso 2, sólo falta decidir en que parte de la tabla de números aleatorios se empieza a hacer la selección. Por ejemplo, por arriba a la izquierda y se procederá de arriba abajo y de izquierda a derecha. Los casos seleccionados se marcan en la columna *Muestra* sumando una unidad (eliminando el retraso anterior). Si se repite algún número (caso del 30+1, 51+1, 80+1, 53+1) se salta al siguiente. Los números seleccionados son los marcados en la columna de *Muestra.*

Este ejemplo se hace por criterio didáctico y pedagógico. El mejor procedimiento es un programa estadístico.

Población	Muestra	Población	Muestra	Población	Muestra	Población	Muestra	Población	Muestra
1		21		41		61	60+1=61	81	80+1=81
2		22		42	41+1=42	62		82	81+1=82
3		23		43		63	62+1=63	83	
4		24	23+1=24	44	43+1=44	64		84	
5		25		45		65		85	
6		26		46		66		86	
7		27	26+1=27	47		67	66+1=67	87	
8		28		48		68	67+1=68	88	
9		29	28+1=29	49		69		89	
10		30		50		70		90	
11	10+1=11	31	30+1=31	51		71		91	
12		32		52	51+1=52	72	71+1=72	92	
13		33	32+1=33	53		73		93	92+1=93
14		34		54	53+1=54	74		94	
15		35		55		75		95	
16		36		56		76		96	
17		37		57	56+1=57	77		97	
18		38		58		78		98	97+1=98
19		39		59		79		99	
20		40		60		80		100	

MUESTREO ALEATORIO SISTEMÁTICO

El *muestreo aleatorio sistemático* es una derivación del anterior. También tiene el inconveniente de que se debe conocer el listado de la población. Para extraer los *n* individuos de la muestra a partir de los *N* individuos de la población, primero se obtiene el *ce* (coeficiente de elevación), se elige de forma aleatoria un número entre 1 y el *ce,* y al número obtenido se le suma de forma sucesiva el *ce* hasta completar la muestra. Ver el ejemplo de la Tabla 117.

Tabla 117 Ejemplo de Muestreo Aleatorio Sistemático.

Suponiendo una población de 100 individuos numerados del 1 al 100, consiste en extraer una muestra de 20, mediante un procedimiento aleatorio.

1º Se calcula el *ce*. $ce = \frac{N}{n} = \frac{100}{20} = 5$

2º Se extrae un número entre 1 y 5 aleatoriamente utilizando la Tabla de Números Aleatorios del Anexo 5.

3º Por ejemplo, el primero de la primera celda arriba a la izquierda: 4.

4º Sumamos sucesivamente 5 a cada número obtenido. Los números seleccionados son los marcados en la columna de *Muestra.*

Este ejemplo se hace por criterio didáctico y pedagógico. El mejor procedimiento es un programa estadístico.

Población	Muestra	Población	Muestra	Población	Muestra	Población	Muestra	Población	Muestra
1		21		41		61		81	
2		22		42		62		82	
3		23		43		63		83	
4	4+5=9	24	24+5=29	44	44+5=49	64	64+5=69	84	84+5=89
5		25		45		65		85	
6		26		46		66		86	
7		27		47		67		87	
8		28		48		68		88	
9	9+5=14	29	29+5=34	49	49+5=54	69	69+5=74	89	89+5=94
10		30		50		70		90	
11		31		51		71		91	
12		32		52		72		92	
13		33		53		73		93	
14	14+5=19	34	34+5=39	54	54+5=59	74	74+5=79	94	94+5=99
15		35		55		75		95	
16		36		56		76		96	
17		37		57		77		97	
18		38		58		78		98	
19	19+5=24	39	39+5=44	59	59+5=64	79	79+5=84	99	99
20		40		60		80		100	

MUESTREO ALEATORIO ESTRATIFICADO

El *muestreo aleatorio simple* y el *sistemático* garantizan la aleatoriedad del proceso, pero no garantiza la selección de individuos de grupos pequeños. Si la muestra debe ser heterogénea como la población, es necesario que incorpore también a los individuos considerados extremos, como pueden ser los de *clase social muy alta*. Si es necesario incorporar a unidades que cumplan requisitos en base a algún criterio, se estratifica la población en base a ese criterio y se procede de la misma manera con la muestra. Los estratos tienen la característica de que los individuos son homogéneos dentro de ellos pero

heterogéneos entre los estratos.

El proceso consiste en distribuir los *n* elementos de la muestra entre los estratos de la población, y después utilizar algún procedimiento para seleccionar a los individuos. Se va a considerar tres tipos de reparto o distribución que el nombre técnico asignado es *afijación*: *afijación no proporcional, afijación proporcional* y *afijación mixta*. Otro tipo de afijación es la *óptima* (Cea, 2004: 136-137). El procedimiento se realiza con el ejemplo de la Tabla 118, Tabla 119, Tabla 120, Tabla 121, Tabla 122, Tabla 123 y Tabla 124 que además se utiliza para introducir el concepto de *afijación* y *ponderación*.

Tabla 118 Muestreo aleatorio estratificado, afijación y ponderación

Supuesta una población de 10.000 individuos de la que interesa que en la muestra estén representados todos según el criterio de *estatus socio-económico*. Se asume que en la población existen los estratos. *Muy alto, Alto, Medio, Bajo y Muy bajo.* Se asume una muestra de *n* = 1.000. Se procede a la *afijación* o asignación de los individuos de la muestra a los estratos de la población de forma *no proporcional, proporcional* y *mixta*. El ejemplo que se muestra pretende facilitar la introducción de los conceptos de *afijación* y *ponderación*.

					Afijación Mixta					Ponderación	
Estratos de la Población	*N*	Afijación no Proporcional	*fm*	Afijación Proporcional	Asignación Directa	*fm*	Afijación Proporcional	Total	Total 2	Alfa	Total Ponderado
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Muy alto (n_1)	10	200	0,10	1	5	0,0975	0,98	5,98	6	0,167	1
Alto (n_2)	1.000	200	0,10	100	5	0,0975	97,50	102,50	102	0,980	100
Medio (n_3)	6.000	200	0,10	600	5	0,0975	585,00	590,00	590	1,017	600
Bajo (n_4)	2.000	200	0,10	200	5	0,0975	195,00	200,00	200	1,000	200
Muy bajo (n_5)	990	200	0,10	99	5	0,0975	96,53	101,53	102	0,971	99
Población (N)	10.000	1.000		1.000	25		975,00	1.000,00	1.000		1.000
Muestra (n)	1.000				975						

Tabla 119 Afijación no proporcional. Columna (1).

Consiste en dividir los individuos de la muestra entre el número de estratos definidos en la población,

$n_i \mid \frac{n}{5} \mid \frac{1.000}{5} \mid 200$, en cada estrato definido por la población, la muestra tendrá 200 individuos.

El inconveniente de este procedimiento es que se da el mismo peso a todos los estratos de la población y puede ocurrir que haya más individuos en alguno de los estratos de la muestra que en el de la población. Como es el caso del estrato *Muy alto*.

Tabla 120 Afijación proporcional. Columnas (2) (3).

Se reparte proporcionalmente los individuos de la muestra al tamaño de los estratos de la población mediante una regla de tres simple, que después resulta ser la multiplicación de la *fracción de muestreo* por el tamaño de cada estrato de la población.

$N \Downarrow n$ $N_i \Downarrow x_i$	$10.000 \Downarrow 1.000$ $\_\_\_10 \Downarrow x_1$	$10.000 \Downarrow 1.000$ $\_1.000 \Downarrow x_2$	$10.000 \Downarrow 1.000$ $\_6.000 \Downarrow x_3$	$10.000 \Downarrow 1.000$ $\_2.000 \Downarrow x_4$	$10.000 \Downarrow 1.000$ $\_\_990 \Downarrow x_5$
$x_i \mid N_i \Delta \frac{n}{N}$	$x_1 \mid 10 \Delta \frac{1.000}{10.000}$	$x_2 \mid 1.000 \Delta \frac{1.000}{10.000}$	$x_3 \mid 6.000 \Delta \frac{1.000}{10.000}$	$x_4 \mid 2.000 \Delta \frac{1.000}{10.000}$	$x_5 \mid 990 \Delta \frac{1.000}{10.000}$
$x_i \mid N_i \Delta fm$	$x_1 \mid 10 \Delta 0{,}10$	$x_2 \mid 1.000 \Delta 0{,}10$	$x_3 \mid 6.000 \Delta 0{,}10$	$x_4 \mid 2.000 \Delta 0{,}10$	$x_{51} \mid 990 \Delta 0{,}10$
	$x_1 \mid 1$	$x_2 \mid 100$	$x_3 \mid 600$	$x_4 \mid 200$	$x_5 \mid 99$

El inconveniente de este sistema es que en los estratos de la población con pocos individuos, puede ocurrir que en la muestra sólo les corresponda uno o pocos. Si son personas y la seleccionada es *varón*, ¿Quiere decir que todos los de ese estrato son varones? Si fuese mujer ¿Querría decir que todos son mujeres? Si fuesen empresas y la seleccionada fuese de *alta tecnología* ¿Quiere decir que todas las empresas de ese estrato son de *alta tecnología*?

Tabla 121 Afijación mixta. Columnas (4) (5) (6) (7) (8).

Cuando en uno de los estratos de la población, a la muestra le corresponden pocos individuos, se puede proceder por un sistema de *afijación mixto* que consistiría en hacer una *asignación directa* de parte de la muestra y el resto por *afijación proporcional*.

En este caso la *asignación directa* puede ser de 5 individuos. Columna (4).

Del total de 1.000 individuos de la muestra quedan 975 (1.000 – 25) por distribuir, con los que se procede por *afijación proporcional*.

La columna (7) es el total de los casos o unidades de observación que hay en cada estrato de la muestra, que es la suma de las columnas (4) y (6). Cuando el resultado de las operaciones no son números enteros, hay que proceder al redondeo de tal manera que la suma total coincida con la *n* de la muestra. Columna (8).

$N \Downarrow n$ $N_i \Downarrow x_i$	$10.000 \Downarrow 975$ $\_\_\_10 \Downarrow x_1$	$10.000 \Downarrow 975$ $\_1.000 \Downarrow x_2$	$10.000 \Downarrow 975$ $\_6.000 \Downarrow x_3$	$10.000 \Downarrow 975$ $\_2.000 \Downarrow x_4$	$10.000 \Downarrow 975$ $\_\_990 \Downarrow x_5$
$x_i \mid N_i \Delta \frac{n}{N}$	$x_1 \mid 10 \Delta \frac{975}{10.000}$	$x_1 \mid 1.000 \Delta \frac{975}{10.000}$	$x_1 \mid 6.000 \Delta \frac{975}{10.000}$	$x_1 \mid 2.000 \Delta \frac{975}{10.000}$	$x_1 \mid 990 \Delta \frac{975}{10.000}$
$x_i \mid N_i \Delta fm$	$x_1 \mid 10 \Delta 0{,}0975$	$x_2 \mid 1.000 \Delta 0{,}0975$	$x_3 \mid 6.000 \Delta 0{,}0975$	$x_4 \mid 2.000 \Delta 0{,}0975$	$x_{51} \mid 990 \Delta 0{,}0975$
	$x_1 \mid 0{,}98$	$x_2 \mid 97{,}50$	$x_3 \mid 585$	$x_4 \mid 195$	$x_5 \mid 96{,}53$

El inconveniente de este sistema es que algunos estratos van a estar sobredimensionados, como es el estrato de estatus socio-económico *Muy alto*, que le correspondía un caso y se van a seleccionar a 6. Otros estratos pueden quedar infradimensionados, como puede ser *Medio* que de 600, ha pasado a 590.

El siguiente paso es equilibrar el número de individuos que tiene que haber en cada estrato de la muestra. Este paso se denomina *ponderación*.

Tabla 122 Ponderación de la muestra. Columnas (9) (10).

Ponderar es asignar a cada estrato de la muestra el número de unidades que le corresponden según la *afijación proporcional* a los estratos de la población.

Una vez realizado el trabajo de campo, debido a variaciones de *campo,* por el procedimiento de *afijación* utilizado u otras causas, puede ocurrir que la n_i^e (n_i empírica) obtenida difiera de la n_i^t (n_i teórica) que es la que debería haber sido según la *afijación proporcional*. Entonces si $n_i^e \prod n_i^t$ debemos proceder a igualarlas. El hecho de igualar la n_i^e a la n_i^t se llama ponderar, equilibrar, reequilibrar, etc. O lo que es lo mismo, darle a cada individuo de la muestra el peso que le corresponde o que representa de la población.

La ponderación consiste en hacer $n_i^e \mid n_i^t$ [Columna (8) = Columna (3)] multiplicando a la n_i^e por un valor ζ_i de tal manera que $n_i^e \Delta \zeta_i \mid n_i^t$, por lo tanto, $\zeta_i \mid \frac{n_i^t}{n_i^e}$ y obtenemos la columna (9).

$n_i^e \Delta \zeta_i \mid n_i^t$	$n_1^e \Delta \zeta_1 \mid n_1^t$	$n_2^e \Delta \zeta_2 \mid n_2^t$	$n_3^e \Delta \zeta_3 \mid n_3^t$	$n_4^e \Delta \zeta_4 \mid n_4^t$	$n_5^e \Delta \zeta_5 \mid n_5^t$
$\zeta_i \mid \frac{n_i^t}{n_i^e}$	$\zeta_1 \mid \frac{n_1^t}{n_1^e}$	$\zeta_2 \mid \frac{n_2^t}{n_2^e}$	$\zeta_3 \mid \frac{n_3^t}{n_3^e}$	$\zeta_4 \mid \frac{n_4^t}{n_4^e}$	$\zeta_5 \mid \frac{n_5^t}{n_5^e}$
	$\zeta_1 \mid \frac{1}{6} \mid 0{,}167$	$\zeta_2 \mid \frac{100}{102} \mid 0{,}980$	$\zeta_3 \mid \frac{600}{590} 1{,}017$	$\zeta_4 \mid \frac{200}{200} \mid 1{,}000$	$\zeta_5 \mid \frac{99}{102} \mid 0{,}971$

Los coeficientes de ponderación se muestran en la Columna (9). Para devolver a la muestra el peso que le corresponde en cada estrato, se multiplica la n_i^e (columna 8) por el coeficiente de ponderación y se obtiene la Columna (10).

Tabla 123 Efecto del coeficiente de ponderación. Columna (10).	
En el estrato *Muy alto* $n_i^t \mid 1$ y la $n_i^e \mid 6$ para hacer que los seis individuos valgan por uno, los multiplicamos por 0,167, $6 \Delta 0{,}167 \mid 1$, que por la propiedad distributiva, $6 \Delta 0{,}167$ es igual que,	
$6 \Delta 0{,}167 \mid 0{,}167 \Delta / 1\,2\,1\,2\,1\,2\,1\,2\,1\,2\,10 \mid (1 \Delta 0{,}167)\,2\,(1 \Delta 0{,}167)\,2\,(1 \Delta 0{,}167)\,2\,(1 \Delta 0{,}167)\,2\,(1 \Delta 0{,}167)\,2\,(1 \Delta 0{,}167)$	
Las seis unidades están multiplicadas por 0,167 o lo que es lo mismo, cada individuo vale por 0,167 de individuo.	
Efecto de la aplicación de la ponderación.	
Asumiendo que de los 10 individuos de la población de estatus socio-económico *Muy alto*, 6 son varones y 4 mujeres ocurre que el 60% son varones y el 40% mujeres.	
Si se opera con un solo individuo que define la muestra teórica, entonces el 100% son varones o mujeres. Si se opera con seis individuos de la muestra sin ponderar, el estrato estaría sobre representado, pero al aplicar la ponderación y cada individuo valer por 0,167, si asumimos que, por cuotas, de los seis individuos de la muestra, cuatro son varones y dos mujeres, entonces,	
$1(v) \Delta 0{,}167 \mid 0{,}167$ de varón	(0,167+0,167+0,167+0,167=0,668) 66,8% (varones)
$1(v) \Delta 0{,}167 \mid 0{,}167$ de varón	
$1(v) \Delta 0{,}167 \mid 0{,}167$ de varón	
$1(v) \Delta 0{,}167 \mid 0{,}167$ de varón	
$1(m) \Delta 0{,}167 \mid 0{,}167$ de mujer	(0,167+0,167=0,334) 33,4% (mujeres)
$1(m) \Delta 0{,}167 \mid 0{,}167$ de mujer	
El resultado obtenido después de la ponderación se acerca más a la realidad que el 100% varones o el 100% mujeres. El coeficiente de ponderación se aplica a todas las variables en el proceso de tabulación o análisis estadístico.	

Tabla 124 Efecto del coeficiente de ponderación en los ingresos.	
Asumiendo que conocemos los ingresos de las seis personas y que son los que se muestran entre paréntesis, el efecto del coeficiente de ponderación es,	
$1 \Delta (1.000{,}00€) \Delta 0{,}167 \mid 167{,}00€$	Ingresos totales sin ponderar = 2.220,00 € Ingresos totales ponderados = 370,74 € Si no se pondera se produce sesgo.
$1 \Delta (100{,}00€) \Delta 0{,}167 \mid 16{,}70€$	
$1 \Delta (10{,}00€) \Delta 0{,}167 \mid 1{,}67€$	
$1 \Delta (1.000{,}00€) \Delta 0{,}167 \mid 167{,}00€$	
$1 \Delta (100{,}00€) \Delta 0{,}167 \mid 16{,}70€$	
$1 \Delta (10{,}00€) \Delta 0{,}167 \mid 1{,}67€$	

En la Tabla 125, Tabla 126, Tabla 127, Tabla 128, Tabla 129, Tabla 130, Tabla 131 y Tabla 132 se muestra el cálculo del coeficiente de ponderación del Estudio de CIRES de enero de 1996 (ver epígrafe 12.5.1).

Tabla 125 Tabla de la población española por sexo y edad de 18 años o más.					
	18 a 29	30 a 49	50 a 64	más de 64	Total
Varón	4.114.621	5.279.137	3.234.765	2.329.031	14.957.554
Mujer	3.985.231	5.279.137	3.467.668	3.364.156	16.096.192
Total	8.099.852	10.558.274	6.702.433	5.693.187	31.053.746
Fuente: Datos estimados a partir del total de población del CIS para varones y mujeres de 18 años y más (enero de 1991).					

$$fm \mid \frac{n}{N} \mid \frac{1.200}{31.053.746} \mid 0{,}00003864$$

Tabla 126 Fórmulas de la Tabla 127.				
	18 a 29	30 a 49	50 a 64	más de 64
Varón	$n_{v,18a29} \mid 4.114.621 \Delta fm$	$n_{v,30a49} \mid 5.279.137 \Delta fm$	$n_{v,50a64} \mid 3.234.765 \Delta fm$	$n_{v,más64} \mid 2.329.031 \Delta fm$
Mujer	$n_{m,18a29} \mid 3.985.231 \Delta fm$	$n_{m,30a49} \mid 5.279.137 \Delta fm$	$n_{m,50a64} \mid 3.467.668 \Delta fm$	$n_{m,más64} \mid 3.364.156 \Delta fm$
Fuente: Calculado a partir de la Tabla 125. Produce la Tabla 127. Se eliminan los decimales por redondeo.				

Tabla 127 Tabla por sexo y edad. Afijación proporcional (Teórico).					
	18 a 29	30 a 49	50 a 64	más de 64	Total
Varón	159	204	125	90	578
Mujer	154	204	134	130	622
Total	313	408	259	220	1.200
Fuente: Calculado a partir de la Tabla 125 y Tabla 126.					

Tabla 128 Tabla por sexo y edad a enero de 1996 (CIRES) (Empírico).					
	18 a 29	30 a 49	50 a 64	más de 64	Total
Varón	161	201	124	88	574
Mujer	152	207	138	129	626
Total	313	408	262	217	1200
Fuente: Encuesta "Usos del tiempo". Enero de 1996.					

$$\zeta_{ij} \mid \frac{n_{ij}^{t}}{n_{ij}^{e}}$$

Tabla 129 Fórmulas de la Tabla 130.				
	18 a 29	30 a 49	50 a 64	más de 64
Varón	$\zeta_{v,18a29} \mid \frac{159}{161}$	$\zeta_{v,30a49} \mid \frac{204}{201}$	$\zeta_{v,50a64} \mid \frac{125}{124}$	$\zeta_{v,más64} \mid \frac{90}{88}$
Mujer	$\zeta_{m,18a29} \mid \frac{154}{152}$	$\zeta_{m,30a49} \mid \frac{204}{207}$	$\zeta_{m,50a64} \mid \frac{134}{138}$	$\zeta_{m,más64} \mid \frac{130}{129}$
Fuente: Calculado a partir de la Tabla 127 y Tabla 128.				

Tabla 130 Coeficientes de ponderación.					
	18 a 29	30 a 49	50 a 64	más de 64	
Varón	0,987578	1,014925	1,008064	1,022727	
Mujer	1,013158	0,985507	0,971014	1,007752	
Fuente: Calculado a partir de la Tabla 127, Tabla 128 y Tabla 129.					

$$n_{ij}^{p} \mid n_{ij}^{t} \mid n_{ij}^{e} \Delta\zeta_{ij}$$

El tamaño de la muestra ponderado, en el estrato (n_{ij}^{p}) es igual al tamaño de la muestra teórico, del estrato (n_{ij}^{t}) y este a su vez es igual al tamaño de la muestra empírico, del estrato (n_{ij}^{e}) por el coeficiente de ponderación, del estrato (ζ_{ij}).

Tabla 131 Fórmulas de la Tabla 132.				
	18 a 29	30 a 49	50 a 64	más de 64
Varón	$161\Delta 0{,}987578$	$201\Delta 1{,}014925$	$124\Delta 1{,}008064$	$88\Delta 1{,}022727$
Mujer	$152\Delta 1{,}013158$	$207\Delta 0{,}985507$	$138\Delta 0{,}971014$	$129\Delta 1{,}007752$
Fuente: Calculado a partir de la Tabla 128 y Tabla 130.				

Tabla 132 Tabla por sexo y edad. Ponderada.					
	18 a 29	30 a 49	50 a 64	más de 64	Total
Varón	159	204	125	90	578
Mujer	154	204	134	130	622
	313	408	259	220	1200
Fuente: Calculado a partir de la Tabla 128, Tabla 130 y Tabla 131.					

MUESTREO POR CONGLOMERADOS

El muestreo por *conglomerados* se fundamenta en que hay que acceder a poblaciones que en sociología normalmente son grandes y dispersas y a veces de difícil acceso. En el muestreo por *conglomerados* se considera que la población es heterogénea y que puede ser dividida en grupos o conjuntos más pequeños, geográficamente reducidos, que son heterogéneos como la población y homogéneos entre ellos y por lo tanto se considera que cada uno de ellos puede representar a la población. A estos grupos o conjuntos se les denomina *conglomerados*. El conglomerado es la *unidad muestral* en esta técnica de muestreo y está formado por varias unidades de observación. La ventaja que se obtiene es que si todos los *conglomerados* se consideran representativos de la población, seleccionando uno pequeño y entrevistando a todas las unidades, se pueden inferir los resultados sobre la población con un coste bajo y relativamente poco esfuerzo.

El planteamiento realizado es una definición de principios, pero todos los conglomerados no son igualmente representativos del total de la población. El muestreo por conglomerados requiere la aplicación de un muestreo en varias etapas (polietápico) para seleccionar varios conglomerados que representen a los diferentes grupos de la población.

En el caso de la población española, se puede considerar que el *conglomerado* es el municipio[80] de tal manera que cada municipio es heterogéneo como toda la población y homogéneos todos los municipios entre sí, pero al mismo tiempo, los municipios tienen diferentes tamaños y pertenecen a diferentes Comunidades Autónomas. Para favorecer la representatividad de la muestra se procede a seleccionar los municipios aplicando un criterio de estratificación doble por Comunidad Autónoma y tamaño de hábitat. Después se procede a seleccionar los municipios por muestreo aleatorio simple y a continuación a las unidades de observación por rutas aleatorias. De esta manera se ha introducido el muestreo por etapas o polietápico que se vera posteriormente en detalle.

Un municipio de Andalucía de menos de 2.000 habitantes puede ser dudoso que represente a toda la población española, pero se puede considerar que sea representativo de los municipios de Andalucía de menos de 2.000 habitantes. Lo mismo se puede decir de un municipio de Galicia de menos de 2.000 habitantes, se le puede considerar representativo de la población gallega de menos de 2.000 habitantes, pero no será de otras comunidades o incluso de la misma pero de distinto tamaño poblacional.

Se pueden considerar conglomerados a cualquier grupo o subconjunto de la población que cumpla los requisitos anteriores. El municipio es el conglomerado que se utiliza habitualmente con las poblaciones de personas. En este tipo de muestreo se debe definir cual es el conglomerado. Otros ejemplos de conglomerados pueden ser: en un centro de estudios el grupo de alumnos o aula; en un hospital la planta, o la especialidad; en una línea de transporte público el autobús o el convoy de metro. En la Tabla 133 se muestra las diferencias entre las técnicas de muestreo probabilístico.

Tabla 133 Relación entre las técnicas de muestreo probabilísticas.

	Muestreo aleatorio simple	Muestreo aleatorio sistemático	Muestreo aleatorio estratificado	Muestreo por conglomerados
Unidad muestral	Unidad de observación	Unidad de observación	Unidad de observación	El conglomerado (contiene varias unidades de observación).
Heterogeneidad			Entre estratos	Dentro del conglomerado
Homogeneidad			Dentro del estrato	Entre conglomerados

[80] División geográfica española formada por un conjunto de habitantes de un mismo término jurisdiccional, regido por un ayuntamiento (Real Academia Española, 2008).

14.6 Extracción de una muestra

Para extraer una muestra se define la población objetivo, la unidad de observación y el ámbito o delimitación geográfica de la misma. Los datos que definen las características de la muestra se especifican en la Ficha Técnica de la Encuesta. Un modelo de ficha técnica que utiliza el Centro de Investigaciones Sociológica (CIS) se muestra en la Tabla 134. Posteriormente se tratará el cálculo del tamaño de la muestra. Los *comentarios* son indicaciones aclaratorias del autor y no aparecen en la Ficha Técnica.

Tabla 134 Ficha Técnica. Estudio CIS nº 2769. Barómetro de julio.

Ámbito: Nacional.

Comentario: Define los límites geográficos, que normalmente son administrativos o políticos, del territorio que comprende a la Población objeto de estudio.

Universo: Población española de ambos sexos de 18 años y más.

Comentario: Define las características que delimitan a las unidades de observación que van a ser objeto de estudio.

Tamaño de la muestra:
Diseñada: 2.500 entrevistas.
Realizada: 2.468 entrevistas.

Comentario: Es el tamaño de la muestra. A veces se puede diferenciar entre la muestra *diseñada*, que es la que se calcula siguiendo los procedimientos que se indican más adelante y la *realizada* que se indica cuando por diferentes motivos no se han podido realizar todas las entrevistas diseñadas o que al volver del *trabajo de campo* se han invalidado algunas. Se debe tener en consideración que cada unidad de la muestra representa a un número de individuos de la población igual al *ce* (coeficiente de elevación).

Afijación: Proporcional.

Comentario: Reparto o distribución de las unidades de la muestra entre los estratos de la población, en este caso de forma proporcional.

Ponderación: No procede.

Comentario: Cuando el tamaño de los estratos de la muestra no se corresponde proporcionalmente con el tamaño de los estratos de la población, se procede a ponderar. Cuando se consideran iguales, no es necesario ponderar.

Puntos de Muestreo: 238 municipios y 48 provincias.

Comentario: Los puntos de muestreo (municipios) se obtienen al distribuir los cuestionarios dentro de los estratos según el criterio que se indica más adelante.

Procedimiento de muestreo:

Polietápico, estratificado por conglomerados, con selección de las unidades primarias de muestreo (municipios) y de las unidades secundarias (secciones) de forma aleatoria proporcional, y de las unidades últimas (individuos) por rutas aleatorias y cuotas de sexo y edad. Los estratos se han formado por el cruce de las 17 comunidades autónomas con el tamaño de hábitat, dividido en 7 categorías: menor o igual a 2.000 habitantes; de 2.001 a 10.000; de 10.001 a 50.000; de 50.001 a 100.000; de 100.001 a 400.000; de 400.001 a 1.000.000, y más de 1.000.000 de habitantes.

Los cuestionarios se han aplicado mediante entrevista personal en los domicilios.

Error muestral:

Para un nivel de confianza del 95,5% (dos sigmas), y P = Q, el error real es de ±2,0% para el conjunto de la muestra y en el supuesto de muestreo aleatorio simple.

Fecha de realización:

Del 7 al 13 de julio de 2008.

El Universo o Población al que quiere representar la muestra no es el Censo, ya que se delimita a los españoles de ambos sexos de 18 años o más.

La muestra diseñada o calculada es de 2.500 unidades de observación, pero las realizadas son 2.468.

La *afijación* o reparto de la muestra a los estratos de la población es *proporcional*, y no se han debido producir variaciones o variaciones significativas porque no se ha aplicado *ponderación*.

Los *puntos de muestreo* o municipios en los que se han realizado las entrevistas son 238 situados en 48 provincias. Posteriormente se harán indicaciones del proceso de selección de los puntos.

El procedimiento de muestreo es *polietápico* porque se ha diseñado en varias etapas o aplicando varias técnicas de muestreo. Primero se define el *conglomerado* que es el *municipio*. Son las unidades muestrales que se consideran heterogéneos como la población y homogéneos entre ellos y representativos cada uno de ellos de la población (primera etapa). Para favorecer la representatividad de los conglomerados, se procede a estratificarlos (segunda etapa) en base a dos criterios: Comunidad Autónoma y tamaño de hábitat. Los municipios de cada celda se considera que son representativos de ese conjunto de población. Así, como ya se dijo anteriormente, los municipios de Andalucía de 2.000 o menos habitantes son representativos de la población de los municipios de Andalucía de 2.000 o menos habitantes y así sucesivamente para todas las celdas.

La selección de los municipios de cada celda se hace por procedimientos de *muestreo aleatorio simple o sistemático* (tercera etapa). En esta ocasión y como se dispone de las *secciones censales* se utilizan también y se extraen por *muestreo aleatorio simple o sistemático* (se puede considerar dentro de la tercera o define la cuarta etapa).

Los individuos o unidades de observación se extraen o se seleccionan siguiendo *rutas aleatorias* (cuarta o quinta etapa) y se utilizan cuotas de edad y sexo (quinta o sexta y última etapa).

Este proceso persigue la mejor representatividad posible de la muestra sobre la población objetivo.

El proceso para la selección de los elementos de la muestra se sigue en la Tabla 135, Tabla 136, Tabla 137, Tabla 138 y Tabla 139.

Tabla 135 Número de municipios españoles según la Comunidad y el tamaño de hábitat.

Comunidad	Tamaño de hábitat							
	<= 2.000	2.001 a 10.000	10.001 a 50.000	50.001 a 100.000	100.001 a 400.000	400.001 a 1.000.000	>1.000.000	Total
Andalucía	320	316	111	11	10	2		770
Aragón	680	38	11			1		730
Asturias	28	29	18	1	2			78
Baleares	18	32	16		1			67
Canarias	8	43	32	1	3			87
Cantabria	61	31	8	1	1			102
Castilla-León	2.126	99	15	4	5			2.249
Castilla-La Mancha	752	139	23	4	1			919
Cataluña	645	203	77	12	8		1	946
Valenciana	325	130	74	8	3	1		541
Extremadura	278	92	10	2	1			383
Galicia	80	179	49	4	3			315
Madrid	88	53	23	8	6		1	179
Murcia	6	13	23	1	2			45
Navarra	220	45	6		1			272
País Vasco	152	58	33	4	3			250
La Rioja	156	15	2		1			174
Total	5.943	1.515	531	61	51	4	2	8.107

Fuente: INE. Censo de población de 2001. Resultados definitivos a 17 de febrero de 2004.

Tabla 136 Población española de 18 años o más por Comunidad y tamaño de hábitat.

Comunidad	Tamaño de hábitat							
	<= 2.000	2.001 a 10.000	10.001 a 50.000	50.001 a 100.000	100.001 a 400.000	400.001 a 1.000.000	>1.000.000	Total
Andalucía	238.497	1.099.686	1.682.446	518.031	1.259.304	973.176		5.771.140
Aragón	213.683	120.730	164.702			515.522		1.014.637
Asturias	25.165	108.684	309.761	71.388	405.275			920.273
Baleares	16.455	128.052	264.852		271.954			681.313
Canarias	9.603	187.604	540.907	67.430	541.788			1.347.332
Cantabria	51.497	100.935	98.268	47.074	153.381			451.155
Castilla-León	641.486	314.570	259.360	216.135	819.716			2.251.267
Castilla-La Mancha	290.823	435.274	352.796	218.021	116.775			1.413.688
Cataluña	324.550	739.215	1.233.832	641.631	1.035.055		1.288.813	5.263.096
Valenciana	195.444	498.516	1.214.488	365.368	507.565	619.906		3.401.287
Extremadura	184.913	292.561	156.725	104.595	104.041			842.836
Galicia	100.424	693.970	682.425	278.049	527.093			2.281.961
Madrid	51.751	189.879	423.937	464.602	839.394		2.489.313	4.458.877
Murcia	5.893	65.914	369.979	61.814	437.725			941.325
Navarra	84.263	140.349	82.876		154.386			461.875
País Vasco	93.798	245.000	553.369	242.727	635.962			1.770.856
La Rioja	41.040	52.600	27.941		110.198			231.779
Total	2.569.286	5.413.539	8.418.665	3.296.865	7.919.612	2.108.604	3.778.126	33.504.697

Fuente: INE. Censo de población de 2001. Resultados definitivos a 17 de febrero de 2004.

Tabla 137 Afijación proporcional de la muestra a los estratos de la población (Aplicando Fórmula 80).

Comunidad	Tamaño de hábitat							
	<= 2.000	2.001 a 10.000	10.001 a 50.000	50.001 a 100.000	100.001 a 400.000	400.001 a 1.000.000	>1.000.000	Total
Andalucía	18	82	126	39	94	73		431
Aragón	16	9	12			38		76
Asturias	2	8	23	5	30			69
Baleares	1	10	20		20			51
Canarias	1	14	40	5	40			101
Cantabria	4	8	7	4	11			34
Castilla-León	48	23	19	16	61			168
Castilla-La Mancha	22	32	26	16	9			105
Cataluña	24	55	92	48	77		96	393
Valenciana	15	37	91	27	38	46		254
Extremadura	14	22	12	8	8			63
Galicia	7	52	51	21	39			170
Madrid	4	14	32	35	63		186	333
Murcia	0	5	28	5	33			70
Navarra	6	10	6		12			34
País Vasco	7	18	41	18	47			132
La Rioja	3	4	2		8			17
Total	192	404	628	246	591	157	282	2.500

Conocido el número de cuestionarios que hay que realizar en cada estrato, se procede como se indica en la Tabla 138, para extraer el número de puntos de muestreo total, que son municipios (conglomerados).

Tabla 138 Ejemplo de criterio de selección del número de municipios.

Tamaño de hábitat	Criterio de selección de municipios
<= 2.000	1 municipio por cada 4 cuestionarios o fracción
2.001 a 10.000	1 municipio por cada 9 cuestionarios o fracción
10.001 a 50.000	1 municipio por cada 14 cuestionarios o fracción
50.001 a 100.000	1 municipio por cada 19 cuestionarios o fracción
100.001 a 400.000	1 municipio por cada 24 cuestionarios o fracción
400.001 a 1.000.000	Reparto proporcional entre los municipios
+ de 1.000.000	Reparto proporcional entre los municipios

Tabla 139 Municipios seleccionados por cada estrato.								
Comunidad	Tamaño de hábitat							
	<= 2.000	2.001 a 10.000	10.001 a 50.000	50.001 a 100.000	100.001 a 400.000	400.001 a 1.000.000	>1.000.000	Total
Andalucía	5	10	9	3	4	1		32
Aragón	4	1	1			1		7
Asturias	1	1	2	1	2			7
Baleares	1	2	2		1			6
Canarias	1	2	3	1	2			9
Cantabria	1	1	1	1	1			5
Castilla-León	12	3	2	1	3			21
Castilla-La Mancha	6	4	2	1	1			14
Cataluña	6	7	7	3	4		1	28
Valenciana	4	5	7	2	2	1		21
Extremadura	4	3	1	1	1			10
Galicia	2	6	4	2	2			16
Madrid	1	2	3	2	3		1	12
Murcia	1	1	2	1	2			7
Navarra	2	2	1		1			6
País Vasco	2	2	3	1	2			10
La Rioja	1	1	1		1			4
Total	54	53	51	20	32	3	2	215

A partir del listado de municipios del INE y sabiendo cuantos hay que seleccionar en cada estrato, se extraen por muestreo aleatorio simple o sistemático. De los 320 municipios que hay en Andalucía (Tabla 135) de 2.000 habitantes o menos, hay que extraer 5. De los 680 municipios que hay en Aragón (Tabla 135) de 2.000 habitantes o menos, hay que extraer 4, y así sucesivamente se procede con todos los estratos hasta conseguir la lista de municipios final.

Con la lista de municipios se obtienen las provincias y se diseñan las rutas, para optimizar tiempos y costes del personal de campo. El profesional en su gabinete de trabajo establecerá otros criterios que faciliten y optimicen el proceso de trabajo de campo en base a sus conocimientos y experiencia académica y profesional.

Las cuotas por edad y sexo se extraen de las frecuencias absolutas de la población española de 18 años o más y según el sexo (Tabla 140) y se calculan los porcentajes respecto del total de la tabla (Tabla 141).

Tabla 140 Población española por sexo y grupos de edad.					
	18 a 29	30 a 49	50 a 64	+ de 64	Total
Varón	3.923.474	6.189.136	3.179.748	2.948.965	16.241.323
Mujer	3.748.633	6.130.861	3.333.165	4.050.715	17.263.374
Total	7.672.107	12.319.997	6.512.913	6.999.680	33.504.697
Fuente: INE. Censo de población de 2001. Resultados definitivos a 17 de febrero de 2004.					

Tabla 141 Cuotas por sexo y grupos de edad.					
	18 a 29	30 a 49	50 a 64	+ de 64	Total
Varón	11,7%	18,5%	9,5%	8,8%	48,5%
Mujer	11,2%	18,3%	9,9%	12,1%	51,5%
Total	22,9%	36,8%	19,4%	20,9%	100,0%
Fuente: INE. Censo de población de 2001. Resultados definitivos a 17 de febrero de 2004.					

El proceso continúa en *campo*, se tiene que seleccionar los hogares donde se van a entrevistar a las unidades de observación. Se pueden establecer diferentes procedimientos. Por ejemplo, al llegar a un municipio, censar todas las viviendas y seleccionar, por muestreo aleatorio simple o sistemático, tantas como entrevistas hay que realizar. Este es un ejemplo, pero no es útil por el tiempo que le llevaría al encuestador preparar el censo.

En el procedimiento de *rutas aleatorias con punto de partida,* se fija un punto de partida (calle y portal) y a partir de ahí, al salir del portal se sigue el curso de la calle y la primera calle a la derecha, la siguiente calle a la izquierda, a la derecha, a la izquierda, etc. Haciendo entrevistas en aquellos portales que cumplan los requisitos fijados y rellenando la hoja de control de ruta.

En municipios grandes, en los portales puede haber más de una escalera, en cada escalera más de una planta, en cada planta más de una vivienda y en cada vivienda se debe entrevistar a la persona según las cuotas. Este punto se recomienda ampliar con V. G. Manzano (1996).[81]

14.7 Cálculo del tamaño de la muestra

Por la *ley de los grandes números*, a medida que *n* tiende a *N* los estadísticos de la muestra tienden a ser los parámetros de la población, y la diferencia entre el parámetro y el estadístico, el error exacto, tiende a cero. Pero a partir de un momento determinado, el incremento de *n* eleva mucho el coste económico y material del trabajo de campo y no se obtienen reducciones de consideración en el error exacto. Las fórmulas para el cálculo del tamaño de la muestra permiten obtener tamaños reducidos o ajustados de *n* controlando el tamaño del error.

Para el cálculo del tamaño de la muestra se asume que el muestreo es aleatorio simple. La población puede ser considerada finita o infinita y el parámetro a estimar va a ser una proporción o una media. Una población se considera finita si su tamaño (*N*) es inferior a 100.000 unidades de observación y se considera infinita si es mayor de esta cantidad o es desconocida.

Tabla 142 Cálculo del tamaño de muestras asumiendo muestreo aleatorio simple.

Para estimación de %	
Población Finita.	Población Infinita.
$n = \frac{Z^2 \cdot p \cdot q \cdot N}{e^2 \cdot (N-1) + Z^2 \cdot p \cdot q}$	$n = \frac{Z^2 \cdot p \cdot q}{e^2}$
Error muestral	Error muestral
$e = Z \cdot \sqrt{\frac{p \cdot q}{n}} \cdot \sqrt{1 - fm}$	$e = Z \cdot \sqrt{\frac{p \cdot q}{n}}$
Para estimación de $\overline{X}$	
Población finita.	Población Infinita.
$n = \frac{Z^2 \cdot \omega^2 \cdot N}{e^2 \cdot N + Z^2 \cdot \omega^2}$	$n = \frac{Z^2 \cdot \omega^2}{e^2}$
Error muestral	Error muestral
$e = Z \cdot \sqrt{\frac{\omega^2}{n}} \cdot \sqrt{1 - fm}$	$e = Z \cdot \sqrt{\frac{\omega^2}{n}}$

Definición de elementos:

n:	Tamaño de la muestra.
N:	Tamaño de la población.
e:	Error muestral o error absoluto.
Z:	Valor de la variable estandarizada *Z* que define el *Nc* o viceversa.
p:	Probabilidad o porcentaje de éxito.
q:	Probabilidad o porcentaje de fracaso.
ω^2:	Varianza de la población.
$\frac{p \cdot q}{n}$	Varianza de la distribución binomial.[82]
fm:	Fracción de muestreo.

Corrector por poblaciones finitas (cpf)[83]
$\sqrt{\frac{N-n}{N}} = \sqrt{\frac{N}{N} - \frac{n}{N}} = \sqrt{1 - fm}$

El cálculo del tamaño de una muestra consiste en aplicar la fórmula correspondiente y se obtiene el número de unidades de observación a las que hay que entrevistar. El presupuesto

[81] Se recomienda la lectura de este libro para orientar el trabajo de campo.

[82] La media de una distribución binomial es $n \times p$ y la varianza es $n \times p \times q$ (Ver Tabla 70). Aplicando la Propiedad 4 (Pág. 65) y la Propiedad 9 (Pág. 71), si se dividie la variable por *n* la media se divide por *n* y la varianza por n^2. Entonces la media de la distribución binomial es *p* y la varianza (*p* x *q*) / *n*.

[83] Para ampliar la información sobre el *cpf,* confrontar con W. G. Cochram (1974: 47-49).

económico esta condicionado por esta *n* debido a que suele ser el apartado más oneroso de una investigación.

La calidad de la investigación no está influida por el número de observaciones. El número de observaciones afecta al error muestral y por consiguiente al intervalo de confianza dentro del cual estará el parámetro desconocido de la población. Si la *n* es pequeña, el error es grande y por lo tanto el intervalo es grande, pero si la *n* es grande, el error es pequeño y el intervalo pequeño. En ambos casos la investigación puede estar bien o mal hecha. La diferencia es el tamaño del intervalo para estimar el parámetro de la población. Por ejemplo, no es lo mismo decir que la demanda de agua de una población estará en el intervalo de 150 l/habitante a 200 l/habitante que decir un intervalo de 10 l/habitante a 1.000 l/habitante. En el primer caso el resultado puede ser útil para decidir políticas de consumo de agua, en el segundo es un intervalo tan amplio que el resultado puede ser correcto, pero la información no ser útil. El resultado puede estar bien obtenido, pero no ser útil.

Pequeñas variaciones en los términos de las fórmulas, pueden producir variaciones importantes en la *n*. Conocer el significado de cada uno de los términos puede permitir alcanzar el tamaño más adecuado conforme al presupuesto y el error deseado. En última instancia, hay que considerar que se pueden introducir variaciones en todos, algunos o uno de los términos, pero no se pueden dejar todos constantes. Por ejemplo, no se puede obtener una muestra grande que tenga un error pequeño con un bajo presupuesto.

Las fórmulas se interpretan de forma global pero interpretando los términos de uno en uno. Utilizando una metáfora, la fórmula se debe comprender como cuando se ha leído un libro, se sabe la historia, pero hay que contarlo y leerlo secuencialmente. Siguiendo el orden de la Tabla 142:

- *n* es el tamaño de la muestra que se obtiene por fórmula.
- *N* es el tamaño de la población que está dado por la delimitación geográfica y los criterios que deben cumplir los individuos.
- *e* es el error muestral o absoluto. Se define previamente y se puede obtener a partir de la distribución de los valores muestrales (ver Tabla 109 y hojas siguientes). El error absoluto es el error típico multiplicado por el valor de *Z*, que está definida o define el *Nc*. Sumado/restado al estadístico de la muestra, proporciona el intervalo de confianza, que es el intervalo dentro del cual estará el parámetro desconocido de la población.
- *Z* es el valor de la variable estandarizada que tiene una distribución $N_{(0,1)}$ y define el *Nc* que a su vez define el intervalo de confianza. Este nivel de confianza es la probabilidad de que el parámetro de la población esté en el intervalo de confianza. En porcentaje indicaría el porcentaje de muestras que tendrían en su intervalo de confianza el parámetro desconocido de la población (Repasar epígrafe 11).
- *p* es la probabilidad de "éxito", es la probabilidad con la que se encuentra el evento estudiado en la población. También se puede expresar en porcentaje. Este valor es desconocido normalmente. Posteriormente se verá la justificación del valor asignado. También es la probabilidad de "éxito" de una distribución binomial. En la población se busca la probabilidad de un evento, una probabilidad binomial y no una multinomial. Ejemplos: tener o no tener frigorífico; tener o no tener coche.
- *q* es la probabilidad de fracaso y es igual a 1 - *p*.
- ω^2 es la varianza de la población que normalmente es desconocida. No presenta la misma facilidad que *p* para ser estimada.

$\notin \frac{p\Delta q}{n}$ se considera la varianza de la población cuando se quieren estimar porcentajes (Ver nota 82).

Siguiendo con la Ficha Técnica del CIS (Tabla 134), en e*rror muestral* para un nivel de confianza (*Nc*) del 95,5% (según las Tablas es 95,44%, pero en las fichas técnicas lo redondean a 95,5%). Sigma (ω), que es la desviación típica de la población, se representa también por *Z* que son unidades de desviación típica, entonces en la fórmula se especifica como Z^2, y $p = q$, por lo que $p = 0{,}50$ y $q = 0{,}50$, el error real o absoluto es de ±2,0%, considerando el total de la muestra (2.500 casos) y subrayamos, en el supuesto de muestreo aleatorio simple.

Entonces el error se calcula según la Tabla 143, pero hay que recordar que $p + q = 1$, entonces $q = (1 - p)$, así es que los valores de *p* y *q* están tabulados y el valor máximo que puede tomar el producto es 0,25. Como la varianza está en el numerador (ver fórmula para poblaciones infinitas y en el caso de proporciones de la Tabla 142) la relación con el tamaño *n* es directa, a mayor varianza, mayor *n* y a menor varianza menor *n*.

Considerar el caso $p = q$, se le denomina el más desfavorable porque al producir la mayor *n* encarece el estudio, no obstante ser el error menor, y el valor máximo es conocido (0,25), aunque la varianza sea desconocida. En el caso de la varianza poblacional $/\omega^2($, el valor es desconocido normalmente y no se conoce el valor máximo.

Tabla 143 Cálculo del error muestral.

p		*q*		Cálculo del error (proporción)	Cálculo del error (porcentaje)
0,1	x	0,9	= 0,09	$e \mid Z\Delta\sqrt{\frac{p\Delta q}{n}}$	$e \mid Z\Delta\sqrt{\frac{p\Delta q}{n}}$
0,2	x	0,8	= 0,16		
0,3	x	0,7	= 0,21		
0,4	x	0,6	= 0,24	$e \mid 2\Delta\sqrt{\frac{0{,}5\Delta 0{,}5}{2.500}}$	$e \mid 2\Delta\sqrt{\frac{50\Delta 50}{2.500}}$
0,5	x	0,5	= 0,25		
0,6	x	0,4	= 0,24	$e \mid 2\Delta\sqrt{0{,}0001}$	$e \mid 2\Delta\sqrt{1}$
0,7	x	0,3	= 0,21		
0,8	x	0,2	= 0,16	$e \mid 2\Delta 0{,}01$	$e \mid 2\Delta 1$
0,9	x	0,1	= 0,09	$e \mid 0{,}02$	$e \mid 2$

En sociología, las muestras se utilizan para estudiar múltiples dimensiones de una población y estas dimensiones pueden ser categóricas o numéricas. La fórmula del tamaño de la muestra estaría definida por la dimensión principal. Normalmente, como en el caso del Estudio del CIS, se utiliza la fórmula para el cálculo del tamaño de la muestra considerando que se quiere estimar una proporción y en el caso de población infinita. Reúne dos ventajas, que normalmente las dimensiones son categóricas y que se conoce el valor máximo de la varianza.

Considerando muestreo aleatorio simple, el cálculo del tamaño de la muestra en el caso del Estudio del CIS y considerando proporciones, está dado por la fórmula,

$$n \mid \frac{Z^2 \Delta p \Delta q}{e^2} \mid \frac{2^2 \Delta 0{,}5 \Delta 0{,}5}{0{,}02^2} \mid \frac{1}{0{,}0004} \mid 2.500$$

Si se consideran los valores en porcentajes,

$$n \mid \frac{Z^2 \Delta p \Delta q}{e^2} \mid \frac{2^2 \Delta 50 \Delta 50}{2^2} \mid \frac{10.000}{4} \mid 2.500$$

El tamaño de la muestra y el error se calculan y son representativos de la parte del censo (la población) para la que se ha diseñado. El error puede ser recalculado al final del trabajo de campo utilizando la *p* obtenida en la muestra. El error también será recalculado si se utiliza una parte de la muestra en vez de la muestra total.

14.8 Ejemplos de cálculo de tamaño de muestra y de error de muestreo

Ejemplo 1: Supuesta una Comarca formada por tres municipios A, B y C con los tamaños de población que se indican, calcular una muestra representativa de la Comarca (Tabla 144).

Tabla 144 Muestra representativa de la Comarca para estimar una proporción.

Municipio	Población	*fm*	Afijación proporcional	Afijación mixta				Ponderación	
				Asignación directa	*fm*	Afijación proporcional	Total	*Ŭ*	Muestra ponderada
A	500	0,0249	12	50	0,0234	12	62	0,194	12
B	40.000	0,0249	995	50	0,0234	935	985	1,010	995
C	60.000	0,0249	1.493	50	0,0234	1.403	1.453	1,027	1.493
Total	100.500		2.500	150		2.350	2.500		2.500

Asumimos población infinita (*n* > 100.000). Como no facilitan datos, se toma *Nc* = 95,44 %; *p* = *q*, y *e* = ±2,0%.

$$n \mid \frac{Z^2 \Delta p \Delta q}{e^2}; \; n \mid \frac{2^2 \Delta 0{,}5 \Delta 0{,}5}{0{,}02^2}; \; n \mid \frac{1}{0{,}0004} \mid 2.500$$

$$fm \mid \frac{n}{N}; \; fm \mid \frac{2.500}{100.500} \mid 0{,}0249$$

$$n_i \mid fm \Delta N_i; \; n_A \mid 0{,}0249 \Delta 500 \mid 12; \; n_B \mid 0{,}0249 \Delta 40.000 \mid 995; \; n_C \mid 0{,}0249 \Delta 60.000 \mid 1.493$$

Si se estima que *n* = 12 en el municipio A puede ser insuficiente, se puede proceder con la afijación mixta.
Asignación directa = 50.

$$fm \mid \frac{n}{N}; \; fm \mid \frac{2.350}{100.500} \mid 0{,}0234$$

$$n_i \mid fm \Delta N_i; \; n_A \mid 0{,}0234 \Delta 500 \mid 12; \; n_B \mid 0{,}0234 \Delta 40.000 \mid 995; \; n_C \mid 0{,}0234 \Delta 60.000 \mid 1.403$$

Factor de ponderación

$$n_i^e \Delta \zeta_i \mid n_i^t$$

$$\zeta_i \mid \frac{n_i^t}{n_i^e}; \; \zeta_A \mid \frac{12}{62} \mid 0{,}194; \; \zeta_B \mid \frac{995}{985} \mid 1{,}010; \; \zeta_C \mid \frac{1.493}{1.453} \mid 1{,}027$$

La n_i ponderada de cada municipio (estrato) es:

$$n_i^p \mid \zeta_i \Delta n_i^e; \; n_A^p \mid 0{,}2016 \Delta 62 \mid 12; \; n_B^p \mid 1{,}0098 \Delta 985 \mid 995; \; n_C^p \mid 1{,}0272 \Delta 1.453 \mid 1.493$$

Ejemplo 2: Supuesta una Comarca formada por tres municipios A, B y C con los tamaños de población que se indican, calcular una muestra representativa a cada municipio (Tabla 145).

Tabla 145 Muestra representativa de cada municipio para estimar una proporción.

Municipio	Población	muestra
A	500	417
B	40.000	2.353
C	60.000	2.400
Total	100.500	5.170

$$n = \frac{Z^2 \cdot p \cdot q \cdot N}{e^2 \cdot (N-1) + Z^2 \cdot p \cdot q}$$

Asumimos población finita (n < 100.000). Como no facilitan datos, se toma Nc = 95,44 %; $p = q$, y e = ±2,0%.

$$n = \frac{Z^2 \cdot p \cdot q \cdot N}{e^2 \cdot (N-1) + Z^2 \cdot p \cdot q} = \frac{2^2 \cdot 0{,}5 \cdot 0{,}5 \cdot 500}{0{,}02^2 \cdot (500-1) + 2^2 \cdot 0{,}5 \cdot 0{,}5} = 417$$

$$n = \frac{Z^2 \cdot p \cdot q \cdot N}{e^2 \cdot (N-1) + Z^2 \cdot p \cdot q} = \frac{2^2 \cdot 0{,}5 \cdot 0{,}5 \cdot 40.000}{0{,}02^2 \cdot (40.000-1) + 2^2 \cdot 0{,}5 \cdot 0{,}5} = 2.353$$

$$n = \frac{Z^2 \cdot p \cdot q \cdot N}{e^2 \cdot (N-1) + Z^2 \cdot p \cdot q} = \frac{2^2 \cdot 0{,}5 \cdot 0{,}5 \cdot 60.000}{0{,}02^2 \cdot (60.000-1) + 2^2 \cdot 0{,}5 \cdot 0{,}5} = 2.400$$

Una vez obtenida la muestra representativa a cada municipio, se puede utilizar la muestra total para representar a toda la comarca. Un procedimiento puede ser el utilizado en la Tabla 146, ajustando el tamaño de la muestra de cada municipio, al estrato de la comarca, esto es, se tendría que ponderar.

Tabla 146 Ajuste de la muestra representativa del municipio para representar a la comarca, para estimar una proporción.

Municipio	Población (*N*)	Muestra (*n*)	*fm*	Afijación proporcional	*Ŭ*	Muestra ponderada
A	500	417	0,0514	26	0,062	26
B	40.000	2.353	0,0514	2.058	0,875	2.058
C	60.000	2.400	0,0514	3.086	1,286	3.086
Total	100.500	5.170		5.170		5.170

El error de la muestra total sería:

$$e = Z \cdot \sqrt{\frac{p \cdot q}{n}}\ ,\ e = 2 \cdot \sqrt{\frac{0{,}5 \cdot 0{,}5}{5.170}} = 0{,}0139$$

Al utilizar p y q en proporciones el error muestral se obtiene en proporciones y al multiplicarlo por 100 el error muestral es ±1,39%.

Tamaño de muestra, para población finita (N = 1.000), variando términos de la fórmula (Tabla 147).

Tabla 147 Tamaño de muestra para población finita y estimación de proporción.

		N 1.000				
				p		
Z	***e***	0,5	0,4	0,3	0,2	0,1
2,0	0,020	714	706	678	616	474
2,5	0,020	796	790	767	714	585
3,0	0,020	849	844	825	783	670
3,5	0,020	885	880	866	831	734
4,0	0,020	909	906	894	865	783
Si *N* y *e* se mantienen y *p* (*q*) y *Z* varían. La muestra disminuye si disminuye *p* y aumenta si aumenta *Z*.						
		N 1.000				
				p		
Z	***e***	0,5	0,4	0,3	0,2	0,1
2,0	0,020	714	706	678	616	474
2,0	0,025	616	606	574	506	366
2,0	0,030	527	516	483	416	286
2,0	0,035	450	440	407	343	227
2,0	0,040	385	375	344	286	184
Si *N* y *Z* se mantienen y *p* (*q*) y *e* varían. La muestra disminuye si disminuye *p* y disminuye si aumenta *e*.						
		N 1.000				
				p		
Z	***e***	0,5	0,4	0,3	0,2	0,1
2,0	0,020	714	706	678	616	474
2,5	0,025	714	706	678	616	474
3,0	0,030	714	706	678	616	474
3,5	0,035	714	706	678	616	474
4,0	0,040	714	706	678	616	474
Si *N* se mantiene y *p* (*q*), *Z* y *e* varían. La muestra disminuye si disminuye *p* y se mantiene con incrementos de *Z* y *e*.						
NOTA: Si *Z* y *e* mantienen la relación 100/1 (*e* en proporciones) o 1/1 (*e* en porcentajes), el tamaño de la muestra no varía si no se varían *N* y *p*. La explicación es que *Z* está en el numerador y en uno de los sumandos del denominador y *e* está en el otro sumando del denominador.						

Tabla 148 Tamaño de muestra para población infinita, $Z = 2$, variando p y e.

	p									
e	0,50	0,45	0,40	0,35	0,30	0,25	0,20	0,15	0,10	0,05
1,00	10.000	9.900	9.600	9.100	8.400	7.500	6.400	5.100	3.600	1.900
1,10	8.264	8.182	7.934	7.521	6.942	6.198	5.289	4.215	2.975	1.570
1,20	6.944	6.875	6.667	6.319	5.833	5.208	4.444	3.542	2.500	1.319
1,30	5.917	5.858	5.680	5.385	4.970	4.438	3.787	3.018	2.130	1.124
1,40	5.102	5.051	4.898	4.643	4.286	3.827	3.265	2.602	1.837	969
1,50	4.444	4.400	4.267	4.044	3.733	3.333	2.844	2.267	1.600	844
1,60	3.906	3.867	3.750	3.555	3.281	2.930	2.500	1.992	1.406	742
1,70	3.460	3.426	3.322	3.149	2.907	2.595	2.215	1.765	1.246	657
1,80	3.086	3.056	2.963	2.809	2.593	2.315	1.975	1.574	1.111	586
1,90	2.770	2.742	2.659	2.521	2.327	2.078	1.773	1.413	997	526
2,00	2.500	2.475	2.400	2.275	2.100	1.875	1.600	1.275	900	475
2,10	2.268	2.245	2.177	2.063	1.905	1.701	1.451	1.156	816	431
2,20	2.066	2.045	1.983	1.880	1.736	1.550	1.322	1.054	744	393
2,30	1.890	1.871	1.815	1.720	1.588	1.418	1.210	964	681	359
2,40	1.736	1.719	1.667	1.580	1.458	1.302	1.111	885	625	330
2,50	1.600	1.584	1.536	1.456	1.344	1.200	1.024	816	576	304
2,60	1.479	1.464	1.420	1.346	1.243	1.109	947	754	533	281
2,70	1.372	1.358	1.317	1.248	1.152	1.029	878	700	494	261
2,80	1.276	1.263	1.224	1.161	1.071	957	816	651	459	242
2,90	1.189	1.177	1.141	1.082	999	892	761	606	428	226
3,00	1.111	1.100	1.067	1.011	933	833	711	567	400	211
3,10	1.041	1.030	999	947	874	780	666	531	375	198
3,20	977	967	938	889	820	732	625	498	352	186
3,30	918	909	882	836	771	689	588	468	331	174
3,40	865	856	830	787	727	649	554	441	311	164
3,50	816	808	784	743	686	612	522	416	294	155
3,60	772	764	741	702	648	579	494	394	278	147
3,70	730	723	701	665	614	548	467	373	263	139
3,80	693	686	665	630	582	519	443	353	249	132
3,90	657	651	631	598	552	493	421	335	237	125
4,00	625	619	600	569	525	469	400	319	225	119

Tabla 149 Tamaño de muestra para población infinita, $Z = 3$, variando p y e.

	p									
e	0,50	0,45	0,40	0,35	0,30	0,25	0,20	0,15	0,10	0,05
1,00	22.500	22.275	21.600	20.475	18.900	16.875	14.400	11.475	8.100	4.275
1,10	18.595	18.409	17.851	16.921	15.620	13.946	11.901	9.483	6.694	3.533
1,20	15.625	15.469	15.000	14.219	13.125	11.719	10.000	7.969	5.625	2.969
1,30	13.314	13.180	12.781	12.115	11.183	9.985	8.521	6.790	4.793	2.530
1,40	11.480	11.365	11.020	10.446	9.643	8.610	7.347	5.855	4.133	2.181
1,50	10.000	9.900	9.600	9.100	8.400	7.500	6.400	5.100	3.600	1.900
1,60	8.789	8.701	8.438	7.998	7.383	6.592	5.625	4.482	3.164	1.670
1,70	7.785	7.708	7.474	7.085	6.540	5.839	4.983	3.971	2.803	1.479
1,80	6.944	6.875	6.667	6.319	5.833	5.208	4.444	3.542	2.500	1.319
1,90	6.233	6.170	5.983	5.672	5.235	4.675	3.989	3.179	2.244	1.184
2,00	5.625	5.569	5.400	5.119	4.725	4.219	3.600	2.869	2.025	1.069
2,10	5.102	5.051	4.898	4.643	4.286	3.827	3.265	2.602	1.837	969
2,20	4.649	4.602	4.463	4.230	3.905	3.487	2.975	2.371	1.674	883
2,30	4.253	4.211	4.083	3.871	3.573	3.190	2.722	2.169	1.531	808
2,40	3.906	3.867	3.750	3.555	3.281	2.930	2.500	1.992	1.406	742
2,50	3.600	3.564	3.456	3.276	3.024	2.700	2.304	1.836	1.296	684
2,60	3.328	3.295	3.195	3.029	2.796	2.496	2.130	1.697	1.198	632
2,70	3.086	3.056	2.963	2.809	2.593	2.315	1.975	1.574	1.111	586
2,80	2.870	2.841	2.755	2.612	2.411	2.152	1.837	1.464	1.033	545
2,90	2.675	2.649	2.568	2.435	2.247	2.007	1.712	1.364	963	508
3,00	2.500	2.475	2.400	2.275	2.100	1.875	1.600	1.275	900	475
3,10	2.341	2.318	2.248	2.131	1.967	1.756	1.498	1.194	843	445
3,20	2.197	2.175	2.109	2.000	1.846	1.648	1.406	1.121	791	417
3,30	2.066	2.045	1.983	1.880	1.736	1.550	1.322	1.054	744	393
3,40	1.946	1.927	1.869	1.771	1.635	1.460	1.246	993	701	370
3,50	1.837	1.818	1.763	1.671	1.543	1.378	1.176	937	661	349
3,60	1.736	1.719	1.667	1.580	1.458	1.302	1.111	885	625	330
3,70	1.644	1.627	1.578	1.496	1.381	1.233	1.052	838	592	312
3,80	1.558	1.543	1.496	1.418	1.309	1.169	997	795	561	296
3,90	1.479	1.464	1.420	1.346	1.243	1.109	947	754	533	281
4,00	1.406	1.392	1.350	1.280	1.181	1.055	900	717	506	267

15 Estadística Paramétrica

Se considera *Estadística Paramétrica* la que utiliza en sus fórmulas parámetros o estadísticos que representan a parámetros, la $\overline{X}, p$ y S^2 como estimadores de σ, P y ω^2. También es considerada *inferencial* porque los estadísticos obtenidos se utilizan para estimar aspectos de los *parámetros* de la población a través de contrastes de hipótesis o estimación de intervalos.

Los estadísticos considerados en este apartado se utilizan para el cálculo de diferencia de medias o proporciones o comparación de grupos a través de las medias o las proporciones o por descomposición de la varianza y son *Z*, *t* y *F*. Las *tablas de contingencia,* a veces son consideradas como estadística paramétrica y otras no, en cualquier caso, sirven para el análisis de asociación de variables categóricas o distribución conjunta de frecuencias (Ver Epígrafe 12).

En el análisis de Varianza se va a tratar sólo el caso de una vía (ONEWAY)[84] que es una variable numérica cruzada por una variable categórica. En ANOVA (acrónimo de *Analysis of variance*) interviene más de una variable categórica e independiente y excede las pretensiones de este manual. Otras técnicas de análisis de varianza son Análisis de Varianza Múltiple (MANOVA) y Medidas Repetidas, que pueden tener más de una variable dependiente y tener o no variables independientes, son Técnicas Multivariable y tampoco se ven en este manual. La pretensión es mostrar el significado de la descomposición de la varianza, que tiene una pedagogía y didáctica cómoda a través de ONEWAY. Los conceptos de descomposición de la varianza son la base en muchas técnicas multivariable y es útil en la teoría de muestreo.

Tabla 150 Estadísticos paramétricos de comparación de proporciones[85] y medias.

		Estadístico	Operación	Esquema de datos
Diferencia de proporciones	Una proporción	*Z*	Compara una proporción muestral con el parámetro de la población	Una proporción
	Dos proporciones independientes	*Z*	Compara dos proporciones de dos muestras independientes	Dos proporciones
	Dos proporciones emparejadas	*Z*	Compara dos proporciones de dos muestras emparejadas	Dos proporciones
Diferencia de medias	Una muestra	*Z* y *t*	Compara la media de una variable o un grupo de una muestra con el parámetro de la población.	Una variable numérica
	Dos muestras independientes	*Z* y *t*	Compara dos medias de dos grupos de una variable de la muestra.	Una variable numérica considerada la dependiente, agrupada por una variable categórica, considerada la independiente con dos categorías o sólo se usan dos categorías, que es la de agrupamiento.
	Dos muestras emparejas	*Z y t*	Compara dos variables emparejadas de la muestra.	Dos variables numéricas consideradas emparejadas (ver el epígrafe 15.2.3).
Análisis de varianza	Una vía	F_S	Compara más de dos grupos de una variable de la muestra.	Una variable numérica considerada la dependiente, agrupada por una variable categórica, considerada la independiente, con más de dos categorías, que es la de agrupamiento.
	Anova	F_S	Compara más de dos grupos de una variable de la muestra con más de una variable independiente, considerando las interacciones.	Una variable numérica considerada la dependiente, agrupada por más de una variable categórica, consideradas las independientes, con dos o más categorías, que son las de agrupamiento.

[84] El acrónimo por el que se conoce Análisis de Varianza de Una Vía es el término en inglés ONEWAY.

[85] Cuando se habla de proporciones es equivalente a porcentajes sólo que los valores se especificarían multiplicados por 100.

Tabla 151 Comparación de una proporción con la proporción poblacional.[86]		
Fórmula	$z_p = \frac{p - P}{S_P} = \frac{p - P}{\sqrt{\frac{P \Delta Q}{n}}}$	En donde: *p* Es la proporción en la muestra o grupo. *n* Número de casos de la muestra o grupo. *P* Es la proporción en la población. S_P Error típico de la distribución muestral de *p*.
Fórmulas de apoyo	$S_P = \sqrt{\frac{P \Delta Q}{n}}$ $Q = 1 - P$	Esquema: La proporción es de una variable de la muestra o grupo de una variable de la muestra y se compara con la *P* de la población. Requiere que el parámetro *P* de la población sea conocido.

Tabla 152 Comparación de dos proporciones. Muestras independientes (Ver nota 86).		
Fórmula	$z_P = \frac{p_1 - p_2}{S_P} = \frac{p_1 - p_2}{\sqrt{\frac{P' \Delta Q'}{n_1} + \frac{P' \Delta Q'}{n_2}}}$	En donde: p_1 Proporción grupo 1. p_2 Proporción grupo 2. n_1 Número de casos del grupo 1. n_2 Número de cassos del grupo 2. S_P Error típico de la distribución muestral de (p_1 - p_2). *P* Estimador de la proporción en la población.
Fórmulas de apoyo	$S_P = \sqrt{\frac{P' \Delta Q'}{n_1} + \frac{P' \Delta Q'}{n_2}}$ $P' = \frac{p_1 \Delta n_1 + p_2 \Delta n_2}{n_1 + n_2}$ $Q' = 1 - P'$	Esquema: Las proporciones comparadas son de dos submuestras o grupos.

Tabla 153 Comparación de dos proporciones. Muestras emparejadas (Ver nota 86).		
Fórmula	$\lvert z_e \rvert = \left\lvert \frac{p_{-1} - p_{1-}}{\sqrt{\frac{b + d}{n^2}}} \right\rvert = \left\lvert \frac{p_{-2} - p_{2-}}{\sqrt{\frac{b + d}{n^2}}} \right\rvert$	En donde: Ver Epígrafe 15.1.3
Fórmulas de apoyo	Ver Epígrafe 15.1.3	

Tabla 154 Comparación de la media de una muestra con la población.[87]		
Fórmula	$z_{\overline{X}} = \frac{\overline{X} - \sigma}{S'_{\overline{X}}} = \frac{\overline{X} - \sigma}{\sqrt{\frac{S_X^2}{n_X}}}$ $t_{\overline{X}} = \frac{\overline{X} - \sigma}{S''_{\overline{X}}} = \frac{\overline{X} - \sigma}{\sqrt{\frac{S_X^2}{n_X - 1}}}$	En donde: $\overline{X}$ Media de la variable (muestra o grupo). σ Media de la población. n_X Número de casos de la muestra o grupo. $S_{\overline{X}}$ Error típico de la media o desviación típica de las medias muestrales. S_X^2 Varianza de la variable.
Fórmulas de apoyo	$S'_{\overline{X}} = \sqrt{\frac{S_X^2}{n_X}}$ $S''_{\overline{X}} = \sqrt{\frac{S_X^2}{n_X - 1}}$	Esquema: La media es de una variable numérica de la muestra o grupo de una variable de la muestra y se compara con la media de la población (σ). Requiere que el parámetro σ de la población sea conocido.

[86] Cuando se habla de proporciones es equivalente a porcentajes sólo que los valores se especificarían multiplicados por 100.

[87] William S. Gosset diseñó la *t* considerando la cuasi-varianza (a la *n* del denominador se le resta la unidad) y la *Z* actúa con la varianza (a la *n* del denominador no se le resta la unidad). Por cuestiones pedagógicas y por claridad expositiva, se opta por usar la varianza y restar la unidad a la *n* en la fórmula de la *t*, para diferenciarla de la *Z* y justificar el uso de grados de libertad (*gl*). Por ejemplo, en los manuales de SPSS no restan una unidad a la *n* porque SPSS utiliza la cuasi-varianza (Norusis, 1986: B-121, B-124). Se pueden consultar otros manuales para ver más ejemplos.

Tabla 155 Comparación de dos medias (muestras independientes).

Fórmula	$z_{\overline{X}} = \frac{\overline{X}_1 - \overline{X}_2}{S'_{\overline{X}}} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$ $t_{\overline{X}} = \frac{\overline{X}_1 - \overline{X}_2}{S''_{\overline{X}}} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1 - 1} + \frac{S_2^2}{n_2 - 1}}}$	En donde: $\overline{X}_1$ Media de la submuestra o grupo 1. $\overline{X}_2$ Media de la submuestra o grupo 2. S_1^2 Varianza de la submuestra o grupo 1. S_2^2 Varianza de la submuestra o grupo 2. n_1 Número de casos de la submuestra o grupo 1. n_2 Número de casos de la submuestra o grupo 2. $S_{\overline{X}}$ Error típico de la media o desviación típica de las medias muestrales.
Fórmulas de apoyo	$S'_{\overline{X}} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$ $S''_{\overline{X}} = \sqrt{\frac{S_1^2}{n_1 - 1} + \frac{S_2^2}{n_2 - 1}}$	Esquema: Las medias que se comparan son de dos grupos o subgrupos de una variable numérica, definidos por las categorías de otra variable categórica.

Tabla 156 Comparación de dos medias (muestras emparejadas).

Fórmula	$Z_{dif} = \frac{\overline{X}_{dif}}{S'_{\overline{X}}} = \frac{\overline{X}_{dif}}{\sqrt{\frac{S_{dif}^2}{n_X}}}$ $t_{dif} = \frac{\overline{X}_{dif}}{S''_{\overline{X}}} = \frac{\overline{X}_{dif}}{\sqrt{\frac{S_{dif}^2}{n_X - 1}}}$	En donde: $\overline{X}_{dif}$ Es la media de la variable obtenida por la diferencia de las dos variables que se quieren comparar. S_{dif}^2 Es la varianza de la variable obtenida por la diferencia de las dos variables que se quieren comparar. n_X Número de casos de la variable diferencia. $S_{\overline{X}}$ Error típico de la media o desviación típica de las medias muestrales.
Fórmulas de apoyo	$S'_{\overline{X}} = \sqrt{\frac{S_{dif}^2}{n_X}}$ $S''_{\overline{X}} = \sqrt{\frac{S_{dif}^2}{n_X - 1}}$	Esquema: Las medias que se comparan son las de dos variables emparejadas.

Tabla 157 Comparación de grupos a través de las medias por descomposición de la varianza.

Fórmula	$F_S = \frac{MCE}{MCI}$ $MCE = \frac{SCE}{k-1}$ $MCI = \frac{SCI}{N-k}$	En donde: F_S: F de Fisher- Snedecor o sólo *F*. *MCE* Media de cuadrados entre-grupos. Variabilidad media del sistema debida a la dispersión que hay entre los grupos. *MCI* Media de cuadrados intra-grupos. Variabilidad media del sistema debida a la dispersión que hay dentro de los grupos. *SCE* Suma de cuadrados entre-grupos. *SCI* Suma de cuadrados intra-grupos. *k* Número de grupos. *N* Número total de casos de la muestra. *k-1* Grados de libertad del numerador (*MCE*). *N-k* Grados de libertad del denominador (*MCI*). $\overline{X}_j$ Media del grupo *j*. $\overline{X}_T$ Media de la variable total. x_i Valor del caso *i-ésimo*.
Fórmulas de apoyo	$SCE = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\overline{X}_j - \overline{X}_T)^2 = \sum_{j=1}^{k} n_j(\overline{X}_j - \overline{X}_T)^2$ $SCI = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_i - \overline{X}_j)^2$	Esquema: Se comparan los grupos formados en una variable numérica por una variable categórica o de agrupamiento. La comparación se hace relacionando la dispersión media entre los grupos (*MCE)* con la media de la dispersión dentro de los grupos (*MCI*). Se utiliza cuando hay más de dos grupos.

Al entrar en el mundo de las probabilidades se puede utilizar alguna metáfora para destacar, a las personas que se acercan por primera vez, la importancia del cambio de la *Estadística Descriptiva* a la *Estadística de Análisis*. El paso es como la serie de TV *StarGate* que mediante una puerta se cambia de tiempo y a veces parece que cambian de galaxia. En Estadística, se pasa de un mundo de afirmaciones de "tanto" es "cuanto" a un mundo de "puede ser" o es "probable". Es "esto" porque el *nivel de significación* (*Ns*) es "este" pero puede ser lo contrario si el *Ns* es "otro". Entonces las cosas ya no *son* o dejan de *ser*, sino que lo que *son* es en función del nivel de significación utilizado.

Este cambio, por ejemplo, involucra un cambio importante en la forma de pensar que a los alumnos de primeros cursos les supone pasar, de una ciencia fija e inmutable que aprenden en la Enseñanza Primaria y lo sucesivos niveles, a decirles que la ciencia o el conocimiento no es una puerta cerrada, no es un destino, sino una puerta abierta, es un continuo "salir" y no un "llegar".

15.1 Diferencia de proporciones

En las investigaciones y estudios sociológicos se realizan múltiples mediciones y clasificaciones sobre las unidades de observación que constituyen las muestras. El proceso de muestreo persigue la representatividad del Universo, esto es, estimar los parámetros desconocidos de la población a partir de los estadísticos conocidos de la muestra. Este proceso también supone que todas las operaciones realizadas sobre la muestra pueden ser inferidas sobre la población, como son las comparaciones de grupos a través de las proporciones, medias y varianzas.

Comparar una proporción o porcentaje con el parámetro de la población, es averiguar si ese grupo pertenece a esa población en el parámetro analizado por tener el mismo comportamiento (un valor similar) o si por el contrario no pertenece a esa población por no tener el mismo comportamiento (un valor similar).

Se presentan tres casos, comparar si un grupo pertenece a una población, comparar dos grupos independientes y comparar dos grupos emparejados, que es equivalente a decidir si pertenecen a la misma población o a distinta o si tienen comportamientos significativamente distintos.

15.1.1 Comparación de una proporción con el parámetro de la población

Sea la proporción p de la muestra y el parámetro P de la población, para saber si p pertenece a esa población, esto es, si p es igual o distinto a P, o mejor si p es significativamente igual o distinto a P, se hace mediante un contraste de hipótesis y aplicando el cálculo de probabilidades.

El protocolo de contraste de hipótesis se puede fijar como:

1. Propuesta de H_1 $p \prod P$
2. Propuesta de H_0 $p \mid P$
3. Estadístico: z
4. Criterio de aceptación/rechazo de H_0, $Ns = 0{,}05$

Para comparar si dos cosas son iguales o no, se procede por diferencia simple.

$$p \,4\, P$$

Para determinar si la diferencia es grande o pequeña, o mejor, significativamente grande o pequeña se estandariza transformando la diferencia a un valor de variable estandarizada conocida, como es el caso de la z y sobre esta distribución y a través de las probabilidades, se puede determinar si se acepta o rechaza la H_0. Se aplica el criterio z,

$$z_i \mid \frac{x_i \,4\, \overline{X}_X}{S_X}$$

En el caso de z_i, x_i pertenece a la variable X de media $\overline{X}_X$ y de desviación típica S_X. En el caso de la comparación de una proporción muestral con el parámetro de la población utilizamos una z_p en la que p es el valor de un estadístico que pertenece a una distribución de estadísticos muestrales que tiene de media $n\Delta P$ y la desviación típica es $\sqrt{n\Delta P\Delta Q}$ por tratarse de una distribución binomial (ver Epígrafe 10.1). Entonces tenemos que.

La proporción p de éxitos en la muestra es la relación entre los x éxitos y los n ensayos que es el tamaño de la muestra. Entonces,

Tabla 158 Comparación de una proporción con el parámetro de la población.	
p es la proporción de x éxitos sobre n pruebas. Entonces,	si $n \mid \frac{x}{p}$, entonces $x \mid n\Delta p$
Y p pertenece a una distribución binomial. La media de una distribución binomial es,	$n\Delta P$
La desviación típica es,	$\sqrt{n\Delta P\Delta Q}$
Entonces para comparar p con P, partimos de z_e y comparar el número de aciertos x con los parámetros de una distribución binomial.	$z_e \mid \frac{x 4 /n\Delta P0}{\sqrt{n\Delta P\Delta Q}}$
Y sustituyendo x	$z_e \mid \frac{/n\Delta p04 /n\Delta P0}{\sqrt{n\Delta P\Delta Q}}$
dividimos el numerador y el denominador por n	$z_e \mid \frac{\frac{1}{n}\Delta//n\Delta p04 /n\Delta P00}{\frac{1}{n}\Delta\sqrt{n\Delta P\Delta Q}} \mid \frac{\left®©™\frac{n\Delta p}{n}\right\} 4 \left®©™\frac{n\Delta P}{n}\right\}}{\sqrt{\frac{n\Delta P\Delta Q}{n^2}}} \mid \frac{p 4 P}{\sqrt{\frac{P\Delta Q}{n}}}$
Entonces, p Es la proporción en la muestra o grupo. n Número de casos de la muestra o grupo. P Es la proporción en la población. S_P Error típico de la distribución muestral de p.	$z_e \mid \frac{p 4 P}{\sqrt{\frac{P\Delta Q}{n}}} \mid \frac{p 4 P}{S_p}$

Entonces, la fórmula buscada es,

$z_e \mid \frac{p 4 P}{\sqrt{\frac{P\Delta Q}{n}}}$	Fórmula 97

Ejemplo: Contraste de una proporción con el parámetro de la población. En una población el 38,5% de los electores han votado a un partido. En una muestra de 2.500 individuos obtenida a partir del censo electoral los que han votado al mismo partido han sido el 36,5%. Operando al *Ns* = 0,05 ¿Se puede decir si el intervalo de confianza a partir del estadístico de la muestra contiene al de la población o si la muestra procede de esa población o si representa a esa población?

Protocolo de contraste de hipótesis:

1. Propuesta de H_1 $p \prod P$
2. Propuesta de H_0 $p \mid P$
3. Estadístico: z
4. Criterio de aceptación/rechazo de H_0, $Ns = 0{,}05$, $z_c = 1{,}96$.

$$z_e \mid \frac{p 4 P}{\sqrt{\frac{P \Delta Q}{n}}} \mid \frac{36{,}5 4 38{,}5}{\sqrt{\frac{38{,}5 \Delta / 100 4 38{,}50}{2.500}}} \mid \frac{4 2{,}0}{\sqrt{\frac{2.367{,}75}{2.500}}} \mid \frac{4 2{,}0}{0{,}97} \mid 42{,}06$$

Esquema de aceptación – rechazo de H_0.

Se acepta H_0 si: $\mid z_e \mid \{ \mid z_c \mid \sum Nc_e \{ Nc_c \sum Ns_e \} Ns_c$

Se rechaza H_0 si: $\mid z_e \mid \varnothing \mid z_c \mid \sum Nc_e \varnothing Nc_c \sum Ns_e \Omega Ns_c$

Como la $|z_e|$ es mayor que la $|z_c|$, se puede asumir que la diferencia entre la proporción de la muestra y la de la población difieren de forma significativa o que las diferencias no son debidas al azar, por lo tanto podemos asumir rechazar la H_0 y aceptar la H_1. La muestra no ha sido obtenida de esa población o no es representativa de la población.

Pero si el criterio de aceptación/rechazo de H_0, cambia a $Ns = 0{,}01$, entonces $z_c = 2{,}57$. En este caso, como la $|z_e|$ es menor que la $|z_c|$, se puede asumir que la diferencia entre la proporción de la muestra y la de la población no difieren de forma significativa o que las diferencias son debidas al azar, por lo tanto podemos asumir aceptar la H_0. La muestra ha sido obtenida de esa población o es representativa de la población. El cambio del Ns puede suponer que una H_0 sea aceptada o rechazada.

El *intervalo de confianza* a un nivel de confianza (Nc) de 0,99 ($Nc = 1 - Ns = 1 - 0{,}01$), está definido por,

$$P_{// p 4 z \Delta S_P 0 \{ P \{ / p 2 z \Delta S_P 00} \mid 0{,}99 \, , \, P_{\left(\left(\left(p 4 2{,}57 \Delta \sqrt{\frac{P \Delta Q}{n}} \right| \{ P \{ \left(\left(p 2 2{,}57 \Delta \sqrt{\frac{P \Delta Q}{n}} \right| \right| \right.} \mid 0{,}99$$

$$P_{// 36{,}54 / 2{,}57 \Delta 0{,}97 00 \{ P \{ / 36{,}52 / 2{,}57 \Delta 0{,}97 000} \mid 0{,}99 \, , \, P_{// 36{,}54 2{,}50 0 \{ P \{ / 36{,}52 2{,}50 00} \mid 0{,}99 \, , \, P_{// 34{,}0 0 \{ P \{ / 39{,}0 00} \mid 0{,}99$$

Como era de esperar, al haber aceptado la H_0, el intervalo de confianza (34,0% a 39,0%) contiene el parámetro de la población (38,5%), con un Nc de 0,99, esto es, que la probabilidad de que el parámetro de la población esté contenido en el intervalo de confianza es de 0,99 ó 99,0%.

El *intervalo de confianza* a un nivel de confianza (*Nc*) de 0,95 (*Nc* = 1 – 0,05), está definido por,

$$P_{//p4z\Delta S_p0\{P\{/p2z\Delta S_p00} \mid 0{,}95\,,\ P_{\left(\left(\left(\right.\right.\right.\right.p41{,}96\Delta\sqrt{\frac{P\Delta Q}{n}}\left.\left.\left.\left.\right)\right)\right)\{P\{\left(\left(\left(\right.\right.\right.\right.p21{,}96\Delta\sqrt{\frac{P\Delta Q}{n}}\left.\left.\left.\left.\right)\right)\right)} \mid 0{,}95$$

$$P_{//36{,}54/1{,}96\Delta0{,}9700\{P\{/36{,}52/1{,}96\Delta0{,}970000} \mid 0{,}95\,,\ P_{//36{,}541{,}900\{P\{/36{,}521{,}90000} \mid 0{,}95\,;\ P_{//34{,}60\{P\{/38{,}400} \mid 0{,}95$$

Como era de esperar, al haber rechazado la H_0, el intervalo de confianza (34,6% a 38,4%) no contiene el parámetro de la población (38,5%).

15.1.2 Comparación de dos proporciones. Muestras independientes

Sea la proporción p_1 de la muestra n_1 y la proporción p_2 de la muestra n_2. Siendo n_1 y n_2 dos muestras aleatorias e independientes, y siendo p_1 la proporción de x_1 aciertos sobre n_1 ensayos y p_2 la proporción de x_2 aciertos sobre n_2 ensayos, la comparación entre p_1 y p_2 se hace mediante un contraste de hipótesis y aplicando el cálculo de probabilidades.

El protocolo de contraste de hipótesis se puede fijar como:

1. Propuesta de $H_1\ \ p_1\ \Pi p_2$
2. Propuesta de $H_0\ \ p_1 \mid\ p_2$
3. Estadístico: z
4. Criterio de aceptación/rechazo de H_0, Ns = 0,05

Para determinar si la diferencia es grande o pequeña, o mejor, significativamente grande o pequeña se procede a estandarizar la diferencia como en el epígrafe anterior transformando la diferencia a un valor de variable estandarizada conocida, como es el caso de Z y sobre esta distribución y a través de las probabilidades se puede determinar si aceptar o rechazar la H_0. Se aplica el criterio Z (ver Epígrafe 10.1). Entonces,

Tabla 159 Comparación de dos proporciones. Muestras independientes.	
p_1 es la proporción de x_1 éxitos sobre n_1 pruebas. Entonces,	$x_1 \mid\ n_1 \Delta p_1$
p_2 es la proporción de x_2 éxitos sobre n_2 pruebas. Entonces,	$x_2 \mid\ n_2 \Delta p_2$
La media de p_1 es,	$\overline{X}_{p1} \mid\ n_1 \Delta p_1$
La media de p_2 es,	$\overline{X}_{p2} \mid\ n_2 \Delta p_2$
La desviación típica de la distribución de p_1 es,	$S_{p1} \mid\ \sqrt{n_1 \Delta p_1 \Delta q_1}$
La desviación típica de la distribución de p_2 es,	$S_{p2} \mid\ \sqrt{n_2 \Delta p_2 \Delta q_2}$
Entonces el estadístico de contraste para comparar p_1 y p_2 es z_e y simbólicamente,	$z_e \mid\ \frac{p_1 4\ p_2}{\sqrt{\left(\left(\left(\right.\right.\right.\right.P\Delta Q\Delta\left(\left(\left(\right.\right.\right.\right.\frac{1}{n_1} 2\ \frac{1}{n_2}\left.\left.\left.\left.\right)\right)\right)\right)}} \mid\ \frac{p_1 4\ p_2}{\sqrt{\left(\left(\left(\right.\right.\right.\right.\frac{P\Delta Q}{n_1} 2\ \frac{P\Delta Q}{n_2}\left.\left.\left.\left.\right)\right)\right)}}$
Como el parámetro P de la población es desconocido se sustituye por el estimador,	$P \mid\ \frac{/n_1 \Delta p_1 02\ /n_2 \Delta p_2 0}{n_1 2\ n_2}$

Entonces, la fórmula buscada es,

$z_e = \frac{p_1 - p_2}{\sqrt{\left(\frac{P\Delta Q}{n_1} + \frac{P\Delta Q}{n_2}\right)}}$	Fórmula 98

Ejemplo: Sean dos grupos (*A* y *B*) de 200 pacientes cada uno, aquejados de cierta enfermedad. Se suministra un medicamento al *A* pero no al *B* (grupo control), controlando todas las demás posibles variables en los dos grupos. Al final del tratamiento, 85 pacientes del *A* y 70 del *B* se recuperan de la enfermedad. ¿Se puede decir que el medicamento ha influido en la curación al *Ns* = 0,05?

Protocolo de contraste de hipótesis:

1. Propuesta de H_1 $p_1 \neq p_2$ (Se producen diferencias significativas entre la proporción de pacientes curados en el grupo que ha recibido el medicamento y la proporción de pacientes del grupo que no lo ha recibido).
2. Propuesta de H_0 $p_1 = p_2$ (No se producen diferencias significativas entre la proporción de pacientes curados en el grupo que ha recibido el medicamento y la proporción de pacientes del grupo que no lo ha recibido).
3. Estadístico: *Z*
4. Criterio de aceptación/rechazo de H_0, $Ns = 0{,}05$, $z_c = 1{,}96$.

$$P = \frac{(n_1 \cdot p_1) + (n_2 \cdot p_2)}{n_1 + n_2} = \frac{(85 \cdot 42{,}5) + (70 \cdot 35{,}0)}{200 + 200} = 38{,}8$$

$$z_e = \frac{p_1 - p_2}{\sqrt{\left(\frac{P\Delta Q}{n_1} + \frac{P\Delta Q}{n_2}\right)}} = \frac{42{,}5 - 35{,}0}{\sqrt{\left(\frac{38{,}8 \cdot 61{,}2}{200} + \frac{38{,}8 \cdot 61{,}2}{200}\right)}} = 1{,}54$$

Esquema de aceptación – rechazo de H_0.

Se acepta H_0 si: $|z_e| \leq |z_c| \rightarrow Nc_e \in [Nc_c - Ns_e, Nc_c + Ns_e]$

Se rechaza H_0 si: $|z_e| > |z_c| \rightarrow Nc_e < Nc_c - Ns_e \cup Nc_e > Nc_c + Ns_e$

Como la $|z_e|$ es menor que la $|z_c|$, se puede asumir que las dos proporciones no difieren de forma significativa o que las diferencias son debidas al azar, por lo tanto podemos asumir aceptar la H_0 y rechazar la H_1. El tratamiento seguido por los pacientes del grupo *A* no ha producido diferencias significativas, respecto del grupo de pacientes *B*, en la curación a un *Ns* de 0,05.

15.1.3 Comparación de dos proporciones. Muestras emparejadas

El esquema de muestras emparejadas o dependientes se produce cuando para una misma muestra o grupo de individuos existen al menos dos tomas de datos en una misma característica o atributo manteniendo o controlando a cada individuo emparejado en sus mediciones. El ejemplo típico son las pruebas *pre-test pos-test*. El objetivo es comparar lo ocurrido *antes* de un efecto con lo ocurrido *después* del efecto. Las diferencias, si las hay, se asume que son debidas al estímulo intermedio introducido entre las dos tomas, mediciones u observaciones.

El esquema en formato de doble entrada se muestra en la Tabla 160 con una población de individuos que en la *toma 1* tenían una cierta actitud que era *pro* o *anti* construcción de una central nuclear. Se proyecta un documental sobre la contaminación que producen las centrales nucleares. Este acto se considera el *efecto*. Después del efecto se realiza la *toma 2*.

Si no se producen diferencias entre los que están a favor y en contra antes del estímulo (documental) y los que están a favor y en contra después del estímulo, se puede asumir que no ha producido efecto. Pero si hay diferencias, entonces es que el estímulo ha producido efecto y se debe determinar el sentido.

Tabla 160 Tabla de la población antes y después del efecto.

		Después del efecto			
		Pro	*Anti*	Frecuencia	Proporción
Antes del efecto	*Pro*	*A*	*B*	*A+B*	$\frac{A \,2\, B}{N} \mid P_{1\_}$
	Anti	*D*	*C*	*D+C*	$\frac{D \,2\, C}{N} \mid P_{2\_}$
	Frecuencia	*A+D*	*B+C*	*N*	
	Proporción	$\frac{A \,2\, D}{N} \mid P_{\_1}$	$\frac{B \,2\, C}{N} \mid P_{\_2}$		

Si extraemos una muestra por muestreo aleatorio simple de tamaño *n* de la población *N* y la representamos en formato de tabla de datos Tipo I (Tabla 161),

Tabla 161 Tabla de datos Tipo I de muestras emparejadas.

Caso	Toma 1	Efecto	Toma 2
1	a	Documental	a
2	a		b
3	a		c
.	a		.
.	a		d
.	b		a
.	b		b
.	b		c
.	b		.
.	b		d
.	c		a
.	c		b
.	c		c
.	c		.
.	c		d
.	d		a
.	d		b
.	d		c
n-1	d		.
n	d		d

y en formato de tabla de doble entrada, se tiene,

Tabla 162 Tabla de la muestra antes y después del efecto.					
		Después del efecto			
		Pro	*Anti*	Frecuencia	Proporción
Antes del efecto	*Pro*	*a*	*b*	*a+b*	$\frac{a 2 b}{n} \mid p_{1\_}$
	Anti	*d*	*c*	*d+c*	$\frac{d 2 c}{n} \mid p_{2\_}$
	Frecuencia	*a+d*	*b+c*	*n*	
	Proporción	$\frac{a 2 d}{n} \mid p_{\_1}$	$\frac{b 2 c}{n} \mid p_{\_2}$		

Entonces, las personas que han cambiado de actitud después del *efecto,* son las que han pasado de *pro* a *anti* (*b*) y las que han pasado de *anti* a *pro* (*d*). Estas personas *b+d* han tenido que ser extraídas de *B+D* que son las que han cambiado de actitud en la población. Si consideramos que de la población *B+D* se han extraído aleatoriamente y con reposición, las *b+d* personas de la muestra. Si arbitrariamente llamamos *éxito* a las personas que han pasado de *pro* a *anti.* Entonces *b* es el número de éxitos en *b+d* pruebas independientes.

Por otra parte, si proponemos en la población la H_0 de proporción *pro* central nuclear *después* igual a *pro* central nuclear *antes*, viene dada por,

$$H_0 : \frac{A 2 D}{N} \mid \frac{A 2 B}{N}$$

Que si se opera con la igualdad, entonces,

$$N \Delta / A 2 D 0 \mid N \Delta / A 2 B 0$$

$$AN 2 DN \mid AN 2 BN$$

$$AN 4 AN 2 DN \mid BN$$

$$DN \mid BN$$

$$D \mid \frac{BN}{N}$$

$$D \mid B$$

Por lo que

$$\frac{A 2 D}{N} \mid \frac{A 2 B}{N}$$

Es equivalente a

$$D \mid B$$

Asumiendo la H_0 como verdadera resulta que la probabilidad de éxito en cada prueba será,

$$\frac{B}{B \,2\, D} \mid \frac{B}{B \,2\, B} \mid 0{,}50$$

Y esta probabilidad se mantendrá constante en las $d+b$ pruebas, puesto que el muestreo es con reposición y por lo tanto son pruebas independientes. Entonces, siendo verdadera la H_0, b se distribuirá binomialmente con las características,

$$\overline{X} \mid n \Delta p \mid /b \,2\, d0\Delta 0{,}50$$

$$S^2 \mid n \Delta p \Delta q \mid /b \,2\, d0\Delta 0{,}50 \Delta /1\,4\, 0{,}500 \mid /b \,2\, d0\Delta 0{,}25$$

El proceso de contraste de H_0 será,

1. Propuesta de H_1

$$d \prod b$$

Que es equivalente a

$$\frac{a \,2\, d}{n} \prod \frac{a \,2\, b}{n}$$

Y por lo tanto

$$p_{\_1} \prod p_{1\_}$$

Por considerarse la muestra representativa, se plantea que H_1

$$D \prod B$$

Que es equivalente a

$$\frac{A \,2\, D}{N} \prod \frac{A \,2\, B}{N}$$

Y por lo tanto

$$P_{\_1} \prod P_{1\_}$$

Que se asume que hay diferencias entre el número o proporción de individuos *pro* central nuclear *antes* del efecto y el número o proporción de individuos *pro* central nuclear *después* del efecto en la muestra y se infiere a la población.

2. Propuesta de H_o

$$d \mid b$$

Que es equivalente a

$$\frac{a \,2\, d}{n} \mid \frac{a \,2\, b}{n}$$

Y por lo tanto

$$p_{\_1} \mid p_{1\_}$$

Por considerarse la muestra representativa, se plantea que H_0

$$D \mid B$$

Que es equivalente a

$$\frac{A \,2\, D}{N} \mid \frac{A \,2\, B}{N}$$

Y por lo tanto

$$P_{\_1} \mid P_{1\_}$$

Que se asume que no hay diferencias entre el número o proporción de individuos *pro* central nuclear *antes* del efecto y el número o proporción de individuos *pro* central nuclear *después* del efecto en la muestra y se infiere a la población.

3. Estadístico: Z
4. Criterio de aceptación/rechazo de H_0, $Ns = 0{,}05$, $z_c = 1{,}96$.

Según McNemar (1947), llamando z_e al *ratio crítico*, tenemos que,

$$z_e \mid \frac{/b \,4\, d0}{\sqrt{/b \,2\, d0}}$$

Si se suma a a cada uno de los términos del numerador la diferencia no varía y si dividimos el numerador y el denominador por la n de la tabla, entonces,

$$z_e \mid \frac{\left\{\frac{/a \,2\, b0 \,4\, /a \,2\, d0}{n}\right\}}{\sqrt{\frac{/b \,2\, d0}{n^2}}} \mid \frac{\frac{/a \,2\, b0}{n} \,4\, \frac{/a \,2\, d0}{n}}{\sqrt{\frac{/b \,2\, d0}{n^2}}} \mid \frac{p_{\_1} \,4\, p_{1\_}}{\sqrt{\frac{/b \,2\, d0}{n^2}}}$$

Por lo tanto, el estadístico de contraste buscado es,

$z_e \mid \frac{p_{\_1} 4 p_{1\_}}{\sqrt{\frac{/b 2 d0}{n^2}}}$	Fórmula 99

Si en vez de seleccionar arbitrariamente como *éxito* a las personas que han pasado de *pro* a *anti* se selecciona como *éxito* las personas que han pasado de *anti* a *pro*, entonces *d* es el número de éxitos en *b+d* pruebas independientes.

Por otra parte, si se propone la H_0 de proporción *anti* central nuclear *después* igual a *anti* central nuclear *antes*, está dada por,

$$\frac{B 2 C}{N} \mid \frac{C 2 D}{N}$$

Que si se opera con la igualdad, entonces,

$$N \Delta / B 2 C 0 \mid N \Delta / C 2 D 0$$

$$BN 2 CN \mid CN 2 DN$$

$$BN 2 CN 4 CN \mid DN$$

$$DN \mid BN$$

$$D \mid \frac{BN}{N}$$

$$D \mid B$$

Por lo que

$$\frac{B 2 C}{N} \mid \frac{C 2 D}{N}$$

Es equivalente a

$$D \mid B$$

Asumiendo la H_0 como verdadera resulta que la probabilidad de éxito en cada prueba será,

$$\frac{D}{B 2 D} \mid \frac{D}{D 2 D} \mid 0{,}50$$

Y esta probabilidad se mantendrá constante en las $d+b$ pruebas, puesto que el muestreo es con reposición y por lo tanto son pruebas independientes. Entonces, siendo verdadera la H_0, d se distribuirá binomialmente con las características,

$$\overline{X} \mid n \Delta p \mid /b 2 d 0 \Delta 0{,}50$$

$$S^2 \mid n \Delta p \Delta q \mid /b 2 d 0 \Delta 0{,}50 \Delta /1 4 0{,}500 \mid /b 2 d 0 \Delta 0{,}25$$

El proceso de contraste de H_0 será,

1. Propuesta de H_1

$$d \Pi b$$

Que es equivalente a

$$\frac{b 2 c}{n} \Pi \frac{c 2 d}{n}$$

Y por lo tanto

$$p_{\_2} \Pi p_{2\_}$$

Por considerarse la muestra representativa, se plantea que H_1

$$D \Pi B$$

Que es equivalente a

$$\frac{B 2 C}{N} \Pi \frac{C 2 D}{N}$$

Y por lo tanto

$$P_{\_2} \Pi P_{2\_}$$

Que se asume que hay diferencias entre el número o proporción de individuos *anti* central nuclear *antes* del efecto y el número o proporción de individuos *anti* central nuclear *después* del efecto en la muestra y se infiere a la población.

2. Propuesta de H_o

$$d \mid b$$

Que es equivalente a

$$\frac{b 2 c}{n} \mid \frac{c 2 d}{n}$$

Y por lo tanto

$$p_{\_2} \mid p_{2\_}$$

Por considerarse la muestra representativa, se plantea que H_1

$$D \mid B$$

Que es equivalente a

$$\frac{B \, 2 \, C}{N} \, | \, \frac{C \, 2 \, D}{N}$$

Y por lo tanto

$$P_{\_2} \, | \, P_{2\_}$$

Que se asume que no hay diferencias entre el número o proporción de individuos *anti* central nuclear *antes* del efecto y el número o proporción de individuos *anti* central nuclear *después* del efecto en la muestra y se infiere a la población.

3. Estadístico: *Z.*
4. Criterio de aceptación/rechazo de H_0, $Ns = 0{,}05$, $z_c = 1{,}96$.

Según McNemar (1947), llamando z_e al *ratio crítico*, tenemos que,

$$z_e \, | \, \frac{/d \, 4 \, b0}{\sqrt{/b \, 2 \, d0}}$$

Si se suma *c* a cada uno de los términos del numerador la diferencia no varía y si dividimos por la *n* de la tabla el numerador y el denominador queda,

$$z_e \, | \, \frac{\left(\frac{/c \, 2 \, d0}{n} 4 \frac{/c \, 2 \, b0}{n} \right)}{\sqrt{\frac{/b \, 2 \, d0}{n^2}}} \, | \, \frac{\frac{/c \, 2 \, d0}{n} 4 \frac{/c \, 2 \, b0}{n}}{\sqrt{\frac{/b \, 2 \, d0}{n^2}}} \, | \, \frac{p_{2\_} \, 4 \, p_{\_2}}{\sqrt{\frac{/b \, 2 \, d0}{n^2}}}$$

Para mantener el mismo criterio que al considerar *b* los aciertos, se compara *anti* de después del efecto con *anti* de antes del efecto, entonces el estadístico de contraste buscado es,

$z_e \, \| \, \frac{p_{\_2} \, 4 \, p_{2\_}}{\sqrt{\frac{/b \, 2 \, d0}{n^2}}}$	Fórmula 100

Como,

$$/p_{1\_} \, 2 \, p_{2\_}(\, | \, 100$$

Y

$$/p_{\_1} \, 2 \, p_{\_2}(\, | \, 100$$

Entonces,

$$/p_{1\_} 2 p_{2\_}(| /p_{\_1} 2 p_{\_2}($$

Por lo tanto,

$/\| p_{\_1} 4 p_{1\_} \|(\| /\| p_{\_2} 4 p_{2\_} \|($	Fórmula 101

Como la suma de las proporciones o porcentajes de las filas es igual que la suma de las proporciones o porcentajes de columnas, entonces el contraste de diferencias entre los *pro* de *antes* y *después* del efecto y el contraste de los *anti* de *antes* y *después* del efecto, es igual pero con distinto signo. Por lo que en valor absoluto son iguales.

$\|z_e\| \| \left\| \frac{p_{\_1} 4 p_{1\_}}{\sqrt{\frac{/b 2 d0}{n^2}}} \right\| \| \left\| \frac{p_{\_2} 4 p_{2\_}}{\sqrt{\frac{/b 2 d0}{n^2}}} \right\|$	Fórmula 102

Ejemplo: Utilizando el estudio de CIRES de enero de 1996, la pregunta, [88]

> P.5. En general, y pensando en todas las cosas que son para Vd. más importantes, y utilizando una escala de 0 a 10 puntos, en la que el 0 significa que la vida le va muy mal, y el 10 significa que la vida le va muy bien, ¿cómo cree Vd. que le van las cosas actualmente? ¿Y como diría Vd. que le iban hace un año? ¿Y como piensa Vd. que le irán dentro de un año? (TARJETA 2)

La pregunta está codificada en una escala de 0 a 10 puntos con el sentido especificado. Para realizar este ejemplo, se ha procedido a la recodificación de "cómo piensa que le iban las cosas hace un año" (variable A23) y "como piensa que le irán dentro de un año" (variable A25), de tal manera que de 0 a 5 se ha considerado *mal* y de 6 a 10 se ha considerado *bien*, y se obtiene la tabla,

		Situación en el futuro			
		Mal	*Bien*	*Total*	*Proporción*
Situación en el pasado	*Mal*	281	146	427	0,44
	Bien	63	475	538	0,56
	Total	344	621	965	
	Proporción	0,36	0,64		

¿Se puede decir que hay diferencia significativas en el cambio de sentimiento al *Ns* = 0,05?

Protocolo de contraste de hipótesis

[88] Este mismo ejemplo se hará más adelante por diferencia de medias de *muestras emparejadas* y tiene la misma interpretación.

1. H_1: Hay diferencias significativas entre la proporción de los que piensan que les va a ir mal en el futuro con la proporción de los que piensan que les iba mal en el pasado. Simbólicamente,

$$p_{\_1} \Pi p_{1\_}$$

2. H_0: No hay diferencias significativas entre la proporción de los que piensan que les va a ir mal en el futuro con la proporción de los que piensan que les iba mal en el pasado. Simbólicamente,

$$p_{\_1} | \ p_{1\_}$$

3. Estadístico: Z
4. Criterio de aceptación/rechazo de H_0, $Ns = 0{,}05$, $z_c = 1{,}96$.

Aplicando la Fórmula 99,

$$z_e | \ \frac{p_{\_1} 4 \ p_{1\_}}{\sqrt{\frac{/n_{21} 2 \ n_{12} 0}{n^2}}} | \ \frac{0{,}36 4 \ 0{,}44}{\sqrt{\frac{/63 2 \ 1460}{965^2}}} | \ \frac{4 \ 0{,}08}{0{,}01} | \ 48{,}00$$

A un $Ns = 0{,}05$, como $|z_e| > |z_c|$ se puede asumir que hay diferencias significativas entre el pensamiento o sentimiento de que las cosas les vayan *mal* dentro de un año a que las cosas les haya ido *mal hace* un año y como es negativo, significa que la proporción de *mal* de *después* (futuro) es significativamente menor que *antes* (pasado).

Si el protocolo se hace con el sentimiento de *bien,* el contraste de hipótesis es,

1. H_1: Hay diferencias significativas entre la proporción de los que piensan que les va a ir bien en el futuro con la proporción de los que piensan que les iba bien en el pasado. Simbólicamente,

$$p_{\_2} \Pi p_{2\_}$$

2. H_0: No hay diferencias significativas entre la proporción de los que piensan que les va a ir bien en el futuro con la proporción de los que piensan que les iba bien en el pasado. Simbólicamente,

$$p_{\_2} | \ p_{2\_}$$

3. Estadístico: Z
4. Criterio de aceptación/rechazo de H_0, $Ns = 0{,}05$, $z_c = 1{,}96$.

Aplicando la Fórmula 100,

$$z_e \mid \frac{p_{\_2} \; 4 \; p_{2\_}}{\sqrt{\frac{/n_{21} \; 2 \; n_{12}0}{n^2}}} \mid \frac{0{,}64 \; 4 \; 0{,}56}{\sqrt{\frac{/63 \; 2 \; 1460}{965^2}}} \mid \frac{0{,}08}{0{,}01} \mid 8{,}00$$

A un *Ns* = 0,05, como $|z_e| > |z_c|$ se puede asumir que hay diferencias significativas entre el pensamiento o sentimiento de que las cosas les vaya *bien dentro* de un año a que las cosas les fue *bien* hace un año y como es positivo, significa que la proporción de *bien* de *después* (futuro) es significativamente mayor que *antes* (pasado).

Según la Fórmula 102, se comprueba que,

$$\left|z_e\right| \mid \left| \frac{p_{\_1} \; 4 \; p_{1\_}}{\sqrt{\frac{/b \; 2 \; d0}{n^2}}} \right| \mid \left| \frac{p_{\_2} \; 4 \; p_{2\_}}{\sqrt{\frac{/b \; 2 \; d0}{n^2}}} \right| \mid \left|4 \; 8{,}00\right| \mid \left|8{,}00\right|$$

15.2 Diferencia de medias

Los estadísticos de *diferencias de medias* comparan dos grupos, subgrupos, muestras o submuestras a través de las medias que los representan. Al comparar las medias, la variable utilizada para medir o clasificar a las unidades de observación debe soportar la aplicación de la función estadística media. Debe ser numérica o considerada numérica.

Los casos que se presentan son: comparación de una media con el parámetro de una población, comparación de dos medias de muestras independientes, comparación de dos medias de muestras emparejadas.

15.2.1 Comparación de una media con el parámetro de una población

En el primer caso, se compara la media de los valores de un grupo o subgrupo de individuos o unidades de observación de una variable, con el parámetro o media poblacional. Los estadísticos utilizados son: *Z* y *t de Student*, el esquema de datos es una variable numérica o la parte de una variable numérica definida por la categoría de una variable categórica. En este caso, como se compara el estadístico de una muestra con el parámetro de una población, es necesario que este parámetro sea conocido.

Tabla 163 Datos iniciales.	
Media de la muestra	$\overline{X}$
Tamaño de la muestra	n
Desviación típica de la muestra	S
Varianza de la muestra	S^2
Parámetro media de la población	σ

El planteamiento es ver si un grupo de individuos con valores en una variable numérica o considerada numérica y supuestamente extraída de una población, pertenece a esa población, que es equivalente a ver si existen diferencias significativas entre la muestra y la población. Como las dos variables que se quieren comparar son numéricas, la comparación se puede hacer a través de sus medias. Para ver si dos cosas son iguales o distintas obtenemos la diferencia simple,

$$\overline{X} 4 \sigma$$

Si la diferencia es cero, significa que las dos medias son iguales,[89] entonces se puede concluir que no hay diferencias entre las medias y por lo tanto asumimos que la muestra ha sido extraída de esa población.

Si la diferencia es muy grande, significa que las dos medias son distintas y por lo tanto que esa muestra no pertenece a esa población y no ha sido extraída de ella. La cuestión ahora es que entre la diferencia igual a cero y la diferencia muy grande, cuando una diferencia que no es cero pero es pequeña ¿También se puede considerar que son iguales? Entonces ¿Hasta qué valor la diferencia se puede considerar pequeña? o ¿Desde que valor la diferencia se puede considerar grande como para que sean distintas? La pregunta entonces es que cuál es el límite en el que la diferencia deja de ser pequeña y empieza a ser grande.

[89] No se cuestiona la posibilidad del error o azar. Lo único que se pretende es seguir el razonamiento del contraste de hipótesis de diferencia de medias de una muestra respecto del parámetro de una población.

Otro problema es que al estar influida la diferencia por la unidad de medida de la variable, se puede alterar la magnitud de la diferencia modificando la unidad de medida. Por ejemplo, si se mide el peso de una muestra de individuos y se obtiene el valor de 65,00 kg. y el valor de la media de la población es de 65,90 *Kg.*, la diferencia es,

$$65{,}00\ \text{Kg} - 65{,}90\ \text{Kg} = -0{,}10\ \text{Kg}$$

Para expresar con un número mayor la diferencia, sólo hay que cambiar la unidad de medida y convertirla en *g*.

$$65.000{,}00\ \text{g} - 65.900{,}00\ \text{g} = -100{,}00\ \text{g}$$

El número 100,00 es mayor que 0,10, pero la diferencia en peso sigue siendo la misma. Para comparar dos valores y que no afecte la unidad de medida, se estandariza o tipifica la diferencia y proporciona la probabilidad asociada a la diferencia. Esta estandarización es *Z* o *t* de *Student*. El criterio de estandarización *Z* es,

$$z_i = \frac{x_i - \overline{X}_X}{S_X}$$

Es la diferencia de un valor *x* de una variable *X,* respecto de la media de la variable y dividido por la desviación típica (ver epígrafe 11).

Al comparar la media de una variable con la media de una población, se asume que la variable se ha extraído de esa población, por lo tanto la media pertenece a una distribución de medias muestrales de esa población y según el *teorema del límite central* (ver epígrafe 14), la media de la distribución de las medias muestrales es igual a la media de la población y la desviación típica es la raíz cuadrada de la varianza de la variable dividido por el número de casos de la variable. Simbólicamente,

$$\overline{X}_{\bar{x}} = \sigma$$

$$S_{\bar{x}} = \sqrt{\frac{S_x^2}{n_x}}$$

Si se sustituyendo en *z* y se llama z_e (*z estimada*), entonces,

$z_e = \dfrac{\overline{X} - \sigma}{\sqrt{\dfrac{S_x^2}{n_x}}}$	Fórmula 103

Si se aplica la *t* de *Student,* sería,

$t_e \mid \frac{\overline{X} 4 \sigma}{\sqrt{\frac{S_x^2}{n_x 4 1}}}$	Fórmula 104

William S. Gosset trabajaba como matemático en las destilerías Guinness. Comparando grupos de muestras pequeñas comprobó que el estadístico *Z* no se comportaba bien y elaboró un estadístico al que denominó *t*. Como la empresa no permitía escribir artículos a sus empleados, por temor a la filtración de información, Karl Pearson los publicaba con el seudónimo *Student*. La amistad de Gosset con Karl Pearson y Ronald A. Fisher influyeron en sus trabajos.

Entonces, como el estadístico *Z* se utiliza para muestras grandes y el estadístico *t* se utiliza con muestras grandes (se comporta igual que *Z,* ver Tabla 86) y con muestras pequeñas, superiores a 30 casos (para menos de 30 casos se usa la Estadística no Paramétrica, no prevista en este manual), entonces, en este manual, igual que los programas estadísticos, se aplicará sólo el estadístico *t*. La interpretación es la misma en ambos casos (ver epígrafe 11.3).

La comparación de la media de una muestra con la media de una población se hace con un contraste de hipótesis según el siguiente protocolo,

1. Hipótesis alternativa, H_1: $\overline{X} \prod \sigma$
2. Hipótesis nula, H_0: $\overline{X} \mid \sigma$
3. Estadístico: *t*
4. Criterio de aceptación/rechazo de H_0, $Ns = 0{,}05$, $gl = n - 1$ t_c = en función de *Ns* y *gl*.

La distribución *t* presenta una distribución variable en función del valor que toma *gl*, hasta que *n* se hace lo suficientemente grande (ver Anexo 3 y Tabla 86) y se comporta como *Z,* entonces la distribución *t* se estabiliza.

Ejemplo: Utilizando la muestra de CIRES de enero de 1996 sobre usos del tiempo, determinar si la muestra pertenece a la población española de 18 años y más en cuanto a la edad se refiere considerando que la media de edad de la población es de 45,05 años. Operar al *Ns* = 0,05. Los estadísticos de la variable edad se muestran en la Tabla 164

Tabla 164 Estadísticos de la variable edad.	
Encuesta CIRES: *Usos del tiempo.* Enero de 1996. Ámbito nacional. Población española de ambos sexos de 18 años o más	
σ	45,05 años
$\overline{X}_X$	44,95 años
S_X	18,33 años
S_X^2	335,97
n_X	1.200
$S_{\overline{X}}$	$\sqrt{\frac{S_X^2}{n_X}}$ años
$S_{\overline{X}}$ (*cpf*)[90]	$\sqrt{\frac{S_X^2}{n_X}} \Delta \sqrt{1 4 fm}$ años
$S_{\overline{X}}$	$\sqrt{\frac{335,97}{1.200}} \mid 0,53$ años

Protocolo de contraste de hipótesis,

1. Hipótesis alternativa, H_1: $\overline{X} \prod \sigma$
2. Hipótesis nula, H_0: $\overline{X} \mid \sigma$
3. Estadístico: *t*
4. Criterio de aceptación/rechazo de H_0, Ns = 0,05, gl = 1.200-1 t_c = 1,9623.

Aplicando el estadístico de contraste *t,*

$$t_e \mid \frac{\overline{X} 4 \sigma}{\sqrt{\frac{S_x^2}{n_x 4 1}}}$$

y sustituyendo,

[90] El *cpf* (corrector por poblaciones finitas) se aplica cuando la población se considera finita (N < 100.000).

$$t_e \mid \frac{44{,}95 \,4\, 45{,}05}{\sqrt{\frac{335{,}97}{1.200 \,4\, 1}}} \mid \; 40{,}19$$

Esquema de aceptación – rechazo de H_0.

Se acepta H_0 si: $\mid t_e \mid \{ \mid t_c \mid \Sigma Nc_e \{ \; Nc_c \; \Sigma Ns_e \; \} \; Ns_c$

Se rechaza H_0 si: $\mid t_e \mid \varnothing \mid t_c \mid \Sigma Nc_e \varnothing Nc_c \; \Sigma Ns_e \; \Omega Ns_c$

Como el valor de la *t* estimada es -0,19 que en valor absoluto es |0,19| y es menor que la *t* crítica (|1,9623|), por lo tanto el Nc_e es menor que el Nc_c y el Ns_e mayor que el Ns_c, entonces se puede asumir aceptar la H_0 al *Ns* de 0,05 de que la media de la muestra es significativamente igual a la media de la población, o que no existen diferencias significativas al *Ns* de 0,05 y por lo tanto se puede asumir que la muestra pertenece a esa población o que es representativa en cuanto a la variable edad se refiere.

Haciendo el contraste con el estadístico *Z* el proceso sería,

Protocolo de contraste de hipótesis,

1. Hipótesis alternativa, H_1: $\overline{X} \prod \sigma$
2. Hipótesis nula, H_0: $\overline{X} \mid \; \sigma$
3. Estadístico *Z*
4. Criterio de aceptación/rechazo de H_0, $Ns = 0{,}05$, $z_c = 1{,}96$.

Por redondeo, para una *n* de 1.200 y un grupo (una variable o lo que es lo mismo una media), el valor de la t_c, es igual al valor de la $z_c = 1{,}96$.

Aplicando el estadístico de contraste *Z,*

$$z_e \mid \frac{\overline{X} \,4\, \sigma}{\sqrt{\frac{S_x^2}{n_x}}}$$

y sustituyendo,

$$z_e \mid \frac{44{,}95 \,4\, 45{,}05}{\sqrt{\frac{335{,}97}{1.200}}} \mid \; 40{,}19$$

Esquema de aceptación – rechazo de H_0.

Se acepta H_0 si: $|t_e| \{ |t_c| \Sigma Nc_e \{ Nc_c \Sigma Ns_e \} Ns_c$

Se rechaza H_0 si: $|t_e| \varnothing |t_c| \Sigma Nc_e \varnothing Nc_c \Sigma Ns_e \Omega Ns_c$

Como el valor de la *z* estimada es -0,19 que en valor absoluto es |0,19| y es menor que la *z* crítica (|1,96|), por lo tanto el Nc_e es menor que el Nc_c y el Ns_e mayor que el Ns_c, entonces se puede asumir aceptar la H_0 al *Ns* de 0,05 de que la media de la muestra es significativamente igual a la media de la población, o que no existen diferencias significativas y por lo tanto se puede asumir que esa muestra pertenece a esa población o que es representativa en cuanto a la variable edad se refiere. El resultado y la interpretación es igual que con la *t*.

Si la media de la muestra hubiese sido de 44,00 años el valor de *t* habría sido de |1,98| que al *Ns* de 0,05 se habría rechazado la H_0 y la diferencia considerada significativa y no se podría asumir que esa muestra hubiese sido extraída de esa población, o que la muestra no sería representativa de la población o que la media de la población no estaría contenida en el intervalo de confianza obtenido a partir de la media de la muestra. Si se disminuye el error de que al rechazar una H_0 sea verdadera (error ζ), expresado como *Ns* = 0,01, la t_c sería 2,58 y entonces se procedería como en el caso anterior, se aceptaría la H_0.

15.2.2 Comparación de dos medias. Muestras independientes

Se denominan muestras independientes cuando una variable numérica o considerada numérica y considerada también como la dependiente, es agrupada por otra variable categórica, que es la de agrupamiento, considerada la independiente y que tiene dos categorías o se dicotomiza o sólo se usan dos categorías de ella. El esquema sería,

Tabla 165 Esquema de muestras independientes.

Muestra total	Submuestras	X	V
$\overline{X}_t, S_t^2, S_t, n_t$	Submuestra v_1 $\overline{X}_1, S_1^2, S_1, n_1$	x_{11}	v_1
		x_{21}	v_1
		x_{31}	v_1
		x_{41}	v_1
		x_{51}	v_1
		x_{61}	v_1
		x_{71}	v_1
		x_{81}	v_1
		x_{91}	v_1
		x_{101}	v_1
	Submuestra v_2 $\overline{X}_2, S_2^2, S_2, n_2$	x_{12}	v_2
		x_{22}	v_2
		x_{32}	v_2
		x_{42}	v_2
		x_{52}	v_2
		x_{62}	v_2
		x_{72}	v_2
		x_{82}	v_2
		x_{92}	v_2
		x_{102}	v_2

Para comparar los dos grupos que se han formado en la variable *X* al cruzarla con la variable *V,* como los subgrupos o submuestras de la variable *X* tienen valores numéricos, se puede hacer a través de la comparación de las medias y con el estadístico *Z* o *t.* Se procede sólo con la *t* considerando que con la *Z* se obtendrían los mismos resultados para *n's* grandes, y además tiene la misma interpretación.

Para comparar si dos cosas son iguales o distintas se procede por diferencia simple,

$$\overline{X}_1 4 \overline{X}_2$$

El estadístico es *t*, que según el *teorema del límite central* y simbólicamente,

$$t_e | \frac{\overline{X} 4 \sigma}{\sqrt{\frac{S_x^2}{n_x 4 1}}}$$

Pero ahora se considera el *teorema del límite central* aplicado a dos distribuciones de medias muestrales.

Sean dos poblaciones de las que extraemos *m* muestras, asumimos que se extraen el mismo número de cada una de ellas. Sean las muestras,

$$m_{x1}, m_{x2}, m_{x3} \ldots m_{xm}$$

de la población 1 que tienen de medias,

$$\overline{x}_1, \overline{x}_2, \overline{x}_3, 2 \overline{x}_m$$

y de tamaño,

$$n_{x1}, n_{x2}, n_{x3} \ldots n_{xm}$$

y sean las muestras,

$$m_{y1}, m_{y2}, m_{y3} \ldots m_{ym}$$

de la población 2 que tienen de medias,

$$\overline{y}_1, \overline{y}_2, \overline{y}_3, 2 \overline{y}_m$$

y de tamaño,

$$n_{y1}, n_{y2}, n_{y3} \ldots n_{ym}$$

Si se representan dos variables *X* e *Y* tal que sus valores sean las medias de las *m* muestras, se obtienen dos variables, $X_{\overline{x}}$ (variable de las medias muestrales de *X*) e $Y_{\overline{y}}$ (variable de las medias muestrales de *Y*), que son las distribuciones de las medias muestrales. Esquemáticamente.

Tabla 166 Variables de medias muestrales

$X_{\overline{x}}$	$Y_{\overline{y}}$
$\overline{x}_1$	$\overline{y}_1$
$\overline{x}_2$	$\overline{y}_2$
$\overline{x}_3$	$\overline{y}_3$
4	4
4	4
$\overline{x}_m$	$\overline{y}_m$

Según el *teorema del límite central* (ver epígrafe 14) para una distribución de medias muestrales, la media y desviación típica de las variables $X_{\bar{x}}$ e $Y_{\bar{y}}$, son,

Tabla 167 Estadísticos de las distribuciones de medias muestrales		
Estadísticos	$X_{\bar{x}}$	$Y_{\bar{y}}$
Media	$\overline{X}_{\bar{x}} \mid \sigma_x$	$\overline{Y}_{\bar{y}} \mid \sigma_y$
Desviación típica de las medias muestrales	$S_{\bar{x}} \mid \sqrt{\frac{S_x^2}{n_x}}$	$S_{\bar{y}} \mid \sqrt{\frac{S_y^2}{n_y}}$

Si se crea una nueva variable que sea la unión de las dos variables de las distribuciones de medias muestrales ($n = n_x + n_y$), la variable de las medias muestrales de X ($X_{\bar{x}}$) y la variable de las medias muestrales de Y ($Y_{\bar{y}}$), la desviación típica de la nueva variable se asume es,

$$S_{\overline{X}\overline{Y}} \mid \sqrt{\frac{\omega_x^2}{n_{\overline{X}}} 2 \frac{\omega_y^2}{n_{\overline{Y}}}}$$

Fórmula 105

Entonces, la desviación típica de la variable que es unión de las variables de la medias muestrales de X y las medias muestrales de Y, es la raíz cuadrada de la varianza de la población de la que se han extraído las muestras de media X_i dividido por el número de muestras de X, más la varianza de la población de la que se han extraído las muestras de media Y_i dividido por el número de muestras de Y.

Aplicando los criterios del Teorema del Límite Central a una sola población, la varianza de las poblaciones son desconocidas y el número de muestras obtenidas es teórico. Por lo tanto se asume que la desviación típica de la variable distribución de las medias muestrales de X e Y es,

$$S_{\overline{X}\overline{Y}} \mid \sqrt{\frac{S_1^2}{n_1} 2 \frac{S_2^2}{n_2}}$$

Fórmula 106

Es la raíz cuadrada de la varianza del grupo 1 dividido por el número de casos del grupo 1, más la varianza del grupo 2 dividido por el número de casos del grupo 2.

Para comparar las medias de dos submuestras, que se propone que pertenecen a dos subdistribuciones de medias muestrales, se puede hacer comparando la diferencia entre ellas con la diferencia entre las medias de las poblaciones, y sustituyendo en el estadístico *t*, entonces,

$$t_e = \frac{(\bar{X}_x - \bar{Y}_y) - (\sigma_x - \sigma_y)}{\sqrt{\frac{S_x^2}{n_x - 1} + \frac{S_y^2}{n_y - 1}}}$$

Fórmula 107

Asumiendo que son dos submuestras de una misma variable, entonces se hace el cambio,

$$t_e = \frac{(\bar{X}_1 - \bar{X}_2) - (\sigma_1 - \sigma_2)}{\sqrt{\frac{S_1^2}{n_1 - 1} + \frac{S_2^2}{n_2 - 1}}}$$

Fórmula 108

Que es la diferencia de dos medias de la muestra de medias muestrales, respecto a la diferencia de las medias poblacionales, dividido por la desviación típica de las medias muestrales. Si se asume la H_0 de que las medias de las poblaciones son iguales, la diferencia de las medias poblacionales es cero y entonces,

$$t_e = \frac{(\bar{X}_1 - \bar{X}_2) - (0)}{\sqrt{\frac{S_1^2}{n_1 - 1} + \frac{S_2^2}{n_2 - 1}}}$$

Fórmula 109

Por lo que el estadístico *t* para contrastar, a través de sus medias, si dos muestras independientes son iguales o lo que es lo mismo, si pertenecen a la misma población por lo que las dos poblaciones serían la misma, es,

$$t_e = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1 - 1} + \frac{S_2^2}{n_2 - 1}}}$$

Fórmula 110

Ejemplo: Se quiere construir una central nuclear en una zona geográfica determinada, entre dos municipios. Se pretende hacer en el término municipal que esté más a favor de la construcción. Se diseña una muestra representativa a los dos municipios y una de las preguntas del cuestionario pretende medir la actitud de los habitantes de los municipios con una escala de intensidad de 0 (está en contra de la construcción de la central nuclear) a 10 (está a favor de la construcción de la central nuclear). Operando con un *Ns* = 0,05 determinar si se puede asumir que un municipio estuviese a favor de construir la central nuclear y cuál sería. Los datos son,

Tabla 168 Estadísticos de la actitud hacia la construcción de una central nuclear.			
Estadísticos	Municipio 1	Municipio 2	Diferencia de grupos
Media	$\overline{X}_1 \mid 7{,}20$	$\overline{X}_2 \mid 6{,}61$	$\overline{X}_{dif} \mid 0{,}59$
Varianza	$S_1^2 \mid 3{,}49$	$S_2^2 \mid 3{,}85$	
Número de casos	$n_1 = 300$	$n_2 = 300$	$n_t = 600$
Desviación típica de las medias muestrales	$S_{\overline{X}1} \mid 0{,}11$	$S_{\overline{X}2} \mid 0{,}11$	$S_{\overline{XY}} \mid 0{,}16$

En donde la media de la variable diferencia y la desviación típica de la diferencia de las medias es,

$$\overline{X}_{dif} \mid \overline{X}_1 4 \overline{X}_2 \mid 7{,}20 4 6{,}61 \mid 0{,}59$$

$$S_{dif} \mid \sqrt{\frac{S_1^2}{n_1} 2 \frac{S_2^2}{n_2}} \mid \sqrt{\frac{3{,}49}{300} 2 \frac{3{,}85}{300}} \mid 0{,}16$$

El error típico de cada grupo es,

$$S_{\overline{X}1} \mid \sqrt{\frac{S_1^2}{n_1}} \mid \sqrt{\frac{3{,}49}{300}} \mid 0{,}11$$

$$S_{\overline{X}2} \mid \sqrt{\frac{S_2^2}{n_2}} \mid \sqrt{\frac{3{,}85}{300}} \mid 0{,}11$$

Para determinar si se presentan diferencias significativas en la valoración de la actitud hacia la construcción de una central nuclear en los municipios, se comparan las medias con el estadístico *t* de diferencias de medias de muestras independientes (Fórmula 110) mediante un protocolo de contraste de hipótesis,

Protocolo de contraste de hipótesis:

1. Hipótesis alternativa, H_1: *Hay diferencias por municipio en la actitud hacia la construcción de una central nuclear*. Simbólicamente,

$$\overline{X}_1 \prod \overline{X}_2$$

Y si la muestra es representativa, inferimos a la población,

$$\sigma_1 \prod \sigma_2$$

2. Relación entre las variables: *Municipio*: la de agrupamiento y supuesta independiente. *Valoración*: la agrupada y supuesta dependiente.
3. Nivel de medida de las variables: Municipio: categórica (2 categorías). *Valoración*: supuestamente numérica.

4. Hipótesis nula, H_0: *No hay diferencias por municipio en la actitud hacia la construcción de una central nuclear*. Simbólicamente,

$$\overline{X}_1 \mid \overline{X}_2$$

Y si la muestra es representativa, inferimos a la población,

$$\sigma_1 \mid \sigma_2$$

5. Estadístico: *t* de muestras independientes.
6. Criterio de aceptación o rechazo de H_0, $Ns = 0{,}05$.

$gl = (n_1 - 1) + (n_2 - 1) = (300\text{-}1) + (300\text{-}1) = 299+299 = 598$ ó $gl = \text{N-2} = 600 - 2 = 598$; $t_c = 1{,}96$.

Aplicando el estadístico de contraste y sustituyendo,

$$t_e \mid \frac{\overline{X}_1 4 \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1 4 1} 2 \frac{S_2^2}{n_2 4 1}}} \mid \frac{7{,}20 4 6{,}61}{\sqrt{\frac{3{,}49}{300 4 1} 2 \frac{3{,}85}{300 4 1}}} \mid \frac{0{,}59}{0{,}16} \mid 3{,}69$$

Esquema de aceptación – rechazo de H_0.

Se acepta H_0 si: $\mid t_e \mid \{ \mid t_c \mid \sum Nc_e \{ Nc_c \sum Ns_e \} Ns_c$

Se rechaza H_0 si: $\mid t_e \mid \not\varnothing \mid t_c \mid \sum Nc_e \varnothing Nc_c \sum Ns_e \Omega Ns_c$

Como la diferencia estandarizada entre las medias de las muestras independientes, esto es, la t_e (|3,69|) es mayor que la tc (|1,96|), entonces se puede asumir rechazar la H_0, de que existen diferencias significativas entre las medias al *Ns* de 0,05 y por lo tanto que hay diferencias en la actitud hacia la construcción de la central nuclear en los términos municipales. Como la media del municipio *A* es mayor que la del municipio *B*, se puede asumir que el municipio *A* es más favorable. Pero no significa que estén a favor o en contra o que el municipio *B* esté en contra. Como la media en los dos municipios es superior a 5,00, por el sentido cultural de la escala de intensidad, se puede entender que en los dos casos están a favor, o por lo menos que no están en contra de la construcción de una central nuclear. Para complementar esta información, también se pueden utilizar otras variables u otra información adicional que ayuden en la toma de la decisión.

15.2.3 Comparación de dos medias. Muestras emparejadas[91]

El esquema de muestras emparejadas se produce cuando para los mismos individuos o controlando los individuos, se tienen dos o más tomas o mediciones en una variable del mismo tipo y de la misma unidad de medida y entre las que media algún estímulo del que se quiere medir su efecto. Al ser el estadístico que se trata ahora de diferencia de medias, sólo se

[91] La diferencia de proporciones de muestras emparejadas se presenta como un contraste de hipótesis "de cambio" y en este manual se va a seguir la misma linea. No obstante, en realidad es un contraste de diferencia de proporciones, porque si el valor de *a* o el de *c* en la Tabla 162 es cero, no se detecta cambio (Glass & Stanley, 1980). No obstante, en las demás ocasiones la diferencia de proporciones detecta cambio.

puede operar con dos variables. El caso de más de dos variables emparejadas se trata con Análisis de Varianza Múltiple (MANOVA) o Medidas Repetidas, que no se ven en este manual.

Supuesto un grupo o muestra de individuos de los que se quiere saber su actitud ante la construcción de una central nuclear, se les administra una escala de intensidad de 0 a 10 en la que el cero es *en contra* y el 10 es *a favor* de la construcción de la central, considerando ésta la *toma 1*. Después se les muestra un documental sobre la contaminación que producen las centrales nucleares. Este acto se considera el *efecto.* Después del efecto se realiza la *toma 2*. El esquema sería,

Tabla 169 Tabla de datos Tipo I de muestras emparejadas.

Caso		Toma 1	Efecto	Toma 2	
1		X_{11}		X_{12}	
2		X_{21}		X_{22}	
3		X_{31}		X_{32}	
.	$\overline{X}_1, S_1^2, n_1$	.	Documental	.	$\overline{X}_2, S_2^2, n_2$
.		.		.	
n-1		X_{n-11}		X_{n-12}	
n		X_{n1}		X_{n2}	

Para cada una de las variables se tiene la media, varianza y tamaño de la muestra, y este es el mismo en las dos variables. Esta es una característica que siempre se cumple en las muestras emparejadas porque si algún caso falla en una de las *tomas*, automáticamente se pierde el otro valor por falta del elemento de comparación.

Para comparar las dos variables que han producido las dos *mediciones* o *tomas,* como las dos variables tienen valores numéricos, se puede hacer a través de la comparación de las medias y el estadístico sería *Z* o *t.* Se procede sólo con la *t* considerando que con la *Z* se obtendrían los mismos resultados para muestras grandes y con la misma interpretación.

Para comparar si dos cosas son iguales se procede por diferencia simple,

$$\overline{X}_1 \, 4 \, \overline{X}_2$$

El estadístico es *t*, simbólicamente,

$$t_e \mid \frac{\overline{X} \, 4 \, \sigma}{\sqrt{\frac{S_x^2}{n_x \, 4 \, 1}}}$$

Fórmula 111

Pero ahora en vez del *teorema del límite central* para una sola distribución de medias muestrales, es el *teorema del límite central* aplicado a dos distribuciones de medias muestrales.

Sean dos poblaciones de las que se extraen *m* muestras, el mismo número de cada una de ellas. Sean las muestras,

$$m_{x1},\ m_{x2},\ m_{x3} \ldots\ m_{xm}$$

de la población 1 que tienen de medias,

$$\bar{x}_1, \bar{x}_2, \bar{x}_3, 2\ \bar{x}_m$$

y de tamaño,

$$n_{x1},\ n_{x2},\ n_{x3} \ldots\ n_{xm}$$

y sean las muestras,

$$m_{y1},\ m_{y2},\ m_{y3} \ldots\ m_{ym}$$

de la población 2 que tienen de medias,

$$\bar{y}_1, \bar{y}_2, \bar{y}_3, 2\ \bar{y}_m$$

y de tamaño,

$$n_{y1},\ n_{y2},\ n_{y3} \ldots\ n_{ym}$$

Asumiendo que existe emparejamiento, si se representan dos variables X e Y tal que sus valores sean las medias de las m muestras, se obtienen dos variables, la variable X de las medias muestrales de X ($X_{\bar{x}}$) y la variable Y de las medias muestrales de Y ($Y_{\bar{y}}$). Si se calcula una tercera variable que sea la diferencia de las dos variables anteriores, caso a caso, se obtiene la distribución de la variable $W_{\overline{dif}}$ que es la diferencia de las variables de las medias muestrales. Esquemáticamente.

Tabla 170 Variable diferencia de dos variables de medias muestrales

$X_{\bar{x}}$	$Y_{\bar{y}}$	$X_{\bar{x}}$ 4 $Y_{\bar{y}}$	$W_{\overline{dif}}$
$\bar{x}_1$	$\bar{y}_1$	$\bar{x}_1$ 4 $\bar{y}_1$	$\bar{w}_1$
$\bar{x}_2$	$\bar{y}_2$	$\bar{x}_2$ 4 $\bar{y}_2$	$\bar{w}_2$
$\bar{x}_3$	$\bar{y}_3$	$\bar{x}_3$ 4 $\bar{y}_3$	$\bar{w}_3$
4	4	4	4
4	4	4	4
$\bar{x}_m$	$\bar{y}_m$	$\bar{x}_m$ 4 $\bar{y}_m$	$\bar{w}_m$

Según el *teorema del límite central* (ver epígrafe 14) para una distribución de medias muestrales, la media y desviación típica de las variables $X_{\bar{x}}$ e $Y_{\bar{y}}$, y aplicado a $W_{\overline{dif}}$, la diferencia de las dos variables, es

Tabla 171 Estadísticos de las distribuciones de medias muestrales

Estadísticos	$X_{\bar{x}}$	$Y_{\bar{y}}$	$W_{\overline{dif}}$
Media	$\overline{X}_{\bar{x}} \mid \sigma_x$	$\overline{Y}_{\bar{y}} \mid \sigma_y$	$\overline{W}_{dif} \mid /\overline{X}_{\bar{x}}$ 4 $\overline{Y}_{\bar{y}}0 \mid /\sigma_x$ 4 $\sigma_y($
Desviación típica de las medias muestrales	$S_{\bar{x}} \mid \sqrt{\frac{S_x^2}{n_x}}$	$S_{\bar{y}} \mid \sqrt{\frac{S_y^2}{n_y}}$	$S_{\overline{dif}} \mid \sqrt{\frac{S_{dif}^2}{n_{dif}}}$

Sustituyendo en la Fórmula 111, la diferencia entre dos medias pertenecientes a dos distribuciones de medias muestrales entre las que existe emparejamiento, sería respecto a la diferencia de las medias poblacionales,

$$t_e = \frac{(\overline{X}_x - \overline{Y}_y) - (\sigma_x - \sigma_y)}{\sqrt{\frac{S^2_{dif}}{n_{dif} - 1}}}$$

Fórmula 112

Que es la diferencia entre las medias de las variables, menos la diferencia de las medias correspondientes a las poblaciones, dividido por la desviación típica. Si se asume la H_0, de que las medias de las poblaciones son iguales, la diferencia de las medias de la poblaciones es cero, y por la propiedad 1 de la media (ver epígrafe 7.1.3.1) al restar la media de la variable *Y* a la media de la variable *X,* como la media sería el valor que tienen todos los casos si todos tuviesen el mismo valor, es como si a todos los valores de *X* se les resta la media de *Y*. La nueva variable obtenida tiene la media de *X* menos la constante media de *Y*. Entonces la media de una variable que es la diferencia de otras dos, es igual a la diferencia de las medias de las dos variables originales, entonces,

$$t_e = \frac{\overline{W}_{dif} - 0}{\sqrt{\frac{S^2_{dif}}{n_{dif} - 1}}}$$

Fórmula 113

Por lo que el estadístico *t* para contrastar, a través de sus medias, si dos muestras emparejadas son iguales o lo que es lo mismo, si pertenecen a la misma población por lo que las dos poblaciones serían la misma es,

$$t_e = \frac{\overline{W}_{dif}}{\sqrt{\frac{S^2_{dif}}{n_{dif} - 1}}}$$

Fórmula 114

Ejemplo: Continuando el ejemplo de la actitud ante la construcción de una central nuclear, antes y después de visionar un documental con los efectos contaminantes que tienen. Para los datos de la *Toma 1,* antes de ver el documental y de la *Toma 2*, después de ver el documental, se obtienen los estadísticos de la Tabla 172.

Tabla 172 Estadísticos de la actitud hacia la construcción de una central nuclear.

Estadísticos	Toma1	Toma 2	Variable diferencia
Media	6,30	6,72	-0,42
Varianza	3,61	2,88	6,78
Número de casos	100	100	100
Desviación típica de las medias muestrales			0,26

El esquema de datos es,

Tabla 173 Tabla de datos de Tipo 1 de muestras emparejadas		
X_1	X_2	X_{dif}
x_{11}	x_{12}	x_{11}-x_{12}
x_{21}	x_{22}	x_{11}-x_{12}
x_{31}	x_{32}	x_{11}-x_{12}
4	4	4
4	4	4
x_{n1}	x_{n2}	x_{n1}-x_{n2}

En donde la media de la variable diferencia y la desviación típica de las medias muestrales son,

$$\overline{X}_{dif} \mid \overline{toma1} \,4\, \overline{toma2} \mid 6{,}30 \,4\, 6{,}72 \mid 40{,}42$$

$$S_{dif} \mid \sqrt{\frac{S_{dif}^2}{n_{dif}}} \mid \sqrt{\frac{6{,}78}{100}} \mid 0{,}26$$

Para determinar la influencia que ha producido el documental, se compara la *toma 1* con la *toma 2,* con el estadístico *t* de diferencias de medias de muestras emparejadas (Fórmula 114) mediante un protocolo de contraste de hipótesis, entendiendo que las diferencias que se detecten entre las dos tomas han sido debidas al visionado del documental, porque todas las demás posibles variables o efectos permanecen constantes (*Céteris páribus*).

Protocolo de contraste de hipótesis:

1. Hipótesis alternativa, H_1: *El documental presentado influye en la actitud ante la construcción de una central nuclear*. Simbólicamente,

   $$\overline{toma1} \prod \overline{toma2} \;\heartsuit\; \overline{X}_1 \prod \overline{X}_2$$

   Y si la muestra es representativa, se infiere a la población,

   $$\sigma_1 \prod \sigma_2$$

2. Nivel de medida de las variables: *Numéricas.*
3. Relación entre las variables: No procede, son muestras emparejadas o relacionadas.
4. Hipótesis nula, H_0: *El documental presentado no influye en la actitud ante la construcción de una central nuclear*. Simbólicamente,

   $$\overline{toma1} \mid \overline{toma2} \;\heartsuit\; \overline{X}_1 \mid \overline{X}_2$$

   Y si la muestra es representativa, se infiere a la población,

   $$\sigma_1 \mid \sigma_2$$

5. Estadístico: *t* de muestras emparejadas.

6. Criterio de aceptación o rechazo de H_0, Ns = 0,05; $gl = n - 1 = 99$; t_c = 1,98.

Aplicando el estadístico de contraste y sustituyendo,

$$t_e \mid \frac{\overline{W}_{dif}}{\sqrt{\frac{S^2_{dif}}{n_{dif}\ 4\ 1}}} \mid \frac{4\ 0{,}42}{\sqrt{\frac{6{,}78}{100\ 4\ 1}}} \mid \frac{4\ 0{,}42}{0{,}26} \mid\ 41{,}62$$

Esquema de aceptación – rechazo de H_0.

Se acepta H_0 si: $\mid t_e \mid \{ \mid t_c \mid \sum Nc_e \{\ Nc_c\ \sum Ns_e\ \}\ Ns_c$

Se rechaza H_0 si: $\mid t_e \mid \varnothing \mid t_c \mid \sum Nc_e\ \varnothing Nc_c\ \sum Ns_e\ \Omega Ns_c$

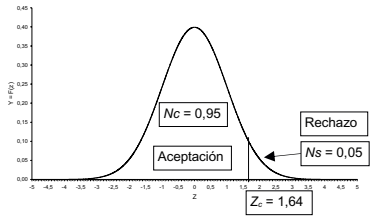
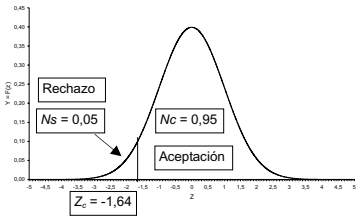
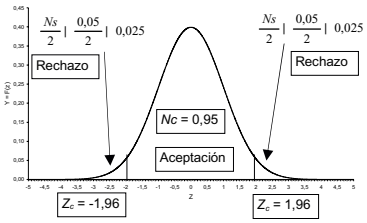
Como la diferencia estandarizada entre las medias de las muestras emparejadas o relacionadas, esto es, la t_e (|1,62|) es menor que la t_c (|1,98|), entonces se puede asumir la aceptación de la H_0, de que no existen diferencias significativas entre las medias al Ns de 0,05 y por lo tanto la actitud hacia la construcción de la central nuclear no ha variado significativamente después de ver el documental. Lo que no se sabe es si la actitud es a favor o en contra. Como la media en las dos tomas es superior a 5,00, por el sentido cultural de la escala de intensidad, se puede entender que están a favor, o por lo menos que no están en contra de la construcción de una central nuclear y la actitud no se ha modificado con el visionado del documental.

15.3 Contraste de hipótesis bilaterales y unilaterales

Un contraste de hipótesis puede ser bilateral o unilateral. El bilateral plantea o propone diferencias por desigualdad sin especificar quién es mayor o menor, el sentido en el que se produce la diferencia se ve después del contraste. Un contraste unilateral propone diferencias por *mayor* o *menor* antes de realizar el cálculo estadístico. En el contraste bilateral con las distribuciones tipificadas Z y t el Ns se reparte por igual entre las dos colas porque su recorrido va de -← a +←, y puede tomar valores positivos o negativos. Si el contraste es unilateral, entonces se plantea que "es mayor que" o "es menor que" y el Ns va en una de las dos colas.

En el caso "es mayor que" la diferencia es positiva y entonces el Ns se deja en la zona derecha de la cola de la distribución. Si el planteamiento es "es menor que" el Ns se especifica en la cola de la izquierda. El planteamiento se ve en la Tabla 174.

Tabla 174 Planteamiento del contraste de hipótesis bilateral y unilateral.

Contraste unilateral a la derecha	Contraste unilateral a la izquierda	Contraste bilateral
		
H_1: $fo \varnothing fe$ H_0: $fo \mid fe$	H_1: $fo \Omega fe$ H_0: $fo \mid fe$	H_1: $fo \prod fe$ H_0: $fo \mid fe$
H_1: $\overline{X}_1 \varnothing \overline{X}_2$ H_0: $\overline{X}_1 \mid \overline{X}_2$	H_1: $\overline{X}_1 \Omega \overline{X}_2$ H_0: $\overline{X}_1 \mid \overline{X}_2$	H_1: $\overline{X}_1 \prod \overline{X}_2$ H_0: $\overline{X}_1 \mid \overline{X}_2$
En donde: En el caso de tablas de contingencia: *fo*: frecuencia observada, *fe*: frecuencia esperada.		
Se acepta H_0 si: $z_e \{ z_c \sum Nc_e \{ Nc_c \sum Ns_e \} Ns_c$ Se rechaza H_0 si: $z_e \varnothing z_c \sum Nc_e \varnothing Nc_c \sum Ns_e \Omega Ns_c$	Se acepta H_0 si: $z_e \} z_c \sum Nc_e \{ Nc_c \sum Ns_e \} Ns_c$ Se rechaza H_0 si: $z_e \Omega z_c \sum Nc_e \varnothing Nc_c \sum Ns_e \Omega Ns_c$	como es bilateral, es para –*z* y +*z* y usamos |z|. Se acepta H_0 si: $\mid z_e \mid \{ \mid z_c \mid \sum Nc_e \{ Nc_c \sum Ns_e \} Ns_c$ Se rechaza H_0 si: $\mid z_e \mid \varnothing \mid z_c \mid \sum Nc_e \varnothing Nc_c \sum Ns_e \Omega Ns_c$

15.4 Análisis de varianza

Cuando se quieren comparar grupos formados en una variable considerada la dependiente, numérica y agrupada, según las categorías de otra variable considerada la independiente, categórica y de agrupamiento, pero que tiene más de dos categorías, se utiliza el análisis de varianza (*Oneway*) (Ver epígrafe 15). El estadístico *t* de diferencia de medias está diseñado para comparar dos medias porque opera con diferencias, y las diferencias aceptan sólo dos términos, de dos grupos. Al existir tres grupos (*a, b* y *c*) se podrían hacer múltiples comparaciones (*a* con *b*, *a* con *c* y *b* con *c*) pero el error que se asume al hacer un contraste de hipótesis, se acumularía con los tres contrastes. Si el número de grupos es mayor, el error acumulado sería mayor.

La comparación entre más de dos grupos se realiza utilizando los estadísticos de cada grupo y de la muestra total ($\overline{X}_i; S_i^2; n_i; \overline{X}_T; S_T^2; N_T$). El procedimiento, indicado como análisis de varianza, quiere decir comparar si existe diferencia significativa entre los grupos, a través de sus medias, por descomposición de la varianza. Como se ha indicado en otro lugar, es una técnica estadística específica de diseños experimentales, pero su inclusión obedece más al interés de desarrollar el significado de *descomposición de la varianza* por ser la base de conceptos de muestreo e intervenir en las Técnicas Multivariable.

En este epígrafe se va a exponer el procedimiento del concepto estadístico *descomposición de la varianza*.

Se considera una variable numérica y una variable categórica de tres categorías. El cruce de la variable numérica con la variable categórica produce tres grupos en la variable numérica, uno por cada categoría de la variable categórica. Como regla general, el cruce de una variable numérica con una variable categórica, produce tantos grupos en la variable numérica, como categorías tiene la variable categórica. Esquemáticamente,

Tabla 175 Esquema de tres muestras independientes.

Muestra total	Submuestras	X	V
$\overline{X}_T, S_T{}^2, S_T, n_T$	Submuestra v_1 $\overline{X}_1, S_1^2, S_1, n_1$	x_{11}	v_1
		x_{21}	v_1
		x_{31}	v_1
		x_{41}	v_1
		x_{51}	v_1
		x_{61}	v_1
		x_{71}	v_1
		x_{81}	v_1
		x_{91}	v_1
		x_{101}	v_1
	Submuestra v_2 $\overline{X}_2, S_2^2, S_2, n_2$	x_{12}	v_2
		x_{22}	v_2
		x_{32}	v_2
		x_{42}	v_2
		x_{52}	v_2
		x_{62}	v_2
		x_{72}	v_2
		x_{82}	v_2
		x_{92}	v_2
		x_{102}	v_2
	Submuestra v_3 $\overline{X}_3, S_3^2, S_3, n_3$	x_{13}	V_3
		x_{23}	v_3
		x_{33}	v_3
		x_{43}	v_3
		x_{53}	v_3
		x_{63}	v_3
		x_{73}	v_3
		x_{83}	v_3
		x_{93}	v_3
		x_{103}	v_3

Cada grupo tiene su media, varianza, tamaño, así como la media, varianza y tamaño de la muestra total que es la suma de los tamaños de los subgrupos. Para comparar los grupos entre sí, se procede mediante el planteamiento de las hipótesis estadísticas, la H_1 (hipótesis alternativa) y la H_0 (hipótesis nula). El protocolo sería,

1. Hipótesis alternativa, H_1: *La variable considerada independiente, influye en la variable considerada dependiente*. Que es lo mismo que decir que, simbólicamente,

   $$\overline{X}_1 \prod \overline{X}_2 \prod \overline{X}_3$$

   Y si la muestra es representativa, se infiere a la población,

   $$\sigma_1 \prod \sigma_2 \prod \sigma_3$$

2. Relación entre las variables: Una supuesta *independiente* y otra supuesta *dependiente.*
3. Nivel de medida de las variables: la variable propuesta como independiente es *categórica* de más de dos categorías, y la propuesta como dependiente es considerada *numérica*.
4. Hipótesis nula, H_0: *La variable considerada independiente, no influye en la variable considerada dependiente*. Que es lo mismo que decir que, simbólicamente,

   $$\overline{X}_1 \mid \overline{X}_2 \mid \overline{X}_3$$

   Y si la muestra es representativa, se infieren a la población,

   $$\sigma_1 \mid \sigma_2 \mid \sigma_3$$

5. Estadístico: F_S. El estadístico es la F de Fisher-Snedecor, que suele ser expresada habitualmente sólo como F.

6. Criterio de aceptación o rechazo de H_0, $Ns = 0{,}05$.

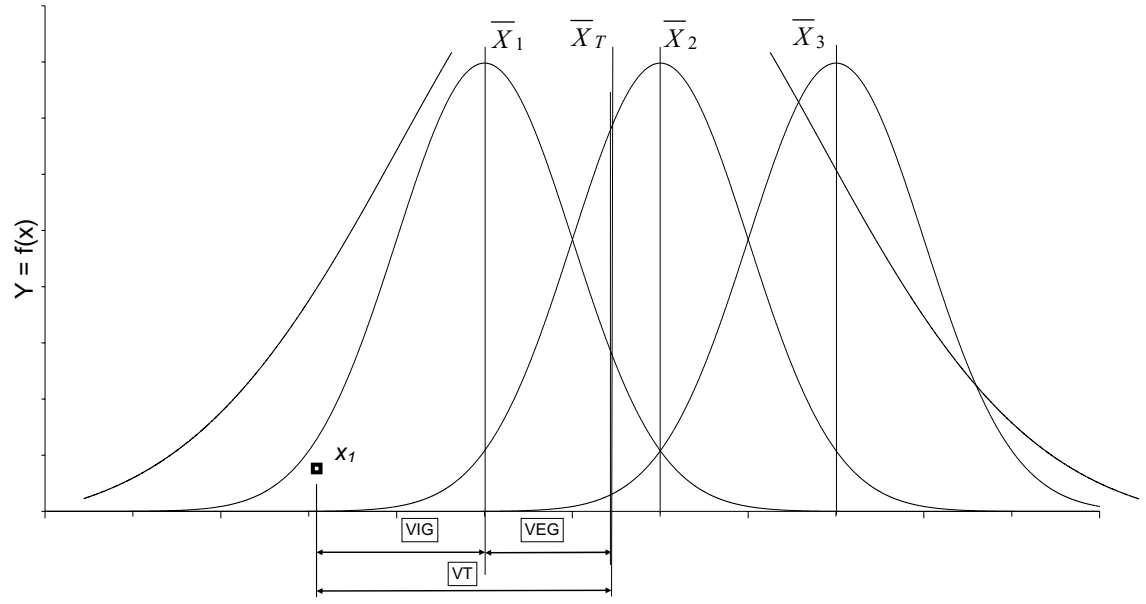
   Par el cálculo de la F crítica (F_c) se utilizan gl del numerador = k-1, y gl del denominador = N-k. Siendo k el número de grupos y N el tamaño de la muestra total.

La finalidad de este protocolo es el mismo que se ha aplicado hasta ahora con los otros estadísticos, la diferencia entre ellos ha sido la redacción de la hipótesis y el estadístico aplicado. Se instrumenta el protocolo para aceptar o rechazar la H_0, y en base a ello aceptar o rechazar la H_1. En esta ocasión, aceptar la H_0, supone que las tres medias son iguales, pero si se rechaza, al aceptar la H_1, al ser tres medias, es necesario precisar si las diferencias significativas se producen porque las tres son distintas o sólo algunas dos de ellas.

Para averiguarlo, se dispone de estadísticos denominados de contraste "*post-hoc*" que permiten comparaciones múltiples por parejas de medias. Accediendo al menú de algún programa estadístico, por ejemplo SPSS, se pueden ver las mencionadas pruebas. La recomendada como una de las más exigentes en detectar diferencias es *Scheffe*. Todas ellas se aplicarían con los mismos criterios, siguiendo el protocolo de contraste de hipótesis ya especificado. Estas pruebas son consideradas *no-paramétricas* lo que evita contemplar los requisitos de las pruebas *paramétricas,* que se verán a continuación.

Para ver el desarrollo de la descomposición de la varianza, si se asume la H_1, de que existen diferencias significativas entre las tres medias y por lo tanto que hay diferencias entre los tres grupos o que son distintos, gráficamente se puede representar (Tabla 176),

Tabla 176 Planteamiento de descomposición de la Varianza.



En donde:

x_1: Valor del caso 1 en la variable numérica *X*.
$\overline{X}_1$: Media del grupo 1.
$\overline{X}_2$: Media del grupo 2.
$\overline{X}_3$: Media del grupo 3.
$\overline{X}_T$: Media de la variable *X*.
VIG: Variación dentro del grupo.
VEG:Variación entre los grupos.
VT: Variación total.

Iniciando el proceso desde el concepto de la *varianza,* simbólicamente,

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{X})^2}{n}$$

Que aplicado a este caso, la *n* se convierte en el número total de casos de la muestra o la suma de las *n's* de las submuestras y se expresa como *N* y la media se refiere a la media total, por lo que la expresión de la *varianza* queda,

$$S^2 = \frac{\sum_{i=1}^{N} (x_i - \overline{X}_T)^2}{N}$$

Si a la distancia o diferencia $(x_i - \overline{X}_T)$ se denomina *variación total* (*VT*) y se aplica al gráfico de la Tabla 176 y considerando sólo el caso x_1, entonces la distancia del caso x_1 a la media total, es la dispersión o variación total de ese caso a la media total de la muestra. La variación total se puede dividir o descomponer en dos partes, la distancia del caso a la media de su grupo, llamada *variación intragrupo* (*VIG*), y la distancia de la media de su grupo a la media total, llamada *variación entregrupos* (*VEG*). De tal manera que para ese caso, la *variación intragrupo* más la *variación entregrupos* es igual a la *variación total*, simbólicamente,

$$VT = VIG + VEG$$

Y para ese caso 1, cada uno de los términos es,

$$VT = (x_1 - \overline{X}_T)$$
$$VIG = (x_1 - \overline{X}_1)$$
$$VEG = (\overline{X}_1 - \overline{X}_T)$$

Sustituyendo,

$$(x_1 - \overline{X}_T) = (x_1 - \overline{X}_1) + (\overline{X}_1 - \overline{X}_T)$$

Se generaliza a todos los casos aplicando sumatorios,

$$\sum_{i=1}^{N} (x_i - \overline{X}_T) = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_i - \overline{X}_j) + (\overline{X}_j - \overline{X}_T)$$

Pero por la propiedad nº 5 de la media (Ver epígrafe 7.1.3.1), el primer término de la igualdad es igual a cero y por lo tanto lo es también el término a la derecha del igual. Para

deshacer la igualdad a cero, se procede como en la *varianza,* se eleva al cuadrado y para mantener la igualdad, se eleva también el otro término,

$$\sum_{i=1}^{N}(x_i - \overline{X}_T)^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_i - \overline{X}_j)^2 + (\overline{X}_j - \overline{X}_T)^2$$

El término de la derecha, es un binomio del tipo $(a+b)^2$, y su desarrollo es $a^2 + b^2 + 2ab$ y queda,

$$\sum_{i=1}^{N}(x_i - \overline{X}_T)^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_i - \overline{X}_j)^2 + (\overline{X}_j - \overline{X}_T)^2 + 2(x_i - \overline{X}_j)(\overline{X}_j - \overline{X}_T)$$

Que se puede expresar como,

$$\sum_{i=1}^{N}(x_i - \overline{X}_T)^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_i - \overline{X}_j)^2 + \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\overline{X}_j - \overline{X}_T)^2 + 2\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_i - \overline{X}_j)(\overline{X}_j - \overline{X}_T)$$

Pero como el tercer sumando no está elevado al cuadrado, su suma es cero. Así es que se puede expresar,

$$\sum_{i=1}^{N}(x_i - \overline{X}_T)^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_i - \overline{X}_j)^2 + 2\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\overline{X}_j - \overline{X}_T)^2$$

En donde llamamos,

$SCT = \sum_{i=1}^{N}(x_i - \overline{X}_T)^2$	*SCT* = Suma de cuadrados total	Fórmula 115

$SCE = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\overline{X}_j - \overline{X}_T)^2$	*SCE* = Suma de cuadrados entregrupos	Fórmula 116

$SCI = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_i - \overline{X}_j)^2$	*SCI* = Suma de cuadrados intragrupos	Fórmula 117

En la *SCE,* la diferencia de la media de cada grupo a la media total se realiza tantas veces como casos hay en el grupo (n_j), entonces se puede expresar como,

$SCE \mid \sum_{j \mid 1}^{k} n_j \Delta(\overline{X}_j 4 \overline{X}_T)^2$	Fórmula 118

El primer término se denomina *suma de cuadrados total* (*SCT*) o variación total. La *suma de cuadrados intragrupos* (*SCI*) es la variabilidad del sistema debida a la dispersión que hay dentro de los grupos. Y la *suma de cuadrados entregrupos* (*SCE*) es la variabilidad del sistema debida a la dispersión que hay entre los grupos.

Entonces la *suma de cuadrados total* es igual a la *suma de cuadrados intragrupos* más la *suma de cuadrados entregrupos,* simbólicamente,

$SCT \mid SCI \; 2 \; SCE$	Fórmula 119

También se puede expresar como que la dispersión total del sistema es igual a la dispersión debida a la que hay dentro de los grupos más la dispersión debida a la que hay entregrupos. La división de cada uno de los términos por sus grados de libertad (*gl)*, se llama *media de cuadrados*, simbólicamente,

$MCE \mid \frac{SCE}{k \; 4 \; 1}$	Fórmula 120

La *media de cuadrados entre grupos* (*MCE*) es la división de la *SCE* entre sus grados de libertad, número de grupos menos 1 (*k*-1), y

$MCI \mid \frac{SCI}{N \; 4 \; k}$	Fórmula 121

Es la *media de cuadrados intragrupos* (*MCI*), que es la división de la *SCI* entre sus grados de libertad, el total de casos de la muestra menos el número de grupos (*N*-*k*).

Pues bien, el estadístico que sirve para ver a que se debe más la variación en un sistema como el expuesto, es la *F* de *Fisher-Snedecor,* simbólicamente F_S o *F* sólo, que es como suele aparecer y simbólicamente,

$F \mid \frac{MCE}{MCI}$	Fórmula 122

Y este estadístico se dice que sigue una distribución de tipo *F* (Ver Anexo 4) con los grados de libertad (*gl*) del numerador (*k*-1) y *gl* del denominador (*N*-*k*). Razonando intuitiva y genéricamente, para su aplicación estadístico-matemática posterior, si *F* es grande, es porque *MCE* es mayor que *MCI,* entonces podemos asumir que la variabilidad del sistema se debe más a la dispersión que hay entre los grupos y por lo tanto que se puede podemos asumir que existen diferencias significativas entre los grupos. Pero si el valor de *F* es pequeño, es porque

MCE es pequeño respecto de *MCI,* entonces podemos asumir que la variabilidad del sistema se debe más a la dispersión que hay dentro de los grupos y por lo tanto que se puede asumir que no hay diferencias significativas entre los grupos. Cómo se aplica *F,* qué es grande y a partir de cuánto se verá con un ejemplo.

Ejemplo 1: Se quiere comprobar la eficacia de dos métodos nuevos de enseñanza. Uno de ellos se aplica utilizando nuevas tecnologías en el aula y otro apoyando la enseñanza presencial con enseñanza a través de Internet fuera del aula. Se diseñan tres grupos seleccionados de forma aleatoria. Uno de ellos sigue la forma de enseñanza tradicional y se le considera *grupo control*, y los otros dos, *grupo experimental 1* y *grupo experimental 2.* Siendo el primero el que sigue el sistema de nuevas tecnologías en el aula y el segundo el asistido por Internet fuera del aula. La valoración al final se realiza a través de una prueba de conocimiento en una escala de 0 a 100, en la que el 0 es ausencia de conocimientos y el 100 el mayor nivel de conocimientos adquiridos. Los datos son,

Tabla 177 Estadísticos de grupo y total.

Grupos	n	Media	Desviación típica	Varianza
Control (1)	200	45,99	2,17	4,71
Nuevas Tecnologías (2)	200	47,11	2,57	6,60
Internet (3)	200	47,60	2,57	6,60
Muestra Total	600	46,90	2,53	6,40

El planteamiento es si el método de enseñanza seguido en cada uno de los tres casos se puede considerar que ha producido diferencias significativas en el conocimiento adquirido por los alumnos al *Ns* = 0,05. El proceso se realiza aplicando el protocolo de contraste de hipótesis,

1. Hipótesis alternativa, H_1: *El método de enseñanza influye en el aprendizaje*. Que es lo mismo que decir que, simbólicamente,

   $$\overline{X}_1 \prod \overline{X}_2 \prod \overline{X}_3$$

   Y si la muestra es representativa, inferimos a la población,

   $$\sigma_1 \prod \sigma_2 \prod \sigma_3$$

2. Relación entre las variables: *Método de enseñanza* considerada independiente y *valoración* considerada dependiente.
3. Nivel de medida de las variables: El *método de enseñanza* es *categórica* de más de dos categorías, y la *valoración* es considerada *numérica*.
4. Hipótesis nula, H_0: *El método de enseñanza no influye en el aprendizaje*. Que es lo mismo que decir que, simbólicamente,

   $$\overline{X}_1 \mid \overline{X}_2 \mid \overline{X}_3$$

   Y si la muestra es representativa, inferimos a la población,

   $$\sigma_1 \mid \sigma_2 \mid \sigma_3$$

5. Estadístico: F_S. El estadístico es la *F* de Fisher-Snedecor, que suele ser expresada habitualmente sólo como *F*.
6. Criterio de aceptación o rechazo de H_0, *Ns* = 0,05.

Con *gl* del numerador = *k*-1 (3-1 = 2), y *gl* del denominador = *N*-*k* (600-3=597) y el *Ns* = 0,05, la *F* crítica es 3,01 (F_c = 3,01).

El proceso de cálculo de la *F estimada* (F_e) es,

(Aplicando Fórmula 118)

$$SCE = \sum_{j=1}^{k} n_j (\overline{X}_j - \overline{X}_T)^2 = n_1 (\overline{X}_1 - \overline{X}_T)^2 + n_2 (\overline{X}_2 - \overline{X}_T)^2 + n_3 (\overline{X}_3 - \overline{X}_T)^2 =$$

$$200 (45{,}99 - 46{,}90)^2 + 200 (47{,}11 - 46{,}90)^2 + 200 (47{,}60 - 46{,}90)^2 = 272{,}44$$

(Aplicando Fórmula 117)

$$SCI = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_i - \overline{X}_j)^2 = (x_1 - \overline{X}_1)^2 + (x_2 - \overline{X}_1)^2 + (x_3 - \overline{X}_1)^2 + \ldots + (x_{n-1} - \overline{X}_1)^2 + (x_n - \overline{X}_1)^2 +$$

$$+ (x_1 - \overline{X}_2)^2 + (x_2 - \overline{X}_2)^2 + (x_3 - \overline{X}_2)^2 + \ldots + (x_{n-1} - \overline{X}_2)^2 + (x_n - \overline{X}_2)^2 +$$

$$+ (x_1 - \overline{X}_3)^2 + (x_2 - \overline{X}_3)^2 + (x_3 - \overline{X}_3)^2 + \ldots + (x_{n-1} - \overline{X}_3)^2 + (x_n - \overline{X}_3)^2 = 3.570{,}56$$

(Aplicando Fórmula 120)

$$MCE = \frac{SCE}{k - 1} = \frac{272{,}44}{3 - 1} = \frac{272{,}44}{2} = 136{,}22$$

(Aplicando Fórmula 121)

$$MCI = \frac{SCI}{N - k} = \frac{3.570{,}56}{600 - 3} = \frac{3.570{,}56}{597} = 5{,}98$$

(Aplicando Fórmula 122)

$$F = \frac{MCE}{MCI} = \frac{136{,}22}{5{,}98} = 22{,}78$$

Algunos programas estadísticos muestran los resultados en el formato de la Tabla 178.

Tabla 178 Análisis de varianza.

	Suma de cuadrados	*gl*	Media de cuadrados	*F*
Entregrupos	272,44	2	136,22	22,78
Intragrupos	3.570,56	597	5,98	
Total	3.842,00	599		

Esquema de aceptación – rechazo de H_0.

Se acepta H_0 si: $F_e \{ F_c \sum Nc_e \{ Nc_c \sum Ns_e \} Ns_c$

Se rechaza H_0 si: $F_e \varnothing F_c \sum Nc_e \varnothing Nc_c \sum Ns_e \Omega Ns_c$

Como la F_e (22,78) es mayor que la F_c (3,01), entonces se puede asumir, al Ns = 0,05, que la variación del sistema se debe más a la dispersión entre los grupos, por lo que se puede asumir que existen diferencias significativas entre los grupos. Entonces se puede considerar que el sistema de enseñanza ha influido en el nivel de aprendizaje de los alumnos. Las medias mayores de las calificaciones corresponden al grupo de *nuevas tecnologías* e *Internet* (47,11 y 47,60, respectivamente), mientras que el *grupo control* ha obtenido una puntuación media de 45,99. Para detectar entre qué grupos se producen las diferencias significativas se utiliza alguno de los test de *comparaciones múltiples* como *Scheffe,* que permite saber, en el análisis de varianza, entre que grupos se producen las diferencias significativas. No se muestra el cálculo por exceder el objetivo de este manual, pero se muestra el cuadro-resultado de SPSS.

Tabla 179 Múltiples comparaciones con Scheffe.

Grupo i	Grupo j	Diferencia de medias	Error típico	Ns_e
Control	Nuevas Tecnologías	-1,12*	0,24	0,00
	Internet	-1,60*	0,24	0,00
Nuevas Tecnologías	Control	1,12*	0,24	0,00
	Internet	-0,49	0,24	0,14
Internet	Control	1,60*	0,24	0,00
	Nuevas Tecnologías	0,49	0,24	0,14

* La diferencia de las medias es significativa para Ns = 0,05

A un Ns = 0,05 se detectan diferencias significativas entre el *grupo control* y *nuevas tecnologías* y *grupo control* e *Internet*. Pero no así entre *nuevas tecnologías* e *Internet*, que tienen medias homogéneas. Entonces se puede asumir que los medios utilizados en los grupos experimentales han producido mejores resultados, entendiendo como tales que los alumnos han obtenido mejores calificaciones en las pruebas administradas.

15.5 Requisitos para aplicar la Estadística Paramétrica

Para aplicar los test que utilizan medias de grupos, es necesario que cumplan ciertos requisitos. Si las comparaciones se hacen con las medias es necesario que sean representativas de las muestras, submuestras o subpoblaciones y es necesario comprobar que se cumple que,

1. Las subdistribuciones de los grupos son normales, supuestamente normales o significativamente normales.
2. Las varianzas de las subdistribuciones son homogéneas.
3. Que las *n*'s de los grupos sean iguales.
4. Que las *n*'s de los grupos sean mayores de 30

El primer requisito se comprueba con estadísticos no paramétricos mediante un protocolo de contraste de hipótesis. Estos estadísticos son: *Kolmogorov-Smirnov*, Chi-

cuadrado y *Shapiro-Wilk.* El programa estadístico SPSS dispone del *gráfico de probabilidad* (*probability plot* (*P-P*)) para ampliar la información de este contraste.

El segundo requisito es mediante otro contraste de hipótesis que SPSS denomina *test de Levene* con el estadístico y distribución *F*.

El tercer y cuarto requisito se comprueban por observación simple, las *n*'s deben ser iguales y mayores de 30.

La aplicación de un test paramétrico debe ir acompañado de la comprobación de los requisitos, frecuentemente llamado diagnóstico del modelo. Algunos requisitos pueden aceptar ciertas variaciones como por ejemplo la igualdad de las *n's*. Si se asume que el incumplimiento de uno o varios de los requisitos no permiten la aplicación del test de contraste de medias, entonces se debe recurrir a la Estadística No Paramétrica, que no necesita requisitos o estos son menos restrictivos. La estadística No Paramétrica no es objetivo de este manual, pero su aplicación sigue los mismos criterios de protocolo de contraste de hipótesis de la Estadística Paramétrica (Ver el software estadístico SPSS).

16 Asociación lineal (covarianza y correlación)

La *asociación lineal* se aplica cuando las variables son numéricas y se cruzan de dos en dos. Se puede considerar en el grupo de la Estadística Descriptiva Bivariable para variables numéricas (Ver Tabla 45), aunque incluye contraste de hipótesis y es la antesala del Análisis de Regresión Lineal y otras técnicas multivariable.

Al ser las variables numéricas, primero procede ver su relación a través de un *gráfico de dispersión* o *X-Y*. Después se calcula la *covarianza* (S_{xy}) de las variables y su estandarización es el coeficiente de correlación de Pearson (r ó r_{xy}). La interpretación del coeficiente *r* implica un contraste de hipótesis por lo que se puede considerar análisis además de descripción. La interpretación de *r* se debe hacer acompañada del gráfico de dispersión puesto que la relación debe ser considerada lineal. Indica o mide la asociación o dispersión lineal de los puntos respecto de una línea imaginaria (la recta de *regresión lineal* o la recta ajustada por *mínimos cuadrados ordinarios* (*MCO*)).

La asociación en este caso es lineal en el sentido que se acaba de mencionar, mientras que la asociación de *Chi-cuadrado* es una asociación de frecuencias o frecuencista, cómo están distribuidos los casos entre las celdas, o sea, las frecuencias absolutas.

Entonces el proceso será,

- Primero el *gráfico de dispersión* o *X-Y*
- Segundo se calcula la *covarianza* y
- Tercero el cálculo del coeficiente *r* que es la estandarización de la *covarianza*.

16.1 Gráfico de dispersión de dos ejes

Para relacionar dos variables numéricas se puede empezar por obtener el *gráfico de dispersión* o *gráfico X-Y*, para representar las dos variables en un sistema de *coordenadas cartesianas* de dos dimensiones (Ver epígrafe 7.6).

La *covarianza* y la *correlación*, al no implicar causalidad, no es necesario definir la relación de dependencia e independencia entre las variables. No obstante, una alta asociación entre las variables puede ser indicativo de la existencia de relación entre las variables, mientras que la falta de asociación puede suponer la no existencia de relación. Siempre se contemplará la posibilidad del azar, tanto en un caso como en otro y en el segundo supuesto, puede ocurrir que la influencia de terceras variables oculte la correlación de otras dos.[92] El análisis de asociación se debe realizar cuando la relación entre las variables sea lineal o considerada lineal aunque sea dispersa, y no se puede realizar en cualquier otro caso. La relación entre las variables, aunque sea dispersa y no funcional, debe responder a la ecuación,

$y \mid a 2 bx$	Fórmula 123

En base a lo anterior, no es necesario definir o proponer la variable independiente y dependiente. Pero como este proceso suele ser previo de otros procedimientos en los que sí se considera la relación de dependencia e independencia entre las variables, para la representación en el gráfico, es preciso considerar cual es la variable *dependiente* y la

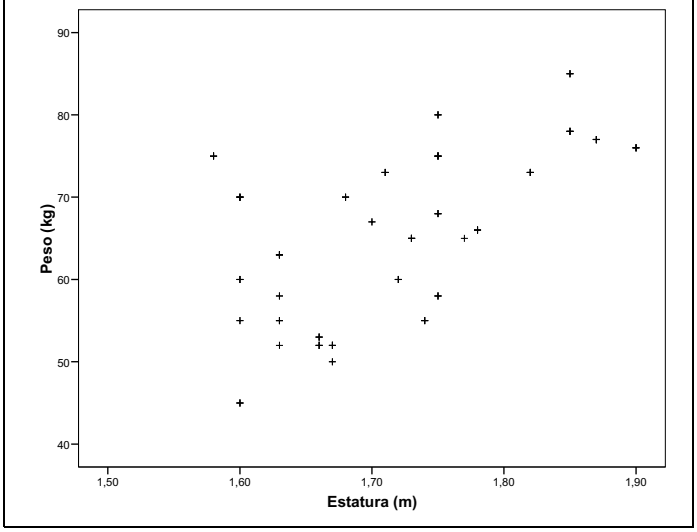
[92] A veces se ha detectado y puede ser generalizado, que la Tasa de Natalidad (*TN*) tenga una correlación alta con el Incremento de la Población (*IP*), y que la Tasa de Mortalidad (*TM*) tenga una correlación nula con el *IP*, pero cuando se correlacionan *IP* y *TM*, eliminando el efecto de la *TN*, se encuentra la verdadera relación entre el *IP* y la *TM*.

independiente.

En los *gráficos de dispersión*, la variable considerada *dependiente* (*y*), se representa en el eje de *ordenadas* o *vertical* y la variable considerada *independiente* (*x*), en el eje de *abscisas* u *horizontal*. A veces no es posible esta distinción y la colocación se hará en función del interés de la representación del gráfico.

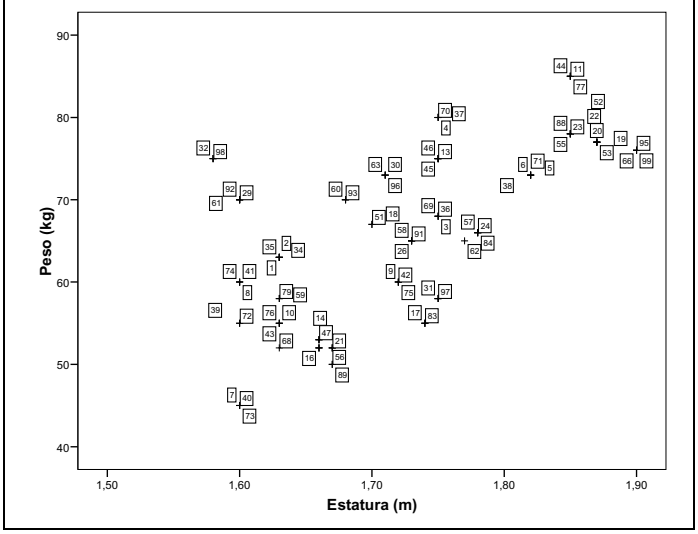
Para ver la relación entre las variables *estatura* y *peso* de la matriz de datos de la Tabla 16, se representa a cada caso por su par de valores *x-y,* siendo el *peso* la variable representada en el eje *Y* y la *estatura* la representada en el eje *X*, se obtiene el Gráfico 20.

Gráfico 20 Gráfico *X-Y* de peso y estatura.



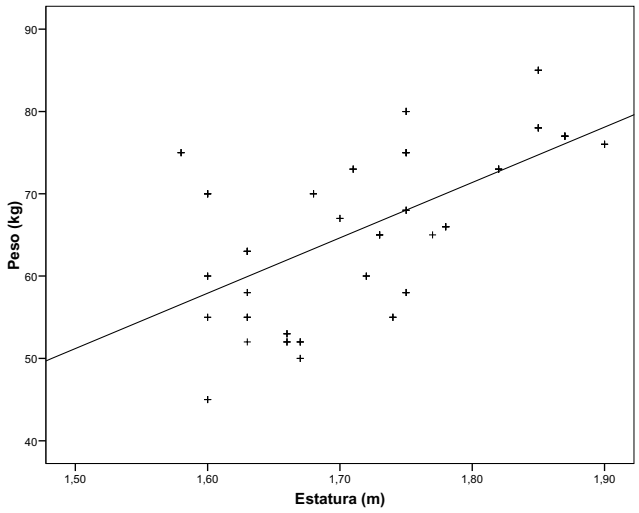
Cada punto puede representar a uno o varios casos. El Gráfico 21 muestra el código de los casos representados en cada punto.

Gráfico 21 Gráfico *X-Y* de peso y estatura, con identificador de caso.



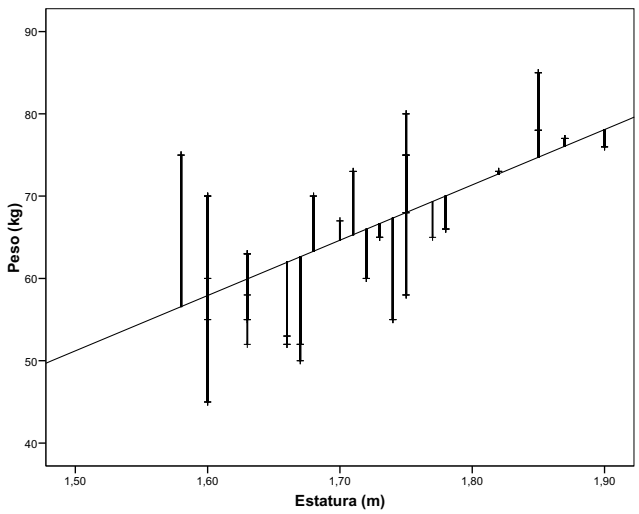
La representación gráfica de la asociación o dispersión de los casos respecto a una línea imaginaria (recta de regresión por mínimos cuadrados ordinarios (*MCO*)) se muestra en el Gráfico 22.

Gráfico 22 Gráfico *X-Y* de peso y estatura, con la recta de regresión.



La covarianza es la medida de la dispersión de los puntos a la línea de regresión y la correlación es la misma medida estandarizada. La dispersión se representa en el Gráfico 23 con las líneas verticales de cada punto a la recta.

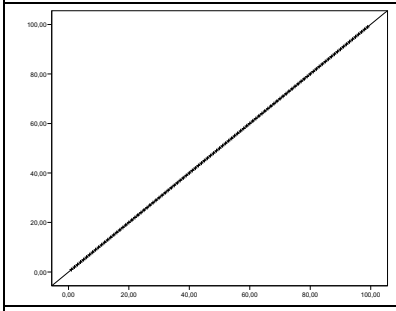
Gráfico 23 Gráfico *X-Y* de peso y estatura, con recta de regresión y representación de la dispersión.



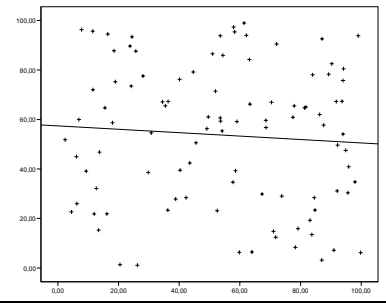
16.2 Cálculo de la covarianza

La forma de la relación entre dos variables numéricas o consideradas numéricas, se puede considerar entre tres modelos que tienen relación de continuidad entre ellos. La relación puede ser considerada *relación lineal directa* (*RLDr*) (Gráfico 24 a)*, relación lineal dispersa* (*RLDs*) (Gráfico 24 b) y *relación lineal inversa* (*RLI*) (Gráfico 24 c).

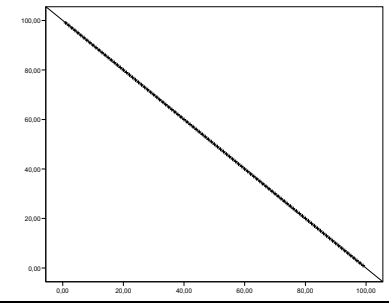
Gráfico 24 Modelos de asociación lineal.



(a) Relación Lineal Directa. (Funcional).



(b) Relación Lineal Dispersa. (No funcional).



(c) Relación Lineal Inversa. (Funcional).

La característica de estas relaciones es que en la *RLDr,* a valores bajos de *x* le corresponden valores bajos de *y*, a valores medios de *x* le corresponden valores medios de *y* y a valores altos de *x* le corresponden valores altos de *y*. La relación directa significa que cuando la variable *x* crece la variable *y* también crece y viceversa, cuando la variable *x* decrece, también lo hace la *y*.

En la *RLI,* a valores bajos de *x* le corresponden valores altos de *y*, a valores medios de *x* le corresponden valores medios de *y* y a valores altos de *x* le corresponden valores bajos de *y*. La relación inversa significa que cuando la variable *x* crece la variable *y* decrece y viceversa, cuando la variable *x* decrece, la *y* crece. En los dos casos, a valores medios de una variable le corresponde valores medios en la otra, esta característica es propia de la relación lineal.

En la *RLDs* a valores bajos en *x* le corresponden valores bajos, medios y altos en *y*; a valores medios en *x* le corresponden valores bajos, medios y altos en *y*, y a valores altos en *x* le corresponden valores bajos, medios y altos en *y*. al ser la relación dispersa, se puede asumir que es lineal.

Ahora se procede a calcular el valor numérico de esta relación y a interpretarlo. El estadístico base es la *varianza*.

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{X})^2}{n}$$

Fórmula 124

Pero la *varianza* también se puede representar como,

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{X})(x_i - \overline{X})}{n}$$

Fórmula 125

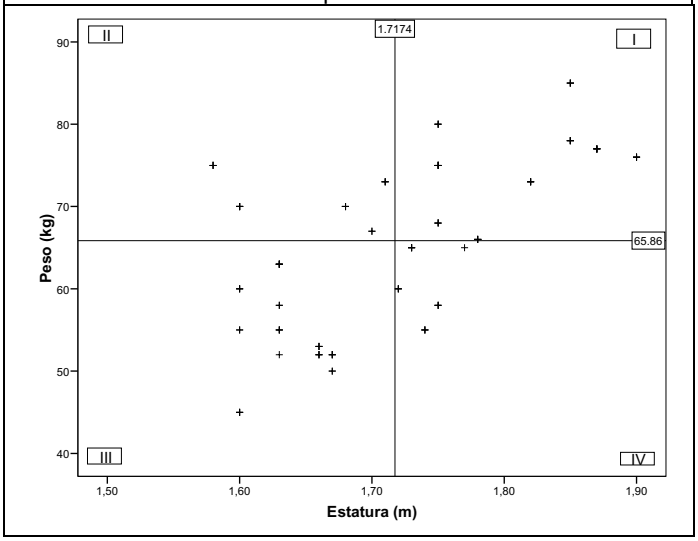
El sumatorio de la distancia del valor de *x* del caso *i-ésimo* respecto a la media por la distancia del mismo valor de *x* otra vez a la media. Como el Gráfico 20 tiene dos variables, se puede aplicar considerando para el caso *i-ésimo* las distancias de los valores en *x* e *y* respecto de sus correspondientes medias y la Fórmula 125 tomaría la forma,

$$S_{xy} = \frac{\sum_{i=1}^{n} (x_i - \overline{X})(y_i - \overline{Y})}{n}$$

Fórmula 126

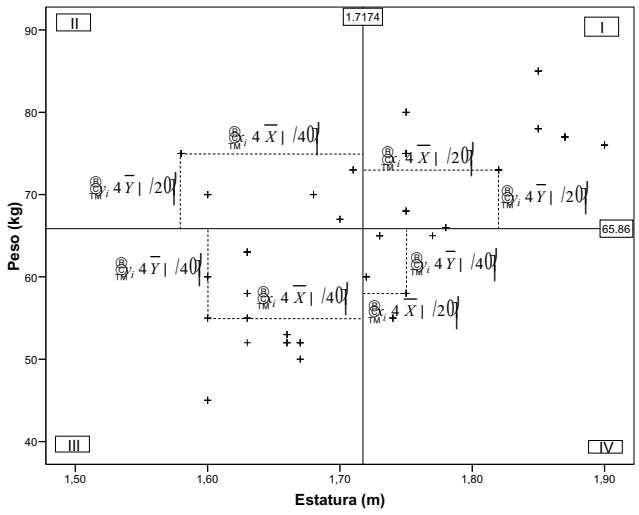
Y recibe el nombre de covarianza. El numerador es el producto cruzado ("*cross-product*") de la distancia, para cada caso, de *x* respecto a su media por el valor de *y* a la suya, dividido por el total de casos. Si se representan las medias de *X* e *Y* con vectores en el Gráfico 20, se obtiene el Gráfico 25,

Gráfico 25 Gráfico X-Y de peso y estatura, con líneas de referencia por las medias.



De esta manera el gráfico representado en el primer cuadrante del sistema de *coordenadas cartesianas*, queda dividido a su vez en otros cuatro cuadrantes: *I*, *II*, *III* y *IV*. Cada caso, a su vez, según la Fórmula 126, tiene una distancia a la media de *X* y a la media de *Y*. En el cuadrante *I* todos los casos tienen la distancia a la media de *X* y a la de *Y* positiva. En el cuadrante *II* la distancia a la media de *X* es negativa pero a la media de *Y* es positiva. En el cuadrante *III* las dos distancias son negativas, y en el cuadrante *IV* la distancia de *X* es positiva y la de *Y* negativa, gráficamente (Gráfico 26),

Gráfico 26 Gráfico *X-Y* de peso y estatura con el signo de las diferencias a las medias.



Si se aplica la Fórmula 126 según el Gráfico 26 y se asume la relación del tipo del Gráfico 24 a, pero sin ser funcional, los casos tenderán a caer en los cuadrantes *I* y *III*, aunque habrá alguno en los cuadrantes *II* y *IV*. Entonces el producto cruzado del numerador es muchos *positivo* x *positivo* y muchos *negativo* x *negativo* que ambos dan un resultado *positivo*. La suma de todos ellos dará un *positivo grande*. Por lo tanto, cuando la relación de las dos variables es del tipo Gráfico 24 a, la *covarianza* es *grande* y *positiva*.[93] Simbólicamente,

$$C \mathbin{4} I \Downarrow \; \ldots 2 \ldots\ldots \Delta \ldots\ldots 2 \ldots\ldots \mid \; 2$$

$$C \mathbin{4} III \Downarrow \; .4 \ldots\ldots \Delta \ldots\ldots 4 \ldots\ldots \mid \; 2$$

$$S_{xy} \mid \frac{\overset{n}{\underset{i \mid 1}{\sum}} (x_i - \overline{X})(y_i - \overline{Y})}{n} \mid \; 2 Grande$$

Si la relación es lineal inversa (Gráfico 24 c), pero sin ser funcional, los casos tenderán a caer en los cuadrantes *II* y *IV*, aunque habrá alguno en los cuadrantes *I* y *III*. Entonces el producto cruzado del numerador es muchos *negativo* x *positivo* y muchos *positivo* x *negativo* que ambos dan un resultado *negativo*. La suma de todos ellos dará un *negativo grande*. Por lo tanto, cuando la relación de las dos variables es del tipo Gráfico 24 c, la *covarianza* es *grande* y *negativa* (Ver nota 93). Simbólicamente,

$$S_{xy} \mid \frac{\overset{n}{\underset{i \mid 1}{\sum}} (x_i - \overline{X})(y_i - \overline{Y})}{n} \mid \; 4 Grande$$

$$C \mathbin{4} II \Downarrow \; \ldots 4 \ldots\ldots \Delta \ldots\ldots 2 \ldots\ldots \mid \; 4$$

$$C \mathbin{4} IV \Downarrow \; ..2 \ldots\ldots \Delta \ldots\ldots 4 \ldots\ldots \mid \; 4$$

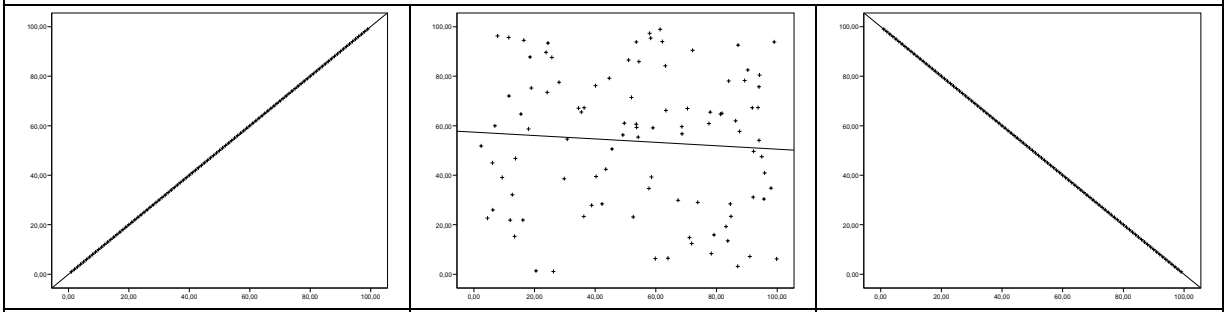
[93] Estas imprecisiones son las que se resuelven con la estandarización de la *covarianza*.

Si la relación es lineal dispersa (Gráfico 24 b), que no es funcional, los casos tenderán a caer en los cuadrantes *I, II, III* y *IV*. Entonces el producto cruzado del numerador es muchos *positivo* x *positivo* y muchos *negativo* x *negativo* que ambos dan un sumatorio *positivo* y muchos *negativo* x *positivo* y muchos *positivo* x *negativo* que ambos dan un sumatorio *negativo.* La suma de todos ellos será ∂0. Por intuición, si de la relación lineal directa que es *grande* y *positivo* hasta la relación lineal inversa que es *grande* y *negativo*, hay solución de continuidad, entonces tiene que pasar por el cero, que es el punto intermedio entre los dos extremos anteriores. Por lo tanto, cuando la relación de las dos variables es del tipo Gráfico 24 b, la *covarianza* es ∂0. Simbólicamente,

$$C\,4\,I \Downarrow \;....2.....\Delta.....2.....\mid\; 2$$
$$C\,4\,III \Downarrow \;.4.....\Delta.....4.....\mid\; 2$$
$$S_{xy} \mid \frac{\sum_{i\mid 1}^{n} x_i 4 \overline{X}\,\Delta\, y_i 4 \overline{Y}}{n} \mid \partial 0$$
$$C\,4\,II \Downarrow \;...4.....\Delta.....2.....\mid\; 4$$
$$C\,4\,IV \Downarrow \;..2.....\Delta.....4.....\mid\; 4$$

Los valores que toma la *covarianza* en las relaciones tipo del Gráfico 24 se muestran en el Gráfico 27,

Gráfico 27 Valores de la covarianza en los modelos de asociación lineal.



(a) Relación Lineal Directa. S_{xy} = +*grande*	(b) Relación Lineal Dispersa. S_{xy} = ∂0	(c) Relación Lineal Inversa. S_{xy} = -*grande*

El *cross-product* del numerador es una abstracción más difícil de comprender que el numerador de la *varianza,* ya que en la *covarianza* se multiplican unidades de medida diferentes entre sí. En este caso, la estandarización o tipificación, requiere no sólo eliminar la unidad de medida, sino que los dos términos del producto pasen a la misma unidad de medida una vez estandarizado. Esta operación se consigue tipificando según el criterio *Z*, simbólicamente,

$z_i \mid \frac{x_i 4 \overline{X}_x}{S_x}$	Fórmula 127

Para aplicarlo a la *covarianza*, se divide cada factor del numerador por su desviación típica y se obtiene el denominado coeficiente de correlación de Pearson, simbólicamente,

$$r_{xy} = \frac{\sum_{i=1}^{n} \left(\frac{x_i - \bar{X}_x}{S_x}\right)\left(\frac{y_i - \bar{Y}_y}{S_y}\right)}{n}$$

El desarrollo de la fórmula anterior es,

$$r_{xy} = \frac{\sum_{i=1}^{n} \left(\frac{x_i - \bar{X}_x}{S_x}\right)\left(\frac{y_i - \bar{Y}_y}{S_y}\right)}{n} = \frac{\left[\left(\frac{x_1 - \bar{X}_x}{S_x}\right)\left(\frac{y_1 - \bar{Y}_y}{S_y}\right)\right] + \left[\left(\frac{x_2 - \bar{X}_x}{S_x}\right)\left(\frac{y_2 - \bar{Y}_y}{S_y}\right)\right] + \cdots + \left[\left(\frac{x_n - \bar{X}_x}{S_x}\right)\left(\frac{y_n - \bar{Y}_y}{S_y}\right)\right]}{n}$$

Sacando factor común las desviaciones típicas de las variables,

$$\frac{\frac{1}{S_x}\frac{1}{S_y}\left[(x_1 - \bar{X}_x)(y_1 - \bar{Y}_y) + (x_2 - \bar{X}_x)(y_2 - \bar{Y}_y) + \cdots + (x_n - \bar{X}_x)(y_n - \bar{Y}_y)\right]}{n}$$

Y representándolo en formato de sumatorio,

$$\frac{\frac{1}{S_x}\frac{1}{S_y}\sum_{i=1}^{n}(x_1 - \bar{X}_x)(y_1 - \bar{Y}_y)}{n} = \frac{1}{S_x}\frac{1}{S_y}\frac{\sum_{i=1}^{n}(x_1 - \bar{X}_x)(y_1 - \bar{Y}_y)}{n} = \frac{1}{S_x}\frac{1}{S_y}S_{xy} = \frac{S_{xy}}{S_x S_y}$$

Y Karl Pearson lo llamó coeficiente de correlación r_{xy} o simplemente r. Simbólicamente,[94]

$r = \frac{S_{xy}}{S_x S_y}$	Fórmula 128

Entonces la estandarización de la *covarianza* (S_{xy}) se llama coeficiente de correlación r de Pearson y es igual a la *covarianza* dividido por el producto de las desviaciones típicas de las variables.

Los valores extremos que puede tomar el coeficiente r se puede ver si se aplica al Gráfico 24 a, b y c, que anteriormente tenían valores imprecisos.

[94] Hay otros procesos de cálculo del coeficiente *r* pero este parece más breve y sencillo.

Si una relación lineal directa es una relación funcional y por lo tanto $y = a + bx$. Siendo $a = 0$ y $b = 1$, entonces la función anterior queda que $y = x$. Entonces r,

$r_{xy} \mid \frac{S_{xy}}{S_x S_y} \mid r_{xx} \mid \frac{S_{xx}}{S_x S_x} \mid \frac{S_x^2}{S_x^2} \mid 1$ ($S_{xx} \mid S_x^2$, ver Fórmula 132)	Fórmula 129

La correlación de una variable consigo misma o la correlación de dos variables que tienen una relación lineal funcional es igual a la unidad y es el valor máximo que puede tomar r (Gráfico 28 a).

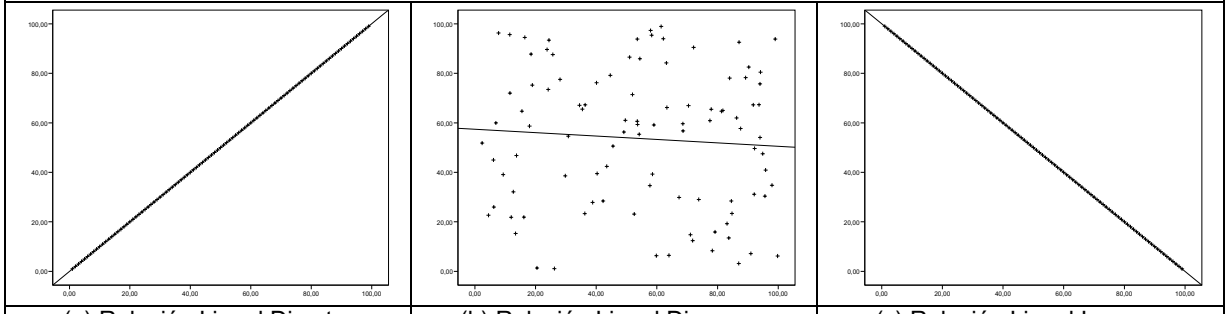
Si dos variables tienen una relación lineal inversa funcional, $y = a + (-b)x$, de tal manera que $a = máx(x+1)$ y $b = -1$, entonces $y = máx(x+1) + (-1)x$, entonces la *covarianza* de las dos variables es igual a la *varianza* de cualquiera de ellas pero con el signo negativo, y $S_x = S_y$, por lo tanto, simbólicamente,

$r_{xy} \mid \frac{4\, S_{xy}}{S_x S_y} \mid r_{xx} \mid \frac{4\, S_{xx}}{S_x S_x} \mid \frac{/40 S_x^2}{S_x^2} \mid 41$ $\lvert 4\, S_{xy} \rvert \mid S_x^2$ ♥ si $y \mid máx(x\,2\,1)\,2\,/4\,10x$, (ver Fórmula 132)	Fórmula 130

La correlación de una variable con otra igual pero inversa o la correlación de dos variables que tienen una relación lineal funcional inversa es igual a -1 y es el valor máximo negativo que puede tomar r (Gráfico 28 c).

Si entre dos variables hay una relación lineal dispersa con una ecuación del tipo $y = a + bx$, de tal manera que $a - \overline{Y}$ y $b - 0$, entonces la *covarianza* está próxima a cero y r es también próxima a cero (Gráfico 28 b).

Gráfico 28 Valores de la covarianza en los modelos de asociación lineal.



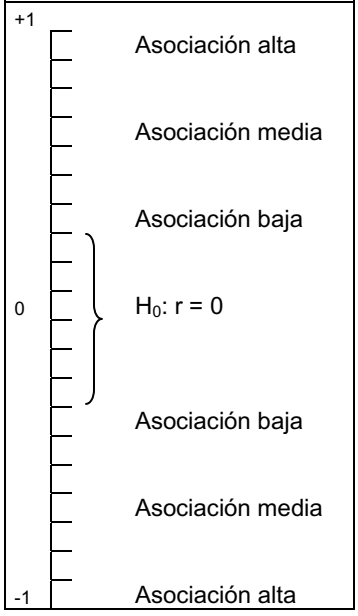
(a) Relación Lineal Directa.	(b) Relación Lineal Dispersa.	(c) Relación Lineal Inversa.
$r_{xy} = +1$	$r_{xy} = \partial 0$	$r_{xy} = -1$

Entonces r toma valores en el rango de -1 ɜ +1, ambos inclusive. En el caso del Gráfico 28 a, que $r = 1$, la dispersión de los casos respecto de la recta imaginaria o de regresión es nula pero la relación es directa. En el caso del Gráfico 28 c, que $r = -1$, la dispersión de los casos respecto de la recta imaginaria o de regresión es nula pero la relación es inversa. Y en el Gráfico 28 b, se produce una relación dispersa que se puede considerar lineal, aunque en este caso, esta consideración es trivial o convencional. No tiene

trascendencia. La concentración de los casos respecto a la recta imaginaria supone dispersión de los casos respecto de la medias. Repasar la Fórmula 126, cuanto mayor es la distancia de los casos a las medias, mayor es la *covarianza* y supone menor dispersión respecto de la recta de regresión.

Al estandarizar la *covarianza* y transformarla en el coeficiente *r*, se pone límite a los valores que puede tomar. El coeficiente de correlación *r*, a su vez, se puede transformar en una variable de tipo *t* con distribución de densidad de probabilidad conocida y sus valores se pueden interpretar en términos de probabilidad. La interpretación del valor de *r* es,

Tabla 180 Interpretación de los valores de r.



Para saber si *r* tiene un valor de cero o significativamente cero, se puede hacer una transformación de *r* en una distribución *t* y hacer el contraste de hipótesis. Si es significativamente distinta de cero, entonces puede ser que tenga una asociación baja, media o alta, bien positiva (asociación lineal directa) o negativa (asociación lineal inversa). La interpretación de una asociación baja, media o alta, depende de la experiencia del investigador con el estadístico, de la información "a mano" y del conocimiento de la materia en estudio.

Igual que otros estadísticos (*coeficiente de contingencia, V de Cramer, fi, lambda, etc.*) la interpretación de *r* está sujeta también a las variables que estamos correlacionando. Si la relación de las variables es funcional, la correlación debe ser muy alta y pequeñas variaciones pueden indicar la ocurrencia de algún evento ajeno al proceso. Pero en el caso de variables sociales, correlaciones moderadas, pueden ser indicativo de que "algo pasa" en la relación entre las variables. Por ejemplo, en el supuesto de que dos variables como *salario* y *sexo* tengan una correlación significativa, aunque sea considerada baja, está indicando que entre el *salario* y *sexo* hay alguna relación cuando no debería existir relación, salvo que haya otras variables intervinientes (categoría profesional, estudios, etc.).

El contraste de igualdad a cero de *r* se hace con un protocolo de contraste de hipótesis,

1. Hipótesis alternativa H_1: $r \prod 0$.
2. Hipótesis nula H_0: $r = 0$.
3. Estadístico: *t*.
4. Criterio de aceptación o rechazo de H_0, $Ns = 0{,}05$.

La transformación de *r* en *t*, es

$t \mid \frac{r}{\sqrt{\frac{1}{n 4 1}}}$	Fórmula 131

Esta transformación convierte *r* en un valor que tiene distribución *t* de *Student* y se puede calcular si el valor es significativamente distinto de cero.

Esquema de aceptación – rechazo de H_0.

Se acepta H_0 si: $\mid t_e \mid \{ \mid t_c \mid \sum Nc_e \{ Nc_c \sum Ns_e \} Ns_c$

Se rechaza H_0 si: $\mid t_e \mid \varnothing \mid t_c \mid \sum Nc_e \varnothing Nc_c \sum Ns_e \Omega Ns_c$

Para aplicar este contraste es necesario considerar que la población de la que se obtienen las muestras debe tener una distribución normal, las muestras deben estar seleccionadas aleatoriamente y los casos de las muestras deben ser independientes.

16.3 Propiedades y características de la covarianza y el coeficiente *r*

La *covarianza* y el coeficiente *r* tienen las siguientes propiedades o características y toman los siguientes valores,

Característica 1

Para aplicar el coeficiente de correlación *r* de Pearson, las variables tienen que ser numéricas o consideradas numéricas. Una variable ordinal se puede considerar numérica, por lo que con una variable numérica y una ordinal también se puede aplicar el coeficiente *r*, aunque lo apropiado es el coeficiente de correlación de *Spearman*.

Cuando se calcula la asociación entre una variable numérica y una binaria, también se puede aplicar el coeficiente de correlación *r*, aunque lo apropiado es el *biserial*.

Si las dos variables son binarias se puede aplicar *r,* pero el coeficiente es el *biserial puntual*. Las variables binarias tienen media, que es la proporción de 1's y se pueden calcular el resto de estadísticos. También se puede considerar el uso de variables dicotómicas (Norusis, 1986).

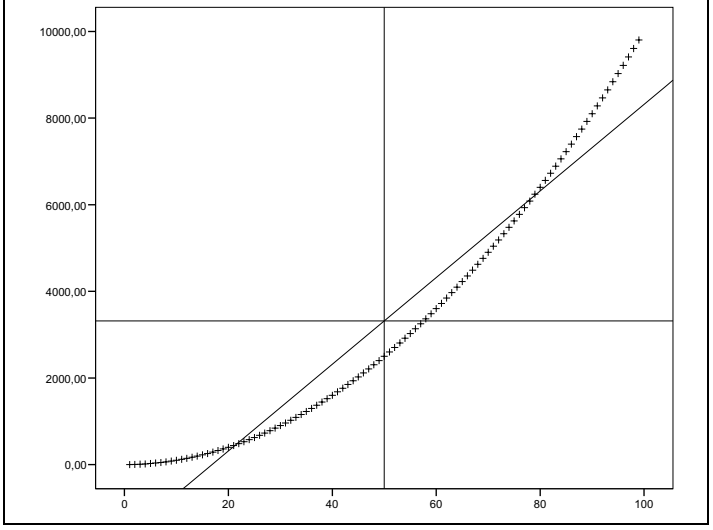
La interpretación de *r* en todos los casos es la misma. Con variables binarias y

dicotómicas, al tener un rango restringido imprimen poca inercia al sistema por lo que la correlación tenderá a ser baja.

Característica 2

El coeficiente de correlación *r* de Pearson se debe calcular cuando las variables tienen una relación lineal o considerada lineal, o por lo menos que su relación sea considerada del tipo $y = a + bx$. La correlación entre dos variables no lineales puede ser alta, lo que indica poca dispersión, pero el gráfico muestra que la dispersión se distribuye desigualmente a lo largo de la curva. El Gráfico 29 es una relación no lineal del tipo $y = a + x^2$ y la correlación es de 0,969.

Gráfico 29 Relación no-lineal.

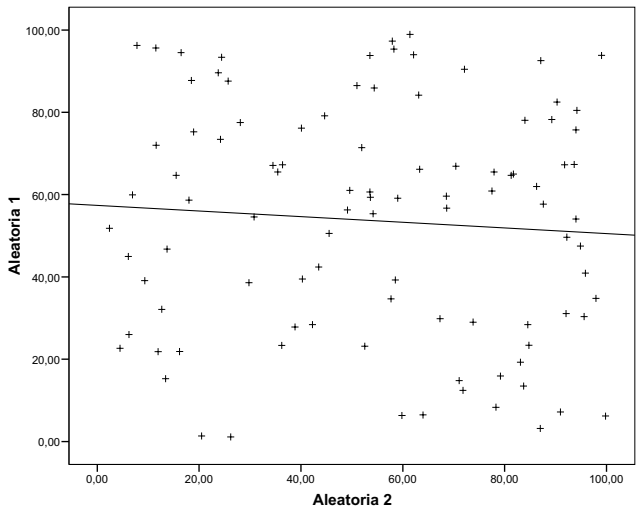


Característica 3

Por el proceso de cálculo de la *covarianza* y *r*, los valores de las variables que tienen valores altos respecto de sus medias (distribuciones con asimetría positiva) confieren mucha inercia (equivalente a la ley de la palanca, cuanto más largo es el brazo más grande resulta la misma fuerza) y tiende a dar coeficientes altos, ocultando la verdadera asociación. Ejemplo: sean dos variables aleatorias uniformemente distribuidas en el rango 1 э 100, generadas con la función específica de SPSS. Los estadísticos descriptivos univariables, la correlación y el gráfico de dispersión entre las dos variables se muestran en la Tabla 181.

Tabla 181 Estadísticos descriptivos, correlación y gráfico de dos variables aleatorias uniformemente distribuidas entre 1 y 100.

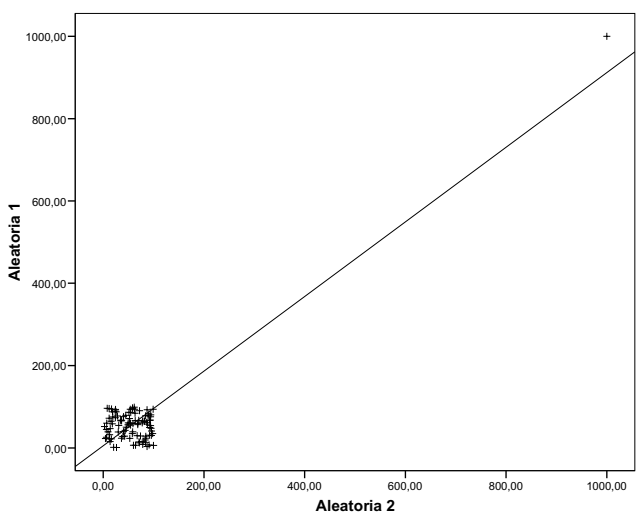
	Aleatoria 1	Aleatoria 2	Gráfico
N	99	99	
Rango	97,82	97,41	
Mínimo	1,11	2,40	
Máximo	98,93	99,81	
Media	53,66	54,49	
Desviación Típica	27,96	29,52	
Varianza	781,55	871,16	
Sesgo	-0,18	-0,16	
Apuntamiento	-1,06	-1,27	
r	-0,07		



Sean esas dos mismas variables añadiendo un caso con valor 1.000 en cada una de las variables. Los estadísticos descriptivos univariables, la correlación y el gráfico de dispersión entre las dos variables se muestran en la Tabla 182.

Tabla 182 Estadísticos descriptivos, correlación y gráfico de dos variables aleatorias uniformemente distribuidas entre 1 y 100, añadido un caso en cada variable de valor 1.000.

	Aleatoria 1	Aleatoria 2	Gráfico
n	100	100	
Rango	998,89	997,60	
Mínimo	1,11	2,40	
Máximo	1000,00	1000,00	
Media	63,12	63,95	
Desviación Típica	98,64	99,01	
Varianza	9729,26	9802,22	
Sesgo	8,80	8,68	
Apuntamiento	84,28	82,68	
r	0,91		



La correlación pasa de -0,07 a 0,91 y el sesgo de -0,18 y -0,16 a 8,80 y 8,68 respectivamente. El resto de los estadísticos varían en consecuencia.

Característica 4

El coeficiente de correlación no se expresa en ninguna unidad de medida, y no se ve afectado por transformaciones lineales tales como sumar, restar, multiplicar o dividir todos los valores de una variable por una constante.

Característica 5

La correlación entre dos variables sin estandarizar es igual que la correlación entre dos variables estandarizadas.

Sean dos variables z_x y z_y estandarizadas según el criterio *Z*. La correlación entre estas dos variables es,

$$r_{z_x z_y} = \frac{S_{z_x z_y}}{S_{z_x} \cdot S_{z_y}}$$

Pero como la desviación típica de una variable estandarizada es la unidad, el resultado es,

$$r_{z_x z_y} = S_{z_x z_y}$$

La correlación de dos variables estandarizadas es igual a la covarianza de las dos variables estandarizadas.

Y la covarianza de dos variables estandarizadas es,

$$S_{z_x z_y} = \frac{\sum_{i=1}^{n} (z_{xi} - \overline{Z}_x)(z_{yi} - \overline{Z}_y)}{n}$$

La media de una variable estandarizada es igual a cero, entonces,

$$S_{z_x z_y} = \frac{\sum_{i=1}^{n} z_{xi} z_{yi}}{n}$$

Y como z_{xi} y z_{yi} son igual a,

$$z_{xi} = \frac{x_i - \overline{X}_x}{S_x}, \quad z_{yi} = \frac{y_i - \overline{Y}_y}{S_y}$$

Y sustituyendo,

$$S_{z_x z_y} = \frac{\sum_{i=1}^{n} \left(\frac{x_i - \overline{X}_x}{S_x} \right) \left(\frac{y_i - \overline{Y}_y}{S_y} \right)}{n} = r_{xy}$$

Por lo tanto, la covarianza de dos variables estandarizadas es igual a la correlación entre las variables sin estandarizar.

$$S_{z_x z_y} = r_{xy}$$

Y consecuentemente, la correlación de dos variables estandarizadas es igual que la correlación de dos variables sin estandarizar, simbólicamente,

$$r_{z_x z_y} = r_{xy}$$

Característica 6

La covarianza de una variable consigo misma es igual a la varianza de la variable. Si calculamos la *covarianza* de una variable consigo misma, la distancia de *y* respecto de su media es la distancia de *x* respecto de su media. Simbólicamente,

$$S_{xy} = \frac{\sum_{i=1}^{n} (x_i - \overline{X})(y_i - \overline{Y})}{n} = S_{xx} = \frac{\sum_{i=1}^{n} (x_i - \overline{X})(x_i - \overline{X})}{n} = \frac{\sum_{i=1}^{n} (x_i - \overline{X})^2}{n} = S_x^2 \text{ ó } S^2$$

Fórmula 132

Por este motivo, una matriz cuadrada de *covarianzas* se denomina matriz de *varianzas-covarianzas*. La diagonal principal tiene las *varianzas* y la matriz es simétrica, por lo que la mitad inferior es igual a la mitad superior y contiene las *covarianzas*.

La *covarianza* y *r* son estadísticos simétricos. Son independientes del orden de las variables, aunque gráficamente, se tiende a poner la variable considerada *independiente* en el eje *X* y la dependiente en el eje *Y*, porque, por ejemplo, la regresión no es simétrica.

En la Tabla 183 se muestra la matriz de *varianzas-covarianzas* de las variables *peso, estatura y edad* de la matriz de datos de la Tabla 16,

Tabla 183 Matriz de varianzas-covarianzas.

	Peso (kg)	Estatura (m)	Edad (años)
Peso (kg)	102,67	0,58	4,40
Estatura (m)	0,58	0,01	0,08
Edad (años)	4,40	0,08	7,91

En esta tabla, la varianza de *peso* es 102,67 kg^2 y la *covarianza* de *peso* y *estatura* es 0,58 kg. x m. La interpretación o lectura de estos valores resulta difícil por la falta de referentes. La zona sombreada más oscura son las *varianzas* y la zona sombreada en gris claro

son las *covarianzas* y la mitad superior tiene los mismos valores que la mitad inferior por ser la matriz simétrica.

Ejemplo:

Desde los valores de la Tabla 183 se puede calcular los coeficientes de correlación de la Tabla 184,

La correlación de una variable consigo misma es la unidad (ver Fórmula 129).

Correlación entre la variable *peso* y *estatura*,

$$r \mid \frac{S_{xy}}{S_x S_y} \mid \frac{0,58}{\sqrt{102,67}\,\Delta\sqrt{0,01}} \mid 0,57$$

Correlación entre la variable *peso* y *edad*,

$$r \mid \frac{S_{xy}}{S_x S_y} \mid \frac{4,40}{\sqrt{102,67}\,\Delta\sqrt{7,91}} \mid 0,15$$

Correlación entre la variable *estatura* y *edad*,

$$r \mid \frac{S_{xy}}{S_x S_y} \mid \frac{0,08}{\sqrt{0,01}\,\Delta\sqrt{7,91}} \mid 0,29$$

Tabla 184 Matriz de correlaciones.

	Peso (kg)	Estatura (m)	Edad (años)
Peso (kg)	1,00	0,57**	0,15
Estatura (m)	0,57**	1,00	0,29**
Edad (años)	0,15	0,29**	1,00
n = 95			
** Correlación significativa			

El protocolo de contraste de hipótesis de igualdad a cero de los coeficientes de correlación es,

Contraste de hipótesis de las variables *peso* por *estatura*,

1. Hipótesis alternativa H_1: 0,57 ∏0.
2. Hipótesis nula H_0: 0,57 = 0.
3. Estadístico *t*.
4. Criterio de aceptación o rechazo de H_0: *Ns* = 0,05; *gl* = n-1 = 95-1 = 94; t_c = 1,9867.

La transformación de *r* en *t*, es

$$t_e \mid \frac{r}{\sqrt{\frac{1}{n \,4\, 1}}} \mid \frac{0{,}57}{\sqrt{\frac{1}{95 \,4\, 1}}} \mid \frac{0{,}57}{0{,}10} \mid 5{,}7$$

Esquema de aceptación – rechazo de H_0.

Se acepta H_0 si: $\mid t_e \mid \{ \mid t_c \mid \sum Nc_e \{ Nc_c \sum Ns_e \} Ns_c$

Se rechaza H_0 si: $\mid t_e \mid \varnothing \mid t_c \mid \sum Nc_e \varnothing Nc_c \sum Ns_e \Omega Ns_c$

Como la $|t_e|$ es mayor que la $|t_c|$ (5,7 > 1,9867) entonces se puede asumir rechazar la H_0 y aceptar la H_1 por lo tanto existe una asociación lineal significativa entre la variable *peso* y *estatura* para el grupo de la matriz de datos, y por el valor de *r* se puede decir que es una correlación media-alta. Es lo esperado, que en un grupo de jóvenes, el *peso* y la *estatura* correlacionen.

Contraste de hipótesis de las variables *peso* por *edad*,

1. Hipótesis alternativa H_1: 0,15 $\prod$0.
2. Hipótesis nula H_0: 0,15 = 0.
3. Estadístico *t*.
4. Criterio de aceptación o rechazo de H_0: Ns = 0,05; gl = n-1 = 95-1 = 94; t_c = 1,9867

La transformación de *r* en *t*, es

$$t_e \mid \frac{r}{\sqrt{\frac{1}{n \,4\, 1}}} \mid \frac{0{,}15}{\sqrt{\frac{1}{95 \,4\, 1}}} \mid \frac{0{,}15}{0{,}10} \mid 1{,}5$$

Esquema de aceptación – rechazo de H_0.

Se acepta H_0 si: $\mid t_e \mid \{ \mid t_c \mid \sum Nc_e \{ Nc_c \sum Ns_e \} Ns_c$

Se rechaza H_0 si: $\mid t_e \mid \varnothing \mid t_c \mid \sum Nc_e \varnothing Nc_c \sum Ns_e \Omega Ns_c$

Como la $|t_e|$ es menor que la $|t_c|$ (1,5 < 1,9867) entonces se puede asumir aceptar la H_0 y rechazar la H_1 por lo tanto no existe una asociación lineal significativa entre la variable *peso* y *edad* para el grupo de la matriz de datos, y se puede decir que el valor de *r* es

significativamente cero, aunque no sea exactamente cero. Es lo esperado, que en un grupo de jóvenes, el *peso* y la *edad* no correlacionen.

Contraste de hipótesis de las variables *estatura* por *edad*,

1. Hipótesis alternativa H_1: 0,28 ∏0.
2. Hipótesis nula H_0: 0,28 = 0.
3. Estadístico *t*.
4. Criterio de aceptación o rechazo de H_0: Ns = 0,05; gl = n-1 = 95-1 = 94; t_c = 1,9867

La transformación de *r* en *t*, es

$$t_e \mid \frac{r}{\sqrt{\frac{1}{n\,4\,1}}} \mid \frac{0{,}28}{\sqrt{\frac{1}{95\,4\,1}}} \mid \frac{0{,}28}{0{,}10} \mid 2{,}8$$

Esquema de aceptación – rechazo de H_0.

Se acepta H_0 si: $\mid t_e \mid \{ \mid t_c \mid \sum Nc_e \{ Nc_c \sum Ns_e \} Ns_c$

Se rechaza H_0 si: $\mid t_e \mid \varnothing \mid t_c \mid \sum Nc_e \varnothing Nc_c \sum Ns_e \Omega Ns_c$

Como la $|t_e|$ es mayor que la $|t_c|$ (2,8 > 1,9867) entonces podemos asumir rechazar la H_0 y aceptar la H_1 por lo tanto existe una asociación lineal significativa entre la variable *estatura* y *edad* para el grupo de la matriz de datos, y por el valor de *r* se puede decir que es baja.

17 Análisis de Regresión Lineal Simple

Con el Análisis de Regresión Lineal Simple (ARLS) se puede considerar que se inicia la Estadística Multivariable o las Técnicas de Análisis Multivariable. El ARLS es una técnica considerada de *dependencia* y *exploratoria* partiendo de la asunción de la ecuación de la línea recta, a la que se denomina *modelo explicativo-predictivo* (Fórmula 133).

$y \mid a 2 bx$	Fórmula 133

Recibe el nombre de Regresión por el experimento que Francis Galton realizó con "guisantes dulces". Comprobó que el tamaño de las plantas parecía *revertir* al tamaño medio y la llamó *ley de reversión*. Después sustituyo el nombre por *regresión* que ya había utilizado anteriormente. El nombre de *regresión* no es apropiado para la técnica estadística, aunque es el que prevalece y el que se utilizará. El nombre apropiado debe ser el de recta por Mínimos Cuadrados Ordinarios (MCO) por ser el criterio estadístico que se utilizará después para construir el modelo.

Se denomina *Simple* porque sólo tiene una variable *independiente* (x). La otra opción es *Múltiple*, cuando tiene más de una variable *independiente* (x_1, x_2, ..., x_n) (Fórmula 134).

$y \mid a 2 b_1x_1 2 b_2x_2 2 3 2 b_nx_n$	Fórmula 134

Se considera de *dependencia* por tener una variable considerada o propuesta como dependiente (y) y otra variable considerada o propuesta como *independiente* (x).

Es exploratorio porque se trata de seleccionar o encontrar la mejor variable *independiente*.

Es *explicativo-predictivo*, porque se trata de explicar y/o predecir la variable dependiente considerada también como explicada o predicha a través de la variable *independiente,* considerada también como explicativa o predictora. Supone admitir una relación de *causa-efecto* entre las variables.

17.1 Conceptos previos

Antes de introducir los fundamentos del ARLS es preciso exponer el significado de las constantes *a* y *b* de la Fórmula 133, según cuatro modelos diferentes (Tabla 185, Tabla 186, Tabla 187 y Tabla 188).

Tabla 185 Modelo de regresión 1.

Si asumimos que *a* =0 y *b* = 1, la ecuación de la Fórmula 133 se expresa,

$y \mid 0\,2\,1x$ y simplificando,

$y \mid x$

Asignando valores a *x* obtenemos valores para *y*

y	*x*
0	0
1	1
2	2
3	3
4	4

Por cada valor asignado a *x* se obtiene el mismo valor en *y*. Con los pares de coordenadas obtenemos una bisectriz en el gráfico al unir los puntos.

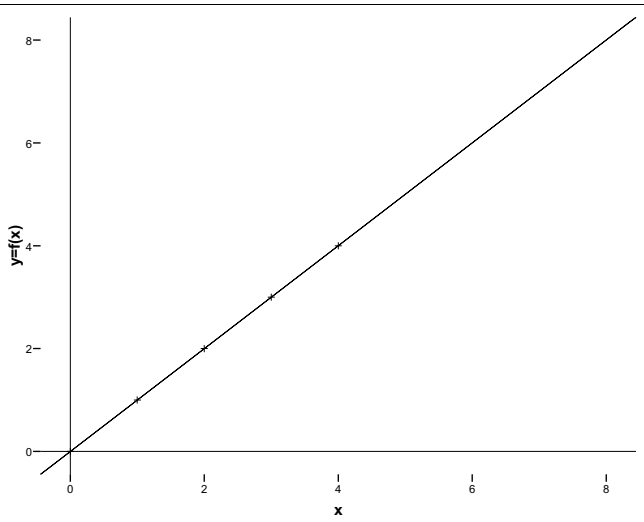


Tabla 186 Modelo de regresión 2.

Si asumimos que *a* = 1 y *b* = 1, la ecuación de la Fórmula 133 se expresa,

$y \mid 1\,2\,1x$

Asignando valores a *x* obtenemos valores para *y*

y	*x*
1	0
2	1
3	2
4	3
5	4

Por cada valor asignado a *x* se obtiene un valor en *y*. Con los pares de coordenadas obtenemos la recta del gráfico.

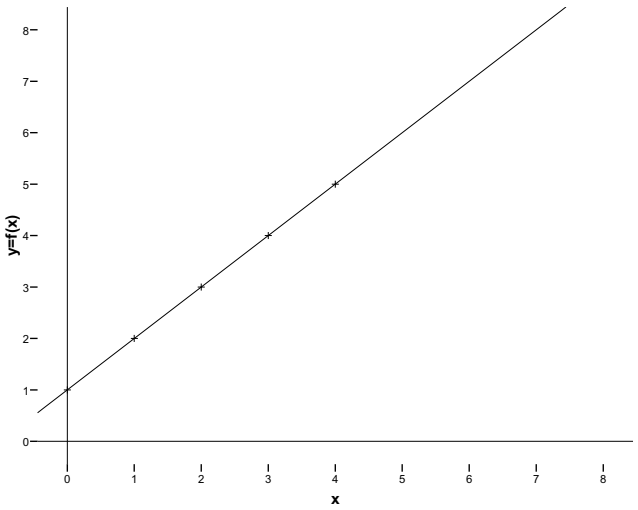


Tabla 187 Modelo de regresión 3.

Si asumimos que $a = 0$ y $b = 2$, la ecuación de la Fórmula 133 se expresa,

$$y \mid 0\,2\,2x$$

Asignando valores a *x* obtenemos valores para *y*

y	*x*
0	0
2	1
4	2
6	3
8	4

Para cada valor asignado a *x* se obtiene un valor en *y*. Con los pares de coordenadas obtenemos la recta del gráfico.

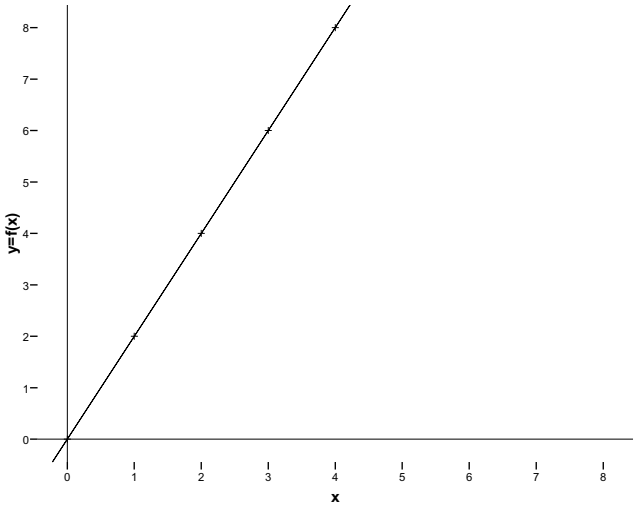


Tabla 188 Modelo de regresión 4.

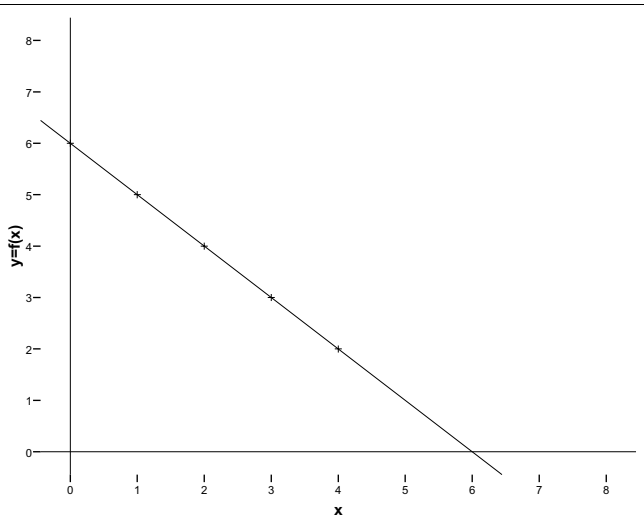
Si asignamos valores para $a = 6$ y $b = -1$, la ecuación de la Fórmula 133 se expresa,

$$y \mid 6\,2/4\,10x$$

Asignando valores a *x* obtenemos valores para *y*

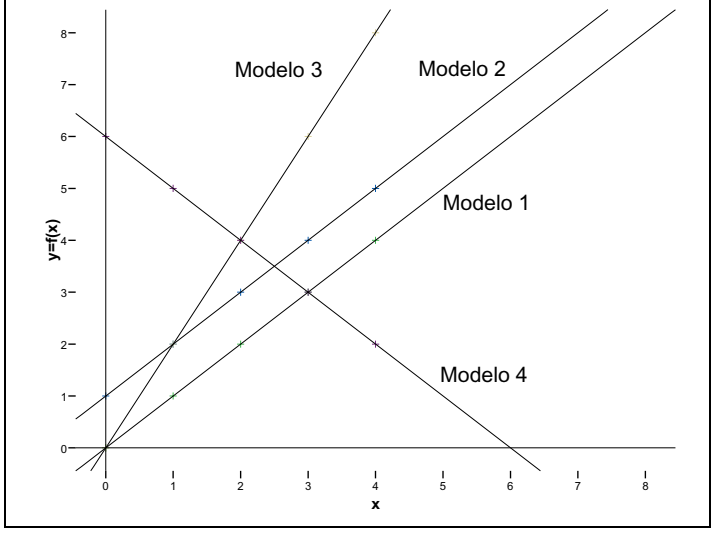
y	*x*
6	0
5	1
4	2
3	3
2	4

Para cada valor asignado a *x* se obtiene un valor en *y*. Con los pares de coordenadas obtenemos la recta del gráfico.



El *modelo 1* y el *modelo 2* son líneas paralelas porque la constante *b* tiene el mismo valor e igual a uno, mientras que el *modelo 3* presenta mayor pendiente que los anteriores. Los modelos 1 y 3 cortan al eje *Y* en la coordenada cero, que es el valor de la constante *a*, mientras que el modelo 2, corta en la coordenada $y = 1$, que es el valor de *a*. El *modelo 4* tiene el coeficiente *b* negativo y la relación es inversa. La relación entre los cuatro modelos se puede observar en el Gráfico 30.

Gráfico 30 Superposición de los cuatro modelos.



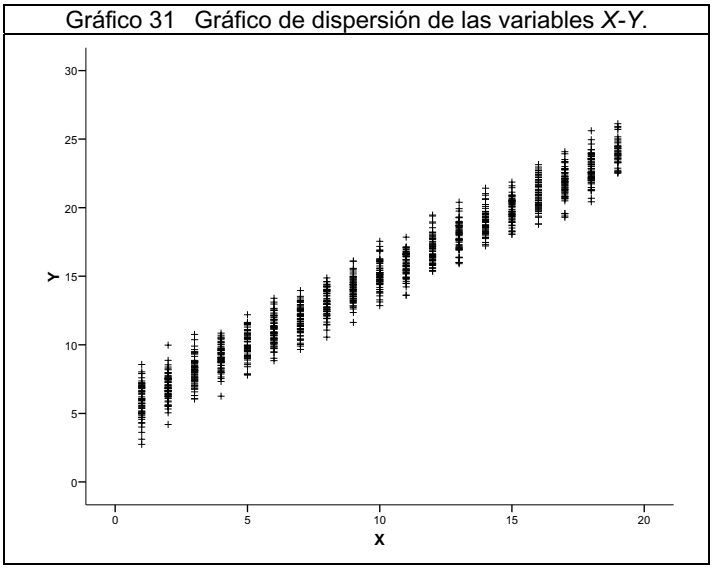
La constante *a* es *la distancia al origen o punto por donde la recta corta al eje Y*. La constante *b* es *la pendiente de la recta*, que significa las unidades en que varía *Y* por cada unidad que varía *X*. El carácter *predictivo* del ARLS se puede ver en el modelo $y = a + bx$, si atribuimos valores a *x* obtenemos valores en *y*, y es predictiva en este sentido. El carácter *explicativo* significa que por cada unidad que varía *X*, la variable *Y* varia *b* unidades y es explicativo en este sentido.

En el *modelo 1* y *modelo 2*, *b* vale uno, por lo tanto por cada unidad que varía *x* la variable *Y* varía en una unidad. Si *x* aumenta, *y* aumenta y si *x* disminuye, *y* disminuye, porque *b* es positivo. En el *modelo 3* la variación en *y* es de 2 unidades.

El *modelo 4* tiene el coeficiente *b* negativo (-1). La relación es que por cada unidad que varia *X* la variable *Y* varía en una unidad, pero cuando *X* aumenta, *Y* disminuye y viceversa, cuando *X* disminuye, *Y* aumenta. La recta corta al eje *Y* en el valor 6, que es el valor de la constante *a.*

Para desarrollar la aplicación y cálculo del Análisis de Regresión Lineal Simple, se va a utilizar un modelo que tiene dos variables generadas experimentalmente. La *X* que se considera una variable no aleatoria y la *Y* obtenida con una función generadora de números aleatorios normalmente distribuidos, es una variable aleatoria que tiene una subdistribución de valores de *Y* por cada valor de *X*. Posteriormente se harán ejemplos con datos reales. La relación entre ambas variables se muestra en el Gráfico 31,

Gráfico 31 Gráfico de dispersión de las variables *X*-*Y*.



Para explicar o predecir los valores de la variable *Y* a partir de la variable *X*, una forma rápida puede ser hacer la predicción de *Y* a partir de su propia media. Asumiendo que la relación entre las dos variables es lineal, el modelo seleccionado para hacer la explicación-predicción sería el de la línea recta, simbólicamente (Aplicando Fórmula 133),

$$y \mid \; a \; 2 \; bx$$

Pero al hacer la predicción a partir de la media de *Y*, queda

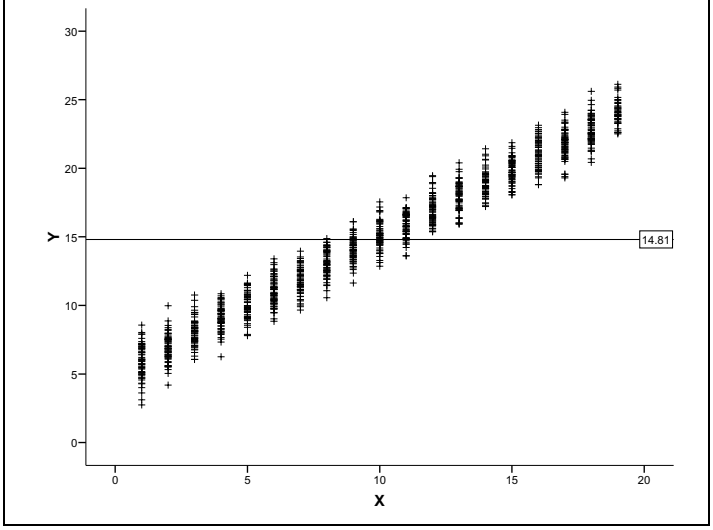
$$y \mid \; \overline{Y} \; 2 \; 0x$$

Y como *x* está multiplicada por cero la expresión es,

$$y \mid \; \overline{Y}$$

La representación es la que se ve en el Gráfico 32,

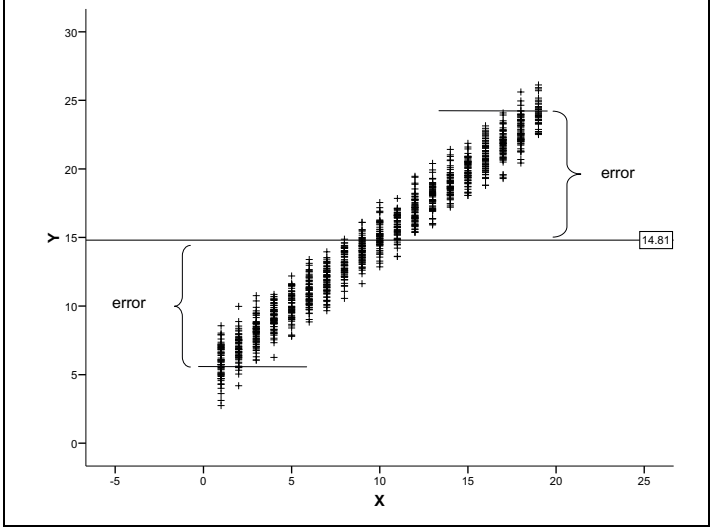
Gráfico 32 Predicción de *y* por su media.



El modelo se ha construido muy rápido. Ahora cada vez que se quiera predecir un valor de *y* para cualquier valor de *x*, sólo hay que aplicar el modelo que multiplica a *x* por cero y la anula, por lo que para cualquier *x* la variable *Y* toma siempre el mismo valor, que es su propia media.

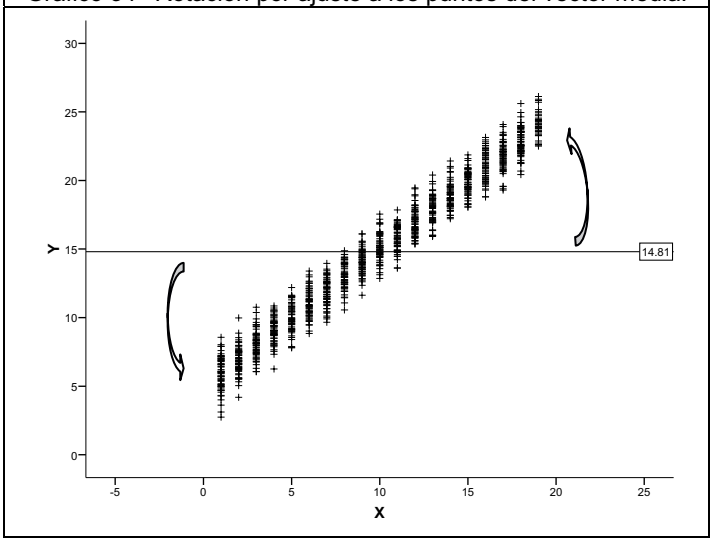
Este modelo se ajusta a los valores de *y* en la zona donde coinciden los vectores de las medias de las dos variables, pero tiene el inconveniente de que el error se incrementa a medida que se separa de los valores medios (Gráfico 33).

Gráfico 33 Error al hacer la predicción de *y* por la media de *Y*.



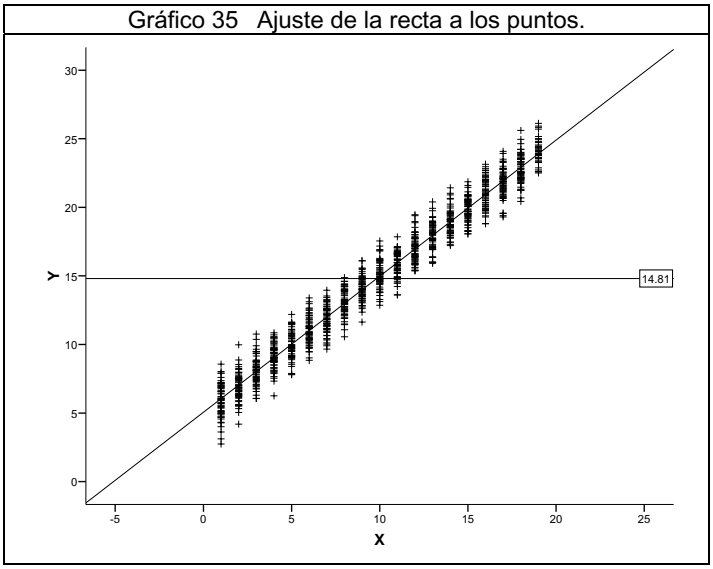
La reducción del error se consigue rotando el vector de la media de *Y* de tal manera que se ajuste lo mejor posible a la nube de puntos, pasando lo más cerca posible de todos ellos (Gráfico 34 y Gráfico 35),

Gráfico 34 Rotación por ajuste a los puntos del vector media.



Y el gráfico obtenido es,

Gráfico 35 Ajuste de la recta a los puntos.



Esta recta se obtendrá por *Mínimos Cuadrados Ordinarios*, y se llama *recta de regresión lineal.* Tiene la característica de pasar lo más cerca posible de todos los puntos tendiendo a hacer mínima la distancia elevada al cuadrado de todos los puntos respecto de la recta. Los errores del Gráfico 33 se han reducido, aunque no se han hecho cero debido a que la relación de X e Y no es funcional.

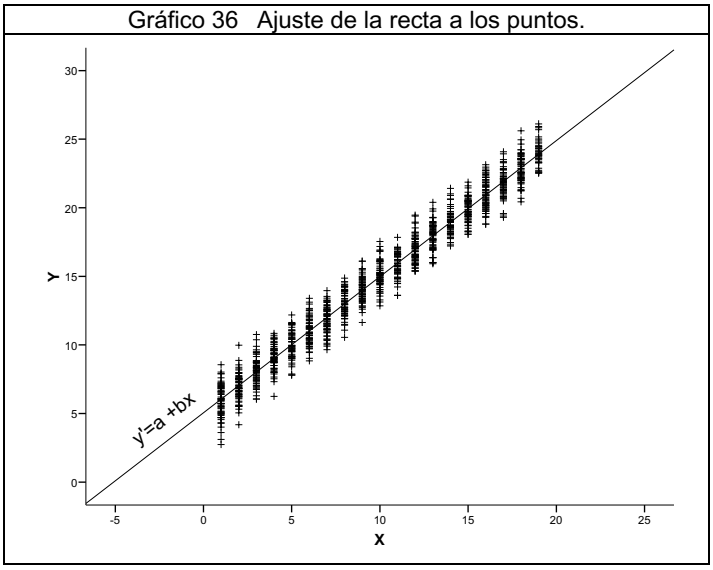
17.2 Ajuste de una recta a una nube de puntos por mínimos cuadrados ordinarios

Entre dos variables supuestamente numéricas y asumiendo que la relación es lineal de la forma que se ha visto en la Fórmula 133, se puede aplicar para estudiar la relación entre las variables y construir un modelo explicativo-predictivo. Si la relación entre las variables fuese marcadamente no lineal, se debería aplicar una ecuación o modelo apropiado. El inconveniente es que se tiene que conocer o buscar el mencionado modelo. En base a este inconveniente y a que cuando la relación no lineal se puede transformar en lineal por la aplicación de inversos, logaritmos, o cualquier otra transformación que lo consiga, se procurará operar siempre con la relación lineal,

$y' \mid a 2 bx$	Fórmula 135

Del Gráfico 36. Como en Ciencias Sociales la relación entre *x* e *y* no es funcional, el resultado de la ecuación no coincide con los valores empíricos de *y* en todos los casos, entonces a los valores obtenidos a través de la ecuación se consideran valores *teóricos* y se denominarán como *y'*.

Gráfico 36 Ajuste de la recta a los puntos.



Para obtener la ecuación que proporciona la recta que mejor se ajusta a la nube de puntos, es necesario calcular los valores de las constantes *a* y *b* y proporcionan la recta. Para obtener las constantes se utiliza el método considerado de *mínimos cuadrados ordinarios*, que es el que tiende a hacer mínimo el sumatorio de la distancia de los valores *empíricos (obsevados o reales)* (y_i) a los *teóricos (estimados)* (y'_i) elevada al cuadrado, simbólicamente,

$$\sum_{i \mid 1}^{n} y_i 4 y_i' \mid 0$$

Entonces sustituyendo *y'* según la Fórmula 135,

$$\sum_{i=1}^{n} (y_i - (a + bx_i)) = 0$$

$$\sum_{i=1}^{n} (y_i - a - bx_i) = 0$$

Calculando las derivadas parciales respecto de *a* y *b*,

$$\frac{\partial \sum_{i=1}^{n} (y_i - a - bx_i)}{\partial a} = -2\sum_{i=1}^{n} (y_i - a - bx_i) = 0$$

$$\frac{\partial \sum_{i=1}^{n} (y_i - a - bx_i)}{\partial b} = -2\sum_{i=1}^{n} (y_i - a - bx_i)x_i = 0$$

Como las dos igualdades son cero, entonces son iguales entre sí y por lo tanto eliminando el -2, la igualdad se mantiene y se puede desarrollar,

$$\sum_{i=1}^{n} (y_i - a - bx_i) = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} a - \sum_{i=1}^{n} bx_i = 0 \Rightarrow \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} a + \sum_{i=1}^{n} bx_i \Rightarrow \sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} (y_i - a - bx_i)x_i = \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} ax_i - \sum_{i=1}^{n} bx_i^2 = 0 \Rightarrow \sum_{i=1}^{n} y_i x_i = \sum_{i=1}^{n} ax_i + \sum_{i=1}^{n} bx_i^2 \Rightarrow$$

$$\sum_{i=1}^{n} y_i x_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2$$

Entonces a la Fórmula 136 y Fórmula 137, se consideran las ecuaciones normales,

$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i$	Fórmula 136

$\sum_{i=1}^{n} y_i x_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2$	Fórmula 137

Si se divide Fórmula 136 por *n,*

$$\frac{\sum_{i=1}^{n} y_i}{n} = \frac{na}{n} + \frac{b\sum_{i=1}^{n} x_i}{n}$$

Se observa que la recta ajustada por *mínimos cuadrados ordinarios* siempre pasa por el punto donde se cruzan las medias de las variables (Fórmula 138),

$\overline{Y} = a + b\overline{X}$	Fórmula 138

Y el valor de *a* buscado es,

$a = \overline{Y} - b\overline{X}$	Fórmula 139

Si la Fórmula 138 se multiplica por $n\overline{X}$ y se el resta a Fórmula 137,

$$\sum_{i=1}^{n} y_i x_i - n\overline{Y}\,\overline{X} = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 - \left(an\overline{X} + bn\overline{X}^2\right)$$

$$\sum_{i=1}^{n} y_i x_i - n\overline{Y}\,\overline{X} = a\sum_{i=1}^{n} x_i - an\overline{X} + b\sum_{i=1}^{n} x_i^2 - bn\overline{X}^2$$

$$\sum_{i=1}^{n} y_i x_i - n\overline{Y}\,\overline{X} = a\left(\sum_{i=1}^{n} x_i - n\overline{X}\right) + b\left(\sum_{i=1}^{n} x_i^2 - n\overline{X}^2\right)$$

$$a\left(\sum_{i=1}^{n} x_i - n\overline{X}\right) = a\left(\sum_{i=1}^{n} x_i - n\frac{\sum_{i=1}^{n} x_i}{n}\right) = a\left(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i\right) = 0$$

$$\sum_{i=1}^{n} y_i x_i - n\overline{Y}\,\overline{X} = b\left(\sum_{i=1}^{n} x_i^2 - n\overline{X}^2\right)$$

$b = \dfrac{\sum_{i=1}^{n} y_i x_i - n\overline{Y}\,\overline{X}}{\sum_{i=1}^{n} x_i^2 - n\overline{X}^2}$	Fórmula 140

Pero la constante *b* también es la relación entre la covarianza (S_{xy}) y la varianza de la variable independiente (S_x), simbólicamente,

$b \mid \frac{S_{xy}}{S_x^2}$	Fórmula 141

Sustituyendo en la Fórmula 141,

$$b \mid \frac{\frac{\sum_{i\mid 1}^{n} (x_i - \overline{X})(y_i - \overline{Y})}{n}}{\frac{\sum_{i\mid 1}^{n} (x_i - \overline{X})^2}{n}} \mid \frac{\frac{\sum_{i\mid 1}^{n} x_i y_i}{n} - \overline{X}\,\overline{Y}}{\frac{\sum_{i\mid 1}^{n} x_i^2}{n} - \overline{X}^2} \mid \frac{\frac{\sum_{i\mid 1}^{n} x_i y_i}{n} - \overline{X}\,\overline{Y}}{\frac{\sum_{i\mid 1}^{n} x_i^2}{n} - \frac{n\overline{X}^2}{n} - \overline{X}\frac{\sum_{i\mid 1}^{n} x_i}{n}} \mid$$

$$\frac{\frac{\sum_{i\mid 1}^{n} x_i y_i}{n} - \overline{X}\,\overline{Y}}{\frac{\sum_{i\mid 1}^{n} x_i^2}{n} - \overline{X}^2 - \overline{X}^2} \mid \frac{\frac{\sum_{i\mid 1}^{n} x_i y_i}{n} - \overline{X}\,\overline{Y}}{\frac{\sum_{i\mid 1}^{n} x_i^2}{n} - \overline{X}^2}$$

Si se multiplican todos los términos por *n,*

$$\frac{n\frac{\sum_{i\mid 1}^{n} x_i y_i}{n} - n\overline{X}\,\overline{Y}}{n\frac{\sum_{i\mid 1}^{n} x_i^2}{n} - n\overline{X}^2} \mid \frac{\sum_{i\mid 1}^{n} x_i y_i - n\overline{X}\,\overline{Y}}{\sum_{i\mid 1}^{n} x_i^2 - n\overline{X}^2}$$

Entonces se demuestra que la Fórmula 141 es igual a la Fórmula 140,

$$\frac{S_{xy}}{S_x^2} \mid \frac{\sum_{i\mid 1}^{n} x_i y_i - n\overline{X}\,\overline{Y}}{\sum_{i\mid 1}^{n} x_i^2 - n\overline{X}^2}$$

Entonces para el cálculo de la constante *b* podemos utilizar la Fórmula 140 o la Fórmula 141.

(*) Como nota aclaratoria se muestra el desarrollo de la varianza en el desarrollo de la Fórmula 141.

$$\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y}) = \sum_{i=1}^{n}(x_i y_i - x_i\bar{Y} - \bar{X}y_i + \bar{X}\bar{Y}) = \sum_{i=1}^{n}x_i y_i - \bar{Y}\sum_{i=1}^{n}x_i - \bar{X}\sum_{i=1}^{n}y_i + n\bar{X}\bar{Y} =$$

$$= \sum_{i=1}^{n}x_i y_i - \bar{Y}\sum_{i=1}^{n}x_i - \bar{X}\sum_{i=1}^{n}y_i + n\bar{X}\frac{\sum_{1=1}^{n}y_i}{n} =$$

$$= \sum_{i=1}^{n}x_i y_i - \bar{Y}\sum_{i=1}^{n}x_i - \bar{X}\sum_{i=1}^{n}y_i + \bar{X}\sum_{1=1}^{n}y_i = \sum_{i=1}^{n}x_i y_i - \bar{Y}\sum_{i=1}^{n}x_i$$

Y como está dividido por *n*,

$$\frac{\sum_{i=1}^{n}x_i y_i - \bar{Y}\sum_{i=1}^{n}x_i}{n} = \frac{\sum_{i=1}^{n}x_i y_i}{n} - \bar{Y}\frac{\sum_{i=1}^{n}x_i}{n} = \frac{\sum_{i=1}^{n}x_i y_i}{n} - \bar{Y}\bar{X}$$

Según el caso considerado, el cuadro de la Tabla 189 muestra los valores de los estadísticos necesarios para calcular las constantes *a* y *b* de la recta de regresión.

Tabla 189 Estadísticos para el cálculo de las constantes *a* y *b*.

	X	Y
media	9,84	14,82
Varianza	29,06	29,63
Covarianza	28,83	
n	1.120	
b	0,99	
a	5,08	

$$b = \frac{S_{xy}}{S_x^2} = \frac{28,83}{29,06} = 0,99$$

$$a = \bar{Y} - b\bar{X} = 14,82 - 0,99 \cdot 9,84 = 5,08$$

La ecuación buscada es,

$$y' = 5,08 + 0,99 \cdot x$$

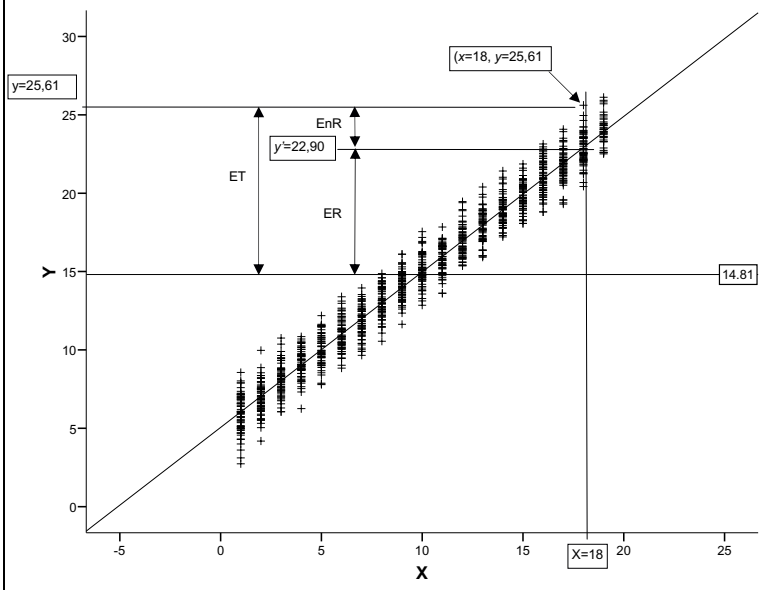
Esta recta es la que hace mínimo el sumatorio de la distancia de todos los puntos a la recta elevados al cuadrado y además es única y para cualquier *x* podemos saber su *y'*. Sea un

punto que tiene el par de coordenadas (x = 18, y = 25,61), entonces la y' será,

$$22{,}90 \mid 5{,}08 \,2\, 0{,}99 \Delta 18$$

Y gráficamente,

Gráfico 37 Representación de la resolución de un caso con el modelo de regresión.



En donde:

EnR: Error no reducido, error no explicado o variación intragrupos.
ER: Error reducido, error explicado o variación entre grupos.
ET. Error total o variación total.

En este gráfico, para este caso, el *error total* es igual a la diferencia $y \,4\, \overline{Y}$ que es igual a 25,61-14,81=10,80. Al ajustar la recta de regresión, el *error total* también llamado *variación total* se descompone (concepto de análisis de varianza) en el *error reducido* (error explicado o variación entregrupos) más el *error no reducido* (error no explicado o variación intragrupos). El *error reducido* es $y' 4 \overline{Y}$, que es 22,90-14,81=8,09, y el *error no reducido* es $y \,4\, y'$, que es 25,61-22,90=2,71. Entonces el *error total* es igual al *error reducido* más el *error no reducido*, al hacer la regresión de y sobre x, simbólicamente,

$$ET \mid ER \,2\, EnR$$

$$/y \,4\, \overline{Y}(\mid /y' \,4\, \overline{Y}(2 /y \,4\, y'0$$

$$10{,}80 \mid 8{,}09 \,2\, 2{,}71$$

Si se generaliza a todos los casos y se aplica el concepto de *descomposición de la varianza*, entonces según epígrafe 15.4, la *suma de cuadrados total* es igual a la *suma de cuadrados intragrupos* más la *suma de cuadrados entregrupos*, simbólicamente,

$$SCT \mid SCE \,2\, SCI$$

Y sustituyendo,

$$\sum_{i=1}^{N}(y_i - \bar{Y})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_i - y'_j)^2 + \sum_{j=1}^{k} n_j (y'_j - \bar{Y})^2$$

Se debe considerar que y'_j, es equivalente a la media del grupo. En el caso del ejemplo propuesto, los valores serían,

$$ET = SCT = \sum_{i=1}^{N}(y_i - \bar{Y})^2 = 33.150{,}23$$

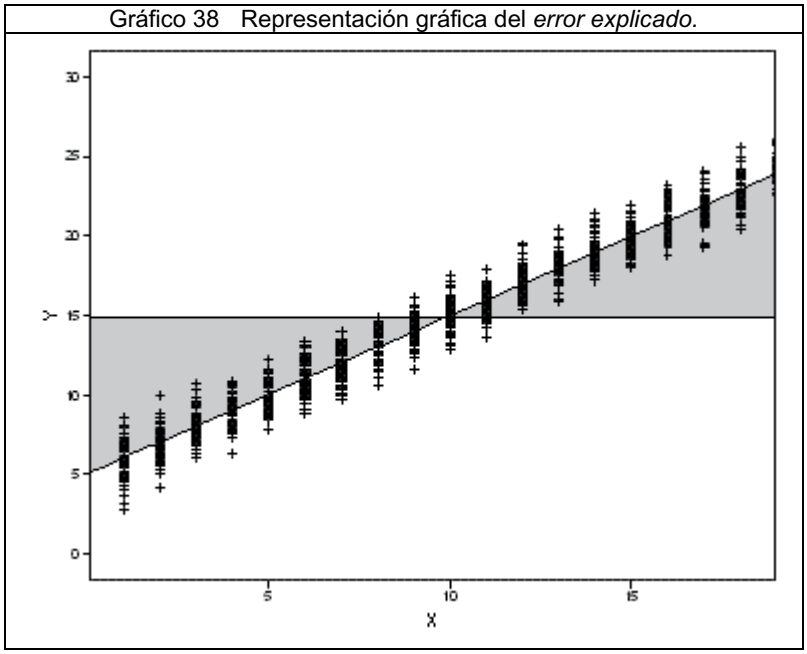
$$ER = SCE = \sum_{j=1}^{k} n_j (y'_j - \bar{Y})^2 = 32.041{,}09$$

$$EnR = SCI = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_i - y'_j)^2 = 1.109{,}14$$

$$33.150{,}23 = 32.041{,}09 + 1.109{,}14$$

El *error reducido* o *error explicado*, se puede representar gráficamente por la zona sombreada del Gráfico 38,

Gráfico 38 Representación gráfica del *error explicado.*



17.3 Calidad del ajuste

La calidad del ajuste de la recta a los puntos se mide por la distancia de estos a la recta. El estadístico que lo mide es r^2 o *coeficiente de determinación*, que es el coeficiente de correlación de Pearson elevado al cuadrado y mide la proporción de la *variación explicada* o el *error reducido* sobre el *error total* o *variación total* al hacer la regresión de la variable *Y* sobre la variable *X,* representada en sombreado en el Gráfico 38, simbólicamente.

$$r = \frac{S_{XY}}{S_X \cdot S_Y} = \frac{28{,}83}{\sqrt{29{,}06} \cdot \sqrt{29{,}63}} = 0{,}98$$

$$r^2 = (r)^2 = (0{,}98)^2 = \frac{\sum_{j=1}^{k} n_j (y'_j - \bar{Y})^2}{\sum_{i=1}^{N} (y_i - \bar{Y})^2} = \frac{32.041{,}09}{33.150{,}23} = 0{,}96 \cdot 100 = 96{,}0\%$$

Entonces, el *error reducido* o *variación explicada* al hacer la regresión de la variable *Y* sobre la variable *X* es de 0,96 o del 96,0%.

17.4 Requisitos para la aplicación de Análisis de Regresión Lineal Simple

Los requisitos para la aplicación del *análisis de regresión lineal simple* son:

1. El número de casos con el que es recomendable operar. Existen dos puntos de vista. El geométrico y el sociológico. En el primero se puede calcular una recta de regresión con dos puntos; dos casos o dos unidades de observación permiten definir una recta en el plano. Sociológicamente, se pretende que los resultados sean representativos y se puedan inferir a la población. Entonces, por los criterios expuestos en el Epígrafe 15.5 y los puntos a continuación, se debería tener, al menos, 30 casos por cada valor o categoría de la variable independiente (*X*). Existen otras opciones (J. F. Hair, 1999: 160; M. A. Cea, 2002: 15). No obstante, aunque en los estudios sociológicos las muestras tienen tamaños grandes y permiten garantizar este requisito, se pueden realizar con muestras menores, asumiendo el riesgo que comporta y dependiendo del tema investigado.
2. Las variables deben ser numéricas o supuestamente numéricas. En el caso de la variable *Y* considerada dependiente debe ser numérica supuestamente continua. La variable independiente *X* numérica y no necesariamente continua. Esta característica permite que la variable *X* pueda ser escalar, ordinal o dicotómica. Estas últimas cumplen requisitos para ser consideradas numéricas. En los manuales de SPSS utilizan variables dicotómicas codificadas como 1 y 0 (binarias o pseudobinarias, que llaman *indicadores)* como variables independientes en el análisis de regresión lineal múltiple (Norusis, 1986: B-217). No obstante, se puede codificar de diferente manera ya que el resultado no varía, aunque los coeficientes de regresión obtienen un valor distinto. Las variables de nivel de medida ordinal, cumplen requisitos de ser no aleatorias, son discretas y aunque no tienen distancia entre sus valores, si tienen orden y se pueden considerar en el ARL.

3. La variable dependiente *Y* debe ser aleatoria.
4. La variable independiente *X* debe ser no aleatoria.
5. Los dos puntos anteriores suponen que por cada valor de *X* debe haber una subdistribución de valores en *Y* (Ver, por ejemplo, el Gráfico 37).
6. Cada una de estas subdistribuciones debe ser normal, supuestamente normal o marcadamente normal.
7. Las subdistribuciones deben tener varianzas homogéneas (*Homocedasticidad)*.
8. Las predicciones de *Y* a partir de *X* deben ser en el rango conocido de *X*. Se conoce el comportamiento de *X* e *Y* en el rango de éstas, pero fuera de ese rango se desconoce si la relación sigue siendo lineal.
9. La diferencia *y-y'* (*valor empírico* menos el *valor teórico*), es el *residuo* o *error*. El *residuo* es una nueva variable. Pues bien, esta nueva variable debe tener distribución normal de media cero y desviación típica *S* de los residuos o *error típico de la estimación.* Simbólicamente $N_{(0,S)}$ (Ver epígrafe 15.5).

El cumplimiento de todos los requisitos puede hacer parecer que la aplicación del análisis de regresión sea una tarea casi imposible. Entonces es necesario conocer no sólo si se cumplen los requisitos, sino en que medida se incumplen o se violan, porque a veces ciertas violaciones pueden ser asumidas y no impedir su aplicación.

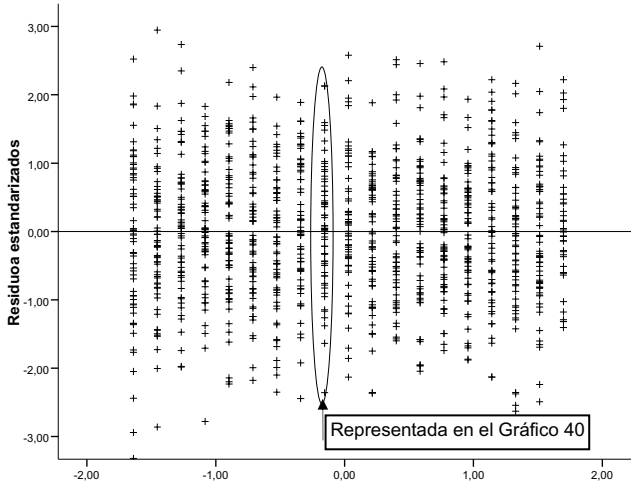
Para ver la violación de los requisitos se utilizan el *gráfico de los residuos*. Es un gráfico en el que la variable *residuos* se presenta en el eje de la variable dependiente *Y*, y en el eje de la variable independiente *X* se presenta la variable pronosticada *Y'*. La unidad de medida utilizada es *unidades de desviación típica* o *unidades Z*. El criterio de tipificación o estandarización es *Z*. Esta unidad de media permite ver variaciones significativas. Simbólicamente,

$$z_{residuo} \mid \frac{residuo \,4\, 0}{S_{residuo}}$$

$$z_{y'} \mid \frac{y' 4 \overline{Y'}}{S_{y'}}$$

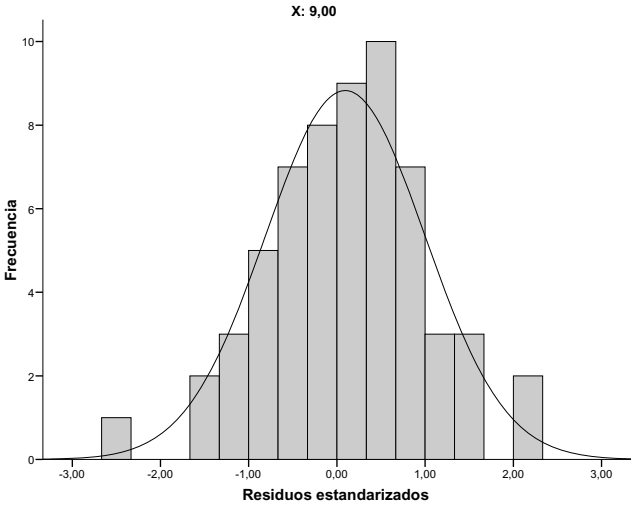
El ejemplo expuesto se había generado experimentalmente para cumplir todos los requisitos, por lo que el Gráfico 39 *de los residuos* es un modelo ideal con el que se cumplen todos los requisitos. Los puntos aparecen distribuidos alrededor del valor de la media de $z = 0$,

Gráfico 39 Gráfico de los residuos.



El Gráfico 40 muestra la normalidad[95] de la subdistribución de los *residuos* para $X = 9$ ($z_{y'} = -0{,}1556$), como ejemplo de los otros 18 gráficos de los restantes valores de X.

Gráfico 40 Gráfico de los residuos de la subdistribución $X = 9$ ($z_{y'} = -0{,}1556$).

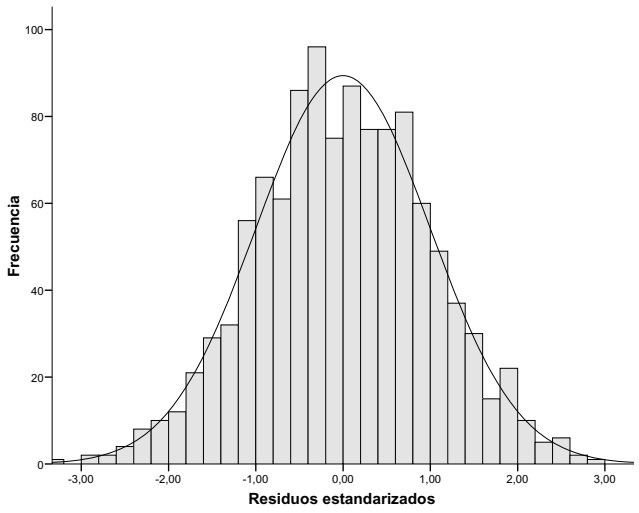


Y el Gráfico 41 muestra la normalidad[96] de la distribución de la variable de los residuales,

[95] Para confirmar la normalidad de la subdistribución se ha realizado un contratse de hipótesis de Kolmogorov-Smirnov (Ns=1,00).

[96] Para confirmar la normalidad de la subdistribución se ha realizado un contratse de hipótesis de Kolmogorov-Smirnov (Ns=0,89).

Gráfico 41 Gráfico de la normalidad de la distribución de los residuos.



17.5 Violación de requisitos en el Análisis de Regresión Lineal Simple

La violación de algunos requisitos deben presentar las formas que se muestran en el Gráfico 42,

Gráfico 42 Gráficos tipo de violación de requisitos

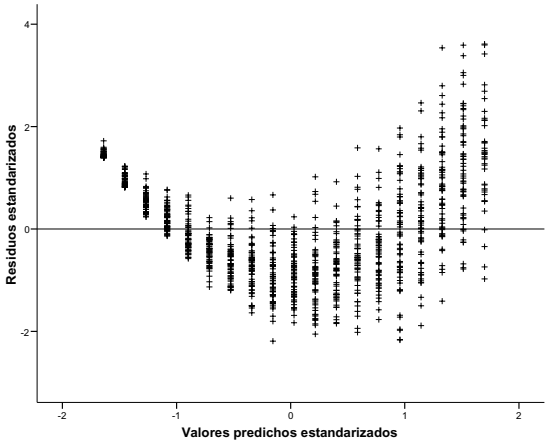
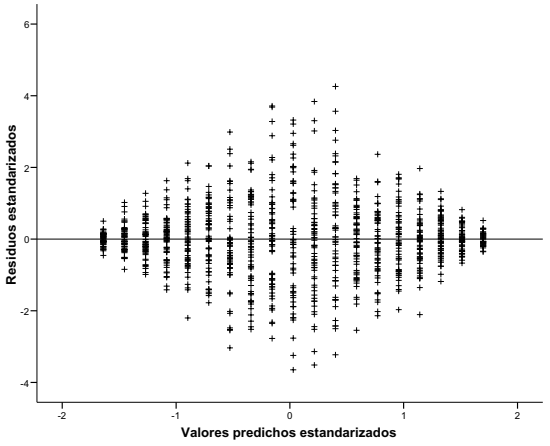
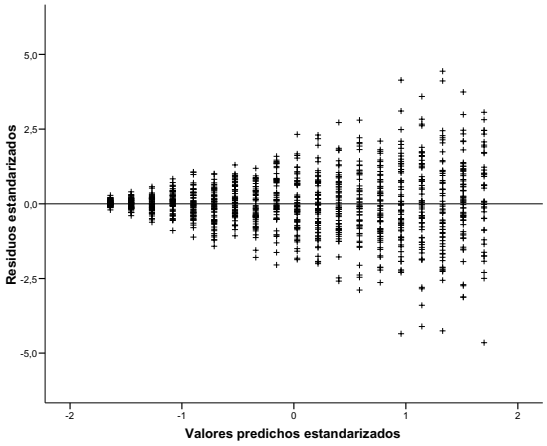
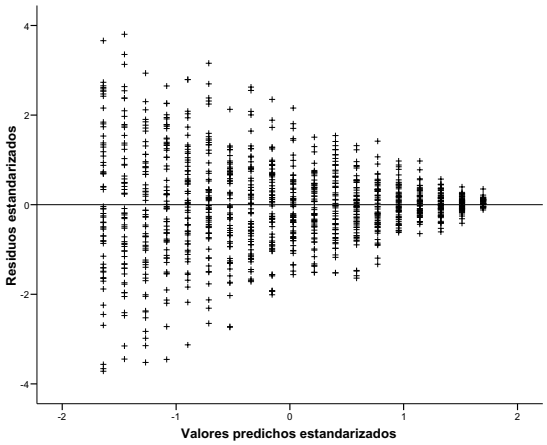
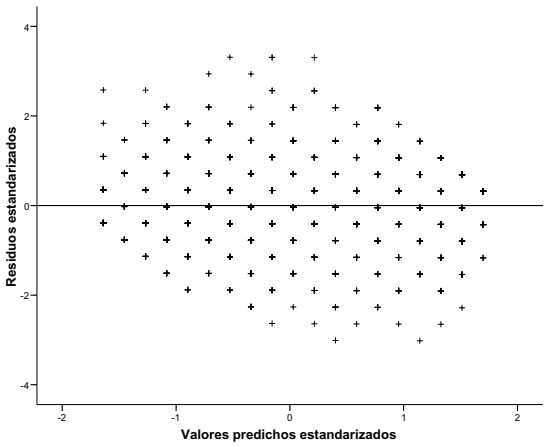
Descripción	Gráfico de residuales
Violación de linealidad por la curvatura de la nube de puntos. Violación de homogeneidad de varianza por la diferencia de dispersión en *Y* en los extremos de *X* (no homocedasticidad)	
Violación de homogeneidad de varianza. Dispersión en *Y* en los valores centrales de *X*.	

Gráfico 42	Gráficos tipo de violación de requisitos
Violación de homogeneidad de varianza. Dispersión en *Y* en los valores altos de *X*.	
Violación de homogeneidad de varianza. Dispersión en *Y* en los valores bajos de *X*.	
Violación de normalidad de las subdistribuciones. Las subdistribuciones de *Y* presentan oblicuidad negativa en valores bajos de *X* y positiva en los valores altos. En el centro hay una probable normalidad de las subdistribuciones.	

17.6 Predicción por intervalo

Si el modelo es aceptado o aceptable, porque no hay violación de requisitos o estas se pueden asumir, se puede estimar o predecir un valor de *y* por intervalo. En la Fórmula 135, *y'* es un valor teórico y considerado la media de la subdistribución. La *y* es el valor empírico. Debido al requisito 3 (Pág. 291) de la regresión, el valor de *y* empírico puede estar por encima de *y'* (la media de la subdistribución) o por debajo. La distancia o el comportamiento de los valores empíricos es la variable *residuo* o *error* y como se distribuye normalmente con media igual a cero y desviación típica conocida ($N_{(0,S)}$), se puede calcular el intervalo de confianza para un determinado *Nc* dentro del cual estará el valor empírico buscado, simbólicamente,

$$y \mid y' \partial e$$

$$y' \mid a 2 bx$$

$$y \mid (a 2 bx) \partial e$$

Y como el error (los residuos) tienen una distribución $N_{(0,S)}$, se puede definir un intervalo de confianza para un determinado *Nc,* dentro del cual estará comprendido el valor de *e*,

$$/0 4 z \Delta S_{residuos} 0\{ 0 \{ /0 2 z \Delta S_{residuos} 0$$

Y sustituyendo,

$y \mid (a 2 bx) \partial / z \Delta S_{residuos} 0$	Fórmula 142

Estando el valor de *z* definido por el *Nc*. Entonces, para un *Nc* = 0,9544, *z* = 2,00, el valor de *y* estimado o predicho, estará en el intervalo,

$y \mid (a 2 bx) \partial / 2{,}00 \Delta S_{residuos} 0$	Fórmula 143

En el ejemplo de la Tabla 189, el intervalo dentro del cual estará un valor estimado de *y*, para *x* = 19 y *Nc* = 0,9544 (*z* = 2) o el intervalo dentro del cual estarán el 95,44% de los casos, está definido por ($S_{residuos}$ = 1,00),

$$y \mid 5{,}06 2 0{,}99 \Delta 19 \partial / 2{,}00 \Delta 1{,}00 0$$

$$y \mid 23{,}87 \partial 2{,}00$$

$$/23{,}87 4 2{,}000\{ y \{ /23{,}87 2 2{,}000$$

$$/21{,}870\{ y \{ /25{,}870$$

Para un valor de *x* = 19, en el intervalo de confianza de *y* = 21,87 ÷ 25,87 estarán el 95,44% de los casos.

17.7 Ejemplo de Análisis de Regresión Lineal Simple

El ejemplo es de Análisis de Regresión Lineal Simple, por ser el tema tratado en este manual. Lo más adecuado sería un Análisis de Regresión Lineal Múltiple porque se utilizaría más de una variable independiente para explicar o predecir la variable dependiente. Una vez vistos los supuestos básicos se puede seguir el Análisis de Regresión Lineal Múltiple por C. De la Puente (1995), J. F. Hair (1999) y M. A. Cea (2002).

Supuesto. Un cervecero español quiere introducir la cerveza que fabrica en España, en el mercado de los EEUU de Norteamérica.

Problema. El cervecero necesita saber cuál es el precio de mercado de su cerveza y se puede hacer el siguiente planteamiento,

- Fabricar la cerveza y llevarla a los EE. UU. para venderla. Para calcular el precio de venta de la cerveza, debe considerar los gastos de producción, gastos financieros, gastos de distribución y posibles aranceles. Si el precio al que debe vender la cerveza para cubrir todos los gastos más los beneficios es elevado, tiene el riesgo de no venderlas y tener que traerlas a España otra vez, lo que le supone un gasto mayor. Pero no puede bajar mucho el precio porque debe cubrir gastos.
- Si el precio al que debe vender es suficiente para obtener beneficios entonces puede llevar a cabo la operación. Pero si el precio no le reporta beneficios y considerando que el precio de la distribución puede ser el más elevado, entonces puede plantearse el instalar una fábrica de cervezas en los EE. UU. o en un país próximo que tenga relación comercial.

El punto de partida puede ser solicitar un estudio del mercado de las cervezas en los EE. UU. para que en base al precio de las cervezas que se venden, hacer una primera aproximación al precio que puede vender la cerveza y tomar decisiones.

Resolución: Entonces, se quiere calcular el precio que se va a recomendar al cervecero español que le debe poner a su unidad de producto, si quiere tener una cierta probabilidad de éxito en el mercado en el que quiere entrar. Se va a llamar al precio la variable *y* o variable dependiente y se va a calcular en base a la característica (variable *x*) que mejor permita explicar y predecir ese precio. Se opta entonces por un modelo generado según *mínimos cuadrados ordinarios* (recta de regresión) que según la Fórmula 133 es.

$$y \mid a 2 bx$$

Pero como el valor de la ecuación es teórico, se opta por la Fórmula 135

$$y' \mid a 2 bx$$

Y en el caso de las cervezas sería,

$$\text{precio unitario}' \mid a 2 b\Delta/\text{característica de la cerveza}0$$

Para conocer el modelo hay que seleccionar primero cuál es la "característica" que mejor permite explicar y predecir el precio unitario de la cerveza y después hallar los valores de los coeficientes *a* (punto de origen o punto donde la recta de regresión corta al eje *Y*) y *b* (coeficiente de regresión o pendiente de la recta).

El problema requiere "conocer" algo, que es el mercado de las cervezas de los EE. UU., y a con este conocimiento elaborar el modelo que permita estimar el precio unitario de la cerveza. Entonces se aplica el Método Científico (ver Epígrafe 1).

1. Diseño Teórico

1.1. Definición del problema: recomendación del precio de mercado a un cervecero español para introducir su producto en el mercado de los EE. UU. de Norteamérica. Definición de conceptos: Por ejemplo pueden ser: significado del contenido en sodio de las cervezas, significado de las calorías en las cervezas, significado del contenido en alcohol de las cervezas, etc. Justificación de la investigación: Ha sido solicitada por un cliente.

1.2. Marco teórico: Conocimientos necesarios sobre el producto (cerveza) y de estudios de mercado, así como de modelos estadísticos.

1.3 Objetivos: Elaborar un modelo lineal para estimar el precio unitario de la cerveza, por intervalo. Hipótesis: No hay.

1.4. Las variables: Se van a utilizar Datos Secundarios de un estudio sobre cervezas de los EE UU de Norteamérica obtenido de los manuales de SPSS (Norusis, 1986: B-1), entonces las variables son (son palabras clave y se omiten reglas ortográficas): *calidad, nombre, origen, area, pre6, pre1, calorias, sodio, alcohol, clase* y *fuerza* (Ver Tabla 190, Tabla 191 y Tabla 192).

1.5. Los indicadores: se pueden considerar indicadores: *calorias, sodio* y *alcohol*.

2. Diseño Técnico

2.1. El Universo es el Censo de las cervezas ($N = 35$) que se venden en los EE. UU. de Norteamérica. La unidad de observación es la "cerveza".

2.2. La muestra: No procede el diseño de muestra porque se opera con el Censo y se utilizan datos secundarios.

2.3. Técnica de Investigación. No procede, porque se recoge la información de todo el Universo y se utilizan datos secundarios.

2.4. Instrumento de obtención de Datos: No procede porque se utilizan Datos Secundarios.

2.5. Codificación, grabación, tabulación y análisis: Al utilizar datos secundarios no es necesaria la codificación. La grabación se ha realizado desde los datos ofrecidos en el manual de SPSS (Norusis, 1986: B-1). La tabulación no se va a aplicar y se elaborará un modelo de análisis de regresión lineal simple.

La información de la matriz de datos, según la produce SPSS es (Tabla 190),

Tabla 190 Información de las variables.

Variable	Etiqueta	Códigos	Etiquetas de códigos
calidad	Calidad de la cerveza	1	Muy Buena
		2	Buena
		3	Regular
nombre	Nombre de la Cerveza		
origen	Lugar donde se fabricó la cerveza	1	USA
		2	Canadá
		3	Francia
		4	Holanda
		5	México
		6	Alemania
		7	Japón
area	Zona de disponibilidad de la cerveza	1	Nacional
		2	Regional
pre6	Precio de la cerveza por seis unidades		
pre1	Precio de la cerveza por unidad		
calorias	Contenido en calorías de la cerveza		
sodio	Contenido en sodio de la cerveza		
alcohol	Contenido en alcohol en % de la cerveza		
clase	Clase de la cerveza	0	Sin Clase
		1	Super-Premium
		2	Premium
		3	Popular
fuerza	Cerveza regular o light	0	Regular
		1	Light

La matriz de datos se presenta completa en la Tabla 191 por ser un ejemplo de $N = 35$. Se considera censo porque se supone que contiene todas las marcas de cervezas que se comercializaban en los EE. UU. en el momento de realizar el estudio.

Tabla 191 Matriz de datos con códigos y valores.

calidad	nombre	origen	area	Pre6	pre1	Calorias	sodio	alcohol	clase	fuerza	origen_R
1	Miller High Life	1	1	2,49	0,42	149	17	4,7	2	0	0
1	Budweiser	1	1	2,59	0,43	144	15	4,7	2	0	0
1	Schlitz	1	1	2,59	0,43	151	19	4,9	2	0	0
1	Lowenbrau	1	1	2,89	0,48	157	15	4,9	1	0	0
1	Michelob	1	1	2,99	0,50	162	10	5,0	1	0	0
1	Labatts	2	2	3,15	0,53	147	17	5,0	0	0	1
1	Molson	2	2	3,35	0,56	154	17	5,1	0	0	1
1	Henry Weinhard	1	2	3,65	0,61	149	7	4,7	1	0	0
1	Kronenbourg	3	2	4,39	0,73	170	7	5,2	0	0	1
1	Heineken	4	1	4,59	0,77	152	11	5,0	0	0	1
1	Anchor Steam	1	2	7,19	1,20	154	17	4,7	1	0	0
2	Old Milwaukee	1	2	1,69	0,28	145	23	4,6	3	0	0
2	Schmidts	1	2	1,79	0,30	147	7	4,7	3	0	0
2	Pabst Blue Ribbon	1	1	2,29	0,38	152	8	4,9	2	0	0
2	Augsberger	1	2	2,39	0,40	175	24	5,5	1	0	0
2	Strohs Bohemian Style	1	2	2,49	0,42	149	27	4,7	2	0	0
2	Miller Light	1	1	2,55	0,43	99	10	4,3	0	1	0
2	Budweiser Light	1	1	2,63	0,44	113	8	3,7	0	1	0
2	Coors	1	2	2,65	0,44	140	18	4,6	2	0	0
2	Olympia	1	2	2,65	0,44	153	27	4,6	2	0	0
2	Coors Light	1	2	2,73	0,46	102	15	4,1	0	1	0
2	Michelob Light	1	1	2,99	0,50	135	11	4,2	0	1	0
2	Dos Equis	5	2	4,22	0,70	145	14	4,5	0	0	1
2	Becks	6	2	4,55	0,76	150	19	4,7	0	0	1
2	Kirin	7	2	4,75	0,79	149	6	5,0	0	0	1
3	Scotch Buy (Safeway)	1	2	1,59	0,27	145	18	4,5	0	0	0
3	Blatz	1	2	1,79	0,30	144	13	4,6	3	0	0
3	Rolling Rock	1	2	2,15	0,36	144	8	4,7	2	0	0
3	Pabst Extra Light	1	1	2,29	0,38	68	15	2,3	0	1	0
3	Hamms	1	2	2,59	0,43	136	19	4,4	2	0	0
3	Heilemans Old Style	1	2	2,59	0,43	144	24	4,9	2	0	0
3	Tuborg	1	2	2,59	0,43	155	13	5,0	2	0	0
3	Olympia Gold Light	1	2	2,75	0,46	72	6	2,9	0	1	0
3	Schlitz Light	1	1	2,79	0,47	97	7	4,2	0	1	0
3	St Pauli Girl	6	2	4,59	0,77	144	21	4,7	0	0	1

Tabla 192 Matriz de datos con etiquetas de códigos y valores.

calidad	nombre	origen	area	pre6	pre1	calorias	sodio	alcohol	clase	fuerza	origen_R
Muy buena	Miller High Life	USA	Nacional	2,49	,42	149	17	4,7	Premium	Regular	USA
Muy buena	Budweiser	USA	Nacional	2,59	,43	144	15	4,7	Premium	Regular	USA
Muy buena	Schlitz	USA	Nacional	2,59	,43	151	19	4,9	Premium	Regular	USA
Muy buena	Lowenbrau	USA	Nacional	2,89	,48	157	15	4,9	Super-premium	Regular	USA
Muy buena	Michelob	USA	Nacional	2,99	,50	162	10	5,0	Super-premium	Regular	USA
Muy buena	Labatts	Canadá	Regional	3,15	,53	147	17	5,0	Sinclase	Regular	No USA
Muy buena	Molson	Canadá	Regional	3,35	,56	154	17	5,1	Sinclase	Regular	No USA
Muy buena	Henry Weinhard	USA	Regional	3,65	,61	149	7	4,7	Super-premium	Regular	USA
Muy buena	Kronenbourg	Francia	Regional	4,39	,73	170	7	5,2	Sinclase	Regular	No USA
Muy buena	Heineken	Holanda	Nacional	4,59	,77	152	11	5,0	Sinclase	Regular	No USA
Muy buena	Anchor Steam	USA	Regional	7,19	1,20	154	17	4,7	Super-premium	Regular	USA
Buena	Old Milwaukee	USA	Regional	1,69	,28	145	23	4,6	Popular	Regular	USA
Buena	Schmidts	USA	Regional	1,79	,30	147	7	4,7	Popular	Regular	USA
Buena	Pabst Blue Ribbon	USA	Nacional	2,29	,38	152	8	4,9	Premium	Regular	USA
Buena	Augsberger	USA	Regional	2,39	,40	175	24	5,5	Super-premium	Regular	USA
Buena	Strohs Bohemian Style	USA	Regional	2,49	,42	149	27	4,7	Premium	Regular	USA
Buena	Miller Light	USA	Nacional	2,55	,43	99	10	4,3	Sinclase	Light	USA
Buena	Budweiser Light	USA	Nacional	2,63	,44	113	8	3,7	Sinclase	Light	USA
Buena	Coors	USA	Regional	2,65	,44	140	18	4,6	Premium	Regular	USA
Buena	Olympia	USA	Regional	2,65	,44	153	27	4,6	Premium	Regular	USA
Buena	Coors Light	USA	Regional	2,73	,46	102	15	4,1	Sinclase	Light	USA
Buena	Michelob Light	USA	Nacional	2,99	,50	135	11	4,2	Sinclase	Light	USA
Buena	Dos Equis	México	Regional	4,22	,70	145	14	4,5	Sinclase	Regular	No USA
Buena	Becks	Alemania	Regional	4,55	,76	150	19	4,7	Sinclase	Regular	No USA
Buena	Kirin	Japón	Regional	4,75	,79	149	6	5,0	Sinclase	Regular	No USA
Regular	Scotch Buy (Safeway)	USA	Regional	1,59	,27	145	18	4,5	Sinclase	Regular	USA
Regular	Blatz	USA	Regional	1,79	,30	144	13	4,6	Popular	Regular	USA
Regular	Rolling Rock	USA	Regional	2,15	,36	144	8	4,7	Premium	Regular	USA
Regular	Pabst Extra Light	USA	Nacional	2,29	,38	68	15	2,3	Sinclase	Light	USA
Regular	Hamms	USA	Regional	2,59	,43	136	19	4,4	Premium	Regular	USA
Regular	Heilemans Old Style	USA	Regional	2,59	,43	144	24	4,9	Premium	Regular	USA
Regular	Tuborg	USA	Regional	2,59	,43	155	13	5,0	Premium	Regular	USA
Regular	Olympia Gold Light	USA	Regional	2,75	,46	72	6	2,9	Sinclase	Light	USA
Regular	Schlitz Light	USA	Nacional	2,79	,47	97	7	4,2	Sinclase	Light	USA
Regular	St Pauli Girl	Alemania	Regional	4,59	,77	144	21	4,7	Sinclase	Regular	No USA

La variable que mejor explica o predice el precio es la que tiene el coeficiente de correlación mayor en valor absoluto (puede ser positivo o negativo) y se selecciona desde la matriz de correlaciones. Aunque ahora no se contempla, se debe considerar que otras variables que tengan correlaciones bajas pueden estar ocultas por la influencia de terceras variables (ver el ejemplo de la nota 92). Las variables consideradas para obtener la matriz de correlaciones son, la variable dependiente: *pre1* y como posibles variables independientes, explicativas o predictoras: *calidad*, *origen*, *area, calorias, sodio, alcohol, clase* y *fuerza*.

La variable *calidad* se considera por ser ordinal. La variable *origen* no se puede considerar como independiente porque es de nivel de medida nominal y tiene más de dos categorías y en este caso, aunque se pudiese incorporar, no tiene sentido porque no figura la categoría *España* y no se podría sustituir para darle valores a *pre1* (*Y*). No obstante, se estima que la distribución puede ser determinante del precio del producto, entonces interesa su inclusión.

En base a lo anterior, se deben buscar estrategias para incorporar la variable *origen* en la matriz de correlaciones como candidata para ser seleccionada para el modelo.

El primer paso es hacerla binaria (pseudobinaria o dicotomizarla) y asignarle categorías que permitan la inclusión de España con algún criterio teórico-estadístico. Si se recodifica la variable *origen* como "cervezas producidas en los EE. UU." y cervezas "no producidas en los EE. UU." se convierte en dicotómica y además se puede considerar el caso "España". En este supuesto no se contempla diferencias entre "Continente Americano", "Continente Europeo" y "Continente Asiático" que probablemente sea un factor determinante en el precio. Tampoco se contemplan los posibles aranceles.

Las influencias se pueden considerar casi infinitas por lo que se aplica el "*céteris*

páribus", esto es, que se considera todo lo demás constante o que no influye. Otra opción para incorporar variables categóricas de más de dos categorías en el análisis de Regresión Lineal, como variable independiente, es crear tantas variables binarias como categorías tiene menos una, la cual es constante y considerada la categoría de referencia para las demás (ver el procedimiento de Regresión Logística Binaria de SPSS, el subcomando *Variable Categórica* y el manual de ayuda).

La opción seleccionada para hacer el ejemplo es dicotomizar la variable *origen* como "cervezas producidas en los EE. UU." y cervezas "no producidas en los EE. UU.".

La variable *area* se considera por ser dicotómica. Las variables *calorias, sodio* y *alcohol* se consideran numéricas. La variable *clase* no se considera porque la categoría "sin clase" rompe la posible ordinalidad de la variable y *fuerza* se considera por ser binaria. Entonces la matriz de correlaciones es (Tabla 193),

Tabla 193 Matriz de correlaciones.

	pre1	calidad	origen_R	area	calorias	sodio	Alcohol	fuerza
pre1	1,00	-0,37*	0,58**	0,14	0,21	-0,09	0,20	-0,15
calidad	-0,37*	1,00	-0,24	0,28	-0,46**	0,04	-0,44**	0,30
origen_R	0,58**	-0,24	1,00	0,25	0,26	-0,06	0,30	-0,27
area	0,14	0,28	0,25	1,00	0,25	0,30	0,22	-0,39*
calorias	0,21	-0,46**	0,26	0,25	1,00	0,28	0,91**	-0,87**
sodio	-0,09	0,04	-0,06	0,30	0,28	1,00	0,20	-0,36*
alcohol	0,20	-0,44**	0,30	0,22	0,91**	0,20	1,00	-0,76**
fuerza	-0,15	0,30	-0,27	-0,39*	-0,87**	-0,36*	-0,76**	1,00

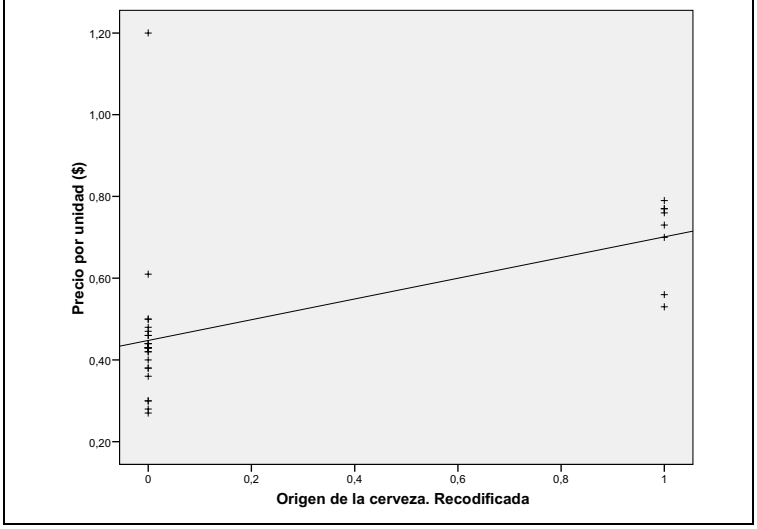
Notas:
*: La correlación es significativa al Ns de 0,05 (bilateral).
**: La correlación es significativa al Ns de 0,01 (bilateral).

La variable independiente que tiene la correlación mayor (0,58) con la variable dependiente (*pre1*) es el origen (*origen_R*). Entonces el modelo buscado es,

$$pre1' \mid a\,2\,b\,\Delta origen\_R$$

Y el gráfico de dispersión o *X-Y*, para comprobar la supuesta linealidad de la relación de las dos variables es (Gráfico 43),

Gráfico 43 Gráfico de dispersión de las variables *pre1* y *origen_R*.



Como el origen de las cervezas puede ser de los EE. UU. o de fuera de los mismos, la variable presenta dos categorías, codificadas como 0 y 1, por lo que sólo hay dos subdistribuciones. Al haber sólo dos subdistribuciones se puede asumir que la relación es lineal. Por supuesto también se podrían asumir otras relaciones, pero no interesan en este caso. Los estadísticos descriptivos son,

Tabla 194 Estadísticos descriptivos.

Variables	Media	Desviación típica	Varianza	Covarianza
pre1 (*Y*)	0,51	0,18	0,03	0,04
origen_R (*X*)	0,23	0,42	0,18	

Los coeficientes *a* y *b* de la recta son,

$$b = \frac{S_{xy}}{S_x^2} = \frac{0,04}{0,18} = 0,25$$

$$a = \overline{Y} - b\overline{X} = 0,51 - 0,25 \cdot 0,23 = 0,45$$

Y sustituyendo en la ecuación tenemos,

$$pre1' = 0,45 + 0,25 \cdot origen\_R$$

Como la cerveza para la que se ha construido el modelo es española y pertenece a la categoría de las "No USA" (valor = 1) se sustituye el valor en la variable *origen_R* y obtenemos,

$$pre1' = 0,45 + 0,25 \cdot \text{No USA} \rightarrow 0,45 + 0,25 \cdot 1$$

$$pre1' = 0,70\$$$

El precio teórico al que se debe vender la cerveza española, asumiendo que las demás variables son constantes o que no varían ("*céteris páribus*"), es de 0,70 $. Pero como el precio real puede estar por encima o por debajo de este valor, el intervalo de confianza, para un *Nc* = 0,9544, dentro del cual estará el precio buscado, según la Fórmula 142 y con la variable residuales $N_{(0;0,15)}$ es,

$$y = (0,45 + 0,25 \cdot 1) \pm 2,00 \cdot 0,150$$

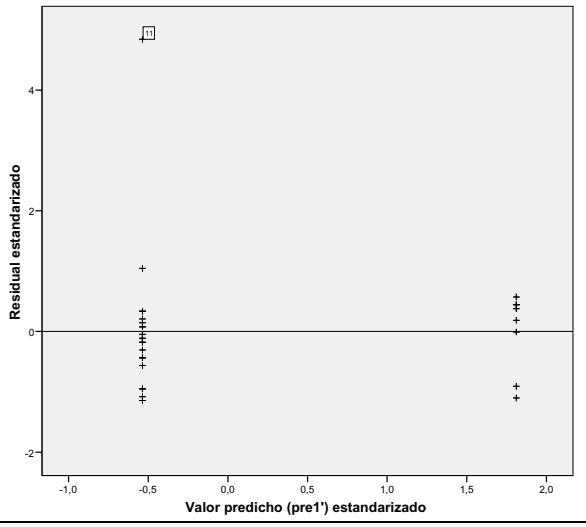
$$y = 0,700 \pm 0,300$$

$$y = 0,40\$ \leq 1,00\$$$

El cervecero español debe poner el precio de la cerveza entre 0,40 $ y 1,00 $. El intervalo se puede considerar que es amplio, por lo que el estudio puede estar bien hecho, pero el intervalo se puede considerar grande.

El gráfico de los residuos es (Gráfico 44),

Gráfico 44 Gráfico de los residuos.



El coeficiente de determinación o de variación explicada (r^2) es 0,33 o lo que es lo mismo, al hacer la regresión de la variable precio por unidad sobre el origen de la cerveza, la variación explicada o error reducido es del 33,0 %.

Para mejorar la explicación del modelo, reducir el error no explicado o aumentar el error reducido e intentar reducir el intervalo de confianza, se puede buscar alguna estrategia. En este caso, la cerveza *Anchor Steam* (caso nº 11. Ver Gráfico 44), tiene un precio por unidad de 1,20 $ y su origen es EE. UU. El valor *z* del residuo es significativamente grande (4,99) por lo que se puede considerar un caso extremo ("*outlier*"). Puesto que está alterando el modelo por tener un precio elevado, no aporta explicación y el modelo tampoco explica bien este caso, se puede elaborar un modelo alternativo eliminando este caso. Si los lectores hacen la prueba, comprobarán que el nuevo modelo es,

$$pre1' \mid 0{,}42\ 2\ 0{,}28\,\Delta\,\text{No USA} \blacktriangledown\ 0{,}42\ 2\ 0{,}28\,\Delta 1$$

$$pre1' \mid 0{,}70\$$$

Por lo tanto el precio teórico con el nuevo modelo no varía y la variable de residuales es $N_{(0;0{,}08)}$. El nuevo coeficiente de correlación de Pearson r = 0,84 y r^2 = 0,70. Se ha mejorado el modelo reduciendo la desviación típica de los residuales y la calidad del modelo, es obvio, al haber eliminado un caso que estaba introduciendo mucha dispersión, y el nuevo intervalo de confianza es,

$$y \mid (0{,}42 \pm 0{,}28 \cdot 1) \cdot 2{,}00 \cdot 0{,}08$$

$$y \mid 0{,}70 \cdot 0{,}16$$

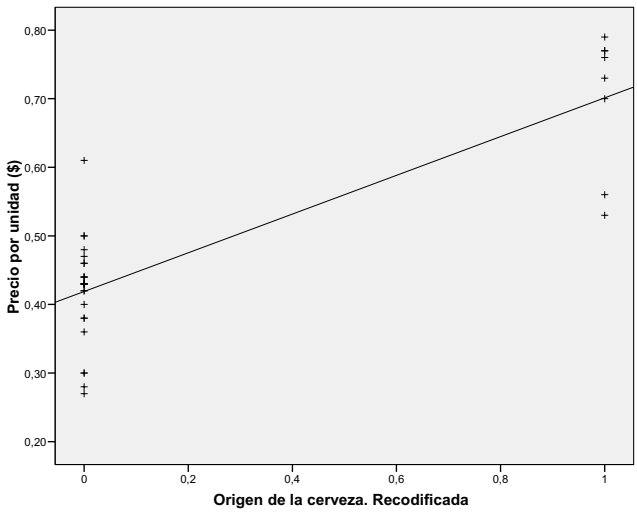
$$y \mid 0{,}54\$ \le 0{,}86\$$$

El intervalo se ha reducido considerablemente y ahora puede ser de mayor utilidad para el cervecero. Esta operación no significa que hay que eliminar casos hasta que el modelo parezca adecuado, sólo se deben eliminar casos en base a algún criterio válido desde un punto de vista teórico-estadístico.

Con esta información, más la información del precio financiero y otra posible información, se dispone de elementos de juicio para tomar una decisión y que el cliente pueda asignar un precio a su unidad de producto e ir al mercado de los EE. UU., construir una fábrica en alguna región periférica, en una región interna o no ir.

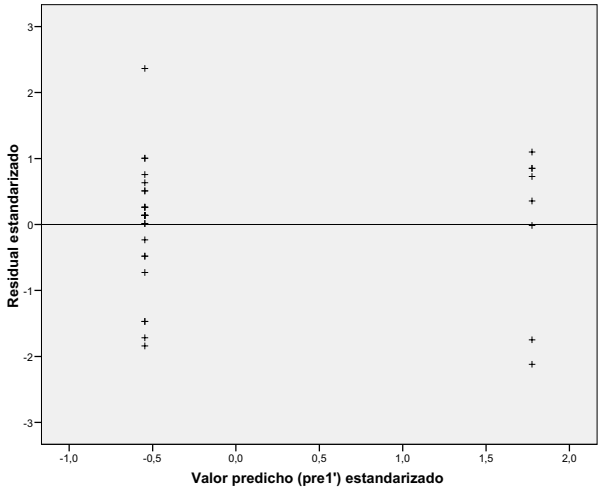
El gráfico de dispersión sin el caso nº 11 es (Gráfico 45),

Gráfico 45 Gráfico de dispersión de las variables *pre1* y *origen_R* sin el caso nº 11.



Y el gráfico de los residuos, después de haber eliminado el caso nº 11 es (Gráfico 46),

Gráfico 46 Gráfico de los residuos sin el caso nº 11.



En este ejemplo, se va a considerar que la nube de puntos permite asumir que se cumplen todos los requisitos, pero teniendo en cuenta que el número de casos es inferior a 30 por cada categoría de la variable independiente, pero considerando que es la información de la que se dispone y que el Censo de cervezas de los EE. UU. no son más, asumimos el modelo. Por lo tanto, por cada categoría de la variable independiente (*origen_R*) hay una subdistribución de valores de la variable dependiente (*pre1*), que por la forma de la nube de puntos (horizontal a lo largo de la media cero de los residuos), asumimos que son subdistribuciones normales y con varianzas homogéneas.

Aunque no es objeto de este manual, según se dijo al principio del ejemplo, lo más adecuado sería un Análisis de Regresión Lineal Múltiple, con el que se buscaría un modelo con más variables independientes para determinar el intervalo del precio por unidad.

18 Números Índice

El *número índice* es un instrumento estadístico utilizado para describir la evolución de una variable en el tiempo. Las variables se corresponden con magnitudes económicas, demográficas, sociológicas, etc. y a veces sintéticas, es decir, originadas por un conjunto de subvariables que la conforman. Algunos ejemplos son: evolución de la población en un período de tiempo, del *IPC*, la producción de un país, etc. Los *números índice* se pueden agrupar según la Tabla 195,

Tabla 195 Clasificación de los números índice.

Simples			
Compuestos	Sin ponderar	Media aritmética Agregativo simple	
	Ponderados	Media aritmética	
		Agregativo compuesto	Laspeyres Paache Fisher

Los *números índice simples* tratan de analizar la evolución de una variable en el tiempo comparándola con el valor que toma en el año considerado inicial, denominado *año base*. Ejemplo: evolución de la población de un país.

Los *números índice compuestos* tratan de analizar la evolución de una variable en un período de tiempo cuando dicha variable está formada por un conjunto de otras variables que se pueden denominar subvariables. Ejemplo: evolución del *IPC* compuesto por cinco variables (alimentación, vestido, vivienda, gastos de casa y gastos generales.

En los *índice compuestos* hay que diferenciar si cada uno de los componentes tiene la misma importancia, en cuyo caso se denominan *índice compuesto sin ponderar*, y cuando cada uno de los componentes tiene diferente peso específico y se denomina *índice compuesto ponderado*.

18.1 Números índice simples

Es el índice que pretende relacionar la cuantía de la variable de interés respecto al valor de la misma variable en un período determinado, que se llama período *base.* El *índice* en el período base, por este procedimiento, siempre tiene el valor 100,00.

$I_t \mid \frac{X_t}{X_0}$ ó $I_t \mid \frac{X_t}{X_0}\Delta 100$	En donde: I_t: Índice en el período *t*. X_t: Valor de la variable en el período *t*. X_0: Valor de la variable en el período base.	Fórmula 144

La *tasa de variación* entre dos períodos es la relación entre el período de interés y el período de referencia multiplicado por 100.

$V_{(t1 4 t2)} \mid \left(\frac{I_{t2}}{I_{t1}} 4 1\right)\Delta 100$	En donde: I_{t2}: Índice en el período *t2*. I_{t1}: Índice en el período *t1* de referencia. $V_{(t1\text{-}t2)}$: Variación en el período 1 a 2.	Fórmula 145

Tabla 196 Ejemplo de números índice y tasa de variación.					
Año	Total (x 10^3)	Fórmula	Índice (%)	Fórmula Variación (I_0 Ÿ I_t)	Tasa de variación a año base (%)
2008	44.468	$I_{2008} \mid \frac{X_{2008}}{X_{2008}}\Delta 100 \mid \frac{44.468}{44.468}*100 \mid$	100,00		
2009	44.906	$I_{2009} \mid \frac{X_{2009}}{X_{2008}}\Delta 100 \mid \frac{44.906}{44.468}*100 \mid$	100,98	$V_{200842009} \mid \left(\frac{I_{2009}}{I_{2008}} 4 1\right)\Delta 100 \mid$	0,98
2010	45.311	$I_{2010} \mid \frac{X_{2010}}{X_{2008}}\Delta 100 \mid \frac{45.311}{44.468}*100 \mid$	101,90	$V_{200842010} \mid \left(\frac{I_{2010}}{I_{2008}} 4 1\right)\Delta 100 \mid$	1,90
2011	45.686	$I_{2011} \mid \frac{X_{2011}}{X2008}\Delta 100 \mid \frac{45.686}{44.468}*100 \mid$	102,74	$V_{200842011} \mid \left(\frac{I_{2011}}{I_{2008}} 4 1\right)\Delta 100 \mid$	2,74
2012	46.055	$I_{2012} \mid \frac{X_{2012}}{X_{2008}}\Delta 100 \mid \frac{46.055}{44.468}*100 \mid$	103,57	$V_{200842012} \mid \left(\frac{I_{2012}}{I_{2008}} 4 1\right)\Delta 100 \mid$	3,57
2013	46.418	$I_{2013} \mid \frac{X_{2013}}{X_{2008}}\Delta 100 \mid \frac{46.418}{44.468}*100 \mid$	104,39	$V_{200842013} \mid \left(\frac{I_{2013}}{I_{2008}} 4 1\right)\Delta 100 \mid$	4,39
2014	46.772	$I_{2014} \mid \frac{X_{2014}}{X_{2008}}\Delta 100 \mid \frac{46.772}{44.468}*100 \mid$	105,18	$V_{200842014} \mid \left(\frac{I_{2014}}{I_{2008}} 4 1\right)\Delta 100 \mid$	5,18
2015	47.118	$I_{2015} \mid \frac{X_{2015}}{X_{2008}}\Delta 100 \mid \frac{47.118}{44.468}*100 \mid$	105,96	$V_{200842015} \mid \left(\frac{I_{2015}}{I_{2008}} 4 1\right)\Delta 100 \mid$	5,96

18.2 Números índice compuestos sin ponderar

Cuando el acontecimiento que se quiere observar depende de un conjunto de variables, se elabora un *índice* sintético o compuesto. Si a cada una de las variables del acontecimiento observado se le da la misma importancia, estamos ante un índice sin ponderar, pero si las variables tienen un peso diferente, se elabora un índice compuesto ponderado. Como ejemplo se presenta una simulación en la que se quiere calcular un índice de producción en un país. Se divide la producción en Sector Primario, Secundario, Terciario y Cuaternario. La evolución de cada sector en el período 2000-2001 es según la Tabla 197,

Tabla 197 Evolución de la producción.*					
Año	Primario	Secundario	Terciario	Cuaternario	Total
2000	316	217	290	499	1.322
2001	320	225	300	502	1.347
2002	325	229	307	504	1.365
2003	329	234	315	510	1.388
Nota: * En millones de euros.					

Los *números índice simples* tomando como base el año 2000 son,

Tabla 198 Números índice simples.					
Año	Primario	Secundario	Terciario	Cuaternario	Total
2000	100,00	100,00	100,00	100,00	100,00
2001	101,27	103,69	103,45	100,60	101,89
2002	102,85	105,53	105,86	101,00	103,25
2003	104,11	107,83	108,62	102,20	104,99

18.2.1 Número índice media aritmética

El *índice media aritmética* está dado por,

Tabla 199 Número índice media aritmética.		
$a_t = \frac{1}{n}\left(I_{1_t} + I_{2_t} + I_{3_t} + \ldots + I_{n_t}\right)$	En donde: a_t: Índice media aritmética (Total producción) en el período *t*. I_{1t}: Componente primero en el año *t*. I_{2t}: Componente segundo en el año *t*. I_{3t}: Componente tercero en el año *t*. I_{nt}: Componente *n-ésimo* en el año *t*. *n*: Número de componentes. *t*: Período de tiempo al que se refiere el índice.	Fórmula 146

En el caso de la Tabla 198, tenemos que para el año 2000,

$$a_{2000} = \frac{1}{4}\left(100{,}00 + 100{,}00 + 100{,}00 + 100{,}00\right) = \frac{400{,}00}{4} = 100{,}00$$

año 2001

$$a_{2001} = \frac{1}{4}\left(101{,}27 + 103{,}69 + 103{,}45 + 100{,}60\right) = \frac{409{,}00}{4} = 102{,}25$$

año 2002

$$a_{2002} = \frac{1}{4}\left(102{,}85 + 105{,}53 + 105{,}86 + 101{,}00\right) = \frac{415{,}24}{4} = 103{,}81$$

año 2003

$$a_{2001} = \frac{1}{4}\left(104{,}11 + 107{,}83 + 108{,}62 + 102{,}20\right) = \frac{422{,}77}{4} = 105{,}69$$

Añadiendo este *índice* a la Tabla 198 se obtiene la Tabla 200,

Tabla 200 Números índice simples y media aritmética sin ponderar.						
Año	Primario	Secundario	Terciario	Cuaternario	Total	a_t
2000	100,00	100,00	100,00	100,00	100,00	100,00
2001	101,27	103,69	103,45	100,60	101,89	102,25
2002	102,85	105,53	105,86	101,00	103,25	103,81
2003	104,11	107,83	108,62	102,20	104,99	105,69

18.2.2 Número índice agregativo simple

El *índice agregativo simple*, no utiliza los *índice simples* como la *media aritmética,* sino que toma los valores absolutos directamente, así tenemos (Tabla 201),

Tabla 201 Número índice agregativo simple.

$$A_t = \frac{x_{1_t} + x_{2_t} + x_{3_t} + \ldots + x_{n_t}}{x_{1_0} + x_{2_0} + x_{3_0} + \ldots + x_{n_0}} \cdot 100$$

Simplificando,

$$A_t = \frac{\sum_{i=1}^{n} x_{i_t}}{\sum_{i=1}^{n} x_{i_0}} \cdot 100$$

En donde:

A_t:	Índice agregativo simple en el período *t*.
x_{1t}:	Valor del primer componente en el período *t*.
x_{2t}:	Valor del segundo componente en el período *t*.
x_{3t}:	Valor del tercer componente en el período *t*.
X_{nt}:	Valor del *n-ésimo* componente en el período *t*.
x_{10}:	Valor del primer componente en el período *base*.
x_{20}:	Valor del segundo componente en el período *base*.
x_{30}:	Valor del tercer componente en el período *base*.
X_{n0}:	Valor del *n-ésimo* componente en el período *base*.
n:	Número de componentes.
t:	Período de tiempo al que se refiere el índice.

Fórmula 147

Aplicándolo a la Tabla 197, se obtiene,

Para el año 2000,

$$A_{2000} = \frac{\sum_{i=1}^{n} x_{i_t}}{\sum_{i=1}^{n} x_{i_0}} \cdot 100 = \frac{x_{1_t} + x_{2_t} + x_{3_t} + \ldots + x_{n_t}}{x_{1_0} + x_{2_0} + x_{3_0} + \ldots + x_{n_0}} \cdot 100 = \frac{316 + 217 + 290 + 499}{316 + 217 + 290 + 499} \cdot 100 = \frac{1.322}{1.322} \cdot 100 = 100{,}00$$

Año 2001,

$$A_{2001} = \frac{\sum_{i=1}^{n} x_{i_t}}{\sum_{i=1}^{n} x_{i_0}} \cdot 100 = \frac{x_{1_t} + x_{2_t} + x_{3_t} + \ldots + x_{n_t}}{x_{1_0} + x_{2_0} + x_{3_0} + \ldots + x_{n_0}} \cdot 100 = \frac{320 + 225 + 300 + 502}{316 + 217 + 290 + 499} \cdot 100 = \frac{1.347}{1.322} \cdot 100 = 101{,}89$$

Año 2002,

$$A_{2002} = \frac{\sum_{i=1}^{n} x_{i_t}}{\sum_{i=1}^{n} x_{i_0}} \cdot 100 = \frac{x_{1_t} + x_{2_t} + x_{3_t} + \ldots + x_{n_t}}{x_{1_0} + x_{2_0} + x_{3_0} + \ldots + x_{n_0}} \cdot 100 = \frac{325 + 229 + 307 + 504}{316 + 217 + 290 + 499} \cdot 100 = \frac{1.365}{1.322} \cdot 100 = 103{,}25$$

Año 2003,

$$A_{2003} = \frac{\sum_{i=1}^{n} x_{i_t}}{\sum_{i=1}^{n} x_{i_0}} \cdot 100 = \frac{x_{1_t} + x_{2_t} + x_{3_t} + \ldots + x_{n_t}}{x_{1_0} + x_{2_0} + x_{3_0} + \ldots + x_{n_0}} \cdot 100 = \frac{329 + 234 + 315 + 510}{316 + 217 + 290 + 499} \cdot 100 = \frac{1.388}{1.322} \cdot 100 = 104{,}99$$

Y añadiéndolo a la Tabla 200 se obtiene la Tabla 202,

Tabla 202 Números índice simples, media aritmética sin ponderar y agregativo simple.

Año	Primario	Secundario	Terciario	Cuaternario	Total	a_t	A_t
2000	100,00	100,00	100,00	100,00	100,00	100,00	100,00
2001	101,27	103,69	103,45	100,60	101,89	102,25	101,89
2002	102,85	105,53	105,86	101,00	103,25	103,81	103,25
2003	104,11	107,83	108,62	102,20	104,99	105,69	104,99

18.3 Números índice compuestos ponderados

Los *números índice compuestos ponderados,* se elaboran para determinar la evolución de una variable compuesta por varias variables con distinto peso específico cada una. En el caso anterior la producción se ha dividido en cuatro sectores considerando que tienen el mismo peso específico. El *índice agregativo simple*, puede introducir sesgos que se pueden compensar dando a cada *Sector* una importancia relativa en forma de ponderación y aplicándola al resto de períodos.

La ponderación considerada en el año base sería,

$W_i = \frac{x_{i_0}}{\sum_{i=1}^{n} x_{i_0}}, \forall i = 1\ 2\ n$	En donde: W_i: Factor de ponderación del componente *i-ésimo.* x_{i0}: Valor del componente *i-ésimo* en el período *base*. *t:* Período de tiempo al que se refiere el factor.	Fórmula 148

El *factor de ponderación* de cada *Sector* en el año base es (Tabla 203),

Tabla 203 Factores de ponderación.

Sector Primario	$W_1 = \frac{x_{1_{2000}}}{\sum_{i=1}^{n} x_{i_{2000}}} = \frac{316}{1.322} = 0{,}239$
Sector Secundario	$W_2 = \frac{x_{2_{2000}}}{\sum_{i=1}^{n} x_{i_{2000}}} = \frac{217}{1.322} = 0{,}164$
Sector Terciario	$W_3 = \frac{x_{3_{2000}}}{\sum_{i=1}^{n} x_{i_{2000}}} = \frac{290}{1.322} = 0{,}219$
Sector Cuaternario	$W_4 = \frac{x_{4_{2000}}}{\sum_{i=1}^{n} x_{i_{2000}}} = \frac{499}{1.322} = 0{,}377$

18.3.1 Número índice media aritmética ponderada

Y la suma de los factores de ponderación es la unidad. Entonces el *número índice compuesto de la media aritmética ponderada* es (Tabla 204),

Tabla 204 Número índice compuesto de la media aritmética ponderada.

$$a'_t = \frac{1}{\sum_{i=1}^{n} W_i} \left(I_{1_t} W_1 + I_{2_t} W_2 + \ldots + I_{n_t} W_n \right)$$

$$a'_t = \frac{1}{\sum_{i=1}^{n} W_i} \sum_{i=1}^{n} I_{i_t} W_i$$

En donde:

a'_t:	Índice media aritmética ponderado (Total producción) en el período *t*.
I_{1t}:	Índice del componente primero en el año *t*.
I_{2t}:	Índice del componente segundo en el año *t*.
I_{3t}:	Índice del componente *n-ésimo* en el año *t*.
n:	Número de componentes.
t:	Período de tiempo al que se refiere el índice.
W_1:	Factor de ponderación del componente 1.
W_2:	Factor de ponderación del componente 2.
W_n:	Factor de ponderación del componente *n-ésimo*.

Fórmula 149

Aplicado al ejemplo de la Tabla 200, se obtiene,

Para el año 2000,

$$a'_{2000} = \frac{1}{\sum_{i=1}^{n} W_i} (100{,}00 \cdot 0{,}239 + 100{,}00 \cdot 0{,}164 + 100{,}00 \cdot 0{,}219 + 100{,}00 \cdot 0{,}377) = \frac{1}{1} \cdot 100{,}00 = 100{,}00$$

Año 2001,

$$a'_{2001} = \frac{1}{\sum_{i=1}^{n} W_i} (101{,}27 \cdot 0{,}239 + 103{,}69 \cdot 0{,}164 + 103{,}45 \cdot 0{,}219 + 100{,}60 \cdot 0{,}377) = \frac{1}{1} \cdot 101{,}89 = 101{,}89$$

Año 2002,

$$a'_{2002} = \frac{1}{\sum_{i=1}^{n} W_i} (102{,}85 \cdot 0{,}239 + 105{,}53 \cdot 0{,}164 + 105{,}86 \cdot 0{,}219 + 101{,}00 \cdot 0{,}377) = \frac{1}{1} \cdot 103{,}25 = 103{,}25$$

Año 2003,

$$a'_{2003} = \frac{1}{\sum_{i=1}^{n} W_i} \Delta(104{,}11 \Delta 0{,}239 + 107{,}83 \Delta 0{,}164 + 108{,}62 \Delta 0{,}219 + 102{,}20 \Delta 0{,}377) = \frac{1}{1} \Delta 104{,}99 = 104{,}99$$

En la Tabla 205 se añade a los otros *números índice*,

Tabla 205 Números índice simples, media aritmética sin ponderar, agregativo simple y media aritmética ponderada.

Año	Total	a_t	A_t	a'_t
2000	100,00	100,00	100,00	100,00
2001	101,89	102,25	101,89	101,89
2002	103,25	103,81	103,25	103,25
2003	104,99	105,69	104,99	104,99

18.3.2 Número índice agregativo compuesto ponderado

El *número índice agregativo compuesto ponderado*, se calcula utilizando los valores absolutos de los *Sectores de producción,* y los *factores de ponderación* obtenidos anteriormente, simbólicamente,

$$A'_t = \frac{x_{1_t} \Delta W_1 + x_{2_t} \Delta W_2 + \ldots + x_{n_t} \Delta W_n}{x_{1_0} \Delta W_1 + x_{2_0} \Delta W_2 + \ldots + x_{n_0} \Delta W_n} \Delta 100$$

Simplificando,

$$A'_t = \frac{\sum_{i=1}^{n} x_{i_t} \Delta W_i}{\sum_{i=1}^{n} x_{i_0} \Delta W_i} \Delta 100$$

Fórmula 150

En donde:

A'_t: Índice agregativo compuesto ponderado en el período *t*.
x_{1t}: Valor del primer componente en el período *t*.
x_{2t}: Valor del segundo componente en el período *t*.
x_{nt}: Valor del *n-ésimo* componente en el período *t*.
x_{10}: Valor del primer componente en el período *base*.
x_{20}: Valor del segundo componente en el período *base*.
x_{n0}: Valor del *n-ésimo* componente en el período *base*.
W_i: Factor de ponderación del componente *i-ésimo.*
n: Número de componentes.
t: Período de tiempo al que se refiere el índice.

Entonces aplicando el *número índice agregativo compuesto ponderado* a la Tabla 197, tenemos, para el año 2000,

$$A'_{2000} \mid \frac{\sum_{i=1}^{n} x_{i_{2000}} \Delta W_i}{\sum_{i=1}^{n} x_{i_{2000}} \Delta W_i} \Delta 100 \mid \frac{x_{1_{2000}} \Delta W_1 + x_{2_{2000}} \Delta W_2 + \ldots + x_{n_{2000}} \Delta W_n}{x_{1_{2000}} \Delta W_1 + x_{2_{2000}} \Delta W_2 + \ldots + x_{n_{2000}} \Delta W_n} \Delta 100 \mid$$

$$\frac{316 \Delta 0{,}239 + 217 \Delta 0{,}164 + 290 \Delta 0{,}219 + 499 \Delta 0{,}377}{316 \Delta 0{,}239 + 217 \Delta 0{,}164 + 290 \Delta 0{,}219 + 499 \Delta 0{,}377} \Delta 100 \mid \frac{363{,}12}{363{,}12} \Delta 100 \mid 100{,}00$$

Para el año 2001,

$$A'_{2001} \mid \frac{\sum_{i=1}^{n} x_{i_{2001}} \Delta W_i}{\sum_{i=1}^{n} x_{i_{2000}} \Delta W_i} \Delta 100 \mid \frac{x_{1_{2001}} \Delta W_1 + x_{2_{2001}} \Delta W_2 + \ldots + x_{n_{2001}} \Delta W_n}{x_{1_{2000}} \Delta W_1 + x_{2_{2000}} \Delta W_2 + \ldots + x_{n_{2000}} \Delta W_n} \Delta 100 \mid$$

$$\frac{320 \Delta 0{,}239 + 225 \Delta 0{,}164 + 300 \Delta 0{,}219 + 502 \Delta 0{,}377}{316 \Delta 0{,}239 + 217 \Delta 0{,}164 + 290 \Delta 0{,}219 + 499 \Delta 0{,}377} \Delta 100 \mid \frac{368{,}72}{363{,}12} \Delta 100 \mid 101{,}54$$

Para el año 2002,

$$A'_{2002} \mid \frac{\sum_{i=1}^{n} x_{i_{2002}} \Delta W_i}{\sum_{i=1}^{n} x_{i_{2000}} \Delta W_i} \Delta 100 \mid \frac{x_{1_{2002}} \Delta W_1 + x_{2_{2002}} \Delta W_2 + \ldots + x_{n_{2002}} \Delta W_n}{x_{1_{2000}} \Delta W_1 + x_{2_{2000}} \Delta W_2 + \ldots + x_{n_{2000}} \Delta W_n} \Delta 100 \mid$$

$$\frac{325 \Delta 0{,}239 + 229 \Delta 0{,}164 + 307 \Delta 0{,}219 + 504 \Delta 0{,}377}{316 \Delta 0{,}239 + 217 \Delta 0{,}164 + 290 \Delta 0{,}219 + 499 \Delta 0{,}377} \Delta 100 \mid \frac{372{,}86}{363{,}12} \Delta 100 \mid 102{,}68$$

Para el año 2003,

$$A'_{2003} \mid \frac{\sum_{i=1}^{n} x_{i_{2003}} \Delta W_i}{\sum_{i=1}^{n} x_{i_{2000}} \Delta W_i} \Delta 100 \mid \frac{x_{1_{2003}} \Delta W_1 + x_{2_{2003}} \Delta W_2 + \ldots + x_{n3} \Delta W_n}{x_{1_{2000}} \Delta W_1 + x_{2_{2000}} \Delta W_2 + \ldots + x_{n_{2000}} \Delta W_n} \Delta 100 \mid$$

$$\frac{329 \Delta 0{,}239 + 234 \Delta 0{,}164 + 315 \Delta 0{,}219 + 510 \Delta 0{,}377}{316 \Delta 0{,}239 + 217 \Delta 0{,}164 + 290 \Delta 0{,}219 + 499 \Delta 0{,}377} \Delta 100 \mid \frac{378{,}66}{363{,}12} \Delta 100 \mid 104{,}28$$

Completando la Tabla 205 tenemos la Tabla 206,

Tabla 206 Números índice simples, media aritmética sin ponderar, agregativo simple, media aritmética ponderada y agregativo compuesto ponderado.

Año	Simple	a_t	A_t	a'_t	A'_t
2000	100,00	100,00	100,00	100,00	100,00
2001	101,89	102,25	101,89	101,89	101,54
2002	103,25	103,81	103,25	103,25	102,68
2003	104,99	105,69	104,99	104,99	104,28

Los índice *agregado simple* (A_t) y la *media aritmética ponderada* (a'_t, son iguales porque según la Fórmula 147,

$$A_t = \frac{x_{1_t} + x_{2_t} + x_{3_t} + \ldots + x_{n_t}}{x_{1_0} + x_{2_0} + x_{3_0} + \ldots + x_{n_0}} \cdot 100 = \frac{x_{1_t} + x_{2_t} + x_{3_t} + \ldots + x_{n_t}}{\sum_{i=1}^{n} x_{i_0}} \cdot 100$$

Y según la Fórmula 149,

$$a'_t = \frac{1}{\sum_{i=1}^{n} W_i} \cdot \left(I_{1_t} \cdot W_1 + I_{2_t} \cdot W_2 + \ldots + I_{n_t} \cdot W_n \right)$$

Teniendo en cuenta que según Fórmula 144 y Fórmula 148,

$$I_t = \frac{X_t}{X_0}, \text{ y que } W_i = \frac{x_{i_0}}{\sum_{i=1}^{n} x_{i_0}}$$

Y sustituyendo en el *índice media aritmética ponderada* tenemos que,

$$a'_t = \frac{1}{\sum_{i=1}^{n} W_i} \cdot \left(\frac{x_{1_t}}{x_{1_0}} \cdot \frac{x_{1_0}}{\sum_{i=1}^{n} x_{i_0}} + \frac{x_{2_t}}{x_{2_0}} \cdot \frac{x_{2_0}}{\sum_{i=1}^{n} x_{i_0}} + \ldots + \frac{x_{n_t}}{x_{n_0}} \cdot \frac{x_{n_0}}{\sum_{i=1}^{n} x_{i_0}} \right)$$

Y eliminando los términos iguales en el numerador y el denominador, que es el valor del producto en el período cero (x_{i0}), tenemos que,

$$a'_t \mid 1\Delta\left(\frac{x_{1_t}}{n} 2 \frac{x_{2_t}}{n} 2 \; 3 \; 2 \frac{x_{n_t}}{n}\right) \mid A_t$$

, por lo que se considera demostrado.

En los dos *índice*, el denominador es el sumatorio de todos los componentes en el periodo base. Se puede considerar que el *índice* más apropiado es el *agregado simple* que es el mismo que la *media aritmética ponderada* por los valores de las variables del año base.

18.4 Números índice de precios

El objeto de estos índice es conocer la evolución del precio de un producto o de un conjunto de productos sin necesidad de estudiar las operaciones realizadas en el período objeto de análisis.

Si el objeto de la investigación consiste en conocer la evolución del precio de un artículo, la cuestión se limitará a examinar la evolución del precio en una serie limitada de locales elaborando un *índice compuesto sin ponderar* referente a cada uno de los locales investigados, es decir, sea *n* el número de locales a investigar, el índice de precios de cada local será,

$I_{j_t} \mid \frac{P_t^j}{P_0^j}\Delta 100$	En donde: I_{jt}: Índice en el período *t* del local *j.* P^j_t: *P*recio en el período *t* y en el local *j.* P^j_0: *P*recio en el período *base* y en el local *j.*	Fórmula 151

Si se aplica el *índice de la media aritmética*, el índice general de precios será,

$I_t \mid \frac{1}{n}\Delta/I_1 2 I_2 2 \; 3 \; 2 I_j 2 \; 3 \; 2 I_n 0$ simplificando, $I_t \mid \frac{1}{n}\Delta\sum_{j\mid 1}^{n} I_{j_t}$	En donde: I_{jt}: Índice en el período *t* del local *j.*	Fórmula 152

Denominado *índice de Saürbeck.*

Si se aplica la fórmula del *índice agregativo simple* (A_t Fórmula 147), entonces,

$$I_t \mid \frac{P_t^1 2 P_t^2 2 \; 3 \; 2 P_t^j 2 \; 3 \; 2 P_t^n}{P_0^1 2 P_0^2 2 \; 3 \; 2 P_0^j 2 \; 3 \; 2 P_0^n}\Delta 100 \mid \frac{\sum_{j\mid 1}^{n} P_{j_t}}{\sum_{j\mid 1}^{n} P_{j_0}}$$

El caso más corriente consiste en conocer la evolución del precio de un conjunto de bienes sobre la base del análisis de la evolución de los precios de sus componentes más representativos. En este caso hay que acudir a los *números índice ponderados*, en los que se tiene en cuenta la importancia relativa de la cantidad (Q) de cada uno de los bienes representativos del conjunto investigado.

En esta ocasión el ejemplo que se va a utilizar es una simulación para trabajar con los *números índice de precios* y es,

Tabla 207 Evolución de los precios de productos básicos.

Subíndice	Año	Pan (€/kg)	Gasolina (€/l)	Huevos (€/doc.)	Patatas (€/kg)
0	2000	1,90	0,82	1,00	0,50
1	2001	1,95	0,81	1,00	0,50
2	2002	1,95	0,81	1,10	0,55
3	2003	2,00	0,82	1,20	0,60

Y la estimación del consumo medio por familia y mes es,

Tabla 208 Evolución de las cantidades medias consumidas por familia y mes.

Subíndice	Año	Pan (kg./mes)	Gasolina (l/mes)	Huevos (doc./mes)	Patatas (kg./mes)
0	2000	35	80	6	12
1	2001	34	90	6	12
2	2002	32	100	6,5	13
3	2003	32	100	6,5	13

Los años se representan por el subíndice 0, 1, 2 y 3. Los precios se representan por la letra P y las cantidades por la letra Q. Los precios de la serie del *pan* se representan como P_0, P_1, P_2 y P_3. Las cantidades de la serie del *pan* se representan como Q_0, Q_1, Q_2 y Q_3. Los *índice simples* de todos los precios, tomando como base P_0, y según la Fórmula 144 serán (Tabla 209),

Tabla 209 Números índice simples de los precios.

Año	Pan	Gasolina	Huevos	Patatas
2000	100,00	100,00	100,00	100,00
2001	102,63	98,78	100,00	100,00
2002	102,63	98,78	110,00	110,00
2003	105,26	100,00	120,00	120,00

Los *números índice simples* de las cantidades son (Tabla 210),

Tabla 210 Números índice simples de las cantidades.

Año	Pan	Gasolina	Huevos	Patatas
2000	100,00	100,00	100,00	100,00
2001	97,14	112,50	100,00	100,00
2002	91,43	125,00	108,33	108,33
2003	91,43	125,00	108,33	108,33

El propósito es obtener una sola serie de precios reduciendo las cuatro a una sola. Esta reducción se puede realizar obteniendo la media aritmética de las cantidades de los productos por cada año o bien obteniendo el sumatorio de las cantidades por año (Tabla 211).

Tabla 211 Sumatorios y medias de precios y cantidades y números índice agregados simples.

Año	Suma de precios	Suma de cantidades	Índice de precios	Índice de cantidades
2000	4,22	133	100,00	100,00
2001	4,26	142	100,95	106,77
2002	4,41	151,5	104,50	113,91
2003	4,62	151,5	109,48	113,91
Año	Media de precios	Media de cantidades	Índice de precios	Índice de cantidades
2000	1,06	33,25	100,00	100,00
2001	1,07	35,5	100,95	106,77
2002	1,10	37,88	104,50	113,91
2003	1,16	37,88	109,48	113,91

En la tabla anterior (Tabla 211) se puede observar en las columnas de los índice que se obtienen los mismos resultados a partir de los sumatorios de las cantidades por año, que de las medias. Entonces el método se simplifica trabajando con los sumatorios. Pero los *índice agregados simples* se han obtenido sin considerar el peso relativo de cada uno de los productos en el índice obtenido y además se han sumado productos expresados en una unidad de medida diferente (€/kg, €/l y €/12). Estos aspectos hacen que el *índice agregativo simple* no sea de los más utilizados, simbólicamente,

Para precios:

$$I_0^t \mid \frac{\frac{\sum_{i=1}^{n} P_{i_t}}{n}}{\frac{\sum_{i=1}^{n} P_{i_0}}{n}} \Delta 100$$

Para cantidades:

$$I_0^t \mid \frac{\frac{\sum_{i=1}^{n} Q_{i_t}}{n}}{\frac{\sum_{i=1}^{n} Q_{i_0}}{n}} \Delta 100$$

En donde:

I^t_0: Es el índice del año *t* con base en el año 0.
n: Número de elementos agregados.
P_{it}: Precio del producto *i-ésimo* en el periodo *t.*
P_{i0}: Precio del producto *i-ésimo* en el periodo base.
Q_{it}: Cantidad del producto *i-ésimo* en el periodo *t.*
Q_{i0}: Cantidad del producto *i-ésimo* en el periodo base.

Fórmula 153

Otro procedimiento es el *índice de la media aritmética*. La media proporciona directamente el índice complejo o compuesto buscado y están referidos a la misma base. Simbólicamente,

Para precios: $I_0^t \mid \frac{\frac{^n P_{i_t}}{i \mid 1}}{n} \Delta 100$ Para cantidades: $I_0^t \mid \frac{\frac{^n Q_{i_t}}{i \mid 1}}{n} \Delta 100$	En donde: I^t_0: Es el índice del año *t* con base en el año 0. *n*: Número de elementos agregados. P_{it}: Precio del producto *i-ésimo* en el periodo *t.* P_{i0}: Precio del producto *i-ésimo* en el periodo base. Q_{it}: Cantidad del producto *i-ésimo* en el periodo *t.* Q_{i0}: Cantidad del producto *i-ésimo* en el periodo base.	Fórmula 154

Los índice resultantes son,

Tabla 212 Números índice media aritmética.

Año	Media aritmética simple de precios	Media aritmética simple de cantidades
2000	100,00	100,00
2001	100,35	102,41
2002	105,35	108,27
2003	111,32	108,27

Siendo estos índice diferentes a los agregativos simples.

18.5 Números índice de valores, precios y cantidades

Si llamamos *P* al precio de un bien y *Q* a su cantidad (vendida, producida, consumida, etc.) podemos obtener el valor *V* mediante la multiplicación,

$V \mid P \Delta Q$	Fórmula 155

Si para cada producto tenemos los precios P_0, P_1, P_2 y P_3 y las cantidades Q_0, Q_1, Q_2 y Q_3 entonces la serie de valores será

$$V_0 \mid P_0 \Delta Q_0$$

$$V_1 \mid P_1 \Delta Q_1$$

4

$$V_t \mid P_t \Delta Q_t$$

Esta serie temporal depende de dos variables, el precio y la cantidad. Si en el transcurso del tiempo permanece constante, la cantidad *Q*, la serie de valores refleja la variación en el precio *P*. Si el precio *P* permanece constante, las variaciones en la serie de valores se deben exclusivamente a las variaciones de las cantidades *Q*. Por lo tanto, para estudiar la variación del precio podemos servirnos de una serie de valores con *Q* constante. Si queremos analizar las variaciones de *Q* se usará la *P* constante.

Entonces se pueden escribir las tres series siguientes (Tabla 213),

Tabla 213 Series de valores.		
P: Variable *Q*: Variable	*P*: Constante *Q*: Variable	*P*: Constante *Q*: Variable
$P_0 \Delta Q_0$	$P \Delta Q_0$	$P_0 \Delta Q$
$P_1 \Delta Q_1$	$P \Delta Q_1$	$P_1 \Delta Q$
4	4	4
$P_t \Delta Q_t$	$P \Delta Q_t$	$P_t \Delta Q$

En la primera serie varían el precio y la cantidad, en la segunda el precio *P* se mantiene constante y en la tercera se mantiene constante la cantidad *Q*.

Las tres series anteriores expresan valores y pueden sumarse con las que se obtengan de otros productos o mercancías. Por ejemplo, si el precio es €/kg y entonces las cantidades son kg/mes, se obtiene que,

$$P \mid \frac{€}{kg} \text{ y } Q \mid \frac{kg}{mes}, \text{ entonces } P\Delta Q \text{ es } \frac{€}{kg}\Delta\frac{kg}{mes} \mid \frac{€}{mes}$$

Si el precio es €/l y entonces las cantidades son l/mes, se obtiene que,

$$P \mid \frac{€}{l} \text{ y } Q \mid \frac{l}{mes}, \text{ entonces } P\Delta Q \text{ es } \frac{€}{l}\Delta\frac{l}{mes} \mid \frac{€}{mes}$$

Si el precio es €/doc. y entonces las cantidades son doc/mes, se obtiene que,

$$P \mid \frac{€}{doc.} \text{ y } Q \mid \frac{doc.}{mes}, \text{ entonces } P\Delta Q \text{ es } \frac{€}{doc.}\Delta\frac{doc.}{mes} \mid \frac{€}{mes}$$

Supuestos *n* productos, la suma de los valores puede representarse (Tabla 214),

Tabla 214 Series de sumatorio de valores		
P: Variable *Q*: Variable	*P*: Constante *Q*: Variable	*P*: Constante *Q*: Variable
$\sum P_0 \Delta Q_0$	$\sum P \Delta Q_0$	$\sum P_0 \Delta Q$
$\sum P_1 \Delta Q_1$	$\sum P \Delta Q_1$	$\sum P_1 \Delta Q$
4	4	4
$\sum P_t \Delta Q_t$	$\sum P \Delta Q_t$	$\sum P_t \Delta Q$

La Tabla 213 y Tabla 214 tienen el mismo significado. La primera es para un producto y la segunda para *n* productos y esta puede convertirse en números índice aplicando el criterio de *número índice simple* y es,

$V_t \mid \frac{\sum_{i\mid 1}^{n} P_{i_t} \Delta Q_{i_t}}{\sum_{i\mid 1}^{n} P_{i_0} \Delta Q_{i_0}} \Delta 100$	En donde: P_{it}: Es el precio del producto *i* en el período *t*. P_{i0}: Es el precio del producto *i* en el período base. Q_{it}: Es la cantidad del producto *i* en el período *t*. Q_{i0}: Es la cantidad del producto *i* en el período base. *n*: Número de productos.	Fórmula 156

$Q_t \mid \frac{\sum_{i\mid 1}^{n} P_i \Delta Q_{i_t}}{\sum_{i\mid 1}^{n} P_i \Delta Q_{i_0}} \Delta 100$	En donde: P_i: Es el precio constante del producto *i*. Q_{it}: Es la cantidad del producto *i* en el período *t*. Q_{i0}: Es la cantidad del producto *i* en el período base. *n*: Número de productos.	Fórmula 157

$P_t \mid \frac{\sum_{i\mid 1}^{n} P_{i_t} \Delta Q_i}{\sum_{i\mid 1}^{n} P_{i_0} \Delta Q_i} \Delta 100$	En donde: P_{it}: Es el precio del producto *i* en el período *t*. P_{i0}: Es el precio del producto *i* en el período 0. Q_i: Es la cantidad constante del producto *i*. *n*: Número de productos.	Fórmula 158

La Fórmula 156 es un índice de valor. La Fórmula 157 es un índice de cantidad, al permanecer el precio constante y variar la cantidad, lo que mide son las variaciones en la cantidad y la Fórmula 158 es un índice considerado de precio por permanecer constante la cantidad y medir variaciones en el precio por ser la parte variable.

Estos índices se consideran compuestos porque reúnen en una sola serie todos los productos. Son también agregativos y ponderados por usar las cantidades y los precios. Existen tres soluciones para determinar cual es el período del que se toman las cantidades y los precios constantes.

- Índice LASPEYRES (Economista alemán). Se toma constante el precio o la cantidad del período base (el período 0).
- Índice PASSCHE (Economista alemán). Se toma constante el precio o la cantidad correspondiente al período para el cual se va a calcular el índice (el período *t*).
- El economista americano Fisher propuso como solución ideal la media geométrica de los dos anteriores y se denomina el índice de FISHER.

Tabla 215 Fórmulas de LASPEIRES, PAASCHE y FISHER.		
Índice	Cantidad	Precios
LASPEYRES	$Q_L \mid \frac{\sum_{i\mid 1}^{n} P_{i_0} \Delta Q_{i_t}}{\sum_{i\mid 1}^{n} P_{i_0} \Delta Q_{i_0}} \Delta 100$	$P_L \mid \frac{\sum_{i\mid 1}^{n} P_{i_t} \Delta Q_{i_0}}{\sum_{i\mid 1}^{n} P_{i_0} \Delta Q_{i_0}} \Delta 100$
PAASCHE	$Q_P \mid \frac{\sum_{i\mid 1}^{n} P_{i_t} \Delta Q_{i_t}}{\sum_{i\mid 1}^{n} P_{i_t} \Delta Q_{i_0}} \Delta 100$	$P_P \mid \frac{\sum_{i\mid 1}^{n} P_{i_t} \Delta Q_{i_t}}{\sum_{i\mid 1}^{n} P_{i_0} \Delta Q_{i_t}} \Delta 100$
FISHER	$Q_F \mid \sqrt{Q_L \Delta Q_P}$	$P_F \mid \sqrt{P_L \Delta P_P}$

Las propiedades de estos *números índice* son,

1. Si en el índice de precios de LASPEYRES y de PAASCHE, multiplicamos y dividimos cada sumando del numerador por P_{i0}, tenemos,

$$P_L \mid \frac{\sum_{i\mid 1}^{n} \frac{P_{i_t}}{P_{i_0}} \Delta P_{i_0} \Delta Q_{i_0}}{\sum_{i\mid 1}^{n} P_{i_0} \Delta Q_{i_0}} \Delta 100$$

$$P_P \mid \frac{\sum_{i\mid 1}^{n} \frac{P_{i_t}}{P_{i_0}} \Delta P_{i_0} \Delta Q_{i_t}}{\sum_{i\mid 1}^{n} P_{i_0} \Delta Q_{i_t}} \Delta 100$$

Se puede hacer lo mismo con los índice de cantidades, pero multiplicando y dividiendo por Q_{i0}. Entonces el índice de precios de LASPYRES es la media aritmética de los índice simples (P_t/P_0), ponderado por el producto del precio y la cantidad del año base (P_0 x Q_0). Y el índice de precios de PAASCHE es la media aritmética de los índice simples (P_t/P_0), ponderado por el precio del año base multiplicado por la cantidad del período t (P_0 x Q_t).

2. La compatibilidad que se producía entre los índice de valor (V), precio (P) y cantidad (Q) (Ver Fórmula 155), no se verifica con los índice de LASPEYRES y PAACHE.

$$V \prod P_L \Delta Q_L$$

$$V \prod P_P \Delta Q_P$$

Pero sí se cumple con FISHER,

$$V \mid P_F \Delta Q_F$$

Y con la combinación de LASPEYRES y PAASCHE,

$$V \mid P_L \Delta Q_P$$

$$V \mid P_P \Delta Q_L$$

Se desarrolla este último como demostración,

$$\frac{\sum_{i=1}^{n} P_{i_t} \Delta Q_{i_t}}{\sum_{i=1}^{n} P_{i_0} \Delta Q_{i_t}} \Delta \frac{\sum_{i=1}^{n} P_{i_0} \Delta Q_{i_t}}{\sum_{i=1}^{n} P_{i_0} \Delta Q_{i_0}} \mid \frac{\sum_{i=1}^{n} P_{i_t} \Delta Q_{i_t}}{\sum_{i=1}^{n} P_{i_0} \Delta Q_{i_0}}$$

El último término es el índice de valor de la Fórmula 156.

Ejemplo realizado sobre la Tabla 207 y Tabla 208

Tabla 216 Índice de precios de LASPEYRES. Año 2000 base = 100.

	Precios				Cantidades			
Año	Pan (€/kg)	Gasolina (€/l)	Huevos (€/doc.)	Patatas (€/kg)	Pan (kg./mes)	Gasolina (l/mes)	Huevos (doc./mes)	Patatas (kg./mes)
2000	1,90	0,82	1,00	0,50	35	80	6	12
2001	1,95	0,81	1,0	0,50	34	90	6	12
2002	1,95	0,81	1,10	0,55	32	100	6,5	13
2003	2,00	0,82	1,20	0,60	32	100	6,5	13
P_{i0} x Q_{i0}	66,50	65,60	6,00	6,00				
P_{it} x Q_{i0}	P_{1t} x Q_{10}	P_{2t} x Q_{20}	P_{3t} x Q_{30}	P_{4t} x Q_{40}				
Año	Pan	Gasolina	Huevos	Patatas	Suma año = t	Suma año = 0	P_L	
2000	66,50	65,60	6,00	6,00	144,10	144,10	100,00	
2001	68,25	64,80	6,00	6,00	145,05	144,10	100,66	
2002	68,25	64,80	6,60	6,60	146,25	144,10	101,49	
2003	70,00	65,60	7,20	7,20	150,00	144,10	104,09	

Año 2000,

$$P_{L\,2000} \mid \frac{\sum_{i=1}^{4} P_{i_{2000}} \Delta Q_{i_{2000}}}{\sum_{i=1}^{4} P_{i_{2000}} \Delta Q_{i_{2000}}} \Delta 100 \mid \frac{/P_{1_{2000}} \Delta Q_{1_{2000}} 02 /P_{2_{2000}} \Delta Q_{2_{2000}} 02 /P_{3_{2000}} \Delta Q_{3_{2000}} 02 /P_{4_{2000}} \Delta Q_{4_{2000}} 0}{/P_{1_{2000}} \Delta Q_{1_{2000}} 02 /P_{2_{2000}} \Delta Q_{2_{2000}} 02 /P_{3_{2000}} \Delta Q_{3_{2000}} 02 /P_{4_{2000}} \Delta Q_{4_{2000}} 0} \Delta 100 \mid$$

$$\frac{/1,90 \Delta 3502 /0,82 \Delta 8002 /1,00 \Delta 6,0002 /0,50 \Delta 120}{/1,90 \Delta 3502 /0,82 \Delta 8002 /1,00 \Delta 6,0002 /0,50 \Delta 120} \Delta 100 \mid \frac{144,10}{144,10} \Delta 100 \mid 100,00$$

Año 2001,

$$P_{L\,2001} = \frac{\sum_{i=1}^{4} P_{i2001}\,\Delta Q_{i2000}}{\sum_{i=1}^{4} P_{i2000}\,\Delta Q_{i2000}}\Delta 100 = \frac{(P_{1_{2001}}\,\Delta Q_{1_{2000}}) + (P_{2_{2001}}\,\Delta Q_{2_{2000}}) + (P_{3_{2001}}\,\Delta Q_{3_{2000}}) + (P_{4_{2001}}\,\Delta Q_{4_{2000}})}{(P_{1_{2000}}\,\Delta Q_{1_{2000}}) + (P_{2_{2000}}\,\Delta Q_{2_{2000}}) + (P_{3_{2000}}\,\Delta Q_{3_{2000}}) + (P_{4_{2000}}\,\Delta Q_{4_{2000}})}\Delta 100 =$$

$$\frac{(1{,}95\,\Delta 35) + (0{,}81\,\Delta 80) + (1{,}00\,\Delta 6{,}00) + (0{,}50\,\Delta 120)}{(1{,}90\,\Delta 35) + (0{,}82\,\Delta 80) + (1{,}00\,\Delta 6{,}00) + (0{,}50\,\Delta 120)}\Delta 100 = \frac{145{,}05}{144{,}10}\Delta 100 = 100{,}66$$

Año 2002,

$$P_{L\,2002} = \frac{\sum_{i=1}^{4} P_{i2002}\,\Delta Q_{i2000}}{\sum_{i=1}^{4} P_{i2000}\,\Delta Q_{i2000}}\Delta 100 = \frac{(P_{1_{2002}}\,\Delta Q_{1_{2000}}) + (P_{2_{2002}}\,\Delta Q_{2_{2000}}) + (P_{3_{2002}}\,\Delta Q_{3_{2000}}) + (P_{4_{2002}}\,\Delta Q_{4_{2000}})}{(P_{1_{2000}}\,\Delta Q_{1_{2000}}) + (P_{2_{2000}}\,\Delta Q_{2_{2000}}) + (P_{3_{2000}}\,\Delta Q_{3_{2000}}) + (P_{4_{2000}}\,\Delta Q_{4_{2000}})}\Delta 100 =$$

$$\frac{(1{,}95\,\Delta 35) + (0{,}81\,\Delta 80) + (1{,}10\,\Delta 6{,}00) + (0{,}55\,\Delta 120)}{(1{,}90\,\Delta 35) + (0{,}82\,\Delta 80) + (1{,}00\,\Delta 6{,}00) + (0{,}50\,\Delta 120)}\Delta 100 = -\,101{,}49$$

Año 2003,

$$P_{L\,2003} = \frac{\sum_{i=1}^{4} P_{i2003}\,\Delta Q_{i2000}}{\sum_{i=1}^{4} P_{i2000}\,\Delta Q_{i2000}}\Delta 100 = \frac{(P_{1_{2003}}\,\Delta Q_{1_{2000}}) + (P_{2_{2003}}\,\Delta Q_{2_{2000}}) + (P_{3_{2003}}\,\Delta Q_{3_{2000}}) + (P_{4_{2003}}\,\Delta Q_{4_{2000}})}{(P_{1_{2000}}\,\Delta Q_{1_{2000}}) + (P_{2_{2000}}\,\Delta Q_{2_{2000}}) + (P_{3_{2000}}\,\Delta Q_{3_{2000}}) + (P_{4_{2000}}\,\Delta Q_{4_{2000}})}\Delta 100 =$$

$$\frac{(2{,}00\,\Delta 35) + (0{,}82\,\Delta 80) + (1{,}20\,\Delta 6{,}00) + (0{,}60\,\Delta 120)}{(1{,}90\,\Delta 35) + (0{,}82\,\Delta 80) + (1{,}00\,\Delta 6{,}00) + (0{,}50\,\Delta 120)}\Delta 100 = 104{,}09$$

Tabla 217 Índice de precios de PAASCHE. Año 2000 base = 100.

	Precios				Cantidades						
Año	Pan (€/kg)	Gasolina (€/l)	Huevos (€/doc.)	Patatas (€/kg)	Pan (kg./mes)	Gasolina (l/mes)	Huevos (doc./mes)	Patatas (kg./mes)			
2000	1,90	0,82	1,00	0,50	35	80	6,00	12			
2001	1,95	0,81	1,00	0,50	34	90	6,00	12			
2002	1,95	0,81	1,10	0,55	32	100	6,50	13			
2003	2,00	0,82	1,20	0,60	32	100	6,50	13			
	Pan		Gasolina		Huevos		Patatas				
Año	$P_{1t} \times Q_{1t}$	$P_{10} \times Q_{1t}$	$P_{2t} \times Q_{2t}$	$P_{20} \times Q_{2t}$	$P_{3t} \times Q_{3t}$	$P_{30} \times Q_{3t}$	$P_{4t} \times Q_{4t}$	$P_{40} \times Q_{4t}$	Suma num.	Suma den.	P_P
2000	66,50	66,50	65,60	65,60	6,00	6,00	6,00	6,00	144,10	144,10	100,00
2001	66,30	64,60	72,90	73,80	6,00	6,00	6,00	6,00	151,20	150,40	100,53
2002	62,40	60,80	81,00	82,00	7,15	6,50	7,15	6,50	157,70	155,80	101,22
2003	64,00	60,80	82,00	82,00	7,80	6,50	7,80	6,50	161,60	155,80	103,72

Año 2000,

$$P_{P2000} = \frac{\sum_{i=1}^{4} P_{i_{2000}} \Delta Q_{i_{2000}}}{\sum_{i=1}^{4} P_{i_{2000}} \Delta Q_{i_{2000}}} \Delta 100 = \frac{(P_{1_{2000}} \Delta Q_{1_{2000}}) + (P_{2_{2000}} \Delta Q_{2_{2000}}) + (P_{3_{2000}} \Delta Q_{3_{2000}}) + (P_{4_{2000}} \Delta Q_{4_{2000}})}{(P_{1_{2000}} \Delta Q_{1_{2000}}) + (P_{2_{2000}} \Delta Q_{2_{2000}}) + (P_{3_{2000}} \Delta Q_{3_{2000}}) + (P_{4_{2000}} \Delta Q_{4_{2000}})} \Delta 100 =$$

$$\frac{(1{,}90 \Delta 350) + (0{,}82 \Delta 800) + (1{,}00 \Delta 6{,}000) + (0{,}50 \Delta 120)}{(1{,}90 \Delta 350) + (0{,}82 \Delta 800) + (1{,}00 \Delta 6{,}000) + (0{,}50 \Delta 120)} \Delta 100 = \frac{144{,}10}{144{,}10} \Delta 100 = 100{,}00$$

Año 2001,

$$P_{P2001} = \frac{\sum_{i=1}^{4} P_{i_{2001}} \Delta Q_{i_{2001}}}{\sum_{i=1}^{4} P_{i_{2000}} \Delta Q_{i_{2001}}} \Delta 100 = \frac{(P_{1_{2001}} \Delta Q_{1_{2001}}) + (P_{2_{2001}} \Delta Q_{2_{2001}}) + (P_{3_{2001}} \Delta Q_{3_{2001}}) + (P_{4_{2001}} \Delta Q_{4_{2001}})}{(P_{1_{2000}} \Delta Q_{1_{2001}}) + (P_{2_{2000}} \Delta Q_{2_{2001}}) + (P_{3_{2000}} \Delta Q_{3_{2001}}) + (P_{4_{2000}} \Delta Q_{4_{2001}})} \Delta 100 =$$

$$\frac{(1{,}95 \Delta 340) + (0{,}81 \Delta 900) + (1{,}00 \Delta 6{,}000) + (0{,}50 \Delta 120)}{(1{,}90 \Delta 350) + (0{,}82 \Delta 900) + (1{,}00 \Delta 6{,}000) + (0{,}50 \Delta 120)} \Delta 100 = \frac{151{,}20}{150{,}40} \Delta 100 = 100{,}53$$

Año 2002,

$$P_{P2002} = \frac{\sum_{i=1}^{4} P_{i_{2002}} \Delta Q_{i_{2002}}}{\sum_{i=1}^{4} P_{i_{2000}} \Delta Q_{i_{2002}}} \Delta 100 = \frac{(P_{1_{2002}} \Delta Q_{1_{2002}}) + (P_{2_{2002}} \Delta Q_{2_{2002}}) + (P_{3_{2002}} \Delta Q_{3_{2002}}) + (P_{4_{2002}} \Delta Q_{4_{2002}})}{(P_{1_{2000}} \Delta Q_{1_{2002}}) + (P_{2_{2000}} \Delta Q_{2_{2002}}) + (P_{3_{2000}} \Delta Q_{3_{2002}}) + (P_{4_{2000}} \Delta Q_{4_{2002}})} \Delta 100 =$$

$$\frac{(1{,}95 \Delta 320) + (0{,}81 \Delta 1000) + (1{,}10 \Delta 6{,}500) + (0{,}55 \Delta 130)}{(1{,}90 \Delta 320) + (0{,}82 \Delta 1000) + (1{,}00 \Delta 6{,}500) + (0{,}50 \Delta 130)} \Delta 100 = \frac{157{,}70}{155{,}80} \Delta 100 = 101{,}22$$

Año 2003,

$$P_{P2003} = \frac{\sum_{i=1}^{4} P_{i_{2003}} \Delta Q_{i_{2003}}}{\sum_{i=1}^{4} P_{i_{2000}} \Delta Q_{i_{2003}}} \Delta 100 = \frac{(P_{1_{2003}} \Delta Q_{1_{2003}}) + (P_{2_{2003}} \Delta Q_{2_{2003}}) + (P_{3_{2003}} \Delta Q_{3_{2003}}) + (P_{4_{2003}} \Delta Q_{4_{2003}})}{(P_{1_{2000}} \Delta Q_{1_{2003}}) + (P_{2_{2000}} \Delta Q_{2_{2003}}) + (P_{3_{2000}} \Delta Q_{3_{2003}}) + (P_{4_{2000}} \Delta Q_{4_{2003}})} \Delta 100 =$$

$$\frac{/2{,}00\Delta 320/0{,}82\Delta 1000/1{,}20\Delta 6{,}500/0{,}60\Delta 130}{/1{,}90\Delta 320/0{,}82\Delta 1000/1{,}00\Delta 6{,}500/0{,}50\Delta 130}\Delta 100 \mid \frac{161{,}60}{155{,}80}\Delta 100 \mid 103{,}72$$

Tabla 218 Índice de precios de FISHER.
Año 2000 base = 100.

Año	P_L	P_P	P_F
2000	100,00	100,00	100,00
2001	100,66	100,53	100,60
2002	101,49	101,22	101,36
2003	104,09	103,72	103,91

Para el año 2000,

$$P_{F2000} \mid \sqrt{P_{L2000}\Delta P_{P2000}} \mid \sqrt{100{,}00\Delta 100{,}00} \mid 100{,}00$$

Para el año 2001,

$$P_{F2001} \mid \sqrt{P_{L2001}\Delta P_{P2001}} \mid \sqrt{100{,}66\Delta 100{,}53} \mid 100{,}60$$

Para el año 2002,

$$P_{F2002} \mid \sqrt{P_{L2002}\Delta P_{P2002}} \mid \sqrt{101{,}49\Delta 101{,}22} \mid 101{,}36$$

Para el año 2003,

$$P_{F2003} \mid \sqrt{P_{L2003}\Delta P_{P2003}} \mid \sqrt{104{,}09\Delta 103{,}72} \mid 103{,}91$$

19 Bibliografía

Aboitiz, F. and Montiel, J. (2007). *Origin and Evolution of the Vertebrate Telencephalon, with Special Reference to the Mammalian Neocortex.* Berlin: Springer.

Alvira, F. (2004). *La encuesta: una perspectiva general metodológica*. Madrid: CIS.

Ander-Egg, E. (1982). Técnicas de Investigación Social. Buenos Aires: Humanitas.

Asouti, E. (2006). Beyond the Pre-Pottery Neolithic B interaction sphere. *J. World Prehist.,* Vol. 20, Nos. 2-4, pp. 87-126.

Atmanspacher, H. (2004). "Quantum approaches to consciousness". En Stanford Encyclopedia of Philosophy.

Babbie, E. (1999). Funadamentos de la investigación social. México: Thomson.

Bacon, F. (1620/1984). *Advancement of learning; Novum organum; New atlantis.* Chicago: Enciclopaedia Britannica, Inc.

Baker, T. L. (1988/1999). Doing social research. New York: McGraw-Hill.

Barbancho, A. G. (1964/1992). Estadística elemental moderna. Barcelona: Ariel.

Bar-Yosef, O. (2002). The Upper Paleolithic Revolution. *Annu. Rev. Anthropology*, 31, pp. 363-393. Disponible en: http://www.cameronmsmith.com/courses/EuropeanPrehistory2007/TheUpperPalaeolithicRevolution.pdf.

Bateson, G. (1971). Interacción familiar: aportes fundamentales sobre teoría y técnica. Buenos Aires: Tiempo Contemporáneo.

Bateson, G. (1977). Doble vínculo y esquizofrenia: el síndrome y sus factores patogénicos interpersonales. Buenos Aires: Carlos Lolhé.

Bateson, G. y Ruesch, J. (1984). Comunicación: La matriz social de la psiquiatría. Barcelona: Paidós.

Bear, M. F. et al. (1998). *Neurociencia, Explorando el cerebro*. Barcelona: Masson.

Bednarik, R. G. (2008). The Mythical Moderns. *J. World Prehist.,* Vol. 21, No 2, pp. 85-102.

Bennet, W. J. (1998). "Neuroscience and the human spirit". En National Review, dec 31.

Bingham, W. D. y Moore, B. V. (1973). Cómo *entrevistar.* Madrid: Rialp.

Blalock, H. N. (1966/1994). Estadística Social. México: FCE.

Botella, J. et al, (1993). Análisis de datos en psicología I. Madrid: Pirámide.

Boulding, K. E. (1993). «Teoría General de los Sistemas. El esqueleto de la Ciencia». En C. Ramió. *Teoría de la organización*. vol.1, La evolución histórica del pensamiento organizativo. Los principales paradigmas teóricos. Madrid: Ministerio para las Administraciones Públicas.

Bunge, M. (1981). *La Investigación Científica*. Barcelona: Ariel.

Canavos, G. C. (1984/1988). *Probabilidad y estadística*. México: McGraw-Hill.

Cea D'Ancona, Mª. A. (1996). *Metodología cuantitativa: estrategias y técnicas de investigación*. Madrid: Síntesis.

Cea D'Ancona, Mª. A. (2004). *Métodos de encuesta. Teoría y práctica, errores y mejora*. Madrid: Síntesis.

Cea D'Ancona, Mª. A. (2002). *Análisis multivariable. Teoría y práctica en la investigación social*. Madrid: Sintesis.

Cicourel, A. V. (1968) *The social organization of juvenile justice*. New York: Wiley.

Cochram, W. G. (1974) *Técnicas de muestreo*. México: CECSA.

Cochran, W. G. (1987). Técnicas de muestreo. México: ED. Continental.

Comte, A. (1893/1968). *Cours de Philosophie Positive*. Tome III. Paris: Editions Anthropos.

Converse, J. M. and Presser, S. (1986). *Survey questions. Handcrafting the standardized questionnaire*. Beverly Hills, CA: Sage Publications.

Coombs, C. H. (1979). "Teoría y Métodos de la medición social". En L. Festinger y D. Katz, (compiladores). *Los métodos de investigación en las ciencias sociales.* Buenos Aires: Paidos.

Copérnico, N. Digges, T. y Galilei, G. (1996). *Opúsculos sobre el movimiento de la tierra*. Madrid: Alianza.

Corbetta, P. (2003). *Metodología y técnicas de investigación social*. Madrid: McGraw-Hill.

Crawford, H. (2006). *Sumer and the sumerians*. Cambridge: Cambridge University Press.

Crick, F. (1994). *The Astonishing Hypothesis. The Scientific Search for the Soul*. London: Simon & Schuster Ltd.

Cuadras, C. M. et al, (1991). *Fundamentos de Estadística. Aplicación a las Ciencias Humanas*. Barceona: Promociones y Publicaciones Universitarias.

Dalton, M. (1959) *Men who manage*. New York: Wiley.

Daniel, W. W. (1977/1992). *Bioestadística. Base para el análisis de las ciencias de la salud*. México: Limusa.

Davis, A. et all (1941) *Deep South*. Chicago: University of Chicago Press.

De Kerckhove, D. (1987). Writing Left and Right. *Interchange.* Vol. 18, Nos. 1-2, pp. 60-77.

De La Puente, C. (1995). Spss/pc+. Una guía para la investigación. Madrid: Ed. Complutense.

De la Puente, C. (2006). "Teoría, métodos y técnicas de la Sociología del futuro. ¿Reinterpretar el pasado?". *XVI ISA World Congress of Sociology. The Quality of Social Existence in a Globalising World*. Durban, South Africa, 23-29 july 2006. Research Committee: 07 (The future and sociological theory). Sesion: 10.

De la Puente, C. (2007 a). "Propuesta a la Ontología del Ser. Hipótesis desde el Estructural-Funcionalismo". *IX Congreso Español de Sociología, Poder, Cultura y Civilización*. Barcelona, España, 13, 14 y 15 de septiembre de 2007.

De la Puente, C. (2007 b). Sobre la Medida, Validez y Fiabilidad en Sociología. Una Aplicación de Análisis de Componentes Principales. *NOMADAS. Revista Crítica de Ciencias Sociales y Jurídicas*, 16, julio-diciembre 2007, págs. 353-361.

De la Puente, C. (en revisión). "Evolución de la estructura organizativa de una empresa. La Organización Estructural Neuronal".

Delgado, J. M. y Gutiérrez, J. (coords.) (1994). *Métodos y técnicas cualitativas de investigación en ciencias sociales*. Madrid: Síntesis.

Denzin, N. K. and Lincoln, I. S. (eds.) (1994/2005). *Handbook of qualitative research*. California: Sage.

Descartes, R. (1637/1986). *Discurso del método, dióptrica, meteoros y geometría*. (G. Quintás, Trad.), Madrid: Alfaguara.

Drake, St. C. and Cayton, R. H. (1945) *Black Metropolis. A study of negro life in a northern city*. New York: Harcourt, Brace.

Dunbar, R. I. M. (2002). "The social brain Hypothesis". En J. T. Cacciopo et al. *Foundations in social neuroscience*. Cambridge: The MIT Press.

Durkheim E. (1895/1978). *Reglas del método sociológico*. Madrid: Akal Editor.

Durkheim, E. (2003). *El suicidio*. Madrid: Akal.

Encyclopedia Britannica, Inc (1994). Britannica CD. (1.01). London.

Fernández Elías, C. (1874/2005). *Novísimo tratado completo de Filosofía del derecho o derecho natural.* España: Biblioteca Virtual Miguel de Cervantes, p 16. http://site.ebrary.com/lib/universidadcomplutense/Doc?id=10074215&ppg=16. Edición digital basada en la edición de Madrid, Librería de D. Leocadio López, 1874.

Fernández García, F. R. (1995) *Muestreo en poblaciones finitas: curso básico*. Barcelona: EUB.

Ferner, H., & Staubesand, J. (1974). *Atlas de anatomía humana. Tomo III, Sistema nervioso central, sistema nervioso autónomo, órganos de los sentidos y piel, vias de conducción periféricas* (Vol. III). Barcelona: Toray.

Festinger, L. y Katz, D. (1953). *Los dos métodos de investigación en las ciencias sociales*. Buenos Aires: Paidos.

Feyerabend, P. K. (1989). *Contra el método. Esquema de una teoría anarquista del conocimiento*. Barcelona: Ariel.

Flick, U. (2004 a). *Introducción a la Investigación Cualitativa*. Madrid: Morata.

Flick, U. et al. (2004 b). *A Companion to Qualitative Research*. London: Sage Publications.

French, S. (2000). Identity and individuality in Quantum Theory. En Stanford Encyclopedia of Philosophy.

Galilei, G. (1632/1995). *Diálogo sobre los máximos sistemas del mundo ptolemáico y copernicano.* (A. Beltran, Trad.) Madrid: alianza Editorial, D. L.

Galilei, G. (1981). *Consideraciones y demostraciones matemáticas sobre dos nuevas ciencias*. Madrid: Editora Nacional.

Gambara, H. (1998). *Diseño de investigaciones*. Madrid: McGraw-Hill.

García Ferrando, M., Ibañez, J. y Alvira, F. (Comp.) (2005). *El análisis de la realidad social*. Madrid: Alianza.

García, J. E. Et al, (2005). *Estadística descriptiva y nociones de probabilidad*. Madrid: Thomson.

Gazzaniga, M. S. (1998). *The Mind's Past.* Berkeley: University of California Press.

Glass, G. V. & Stanley, J. C. (1980). *Métodos Estadísticos Aplicados a las Ciencias Sociales*. Madrid: Prentice-Hall Internacional, D. L.

Gobo, G. (2004) "Sampling, representativeness and generalizability", en C. SEALE et al.

(ed.) *Qualitative research practice*. London: Sage Publications.

Goffman, E. (1961/2004) *Internados: ensayos sobre la situación social de los enfermos mentales*. Buenos Aires: Amorrortu.

González Seara, L. (1973). "Informe/respuesta de D. Luis González Seara". En J. M. Rodríguez Delgado. *Planificación cerebral del hombre futuro*. Madrid: Publicaciones de la Fundación Juan March.

Gouldner, A. W. (1954) *Patterns of Industrial Bureaucracy*. New York: Free Press.

Guyton, Artur, C. (1994). *Anatomía y fisiología del sistema nervioso*. Buenos Aires: Editorial Médica Panamericana.

Hair Jr., J. F. et al. (1999). *Análisis Multivariante.* Madrid: Prentice Hall

Hameroff, S. (2005). *Consciousness, neurobiology and quantum mechanics: The case for a connection*. Tucson, Arizona: The University of Arizona.

Harris, J. W. (2001) *Deep Souths: Delta, Piedmont, and Sea Island Society in the Age of Segregation*. Baltimore, MD: The Johns Hopkins University Press. Disponible en: http://site.ebrary.com/lib/universidadcomplutense/.

Hawkins, J. & Blakeslee, S. (2005). *Sobre la inteligencia*. Madrid: Espasa.

Hernández Sampieri, R. (2007). *Fundamentos de Metodología de la Investigación*. Madrid: McGraw Hill.

Kandel, E. R., Schwartz, J. H. And Jessell, T. M. (2001). *Principios de neurociencia*. Madrid: McGraw Hill.

Kant, I. (1787/2003). *Crítica de la razón pura.* (P. Ribas, Trad.), Madrid: Alfaguara.

Kanter, R. M. (1977) *Men and women of the Corporation.* New York: Basic Books

Kuhn, T.S. (1977). *La estructura de las Revoluciones Científicas*. Madrid: Fondo de Cultura Económica.

Lara Peinado, F. (1988). *Himnos sumerios*. Madrid: Tecnos.

Lara peinado, F. (1998). *La Civilización Sumeria*. Madrid: Historia 16.

León, O.G. y Montero, I. (1998). *Diseño de investigaciones. Introducción a la lógica de la investigación en Psicología y Educación*. Madrid: McGraw-Hill.

Lohr, S. S. (1999). *Muestreo: diseño y análisis*. México: Thomson.

López Cachero, M. (1992). *Fundamentos y métodos de estadística*. Madrid: Ediciones Pirámide.

Losada, J. L. y López-Feal R. (2003). *Métodos de investigación en Ciencias Humanas y Sociales*. Madrid: Thomson.

Lynd, R. S. and Lynd, H. M. (1937) *Middletown*. New York: Harcourt, Brace.

Manzano, V. G. et al. (1996) *Manual para encuestadores*. Barcelona: Ariel Practicum.

Martineau, H. (TR). (2000). *The Positive Philosophy of Auguste Comte*. Ontario: Batoche Book. 3 vol.

Maslow, A. H. (1954/1963). *Motivación y Personalidad.* Barcelona: Sagitario.

Masó Ferrer, F. (2007). El legado sumerio: el origen de la historia. *Historia National Geographic.* Julio, 2007, págs. 44-55.

Mateo Rivas, Mª J. (1992). *Estadística en investigación social: ejercicios resueltos*. Madrid: ITP Paraninfo.

Mateo Rivas, Mª J. y García Ferrando, M. (1993). *Estadística Aplicada a las Ciencias Sociales: estadística descriptiva, estadística inferencial*. Madrid: UNED.

Mayntz, R. (1976). *Introducción a los métodos de la sociología empírica.* Madrid: Alianza, D. L.

McGraw-Hill. (2002). Dictionary of Scientific and Technical Terms. (6ª). McGraw-Hill.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, Vol. 12 No. 2, pags. 153-157.

Micronet S. A. (abril de 1999). Enciclopedia Universal Micronet. (9ª). Madrid.

Miller, D. C. (1991). *Handbook of Research Design and Social Measurement*. Newbury Park, CA: Sage.

Mirás, J. (1985) *Elementos de muestreo para poblaciones finitas*. Madrid: INE.

Molina, M. (2000). *La ley más antigua. Textos legales sumerios*. Madrid: Trotta.

Moliner, M. (1996). *Diccionario de uso del español*. Madrid: Novell, Inc.

Mora, F. (1996). "Neurociencias: una nueva perspectiva de la naturaleza humana". En F. Mora (ed.). *El cerebro íntimo. Ensayos sobre neurociencia*. Barcelona: Ariel.

Morales, P. (1988). *Medición de actitudes en psicología y educación : construcción de escalas y problemas metodológicos*. San Sebastián: Ttarttalo.

Naccache, A. F. H. (2003). "Accumulation and Emergence in Cultural Evolution: The case of the Neolithic 'Revolution'", in Proceeding of the 3ICAANE, Paris, April. Disponible en: http://nidal.com/anaccash/Accumulation%20and%20Emergence.pdf.

Newton, I. (1686/1987). *Principios matemáticos de la filosofía natural*. Madrid: Alianza, D. L.

Norman, D. A. (1988) *The psychology of everyday things.* New York: Basic Books. (Traducción española: Norman, D. A. (2006) *La psicología de los objetos cotidianos*. Madrid: Nerea).

Norusis, M. J. (1986) *Base Manual SPSS/PC+ for the IBM PC/XT/AT*. Chicago: SPSS Inc.

Oxford English Dictionary. (2008). *The Oxford English Dictionary*. Recuperado el 10 de 10 de 2008, de http://www.oed.com.

Pardo, A. et al, (1994). *Análisis de datos en psicología II*. Madrid: Pirámide.

Parsons, T. (1999). *El sistema social*. Madrid: Alianza.

Payne, S. L. (1951/1980). *The art of asking questions*. New Jersey: Princenton University Press.

Payne, S. L. (1979). *The art of asking questions*. Princeton: Princeton University Press.

Pinker, S. (2005). *La tabla rasa, el buen salvaje y el fantasma en la máquina*. Barcelona: Paidós Ibérica.

Popper, K. R. (1994). *La lógica de la investigación científica*. Madrid: Técnos.

Real Academia Española. (2008). *Diccionario de la Lengua Española*. Recuperado el 10 de 10 de 2008, de http://www.rae.es/rae.html.

Rodríguez Osuna, J. (1991). *Métodos de muestreo*. Madrid: Centro de Investigaciones Sociológicas.

Rodríguez Osuna, J. (1993). *Métodos de muestreo: casos prácticos*. Madrid: Centro de Investigaciones Sociológicas.

Ruiz Olabuenaga, J. I. e Ispizua, M. A. (1989). *La descodificación de la vida cotidiana: métodos de investigación cualitativa*. Bilbao: Universidad de Deusto.

Sánchez Carrión, J. J. (1998/2005). *Manual de análisis estadístico de los datos*. Madrid: Alianza Editorial.

Sánchez-Crespo, F. A. y J. L. (1986) *Métodos y aplicaciones del muestreo*. Madrid: Alianza Universidad.

Scheaffer, R. L. et al (2007) *Elementos de muestreo*. Madrid: Thomson.

Scheuch, E. K. (1973). "La entrevista en la investigación social". En René König, *Tratado de sociología empírica*. Madrid: Tecnos, pp. 166-229.

Seale, C. et al. (2004) *Qualitative research practice*. London: Sage Publications.

Serbia, J. M. (2007) "Diseño, muestreo y análisis en la investigación cualitativa" *Hologramática*, Año 4, Número 7, V 3, p. 133. Disponible en: http://www.cienciared.com.ar/ra/usr/3/206/n7_vol3pp123_146.pdf.

Siegel, S. (1956/1970). *Estadística no paramétrica aplicada a las ciencias de la conducta*. México: Trillas.

Sobotta, J., Posel, P., & Scheneiderbanger, D. (1996). *Esquemas de anatomía. Vol. 3, SNC, vías y centros nerviosos* (Vol. 3). (P. San Juan Sanz, Trans.) Madrid: Marbán, D. L.

Society for Neuroscience (2008). *Brain Facts. A primer on the Brain and Nervous System.* Washington DC: Society for Neuroscience.

Solanas, A. (2004). *Estadística descriptiva en ciencias del comportamiento*. Madrid: Thomson.

Spiegel, M.R. (1998). *Estadística*. Madrid: McGraw Hill.

Stoetzel, J. y Girard, A. (1973). *Las encuestas de opinión pública*. Madrid: Instituto de la Opinión Pública.

TenHouten, W. D. (1971), "Political leadership in poor communities: aplication of two sampling methodologies", en P. Orleans and W. R. Ellis Jr (eds), *Race, change and urban society*, vol. V. Beverly Hills, CA: Sage.

Tezanos, J. F. (2006). *La explicación sociológica. Una introducción a la sociología*. Madrid: UNED.

Uña Juárez, O. y Hernández Sánchez, A. (2004). *Diccionario de Sociología*. Madrid: ESIC.

Valles, M. S. (1997). *Técnicas Cualitativas de Investigación Social. Reflexión metodológica y práctica profesional*. Madrid: Síntesis.

Van Dijk, T. A. (1983) Cognitive and conversational strategies in he ethnic prejudice, *Text*, 3 (4). 375-404.

Wallace, W. (1980). *La lógica de la ciencia en sociología*. Madrid: Alianza Universidad.

Warner, L. W. (1949) *Democracy in Jonesville.* New York: Harper & Row.

Watkins, T. (2000 a). *The Neolithic revolution and the emergence of humanity: a cognitive*

approach to the first comprehensive world-view. Disponible en: http://www.arcl.ed.ac.uk/arch/watkins/humanity_paper.pdf.

Watkins, T. (Trans.) (2000 b). *The Birth of the Gods and the origins of agriculture.* Cambridge: Cambridge University Press.

Wegener, A (1983). *El origen de los continentes y océanos*. Madrid: Pirámide.

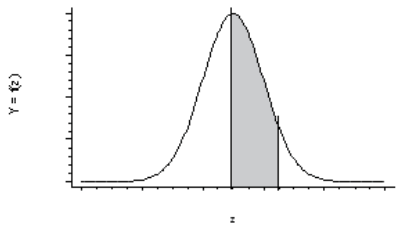
Whyte, W. F. (1943) *Street corner society*. Chicago: University of Chicago Press. (Traducción española: Whyte, W. F. (1971) *La sociedad de las esquinas*. México: Diana).

Williams, P. L. et al. (1998). *Anatomía de Gray:bases anatómicas de la medicina y la cirugía.* Madrid: Harcourt Brace. 2 vol.

Williams, P. L. *et al.* (eds.) (2001). *Anatomía de Gray: bases anatómicas de la medicina y la cirugía.* Madrid: Harcourt, D. L.

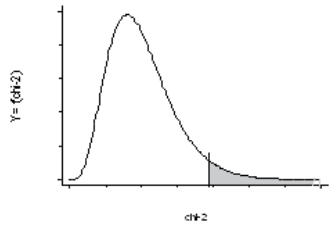
Wittgenstein, L. (1984). *Tractatus Logico-Philosophicus*. Madrid: Alianza Universidad.

1. Anexo. Normal Estandarizada.



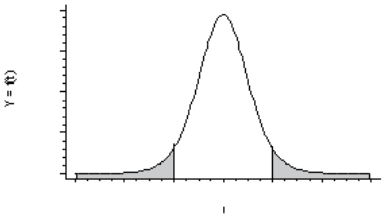
Area bajo la curva normal estandarizada entre 0 y z. (Elaboración propia)											
Valores de z: en cabecera de filas: unidades y décimas, y en cabecera de columnas: centésimas											
	0	**1**	**2**	**3**	**4**		**5**	**6**	**7**	**8**	**9**
0,0	0,0000	0,0040	0,0080	0,0120	0,0160		0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557		0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948		0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331		0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700		0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054		0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389		0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704		0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995		0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264		0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508		0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729		0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925		0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099		0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251		0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382		0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495		0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591		0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671		0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738		0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793		0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838		0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875		0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904		0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927		0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945		0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959		0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969		0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977		0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984		0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988		0,4989	0,4989	0,4989	0,4990	0,4990
3,1	0,4990	0,4991	0,4991	0,4991	0,4992		0,4992	0,4992	0,4992	0,4993	0,4993
3,2	0,4993	0,4993	0,4994	0,4994	0,4994		0,4994	0,4994	0,4995	0,4995	0,4995
3,3	0,4995	0,4995	0,4995	0,4996	0,4996		0,4996	0,4996	0,4996	0,4996	0,4997
3,4	0,4997	0,4997	0,4997	0,4997	0,4997		0,4997	0,4997	0,4997	0,4997	0,4998
3,5	0,4998	0,4998	0,4998	0,4998	0,4998		0,4998	0,4998	0,4998	0,4998	0,4998
3,6	0,4998	0,4998	0,4999	0,4999	0,4999		0,4999	0,4999	0,4999	0,4999	0,4999
3,7	0,4999	0,4999	0,4999	0,4999	0,4999		0,4999	0,4999	0,4999	0,4999	0,4999
3,8	0,4999	0,4999	0,4999	0,4999	0,4999		0,4999	0,4999	0,4999	0,4999	0,4999
3,9	0,5000	0,5000	0,5000	0,5000	0,5000		0,5000	0,5000	0,5000	0,5000	0,5000
4,0	0,5000	0,5000	0,5000	0,5000	0,5000		0,5000	0,5000	0,5000	0,5000	0,5000

2. Anexo. Chi cuadrado.



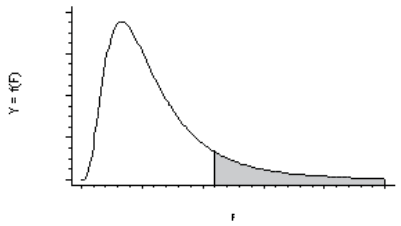
Area bajo la curva de Chi cuadrado, por encima de un valor de Chi cuadrado y gl grados de libertad. (Elaboración propia)										
gl	**Ns**									
	0,09	**0,08**	**0,07**	**0,06**	**0,05**	**0,04**	**0,03**	**0,02**	**0,01**	**0,001**
1	2,8744	3,0649	3,2830	3,5374	3,8415	4,2179	4,7093	5,4119	6,6349	10,8274
2	4,8159	5,0515	5,3185	5,6268	5,9915	6,4377	7,0131	7,8241	9,2104	13,8150
3	6,4915	6,7587	7,0603	7,4069	7,8147	8,3112	8,9473	9,8374	11,3449	16,2660
4	8,0434	8,3365	8,6664	9,0444	9,4877	10,0255	10,7119	11,6678	13,2767	18,4662
5	9,5211	9,8366	10,1910	10,5962	11,0705	11,6443	12,3746	13,3882	15,0863	20,5147
6	10,9479	11,2835	11,6599	12,0896	12,5916	13,1978	13,9676	15,0332	16,8119	22,4575
7	12,3372	12,6912	13,0877	13,5397	14,0671	14,7030	15,5091	16,6224	18,4753	24,3213
8	13,6975	14,0684	14,4836	14,9563	15,5073	16,1708	17,0105	18,1682	20,0902	26,1239
9	15,0342	15,4211	15,8537	16,3459	16,9190	17,6083	18,4796	19,6790	21,6660	27,8767
10	16,3516	16,7535	17,2026	17,7131	18,3070	19,0208	19,9219	21,1608	23,2093	29,5879
11	17,6526	18,0687	18,5334	19,0614	19,6752	20,4120	21,3416	22,6179	24,7250	31,2635
12	18,9395	19,3692	19,8488	20,3934	21,0261	21,7851	22,7418	24,0539	26,2170	32,9092
13	20,2140	20,6568	21,1507	21,7113	22,3620	23,1423	24,1249	25,4715	27,6882	34,5274
14	21,4778	21,9331	22,4408	23,0166	23,6848	24,4854	25,4931	26,8727	29,1412	36,1239
15	22,7319	23,1992	23,7202	24,3108	24,9958	25,8161	26,8480	28,2595	30,5780	37,6978
16	23,9774	24,4564	24,9901	25,5950	26,2962	27,1356	28,1908	29,6332	31,9999	39,2518
17	25,2150	25,7053	26,2514	26,8701	27,5871	28,4449	29,5227	30,9950	33,4087	40,7911
18	26,4455	26,9467	27,5049	28,1370	28,8693	29,7450	30,8447	32,3462	34,8052	42,3119
19	27,6695	28,1813	28,7512	29,3964	30,1435	31,0367	32,1577	33,6874	36,1908	43,8194
20	28,8874	29,4097	29,9910	30,6488	31,4104	32,3206	33,4623	35,0196	37,5663	45,3142
21	30,0998	30,6322	31,2246	31,8949	32,6706	33,5972	34,7593	36,3434	38,9322	46,7963
22	31,3071	31,8494	32,4526	33,1350	33,9245	34,8672	36,0491	37,6595	40,2894	48,2676
23	32,5096	33,0616	33,6754	34,3696	35,1725	36,1310	37,3323	38,9683	41,6383	49,7276
24	33,7077	34,2690	34,8932	35,5989	36,4150	37,3891	38,6093	40,2703	42,9798	51,1790
25	34,9015	35,4721	36,1065	36,8235	37,6525	38,6417	39,8804	41,5660	44,3140	52,6187
26	36,0914	36,6711	37,3154	38,0435	38,8851	39,8891	41,1461	42,8558	45,6416	54,0511
27	37,2777	37,8662	38,5202	39,2593	40,1133	41,1318	42,4066	44,1399	46,9628	55,4751
28	38,4604	39,0577	39,7213	40,4710	41,3372	42,3699	43,6622	45,4188	48,2782	56,8918
29	39,6398	40,2456	40,9187	41,6789	42,5569	43,6038	44,9132	46,6926	49,5878	58,3006
30	40,8161	41,4303	42,1126	42,8831	43,7730	44,8335	46,1600	47,9618	50,8922	59,7022
31	41,9895	42,6120	43,3033	44,0840	44,9853	46,0595	47,4024	49,2263	52,1914	61,0980
32	43,1600	43,7906	44,4909	45,2815	46,1942	47,2817	48,6410	50,4867	53,4857	62,4873
33	44,3278	44,9664	45,6755	46,4759	47,3999	48,5005	49,8759	51,7429	54,7754	63,8694
34	45,4930	46,1395	46,8573	47,6674	48,6024	49,7159	51,1073	52,9953	56,0609	65,2471
35	46,6558	47,3101	48,0364	48,8560	49,8018	50,9281	52,3350	54,2439	57,3420	66,6192
36	47,8163	48,4782	49,2129	50,0420	50,9985	52,1372	53,5596	55,4889	58,6192	67,9850
37	48,9744	49,6440	50,3869	51,2253	52,1923	53,3435	54,7811	56,7304	59,8926	69,3476
38	50,1305	50,8074	51,5586	52,4060	53,3835	54,5470	55,9995	57,9689	61,1620	70,7039
39	51,2845	51,9688	52,7280	53,5845	54,5722	55,7477	57,2151	59,2040	62,4281	72,0550
40	52,4364	53,1280	53,8952	54,7606	55,7585	56,9459	58,4278	60,4361	63,6908	73,4029
41	53,5865	54,2852	55,0603	55,9345	56,9424	58,1415	59,6379	61,6654	64,9500	74,7441

3. Anexo. t-Student.



Area bajo la curva t-Student, por encima de un valor de t y gl grados de libertad (bilateral). (Elaboración propia)										
	Ns									
gl	**0,09**	**0,08**	**0,07**	**0,06**	**0,05**	**0,04**	**0,03**	**0,02**	**0,01**	**0,001**
1	7,0264	7,9158	9,0579	10,5789	12,7062	15,8945	21,2051	31,8210	63,6559	636,5776
2	3,1040	3,3198	3,5782	3,8964	4,3027	4,8487	5,6428	6,9645	9,9250	31,5998
3	2,4708	2,6054	2,7626	2,9505	3,1824	3,4819	3,8961	4,5407	5,8408	12,9244
4	2,2261	2,3329	2,4559	2,6008	2,7765	2,9985	3,2976	3,7469	4,6041	8,6101
5	2,0978	2,1910	2,2974	2,4216	2,5706	2,7565	3,0029	3,3649	4,0321	6,8685
6	2,0192	2,1043	2,2011	2,3133	2,4469	2,6122	2,8289	3,1427	3,7074	5,9587
7	1,9662	2,0460	2,1365	2,2409	2,3646	2,5168	2,7146	2,9979	3,4995	5,4081
8	1,9280	2,0042	2,0902	2,1892	2,3060	2,4490	2,6338	2,8965	3,3554	5,0414
9	1,8992	1,9727	2,0554	2,1504	2,2622	2,3984	2,5738	2,8214	3,2498	4,7809
10	1,8768	1,9481	2,0283	2,1202	2,2281	2,3593	2,5275	2,7638	3,1693	4,5868
11	1,8588	1,9284	2,0067	2,0961	2,2010	2,3281	2,4907	2,7181	3,1058	4,4369
12	1,8440	1,9123	1,9889	2,0764	2,1788	2,3027	2,4607	2,6810	3,0545	4,3178
13	1,8317	1,8989	1,9742	2,0600	2,1604	2,2816	2,4358	2,6503	3,0123	4,2209
14	1,8213	1,8875	1,9617	2,0462	2,1448	2,2638	2,4149	2,6245	2,9768	4,1403
15	1,8123	1,8777	1,9509	2,0343	2,1315	2,2485	2,3970	2,6025	2,9467	4,0728
16	1,8046	1,8693	1,9417	2,0240	2,1199	2,2354	2,3815	2,5835	2,9208	4,0149
17	1,7978	1,8619	1,9335	2,0150	2,1098	2,2238	2,3681	2,5669	2,8982	3,9651
18	1,7918	1,8553	1,9264	2,0071	2,1009	2,2137	2,3562	2,5524	2,8784	3,9217
19	1,7864	1,8495	1,9200	2,0000	2,0930	2,2047	2,3457	2,5395	2,8609	3,8833
20	1,7816	1,8443	1,9143	1,9937	2,0860	2,1967	2,3362	2,5280	2,8453	3,8496
21	1,7773	1,8397	1,9092	1,9880	2,0796	2,1894	2,3278	2,5176	2,8314	3,8193
22	1,7734	1,8354	1,9045	1,9829	2,0739	2,1829	2,3202	2,5083	2,8188	3,7922
23	1,7699	1,8316	1,9003	1,9783	2,0687	2,1770	2,3132	2,4999	2,8073	3,7676
24	1,7667	1,8281	1,8965	1,9740	2,0639	2,1715	2,3069	2,4922	2,7970	3,7454
25	1,7637	1,8248	1,8929	1,9701	2,0595	2,1666	2,3011	2,4851	2,7874	3,7251
26	1,7610	1,8219	1,8897	1,9665	2,0555	2,1620	2,2958	2,4786	2,7787	3,7067
27	1,7585	1,8191	1,8867	1,9632	2,0518	2,1578	2,2909	2,4727	2,7707	3,6895
28	1,7561	1,8166	1,8839	1,9601	2,0484	2,1539	2,2864	2,4671	2,7633	3,6739
29	1,7540	1,8142	1,8813	1,9573	2,0452	2,1503	2,2822	2,4620	2,7564	3,6595
30	1,7520	1,8120	1,8789	1,9546	2,0423	2,1470	2,2783	2,4573	2,7500	3,6460
40	1,7375	1,7963	1,8617	1,9357	2,0211	2,1229	2,2503	2,4233	2,7045	3,5510
50	1,7289	1,7870	1,8516	1,9244	2,0086	2,1087	2,2338	2,4033	2,6778	3,4960
60	1,7232	1,7808	1,8448	1,9170	2,0003	2,0994	2,2229	2,3901	2,6603	3,4602
70	1,7192	1,7765	1,8401	1,9118	1,9944	2,0927	2,2152	2,3808	2,6479	3,4350
80	1,7162	1,7732	1,8365	1,9078	1,9901	2,0878	2,2095	2,3739	2,6387	3,4164
90	1,7138	1,7707	1,8337	1,9048	1,9867	2,0839	2,2050	2,3685	2,6316	3,4019
100	1,7120	1,7687	1,8315	1,9024	1,9840	2,0809	2,2015	2,3642	2,6259	3,3905
110	1,7105	1,7670	1,8297	1,9004	1,9818	2,0784	2,1986	2,3607	2,6213	3,3811
120	1,7092	1,7656	1,8282	1,8987	1,9799	2,0763	2,1962	2,3578	2,6174	3,3734
130	1,7081	1,7645	1,8270	1,8973	1,9784	2,0746	2,1942	2,3554	2,6142	3,3670
200	1,7036	1,7596	1,8217	1,8915	1,9719	2,0672	2,1857	2,3451	2,6006	3,3398
300	1,7009	1,7566	1,8184	1,8879	1,9679	2,0627	2,1805	2,3388	2,5923	3,3232
400	1,6995	1,7551	1,8168	1,8861	1,9659	2,0605	2,1779	2,3357	2,5882	3,3151
500	1,6987	1,7543	1,8158	1,8851	1,9647	2,0591	2,1763	2,3338	2,5857	3,3101
600	1,6981	1,7537	1,8152	1,8844	1,9639	2,0582	2,1753	2,3326	2,5841	3,3068
700	1,6977	1,7532	1,8147	1,8838	1,9634	2,0576	2,1745	2,3317	2,5829	3,3044
800	1,6975	1,7529	1,8143	1,8835	1,9629	2,0571	2,1740	2,3310	2,5820	3,3027
900	1,6972	1,7527	1,8141	1,8832	1,9626	2,0567	2,1735	2,3305	2,5813	3,3014
1000	1,6970	1,7525	1,8139	1,8829	1,9623	2,0564	2,1732	2,3301	2,5807	3,3002
10000	1,6956	1,7509	1,8121	1,8810	1,9602	2,0540	2,1704	2,3267	2,5763	3,2915
20000	1,6955	1,7508	1,8120	1,8809	1,9601	2,0539	2,1702	2,3265	2,5761	3,2911
30000	1,6955	1,7507	1,8120	1,8809	1,9600	2,0538	2,1702	2,3265	2,5760	3,2908
40000	1,6954	1,7507	1,8120	1,8808	1,9600	2,0538	2,1702	2,3264	2,5759	3,2908

4. Anexo. F de Fisher-Snedecor (F_S)



Area bajo la curva de F, por encima de un valor de F y K-1 gl (columnas) y N-K gl (filas). (Elaboración propia)													
		k-1											
N-k	**Ns**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
	0,01	6,8953	4,8239	3,9837	3,5127	3,2059	2,9877	2,8233	2,6943	2,5898	2,5033	2,4302	2,3676
100	**0,05**	**3,9362**	**3,0873**	**2,6955**	**2,4626**	**2,3053**	**2,1906**	**2,1025**	**2,0323**	**1,9748**	**1,9267**	**1,8857**	**1,8503**
	0,10	2,7564	2,3564	2,1394	2,0019	1,9057	1,8339	1,7778	1,7324	1,6949	1,6632	1,6360	1,6124
	0,01	6,7633	4,7128	3,8810	3,4143	3,1100	2,8933	2,7298	2,6012	2,4971	2,4106	2,3375	2,2747
200	**0,05**	**3,8884**	**3,0411**	**2,6498**	**2,4168**	**2,2592**	**2,1441**	**2,0556**	**1,9849**	**1,9269**	**1,8783**	**1,8368**	**1,8008**
	0,10	2,7308	2,3293	2,1114	1,9732	1,8763	1,8038	1,7470	1,7011	1,6630	1,6308	1,6031	1,5789
	0,01	6,7201	4,6766	3,8475	3,3822	3,0787	2,8625	2,6993	2,5709	2,4668	2,3804	2,3073	2,2444
300	**0,05**	**3,8726**	**3,0258**	**2,6347**	**2,4017**	**2,2441**	**2,1288**	**2,0402**	**1,9693**	**1,9112**	**1,8623**	**1,8206**	**1,7845**
	0,10	2,7223	2,3203	2,1021	1,9637	1,8666	1,7938	1,7369	1,6908	1,6525	1,6201	1,5922	1,5679
	0,01	6,6987	4,6586	3,8309	3,3664	3,0632	2,8472	2,6842	2,5559	2,4518	2,3654	2,2923	2,2294
400	**0,05**	**3,8648**	**3,0183**	**2,6272**	**2,3943**	**2,2366**	**2,1212**	**2,0325**	**1,9616**	**1,9033**	**1,8544**	**1,8126**	**1,7764**
	0,10	2,7181	2,3159	2,0975	1,9590	1,8617	1,7889	1,7318	1,6856	1,6472	1,6147	1,5867	1,5623
	0,01	6,6858	4,6479	3,8210	3,3569	3,0540	2,8381	2,6751	2,5469	2,4429	2,3565	2,2833	2,2204
500	**0,05**	**3,8601**	**3,0138**	**2,6227**	**2,3898**	**2,2320**	**2,1167**	**2,0279**	**1,9569**	**1,8986**	**1,8496**	**1,8078**	**1,7716**
	0,10	2,7156	2,3132	2,0948	1,9561	1,8588	1,7859	1,7288	1,6825	1,6441	1,6115	1,5835	1,5590
	0,01	6,6773	4,6407	3,8144	3,3506	3,0478	2,8321	2,6691	2,5409	2,4369	2,3505	2,2773	2,2144
600	**0,05**	**3,8570**	**3,0107**	**2,6198**	**2,3868**	**2,2290**	**2,1137**	**2,0248**	**1,9538**	**1,8955**	**1,8465**	**1,8046**	**1,7683**
	0,10	2,7139	2,3114	2,0929	1,9543	1,8569	1,7840	1,7268	1,6805	1,6420	1,6094	1,5813	1,5568
	0,01	6,6713	4,6356	3,8097	3,3460	3,0434	2,8278	2,6648	2,5367	2,4327	2,3463	2,2731	2,2102
700	**0,05**	**3,8548**	**3,0086**	**2,6176**	**2,3847**	**2,2269**	**2,1115**	**2,0226**	**1,9516**	**1,8932**	**1,8442**	**1,8023**	**1,7660**
	0,10	2,7127	2,3102	2,0916	1,9529	1,8555	1,7825	1,7253	1,6790	1,6405	1,6079	1,5797	1,5552
	0,01	6,6667	4,6318	3,8062	3,3427	3,0402	2,8245	2,6617	2,5335	2,4295	2,3431	2,2699	2,2070
800	**0,05**	**3,8531**	**3,0070**	**2,6160**	**2,3831**	**2,2253**	**2,1099**	**2,0210**	**1,9500**	**1,8916**	**1,8425**	**1,8006**	**1,7643**
	0,10	2,7118	2,3092	2,0906	1,9519	1,8545	1,7815	1,7243	1,6779	1,6394	1,6067	1,5786	1,5540
	0,01	6,6631	4,6288	3,8034	3,3401	3,0376	2,8220	2,6592	2,5310	2,4270	2,3406	2,2674	2,2045
900	**0,05**	**3,8518**	**3,0057**	**2,6148**	**2,3818**	**2,2240**	**2,1086**	**2,0197**	**1,9487**	**1,8903**	**1,8412**	**1,7993**	**1,7629**
	0,10	2,7111	2,3085	2,0899	1,9511	1,8537	1,7807	1,7234	1,6770	1,6385	1,6058	1,5777	1,5531
	0,01	6,6603	4,6264	3,8012	3,3380	3,0356	2,8200	2,6572	2,5290	2,4250	2,3386	2,2655	2,2025
1000	**0,05**	**3,8508**	**3,0047**	**2,6138**	**2,3808**	**2,2231**	**2,1076**	**2,0187**	**1,9476**	**1,8892**	**1,8402**	**1,7982**	**1,7618**
	0,10	2,7106	2,3079	2,0893	1,9505	1,8530	1,7800	1,7227	1,6764	1,6378	1,6051	1,5770	1,5524
	0,01	6,6476	4,6158	3,7914	3,3286	3,0264	2,8110	2,6482	2,5201	2,4162	2,3298	2,2566	2,1936
2000	**0,05**	**3,8461**	**3,0002**	**2,6094**	**2,3764**	**2,2186**	**2,1031**	**2,0142**	**1,9430**	**1,8846**	**1,8354**	**1,7934**	**1,7570**
	0,10	2,7080	2,3052	2,0865	1,9477	1,8502	1,7771	1,7197	1,6733	1,6347	1,6019	1,5737	1,5491
	0,01	6,6433	4,6123	3,7882	3,3254	3,0233	2,8080	2,6453	2,5172	2,4132	2,3268	2,2536	2,1907
3000	**0,05**	**3,8446**	**2,9987**	**2,6079**	**2,3749**	**2,2171**	**2,1016**	**2,0126**	**1,9415**	**1,8830**	**1,8339**	**1,7918**	**1,7554**
	0,10	2,7072	2,3044	2,0856	1,9467	1,8492	1,7761	1,7187	1,6722	1,6336	1,6008	1,5726	1,5480
	0,01	6,6412	4,6105	3,7865	3,3239	3,0218	2,8065	2,6438	2,5157	2,4118	2,3253	2,2522	2,1892
4000	**0,05**	**3,8438**	**2,9980**	**2,6071**	**2,3742**	**2,2163**	**2,1009**	**2,0119**	**1,9407**	**1,8822**	**1,8331**	**1,7910**	**1,7546**
	0,10	2,7068	2,3039	2,0852	1,9463	1,8487	1,7756	1,7182	1,6717	1,6331	1,6003	1,5721	1,5474
	0,01	6,6400	4,6094	3,7855	3,3229	3,0209	2,8056	2,6429	2,5148	2,4109	2,3245	2,2513	2,1883
5000	**0,05**	**3,8433**	**2,9975**	**2,6067**	**2,3737**	**2,2159**	**2,1004**	**2,0114**	**1,9403**	**1,8818**	**1,8326**	**1,7906**	**1,7541**
	0,10	2,7065	2,3036	2,0849	1,9460	1,8484	1,7753	1,7179	1,6714	1,6328	1,6000	1,5718	1,5471
	0,01	6,6391	4,6087	3,7849	3,3223	3,0203	2,8050	2,6423	2,5142	2,4103	2,3239	2,2507	2,1877
6000	**0,05**	**3,8430**	**2,9972**	**2,6064**	**2,3734**	**2,2156**	**2,1001**	**2,0111**	**1,9400**	**1,8814**	**1,8323**	**1,7902**	**1,7538**
	0,10	2,7064	2,3035	2,0847	1,9458	1,8482	1,7751	1,7177	1,6712	1,6326	1,5998	1,5715	1,5469
	0,01	6,6385	4,6082	3,7844	3,3218	3,0199	2,8045	2,6419	2,5138	2,4098	2,3234	2,2502	2,1873
7000	**0,05**	**3,8428**	**2,9970**	**2,6062**	**2,3732**	**2,2154**	**2,0999**	**2,0109**	**1,9397**	**1,8812**	**1,8321**	**1,7900**	**1,7536**
	0,10	2,7063	2,3033	2,0846	1,9457	1,8481	1,7749	1,7176	1,6711	1,6324	1,5996	1,5714	1,5467
	0,01	6,6381	4,6078	3,7841	3,3215	3,0195	2,8042	2,6415	2,5135	2,4095	2,3231	2,2499	2,1870
8000	**0,05**	**3,8426**	**2,9969**	**2,6060**	**2,3730**	**2,2152**	**2,0997**	**2,0107**	**1,9396**	**1,8811**	**1,8319**	**1,7898**	**1,7534**
	0,10	2,7062	2,3032	2,0845	1,9456	1,8480	1,7748	1,7175	1,6710	1,6323	1,5995	1,5713	1,5466
	0,01	6,6377	4,6075	3,7838	3,3212	3,0193	2,8040	2,6413	2,5132	2,4093	2,3229	2,2497	2,1867
9000	**0,05**	**3,8425**	**2,9967**	**2,6059**	**2,3729**	**2,2151**	**2,0996**	**2,0106**	**1,9394**	**1,8809**	**1,8318**	**1,7897**	**1,7532**
	0,10	2,7061	2,3032	2,0844	1,9455	1,8479	1,7748	1,7174	1,6709	1,6322	1,5994	1,5712	1,5465
	0,01	6,6374	4,6073	3,7836	3,3210	3,0191	2,8038	2,6411	2,5130	2,4091	2,3227	2,2495	2,1865
10000	**0,05**	**3,8424**	**2,9966**	**2,6058**	**2,3728**	**2,2150**	**2,0995**	**2,0105**	**1,9393**	**1,8808**	**1,8316**	**1,7896**	**1,7531**
	0,10	2,7060	2,3031	2,0843	1,9454	1,8478	1,7747	1,7173	1,6708	1,6321	1,5994	1,5711	1,5464
	0,01	6,6372	4,6071	3,7834	3,3209	3,0189	2,8036	2,6409	2,5129	2,4089	2,3225	2,2493	2,1864
11000	**0,05**	**3,8423**	**2,9965**	**2,6057**	**2,3727**	**2,2149**	**2,0994**	**2,0104**	**1,9393**	**1,8807**	**1,8316**	**1,7895**	**1,7531**
	0,10	2,7060	2,3031	2,0843	1,9454	1,8478	1,7746	1,7173	1,6708	1,6321	1,5993	1,5710	1,5464

5. Anexo. Tabla de números aleatorios

Tabla de números aleatorios uniformemente distribuidos (Elaboración propia).																									
43	60	06	72	18	07		88	55	85	03	90	02		89	38	18	74	73	55		89	38	18	74	73	55
30	73	84	66	39	14		16	33	51	78	98	79		19	49	59	17	86	92		19	49	59	17	86	92
28	91	80	46	75	07		57	23	73	25	18	73		11	35	90	04	60	16		11	35	90	04	60	16
32	17	29	67	28	69		03	19	85	15	27	43		29	53	22	21	62	29		29	53	22	21	62	29
62	76	11	15	71	64		04	81	28	53	85	29		02	42	96	71	35	27		02	42	96	71	35	27
60	22	33	73	54	77		18	38	36	83	46	71		37	50	80	01	20	55		37	50	80	01	20	55
71	64	77	98	44	24		81	26	20	23	49	05		92	48	97	07	06	58		92	48	97	07	06	58
26	33	29	88	01	94		11	76	39	02	31	40		46	53	05	26	56	45		46	53	05	26	56	45
51	66	86	27	05	57		64	68	12	34	14	69		35	31	70	33	28	50		35	31	70	33	28	50
10	41	33	29	53	55		35	55	78	93	97	37		42	39	59	47	54	51		42	39	59	47	54	51
81	54	89	96	74	89		67	13	21	31	24	80		64	06	76	49	26	76		64	06	76	49	26	76
30	80	94	70	16	19		11	82	58	04	43	65		17	67	48	21	20	24		17	67	48	21	20	24
51	38	22	41	65	34		12	80	98	08	50	80		50	10	52	82	75	87		50	10	52	82	75	87
66	88	18	74	94	39		84	09	17	40	05	36		69	46	06	58	24	44		69	46	06	58	24	44
92	12	67	51	01	26		47	94	13	44	66	70		78	16	51	60	55	91		78	16	51	60	55	91
67	93	03	56	67	60		03	08	30	61	08	49		51	04	46	32	85	17		51	04	46	32	85	17
41	34	29	08	88	50		56	09	13	53	16	70		38	73	10	16	18	98		38	73	10	16	18	98
53	95	43	36	21	68		68	60	02	72	24	86		91	36	09	28	79	12		91	36	09	28	79	12
80	83	14	03	87	66		40	39	41	83	20	32		32	64	13	17	08	14		32	64	13	17	08	14
23	17	53	59	90	50		34	35	22	78	46	37		20	23	34	72	22	76		20	23	34	72	22	76
80	78	34	10	71	61		59	27	23	27	39	40		52	81	02	96	11	69		52	81	02	96	11	69
53	03	37	06	92	93		70	19	58	16	43	20		42	21	10	51	25	15		42	21	10	51	25	15
56	63	73	50	42	03		37	31	03	45	44	45		39	86	87	27	67	50		39	86	87	27	67	50
97	75	71	29	97	73		96	84	44	22	13	98		21	81	96	15	31	48		21	81	96	15	31	48
80	51	90	41	77	10		19	61	74	19	66	36		67	16	43	37	96	03		67	16	43	37	96	03
58	99	87	38	40	69		48	69	07	76	90	65		40	48	99	81	62	53		40	48	99	81	62	53
38	82	32	80	50	85		60	52	87	09	33	63		76	86	31	03	90	36		76	86	31	03	90	36
12	79	24	12	97	15		04	65	29	37	51	73		69	22	88	62	52	63		69	22	88	62	52	63
99	63	64	28	80	33		57	17	50	78	75	14		31	57	73	33	56	30		31	57	73	33	56	30
37	45	32	57	89	77		04	07	74	54	08	61		18	45	24	30	23	20		18	45	24	30	23	20
68	91	58	55	12	80		21	68	63	11	47	71		41	83	38	82	07	13		41	83	38	82	07	13
41	64	36	69	37	94		96	25	73	46	80	45		92	66	46	95	26	57		92	66	46	95	26	57
54	06	38	18	99	83		41	11	39	22	09	50		84	72	52	05	67	01		84	72	52	05	67	01
65	32	34	19	31	50		25	95	47	82	64	26		60	15	38	86	20	51		60	15	38	86	20	51
61	85	29	47	04	84		36	69	68	31	67	42		58	40	13	27	62	92		58	40	13	27	62	92
49	05	27	21	61	33		61	88	26	99	88	60		80	96	98	45	29	38		80	96	98	45	29	38
85	43	10	92	34	82		77	30	82	28	69	05		33	93	13	59	64	09		33	93	13	59	64	09
32	45	86	01	53	20		81	15	33	80	25	38		23	68	73	57	78	06		23	68	73	57	78	06
93	12	54	41	13	12		05	14	96	12	15	39		69	16	34	07	04	98		69	16	34	07	04	98
23	79	69	14	10	91		12	97	07	81	77	04		51	73	84	31	54	57		51	73	84	31	54	57
68	91	58	55	12	80		21	68	63	11	47	71		41	83	38	82	07	13		41	83	38	82	07	13
41	64	36	69	37	94		96	25	73	46	80	45		92	66	46	95	26	57		92	66	46	95	26	57
54	06	38	18	99	83		41	11	39	22	09	50		84	72	52	05	67	01		84	72	52	05	67	01
65	32	34	19	31	50		25	95	47	82	64	26		60	15	38	86	20	51		60	15	38	86	20	51
61	85	29	47	04	84		36	69	68	31	67	42		58	40	13	27	62	92		58	40	13	27	62	92
49	05	27	21	61	33		61	88	26	99	88	60		80	96	98	45	29	38		80	96	98	45	29	38
85	43	10	92	34	82		77	30	82	28	69	05		33	93	13	59	64	09		33	93	13	59	64	09
32	45	86	01	53	20		81	15	33	80	25	38		23	68	73	57	78	06		23	68	73	57	78	06
93	12	54	41	13	12		05	14	96	12	15	39		69	16	34	07	04	98		69	16	34	07	04	98
23	79	69	14	10	91		12	97	07	81	77	04		51	73	84	31	54	57		51	73	84	31	54	57