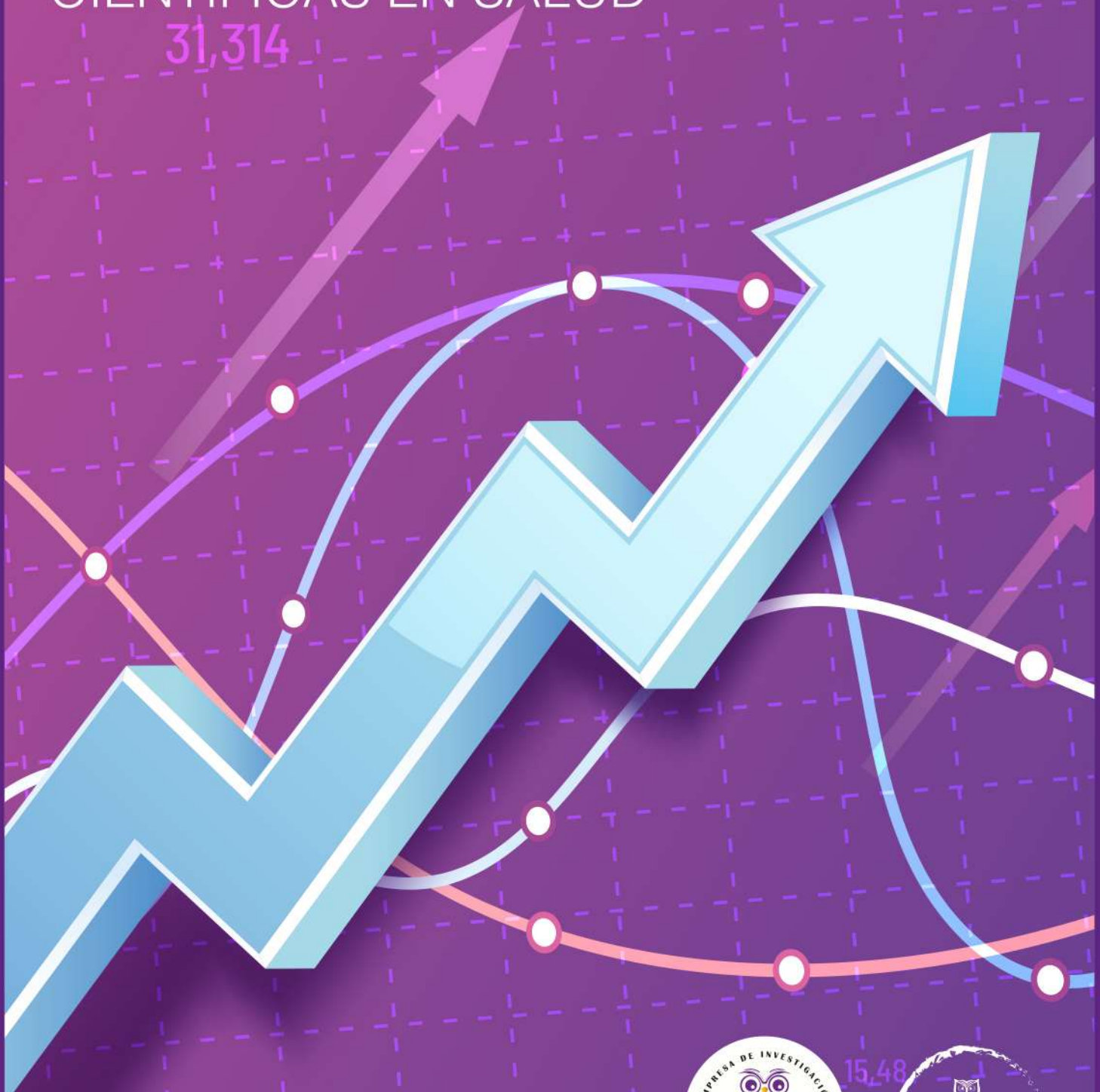
BIOESTADÍSTICA

APLICADA A INVESTIGACIONES CIENTÍFICAS EN SALUD



eBook

BIOESTADÍSTICA APLICADA A INVESTIGACIONES CIENTÍFICAS EN SALUD

1RA Edición

BIOESTADÍSTICA

APLICADA A INVESTIGACIONES
CIENTÍFICAS EN SALUD

Autores Investigadores

Kleber Dionicio Orellana Suarez
José Clímaco Cañarte Vélez

EDICIONES **MAWIL**

1RA Edición

BIOESTADÍSTICA

APLICADA A INVESTIGACIONES
CIENTÍFICAS EN SALUD

Autores Investigadores

Kleber Dionicio Orellana Suarez
Magíster en Contabilidad y Auditoría;
Ingeniero en Administración de Empresas Agropecuarias;
Docente de la Universidad Estatal del Sur de Manabí;
Jipijapa, Ecuador;
kleber.orellana@unesum.edu.ec
https://orcid.org/0000-0002-4202-0435

José Clímaco Cañarte Vélez
Licenciado en Laboratorio Clínico;
Magíster en Gerencia y Administración de Salud;
Profesor Titular en la Universidad Estatal del Sur de Manabí;
Jipijapa, Ecuador;
jose.canarte@unesum.edu.ec
https://orcid.org/0000-0002-3843-1143

1RA Edición

BIOESTADÍSTICA

APLICADA A INVESTIGACIONES
CIENTÍFICAS EN SALUD

Revisores Académicos

Roberto Enrique Alvarado Chacón
PhD. en Enfermería Salud y Cuidado Humano;
Magíster en Enfermería en Salud Reproductiva;
Especialista en Investigación en el Fenómeno de las Drogas;
Licenciado en Enfermería; Abogado;
Docente Titular Agregado 1, de la Universidad Metropolitana,
Carrera de Enfermería Sede Quito; Quito, Ecuador;
r.alvarado@umet.edu.ec
https://orcid.org/0000-0002-8883-3140

Elsa Josefina Albornoz Josefina
PhD en Ciencias de la Educación,
PhD en Gerencia de la Administración Pública,
Magister Scientiarum en Investigación Educativa;
Magister en Ciencias de la Orientación de la Conducta,
Especialista en Docencia Universitaria,
Especialista en Salud Pública; Licenciada en Enfermería;
Tesista de Filosofía Docente Titular de la Universidad Metropolitana;
Carrera de Enfermería Guayaquil; Guayaquil, Ecuador;
ealbornoz@umet.edu.ec
https://orcid.org/0000-0003-1382-0596

Catalogación Bibliografica

AUTORES: Kleber Dionicio Orellana Suarez
José Clímaco Cañarte Vélez

Título: Bioestadística aplicada a investigaciones científicas en Salud
Descriptores: Ciencias de la Vida; Bioestadística; Ciencias de la Salud; Investigación científica
Código UNESCO: 2404.01 Bioestadística
Clasificación Decimal Dewey/Cutter: 570.1/Or343
Área: Investigación
Edición: 1era
ISBN: 978-9942-602-23-7
Editorial: Mawil Publicaciones de Ecuador, 2022
Ciudad, País: Quito, Ecuador
Formato: 148 x 210 mm.
Páginas: 235
DOI: https://doi.org/10.26820/978-9942-602-23-7

Texto para docentes y estudiantes universitarios

El proyecto didáctico **Bioestadística aplicada a investigaciones científicas en Salud,** es una obra colectiva escrita por varios autores y publicada por MAWIL; publicación revisada por el equipo profesional y editorial siguiendo los lineamientos y estructuras establecidos por el departamento de publicaciones de MAWIL de New Jersey.

Director Académico: PhD. Jose María Lalama Aguirre
Dirección Central MAWIL: Office 18 Center Avenue Caldwell; New Jersey # 07006
Gerencia Editorial MAWIL-Ecuador: Mg. Vanessa Pamela Quishpe Morocho
Editor de Arte y Diseño: Lic. Eduardo Flores, Arq. Alfredo Díaz
Corrector de estilo: Lic. Marcelo Acuña Cifuentes

1RA Edición

BIOESTADÍSTICA

APLICADA A INVESTIGACIONES
CIENTÍFICAS EN SALUD

ÍNDICE

ÍNDICE

MAWIL
Publicaciones Impresas
y Digitales

CAPÍTULO II

CAPÍTULO III

CAPÍTULO IV

1RA Edición

BIOESTADÍSTICA

APLICADA A INVESTIGACIONES
CIENTÍFICAS EN SALUD

PRÓLOGO

PRÓLOGO

MAWIL
Publicaciones Impresas
y Digitales

La investigación científica, comprendida como un proceso de generación de nuevos conocimientos o la comprobación de conocimientos ya desarrollados, tiene una estrecha relación con la estadística aplicada (bioestadística en salud).

El propósito es mostrar esta estrecha relación entre la bioestadística y la investigación científica. Asimismo, se presentan de manera general las herramientas que ofrece la bioestadística a los investigadores; debido a su rápido desarrollo es imperiosa su implementación en la academia y en las técnicas de investigación sobre salud.

En nuestros días, la bioestadística se ha convertido en un método efectivo para describir con exactitud los valores de los datos sociales, sanitarios, psicológicos, biológicos y físicos, y sirve como herramienta para relacionar y analizar dichos datos. El actual trabajo del experto estadístico no consiste ya solo en reunir y tabular los datos, sino, sobre todo, en interpretar esa información.

Esta obra presenta, de manera didáctica y con procedimientos de fácil aplicación, las diferentes teorías y conceptos relacionados con el análisis de datos; esto permitirá disminuir de manera significativa el déficit de profesionales y estudiantes de las ciencias de la salud que presentan limitaciones en esta área del conocimiento tan transcendental como es la bioestadística.

Acerca de este libro

Es importante mencionar que este texto es fruto de la experiencia de sus autores en el campo de la docencia y la investigación, las mismas que incentivaron la lectura de una gran cantidad de textos y manuscritos sobre la materia, algunos con un gran aporte y otros complementarios. En este sentido, el propósito principal del texto es acercar a usted, amable lector, los conceptos y las diversas técnicas de análisis de datos de manera amigable, mediante ejemplos prácticos y con el uso de herramientas adecuadas, como el paquete estadístico de IBM

SPSS, versión 25, hoja de cálculo de Microsoft Excel y una adecuada interpretación de sus resultados.

Sobre los autores

Kléber Dionicio Orellana Suárez
De nacionalidad ecuatoriana, profesional en Administración de Empresas Agropecuarias con maestría en Contabilidad y Auditoría, consultor y capacitador en Estadística, docente universitario de grado y posgrado por 13 años en asignaturas como Estadística, Bioestadística en Salud, Metodología de la Investigación y Demografía en la carrera de Laboratorio Clínico, Facultad de Ciencias de la Salud de la Universidad Estatal del Sur de Manabí. Desde el año 2017 a la fecha desempeña funciones como coordinador del área de Investigación de la UNESUM.

José Clímaco Cañarte Vélez
De nacionalidad ecuatoriana, con formación de grado como licenciado en Laboratorio Clínico, y formación de posgrado como magíster en Gerencia y Administración de Salud, profesor titular en la Universidad Estatal del Sur de Manabí (UNESUM), carrera de Laboratorio Clínico de la Facultad de Ciencias de la Salud, donde dicta las asignaturas de Administración en Salud y Salud Pública. Actualmente desempeña funciones como coordinador de la Maestría Ciencias del Laboratorio Clínico de la UNESUM.

1RA Edición

BIOESTADÍSTICA

APLICADA A INVESTIGACIONES
CIENTÍFICAS EN SALUD

CAPÍTULO I

ASPECTOS INTRODUCTORIOS A LA BIOESTADÍSTICA

MAWIL
Publicaciones Impresas y Digitales

1.1. Historia de la estadística

La historia de la estadística es la historia del hombre sobre la tierra. Sin duda que empezó con las interrogantes sobre las distancias recorridas, el conteo de las personas del grupo, volúmenes de cifras relativas a nacimientos, muertes, impuestos, poblaciones, ingresos, deudas, créditos y demás.

Desde los comienzos de la civilización han existido formas sencillas de estadísticas, pues ya se utilizaban representaciones gráficas y otros símbolos en pieles, rocas, palos de madera y paredes de cuevas para contar el número de personas, animales y diversos artículos.

Hacia el año 3000 a. C. los babilonios utilizaban ya pequeñas tablillas de arcilla para recopilar datos sobre la producción agrícola y los géneros vendidos o cambiados mediante trueque. En el antiguo Egipto, los faraones lograron recopilar, alrededor del año 3050 a. C., prolijos datos relativos a la población y la riqueza del país; de acuerdo con el historiador griego Heródoto, dicho registro de la riqueza y la población se hizo con el propósito de preparar la construcción de las pirámides. En el mismo Egipto, Ramsés II hizo un censo de las tierras con el objeto de verificar un nuevo reparto (González, 2005).

Pero fueron los romanos, maestros de la organización política, quienes mejor supieron emplear los recursos de la estadística. Cada cinco años llevaban a cabo un censo de la población, y los funcionarios públicos tenían la obligación de anotar nacimientos, defunciones y matrimonios, sin olvidar los recuentos periódicos del ganado y de las riquezas contenidas en las tierras conquistadas. En la época del nacimiento de Cristo sucedía uno de estos empadronamientos de la población bajo la autoridad del Imperio.

Aunque Carlo Magno, en Francia, y Guillermo el Conquistador, en Inglaterra, trataron de revivir la técnica romana, los métodos estadísticos permanecieron casi olvidados durante la Edad Media. En los siglos XV,

XVI y XVII, hombres como Leonardo de Vinci, Nicolás Copérnico, Galileo Galilei, William Harvey, Francis Bacon y René Descartes hicieron grandes operaciones con base en el método científico, de tal forma que cuando se crearon los Estados nacionales y surgió como fuerza el comercio internacional, había ya un método capaz de aplicarse a los datos económicos.

Debido al temor que Enrique VII tenía de la peste, en el año 1532 empezaron a registrarse en Inglaterra las defunciones causadas por esta enfermedad. En Francia, más o menos por la misma época, la ley exigía a los clérigos registrar los bautismos, fallecimientos y matrimonios (González, 2005).

Durante un brote de peste que apareció a fines del siglo XVI, el gobierno inglés comenzó a publicar estadísticas semanales de los decesos. Esa costumbre continuó por muchos años, y en 1632 las llamadas "bills of mortality" (cuentas de mortalidad) ya contenían datos sobre los nacimientos y fallecimientos por sexo. En 1662, el capitán John Graunt compiló documentos que abarcaban treinta años, mediante los cuales efectuó predicciones sobre el número de personas que morirían de diversas enfermedades, así como de las proporciones de nacimientos de hombres y mujeres que cabía esperar. El trabajo de Graunt, condensado en su obra ***Natural and political observations… made upon the bills of mortality*** (Observaciones políticas y naturales… hechas a partir de las cuentas de mortalidad), fue un esfuerzo de inferencia y teoría estadística (González, 2005).

Después de revisar miles de partidas de defunción, pudo demostrar que en tales años no fallecían más personas que en los demás. Los procedimientos de Neumann fueron conocidos por el astrónomo inglés Halley, descubridor del cometa que lleva su nombre, quien los aplicó al estudio de la vida humana. Sus cálculos sirvieron como base para establecer las tablas de mortalidad que hoy utilizan todas las compañías de seguros.

Godofredo Achenwall, profesor de la Universidad de Gotinga, acuñó en 1760 la palabra estadística, que extrajo del término italiano estatista (estadista). Creía, y con sobrada razón, que los datos de la nueva ciencia serían el aliado más eficaz del gobernante consciente. La raíz remota de la palabra se halla en el término latino ***status***, que significa "estado" o "situación". Esta etimología aumenta el valor intrínseco de la palabra, por cuanto la estadística revela el sentido cuantitativo de las más variadas situaciones.

Durante el siglo XVIII empieza el auge de la estadística descriptiva en asuntos sociales y económicos, y es a finales de ese siglo y comienzos del XIX cuando se comienzan a asentar verdaderamente las bases teóricas de la teoría de probabilidades con los trabajos de Joseph Louis Lagrange y Pierre Simon de Laplace, del brillantísimo y ubicuo matemático y astrónomo alemán Carl Friedrich Gauss, y de Simeón-Denis Poisson. Previamente, cabe destacar el descubrimiento de la distribución normal por Abraham de Moivre, distribución que será posteriormente "redescubierta" por Gauss y Poisson.

Jacques Quételect es quien aplica la estadística a las ciencias sociales. Interpretó la teoría de la probabilidad para su uso en esas ciencias y empleó el principio de promedios y de la variabilidad a los fenómenos sociales.

En el periodo de 1800 a 1820 se desarrollaron dos conceptos matemáticos fundamentales para la teoría estadística: la teoría de los errores de observación, aportada por Laplace y Gauss, y la teoría de los mínimos cuadrados, realizada por Laplace, Gauss y Legendre. A finales del siglo XIX, Sir Francis Galton ideó el método conocido como correlación, que tenía por objeto medir la influencia relativa de los factores sobre las variables. De aquí partió el desarrollo del coeficiente de correlación creado por Karl Pearson y otros cultivadores de la ciencia biométrica, tales como J. Pease Norton, R. H. Hooker y G. Udny Yule, que efectuaron amplios estudios sobre la medida de las relaciones.

Una vez sentadas las bases de la teoría de probabilidades, podemos situar el nacimiento de la estadística moderna y su empleo en el análisis de experimentos en los trabajos de Francis Galton y Kurt Pearson. Este último publicó en 1892 el libro ***The grammar of science*** (La gramática de la ciencia), un clásico en la filosofía de la ciencia, y fue él quien ideó el conocido test de chi-cuadrado. El hijo de Pearson, Egon, y el matemático nacido en Polonia, Jerzy Neyman, pueden considerarse los fundadores de las pruebas modernas de contraste de hipótesis.

Pero es, sin lugar a dudas, Ronald Arnold Fisher la figura más influyente de la estadística, pues la situó como una poderosa herramienta para la planeación y análisis de experimentos. Contemporáneo de Pearson, desarrolló el análisis de varianza y fue pionero en el impulso de numerosas técnicas de análisis multivariante y en la introducción del método de máxima verosimilitud para la estimación de parámetros. Su libro ***Statistical methods for research workers*** (Métodos estadísticos para los investigadores), publicado en 1925, ha sido probablemente el libro de estadística más utilizado a lo largo de muchos años (González, 2005).

En la línea de tiempo que se indica a continuación, se puede observar una parte del desarrollo histórico de la estadística, destacando la contribución de algunos acontecimientos y los nombres de célebres estadísticos:

Años	Acontecimiento
100-44 a. C.	**Julio César** decretó que se realice un censo en el Imperio romano para el cobro de impuestos (Wikipedia, 2019).
1027-1087	**Guillermo I**, rey de Inglaterra, ordenó que se hiciera un registro de todos los bienes para fines tributarios y militares (suttonclonard.com, 2019).
1654	**Pierre de Fermat** desarrolló los principios de la teoría de la probabilidad y un algoritmo de diferenciación, mediante el cual pudo determinar los valores máximos y mínimos de una curva polinómica (Biografias y vidas, 2019).
1662	**John Graunt** realizó predicciones de mortalidad y de proporciones de hombres y mujeres (Wikipedia, 2019).

1718-1730	**Abraham de Moivre** publicó tres obras, entre 1718 y 1730, sobre temas de probabilidad, probabilidad binomial y aproximación para muestras grandes (Biografias y vidas, 2019).
1763	**Thomas Bayes** fue un estadístico. Una obra póstuma publicada en 1763, dos años después de su muerte, denominada *Ensayo sobre la resolución de un problema en la doctrina del azar*, trata el problema de las causas a través de los efectos observados y donde se enuncia el teorema que lleva su nombre (Wikipedia, 2019).
1801	**Carl Gauss**, junto a Arquímedes y Newton, es uno de los tres genios de la historia de las matemáticas. Quizás la obra más importante publicada por Gauss sea *Disquisitiones arithmeticae,* en 1801. Llegó a publicar alrededor de 155 títulos.
1837	**Simeón Denis Poisson** realizó muchas publicaciones, pero es especialmente importante *Recherches sur la probabilité des jugements en matiéres criminelles et matière civile* (1837) que contiene el germen de dos elementos asociados al nombre de Poisson: la ley de probabilidad conocida como distribución de Poisson, y la generalización de la ley de los grandes números de Bernoulli.
1875	**Wilhelmm Lexis** publica su libro *Einleitung in die theorie der bevolkerungsstatistik, donde destaca la* teoría de la dispersión estadística y un nuevo método de elaboración de tablas de mortalidad. Además, inicia un tema importante dentro de la estadística, el de las series de tiempo.
1902	**Karl Pearson**, considerado como uno de los fundadores de la estadística, realizó investigaciones en el campo de la herencia y de la genética, fundando en 1902 el periódico *Biometrika*. Mostró interés en distribuciones probabilísticas asimétricas, en contraposición con las distribuciones normales, simétricas. Introdujo una familia de distribuciones probabilísticas, hoy conocida como Gama. A Karl Pearson se debe también el estadístico ji-cuadrado, introducido en 1900, utilizado para la comparación entre dos tablas de frecuencia, y una de sus aplicaciones es el probar el ajuste de una ley probabilística a un conjunto de datos empíricos (Divulgamat, 2019).
1822-1911	**Francis Galton** estudió la regresión y componentes de la varianza. También utilizó la ley de probabilidad normal, en su versión bivariada, para describir el comportamiento probabilístico de los errores de dos características que varían en forma conjunta (Psicoactiva, 2019).
1890	**Maurice Kendall** sitúa el origen de la estadística teórica moderna. Es uno de los que desarrollaron la estadística no paramétrica (ECO, 2019).
1892	**Frank Wilcoxon** nació en 1892. Hizo aportes de importancia en el campo de la estadística no paramétrica. Recurrió a la simple idea de reemplazar los datos por sus rangos de orden (Estadística, 2019).

1904	**Charles Spearman** propuso el primer modelo factorial (1904). También aportó el coeficiente de correlación ordinal que lleva su nombre, que permite correlacionar dos variables por rangos en lugar de medir el rendimiento separado en cada una de ellas. Sus obras más importantes son *The nature of intelligence and the principles of cognition* (1923) y *The abilities of man* (1927) (Biografias y vidas, 2019).
1934	**Jerzy Neyman** publicó muchos libros relacionados a experimentos y estadísticas. Neyman ideó la forma con la cual la Administración de Alimentos y Fármacos (FDA-USDA) prueba los medicamentos en la actualidad. Introdujo el concepto de los intervalos de confianza (Wikipedia, 2019).
1908	**William Sealy Gosset** publica el artículo "El error probable de una media", bajo el seudónimo de "Student" que constituye un paso importante en la cuantificación de los resultados de la experimentación (wikipedia.org, 2019).
1925	**Ronald Fisher** publicó su metodología estadística en 1925 en *Methods for research workers*. Trasladó sus investigaciones al campo de la genética en *The genetical theory of natural selection* (1930), donde destaca el papel de control que ejercen los genes sobre los caracteres dominantes y considera la selección como la fuerza directriz de la evolución (Wikipedia, 2019).
1934	**George Snedecor** trabajó junto con Ronald Fisher y de dicha colaboración surgieron muchos de los resultados en los que se basa el análisis de varianza. Uno de sus textos más famosos es *Cálculo e interpretación del análisis de varianza y covarianza*, que publicó en 1934 (Estadística inferencial, 2019).
1965	**William Cochran**, nombre ligado a Snedecor, quien hizo aportes al diseño de experimentos y a la teoría del muestreo (Rojas & Rojas, 2000).
1933	**Harold Hotelling** trabajo con Fisher y se interesó en comparar tratamientos agrícolas en función de varias variables descubriendo las semejanzas entre este problema y el planteado por Pearson (EUMED, 2019).

El primer médico que utilizó métodos matemáticos para cuantificar variables de pacientes y sus enfermedades fue el francés Pierre Charles-Alexandre Louis (1787-1872). La primera aplicación del ***método numérico*** (que es como tituló a su obra y llamó a su método) es su clásico estudio de la tuberculosis, que influyó en toda una generación de estudiantes. Sus discípulos, a su vez, reforzaron la nueva ciencia de la epidemiología con el método estadístico. En las recomendaciones de Louis para evaluar diferentes métodos de tratamiento están las bases de los ensayos clínicos que se hicieron un siglo después. En Francia Louis René Villermé (1782-1863) y en Inglaterra William Farr

(1807-1883) —que había estudiado estadística médica con Louis— hicieron los primeros mapas epidemiológicos usando métodos cuantitativos y análisis epidemiológicos. Francis Galton (1822-1911), basado en el darwinismo social, fundó la biometría estadística.

Pierre Simon Laplace (1749-1827), astrónomo y matemático francés, publicó en 1812 un tratado sobre la teoría analítica de las probabilidades, *Théorie analytique des probabilités*, sugiriendo que tal análisis podría ser una herramienta valiosa para resolver problemas médicos.

Los primeros intentos por hacer coincidir las matemáticas de la teoría estadística con los conceptos emergentes de la infección bacteriana tuvieron lugar a comienzos del siglo XX. Tres diferentes problemas cuantitativos fueron estudiados por otros tantos autores. William Heaton Hamer (1862-1936) propuso un modelo temporal discreto en un intento de explicar la ocurrencia regular de las epidemias de sarampión; John Brownlee (1868-1927), primer director del British Research Council, luchó durante veinte años con problemas de cuantificación de la infectividad epidemiológica, y Ronald Ross (1857-1932) exploró la aplicación matemática de la teoría de las probabilidades con la finalidad de determinar la relación entre el número de mosquitos y la incidencia de malaria en situaciones endémicas y epidémicas. Pero el cambio más radical en la dirección de la epidemiología se debe a Austin Bradford Hill (1897-1991) con el ensayo clínico aleatorizado, y en colaboración con Richard Doll (n. 1912), el épico trabajo que correlacionó el tabaco y el cáncer de pulmón (González, 2005).

Los primeros trabajos bioestadísticos en enfermería los realizó, a mediados del siglo XIX, la enfermera inglesa Florence Nightingale. Durante la guerra de Crimea, Florence Nightingale observó que eran mucho más numerosas las bajas producidas en el hospital que en el frente. Por lo tanto, recopiló información y dedujo que la causa de la elevada tasa de mortalidad se debía a la precariedad higiénica existente. Así, gracias a sus análisis estadísticos, se comenzó a tomar conciencia de

la importancia y la necesidad de unas buenas condiciones higiénicas en los hospitales (Wikipedia, 2019).

1.2. Definición de bioestadística y su aplicación en investigaciones en salud

1.2.1. Definición de bioestadística

Como se pudo evidenciar en la cronología histórica, la estadística ha variado su significado a través del tiempo, pasando de ser una herramienta usada solo para la administración del Estado, papel fundamental que lo sigue cumpliendo, a una ciencia con un sinfín de aplicaciones en diferentes disciplinas para el análisis de datos.

La estadística estudia los métodos y procedimientos para recoger, clasificar, resumir, hallar regularidades, analizar e interpretar los datos, siempre y cuando la variabilidad e incertidumbre sean una causa intrínseca de los mismos; así como realizar inferencias a partir de ellos, con la finalidad de ayudar a la toma de decisiones y, en su caso, formular predicciones.

Se entiende como **bioestadística** la aplicación de técnicas estadísticas a las ciencias de la naturaleza, entre las que se encuentran todas las ciencias de la salud. Para que esta definición tenga sentido habremos de entender plenamente qué es la estadística.

La estadística pasa a ser una ciencia básica cuyo objetivo principal es el procesamiento y análisis de grandes volúmenes de datos, resumiéndolos en tablas, gráficos e indicadores (estadísticos), que permiten la fácil comprensión de las características concernientes al fenómeno estudiado.

La investigación científica, comprendida como un proceso de generación de nuevos conocimientos o la comprobación de conocimientos ya desarrollados, tiene una estrecha relación con la estadística aplicada (bioestadística).

El propósito es, por lo tanto, mostrar la relación entre la bioestadística y la investigación científica. Asimismo, se presentan de manera más general las herramientas que ofrece la bioestadística a los investigadores.

Dejando de lado la bioestadística en su contexto puramente teórico y, por el contrario, considerando su dimensión aplicada, la bioestadística puede ser comprendida como un conjunto de herramientas que tienen un doble propósito:

a) Apoyar el proceso de generación de datos de manera organizada, "Base de datos"

El protocolo de observación implica al menos:

- Definición de la población y unidades de observación.
- Definición del tamaño de la muestra.
- Definición del plan de muestreo.
- Definición de las variables.
- Definición del método de colecta de datos.
- Definición de los métodos estadísticos.

El protocolo de experimentación:

- Definición de factores.
- Definición del diseño experimental.
- Definición del modelo matemático.
- Definición de las unidades de experimentación.
- Definición de observaciones.

b) Proponer herramientas pertinentes de análisis de datos colectados

En el análisis de los datos se deben considerar tres etapas:

- **Análisis exploratorio de datos**. En la literatura en inglés se conoce como herramientas EDA (Exploring Data Analysis). El propósito de las herramientas EDA es controlar la calidad de los datos.

- **Análisis descriptivo de los datos.** El propósito de las herramientas de la estadística descriptiva es resumir los datos.
- **Análisis inferencial de los datos.** El propósito de la herramienta de inferencia estadística es generalizar los resultados obtenidos en la muestra hacia la población.

El análisis estadístico realiza un tratamiento de la información procedente de una investigación, abarcando diversos aspectos relativos a su descripción, o a la extracción de conclusiones y generalización de éstas que podamos realizar.

La calidad de las aplicaciones estadísticas realizadas dependerá en gran medida de un manejo correcto de tal información y una identificación adecuada de los diversos elementos estadísticos que la conforman, pudiéndose extraer conclusiones erróneas como consecuencia de una identificación incorrecta.

El desarrollo de software estadísticos cada vez más accesibles, plantea un gran desafío para el profesional estadístico, porque permite poner en evidencia el peligro que representa el uso de estas herramientas sin bases teóricas sólidas en estadística.

La mayor parte de los programas estadísticos comerciales, SPSS, R, MINITAB, Info Stat, etc., operan con menús desplegables muy fáciles de manejar y, por tanto, accesible a cualquier tipo de usuario.

El trabajo estadístico **no consiste ya solo en reunir y tabular los datos, sino, sobre todo, en el proceso de interpretación** de esa información.

1.2.2. Objetivo y aplicaciones de la estadística en salud pública

Entre los objetivos más importantes relacionados con la estadística y que contribuyen al campo de la salud pública y sectores relacionados tenemos los siguientes:

- Permite comprender los fundamentos racionales en que se basan las decisiones en materia de diagnóstico, pronóstico y terapéutica.
- Interpreta las pruebas de laboratorio y las observaciones y mediciones clínicas con un conocimiento de las variaciones fisiológicas y de las correspondientes al observador y a los instrumentos.
- Proporciona el conocimiento y comprensión de la información acerca de la etiología y el pronóstico de las enfermedades, a fin de asesorar a los pacientes sobre la manera de evitar las enfermedades o limitar sus efectos.
- Otorga un discernimiento de los problemas sanitarios para que eficientemente se apliquen los recursos disponibles para resolverlos.

Adicionalmente a los objetivos antes citados, resalta la utilidad de la estadística en el desarrollo del pensamiento crítico, a fin de:

a. Pensar críticamente acerca de los problemas de salud;
b. Evaluar correctamente los datos disponibles para la toma de decisiones; e
c. Identificar las decisiones y conclusiones que carecen de base científica y lógica.

Los principios y conceptos de los métodos estadísticos se aplican en diversos campos de la salud pública, tales como en estudios de variación, diagnóstico de enfermedades y de la salud de la comunidad, predicción del resultado probable de un programa de intervención, elección apropiada de intervención en pacientes o comunidad, administración sanitaria, realización y análisis en las investigaciones de salud pública.

Cuadro 1. Aplicación de la estadística en salud pública.

Área de aplicación	Descripción	Ejemplo
Estudios de variación	La variación de una característica se produce cuando su valor cambia de un sujeto a otro, o de un momento a otro en el mismo sujeto.	Edad, peso, estatura, presión sanguínea, niveles de colesterol, albúmina sérica, recuento de plaquetas.
Diagnóstico de enfermedades y de la salud de la comunidad	Proceso mediante el cual se identifican el estado de salud de un individuo, o de un grupo, y los factores que lo producen.	Valoración de los síntomas declarados o recabados en los individuos para realizar un diagnóstico de salud.
Predicción del resultado probable de un programa de intervención	Programa de intervención nutricional para determinar el impacto de la aplicación de un suplemento alimenticio.	Es la evaluación del resultado de un programa de intervención en una comunidad o de una enfermedad en los pacientes, a la luz de los síntomas, signos y circunstancias existentes.
Elección apropiada de intervención en paciente o comunidad	Evaluación de la eficacia de un fármaco y/u otros métodos de tratamiento.	Se basa en la experiencia anterior con pacientes o comunidades de análogas características que hayan tenido una intervención.
Administración sanitaria y planificación	Determinar el perfil sanitario de la población en términos de distribución de la enfermedad y la utilización de los recursos de salud.	Refiere al empleo de los datos relativos a la enfermedad en la población a fin de hacer un diagnóstico en la comunidad.
Realización y análisis en la investigación en salud pública	Contempla otorgar la validez a investigaciones analíticas o de encuestas descriptivas.	Probabilidad de cáncer de próstata en individuos con edad mayor a 60 años

1.2.3. Clasificación de la estadística

La estadística es una rama de las matemáticas que trata de reunir, organizar, analizar e interpretar datos con el propósito de solucionar los problemas cotidianos, en todos los campos, dando un soporte de racionalidad a las decisiones. El lenguaje de la estadística es la matemá-

tica; en consecuencia, se requiere del conocimiento de los fundamentos matemáticos y de los modelos de representación para comprender los alcances de sus aplicaciones.

Las decisiones en el mundo actual se toman con base en datos. La estadística se aplica en todas las áreas del conocimiento como son: la administración, las ciencias agropecuarias, las ciencias de la salud, las ciencias sociales, las ciencias biológicas y ambientales y la ingeniería en general (Flores Ruiz, Miranda Novales, 2017).

La estadística se puede clasificar en dos grandes ramas:

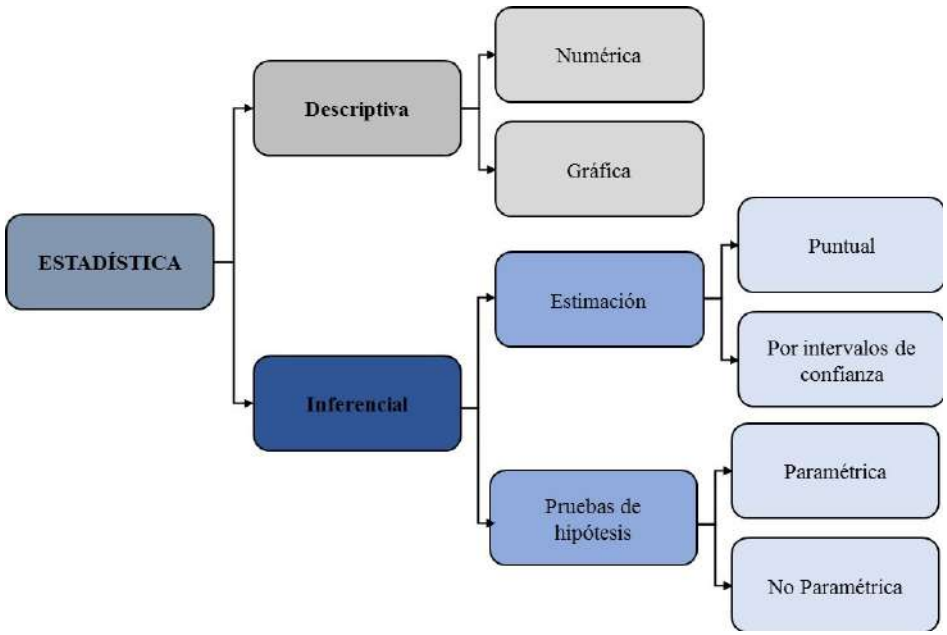


Figura 1. Clasificación de la estadística.

Estadística descriptiva. Rama de la estadística relacionada con las mediciones, la organización de los datos y presentación en indicadores (estadísticos o parámetros), tablas o figuras para describir los atributos, características o propiedades de una muestra estadística o de una población (Molina & Rodríguez, 2019).

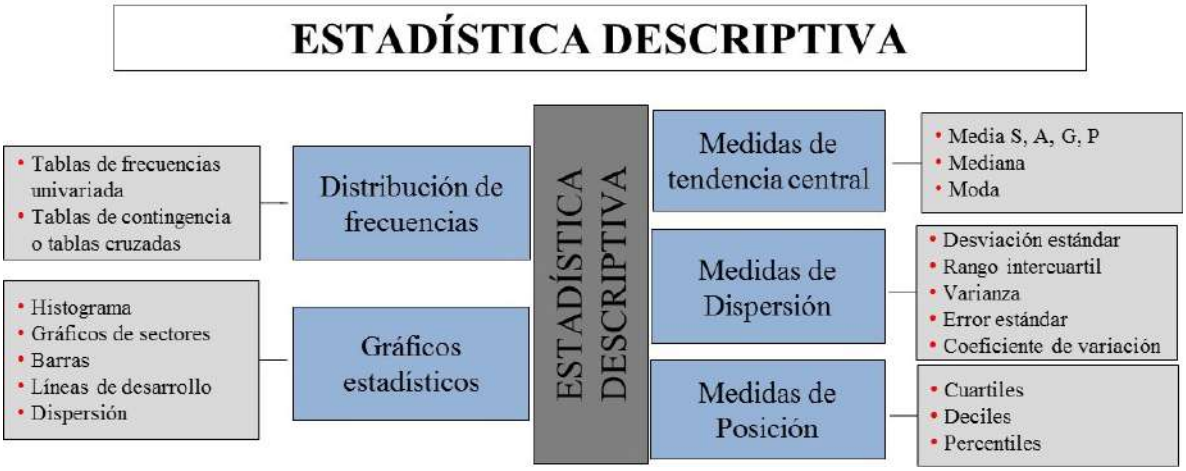


Figura 2. Distribución de la estadística descriptiva.

Estadística inferencial. La inferencia estadística se interesa en dos tipos de problemas: la estimación de los parámetros de la población y las pruebas de hipótesis. La inferencia se determina en términos de probabilidad y sirve para la toma de decisiones, asumiendo un nivel de riesgo y error.

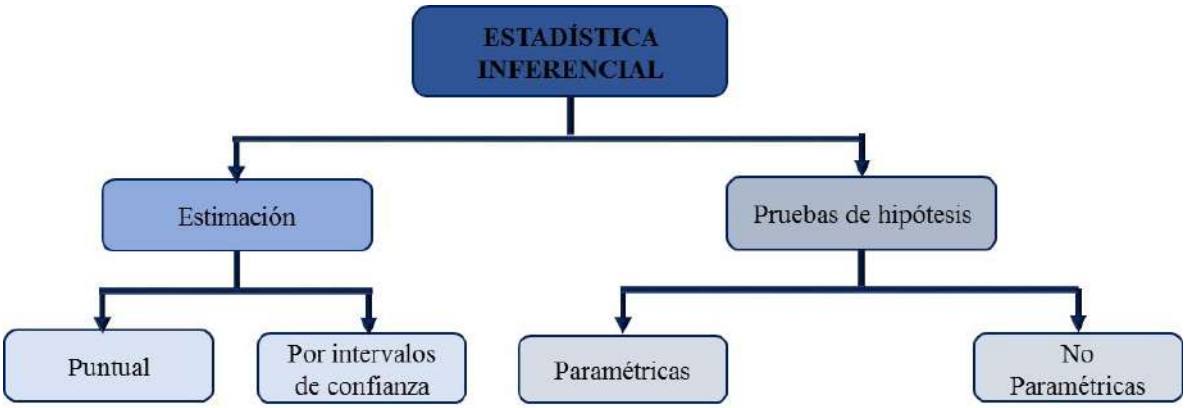


Figura 3. Distribución de la estadística inferencial.

¿Cuáles son las diferencias entre la estadística descriptiva y estadística inferencial?

La estadística descriptiva proporciona las herramientas para conocer las características o atributos de una muestra o población determinada y la estadística inferencial es presunción de lo que probablemente

pueda ocurrir en el comportamiento futuro de las muestras y poblaciones.

Si se usan las técnicas únicamente para reunir, organizar y representar gráficamente los datos de una muestra o población, entonces se trata de la estadística descriptiva; si se usan las técnicas estadísticas para estimar los parámetros de una población a partir de una muestra se está en el campo de la estadística inferencial.

1.3. Población y muestra

Figura 4. Población estadística y muestra.

1.3.1. Población

Es el conjunto de individuos u objetos de los que se desea conocer algo en una investigación. Totalidad de individuos o elementos en los cuales puede presentarse determinada característica susceptible de ser estudiada.

Llamamos población estadística, universo o colectivo al conjunto de referencia del que extraemos las observaciones, es decir, el conjunto de todas las posibles unidades experimentales. Puede estar constituida por personas, animales, registros médicos, los nacimientos, las muestras de laboratorio, los accidentes viales, entre otros, y se suele representar por la letra mayúscula **M**. El universo es el grupo de ele-

mentos al que se generalizarán los hallazgos. Por esto es importante identificar correctamente la población desde el inicio del estudio y hay que ser específicos al incluir sus elementos. Como ejemplo se puede analizar el caso de un estudio de las características de las estudiantes de la carrera de Laboratorio Clínico de la Universidad Estatal del Sur de Manabí.

Población infinita: No se conoce el tamaño y no se tiene la posibilidad de contar o construir un marco muestral (listado en el que encontramos las unidades elementales que componen la población).

Población finita: Se conoce el tamaño, a veces son tan grandes que se comportan como infinitas. Existe un marco muestral donde hallar las unidades de análisis (marcos muestrales = listas, mapas, documentos).

Población de estudio - **blanco o diana:** Población a la que queremos extrapolar los resultados.

Población accesible: Conjunto de casos que satisfacen los criterios predeterminados y que, al mismo tiempo, son accesibles para el investigador.

Población elegible: Determinada por los criterios de selección.

1.3.2. Muestra

Llamamos muestra a un subconjunto de elementos de la población que habitualmente utilizaremos para realizar un estudio estadístico. Se suelen tomar muestras cuando es difícil, imposible o costosa la observación de todos los elementos de la población estadística; es decir, su uso se debe a que frecuentemente la población es demasiado extensa para trabajar con ella. El número de elementos que componen la muestra es a lo que llamamos tamaño muestral y se suele representar por la letra minúscula **n.**

Una muestra probabilística es aquella extraída de una población de tal manera que todo miembro de esta última tenga una probabilidad conocida de estar incluido en la muestra.

En primer lugar, si lo que se busca es estudiar algo en un grupo menor que el total para luego generalizar los hallazgos al todo, esa parte que se estudia tiene que ser representativa del universo, es decir, debe poseer las características básicas del todo.

- Para que sea representativa y útil, debe reflejar las semejanzas y diferencias encontradas en la población, ejemplificar las características y tendencias de la misma.
- Una muestra representativa indica que reúne aproximadamente las características de la población que son importantes para la investigación.

Ejemplo:
Estudio de la hipertensión en personas mayores de 65 años de la ciudad de Jipijapa. Si quisiéramos conocer las características de los hipertensos en cuanto a calidad de vida, edad, sexo, presión arterial sistólica, variables que influyen en la enfermedad, difícilmente podríamos acceder a todos y cada uno de los hipertensos que existen en la ciudad (población en estudio), pero posiblemente podríamos conseguir a través de las unidades de salud pública los datos de una cantidad determinada de pacientes (por ejemplo, **n = 300 enfermos**). Nuestro objetivo no sería conocer las características de esos 300 hipertensos en concreto, pero utilizaríamos el conocimiento sobre estos 300 enfermos para obtener conclusiones sobre todos los enfermos renales (nuestra población de estudio). Este proceso es lo que se conoce como inferencia estadística.

1.4. Unidad muestral

Unidad muestral es el conjunto de elementos extraídos de la población que conforman la muestra.

Las unidades elementales (UE) y las unidades muestrales (UM) pueden no coincidir. Por ejemplo, para estudiar la obesidad en niños en edad escolar, la UE serán los niños, pero en un muestreo probabilístico primero debemos muestrear las escuelas (UM).

Criterios de selección de las unidades de análisis del estudio.

- **Criterios de inclusión:** Define las características que deberán tener los elementos en estudio.
- **Criterios de exclusión:** Definen las características cuya existencia obligue a no incluir a un caso como elemento de estudio, aun cumpliendo los criterios de inclusión (nunca entraron al estudio)
- **Criterios de eliminación:** Definen las características que, al presentarse en los individuos ya incluidos en la población, motivarán su salida del estudio (entraron al estudio y salieron).

1.5. Muestreo

Es la técnica empleada para la selección de elementos (unidades de análisis o de investigación) representativos de la población de estudio que conformarán una muestra y que será utilizada para hacer inferencias (generalización) a dicha población de estudio.

Existen multitud de mecanismos para seleccionar una muestra que sea representativa de la población, y éstos dependen principalmente de los recursos disponibles y de la naturaleza de los elementos que componen la población. Hay dos preguntas fundamentales en la selección de una muestra:

- **¿Cuántos elementos debe tener la muestra?, es decir, ¿cuál debe ser el tamaño de la misma?**
- **¿De qué forma seleccionamos esos elementos?**

En general, en la investigación se trabaja con muestras, y a pesar de que no hay garantía de su representatividad, hay una serie de ventajas que se pueden destacar:

- Permite que el estudio se realice en menor tiempo.
- Se incurre en menos gastos.
- Posibilita profundizar en el análisis de las variables.
- Permite tener mayor control de las variables a estudiar.

No obstante, dado que es una muestra, la misma no puede ser seleccionada arbitrariamente. Los estudiosos de este campo han planteado algunas consideraciones que deben tenerse presentes en el proceso de muestreo:

a. Definir en forma concreta y específica cuál es el universo a estudiar. Tal como se mencionó anteriormente, debe hacerse una delimitación cuidadosa de la población en función del problema, objetivos, hipótesis, variables y tipo de estudio, definiendo cuáles serán las unidades de observación y las unidades de muestreo en caso de que éstas no sean iguales (familias, viviendas, manzanas, estudiantes, escuelas, animales u otros). Por ejemplo, en un estudio la familia o *la vivienda puede ser la unidad de muestreo, pero el jefe de familia puede ser la unidad de observación*. Como se dijo antes, en la mayoría de los casos ambas unidades coinciden.
b. La muestra a seleccionar tiene que ser representativa de esa población para poder hacer generalizaciones válidas. Se estima que una muestra es representativa cuando reúne las características principales de la población en relación con la variable o condición particular que se estudia. Nótese que se dice "características principales", ya que a veces es casi imposible pretender que esa muestra reúna todas las características o particularidades de la población. *La representatividad de una muestra está dada por su tamaño y por la forma en que el muestreo se ha realizado* (Hernández Sampieri, 2014).

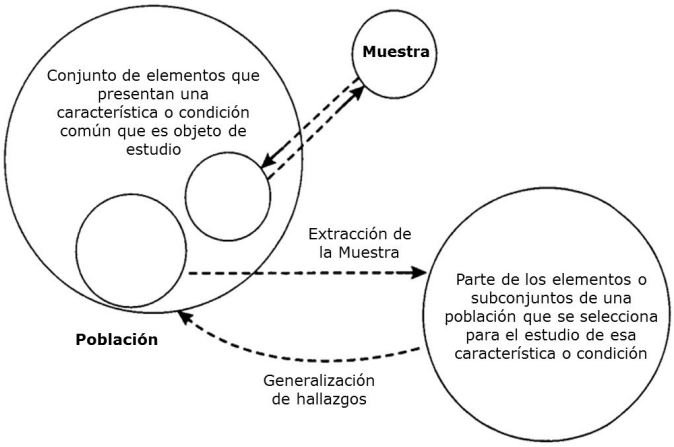


Figura 5. Concepto de universo y muestra y su relación (Pineda, Alvarado, & Canales, 1994).

1.6. Tipos de muestreo

Figura 6. Clasificación del tipo de muestreo.

1.6.1. Muestreo probabilístico

Los diseños probabilísticos son aquellos en los que se utiliza algún sistema de selección aleatoria para garantizar que cada unidad de la población tenga una probabilidad específica de ser seleccionada, por lo que toda unidad tiene una probabilidad de ser elegida y esa probabilidad es conocida de antemano.

Para que un muestreo sea aleatorio es requisito que todos y cada uno de los elementos de la población tengan la misma probabilidad de ser seleccionados. Además, esa probabilidad es conocida.

1.6.1.1. Muestreo aleatorio simple

Este método es uno de los más sencillos y tal vez el más utilizado. Se caracteriza porque cada unidad tiene la probabilidad equitativa de ser incluida en la muestra. En este tipo de muestreo hay varias modalidades. En una de ellas el procedimiento es un tipo de "sorteo" o "rifa" (por ejemplo, colocando en un recipiente fichas o tarjetas que contienen nombres o números que corresponden a cada unidad del universo); se sugiere la siguiente secuencia de acciones:

a. Identifique y defina la población.
b. Establezca el marco muestral, que consiste en la lista real de unidades o elementos de la población.
c. Determine el número que conformará la muestra.
d. Anote cada uno de los números individualmente y en secuencia en pedazos de papel o cartón hasta completar el número que compone el universo y colóquelos en un recipiente.
e. Extraiga una por una las unidades correspondientes a la muestra. Cada número indicará la unidad que formará parte de la muestra.
f. Controle periódicamente el tamaño de la muestra seleccionada, para asegurarse de que tendrá el número de unidades determinado.

Otro procedimiento es mediante el uso del programa de Microsoft Excel; se sugiere la siguiente secuencia de acciones:

Ejemplo: se requiere de un grupo de 100 fichas clínicas, seleccionar mediante el muestreo aleatorio simple 30 fichas, para su selección utilizaremos el programa Microsoft Excel siguiendo los siguientes pasos:

1. Ir a la barra de menú “Datos”.
2. Seleccionar “Análisis de Datos” (***primero se debe activar en complementos de Excel las herramientas para análisis de datos***: archivo, opciones, complementos, complementos de Excel, ir, herramientas de análisis, aceptar).
3. Se presenta ventana y seleccionar “Muestra” y aceptar.
4. Se presenta ventana y seleccionar “rango de entrada” y seleccionar la columna de listado de fichas clínicas, en “aleatorio” colocar el número de muestras a seleccionar (30), seleccionar la celda como “rango de salida” y aceptar.

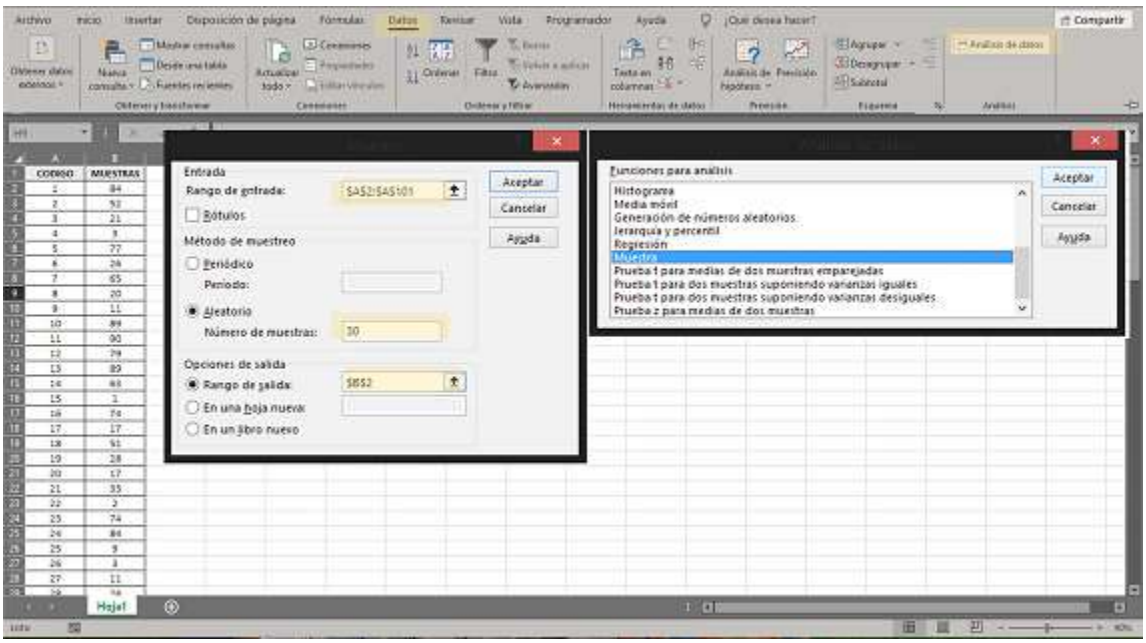


Figura 7. Muestreo en Microsoft Excel.

Como resultado podemos observar que el programa seleccionó aleatoriamente 30 fichas del listado de 100:

84	52	21	3	77	26	65	20	11	89	90	79	89	63	1
74	17	51	28	17	33	2	74	84	9	3	11	24	15	46

Siguiendo el ejemplo anterior, es posible realizar la selección de muestras mediante el uso del programa SPSS versión 25. Para ello se puede seguir la siguiente secuencia de menús y submenús:

- Datos
- Seleccionar casos
- Muestra aleatoria de casos

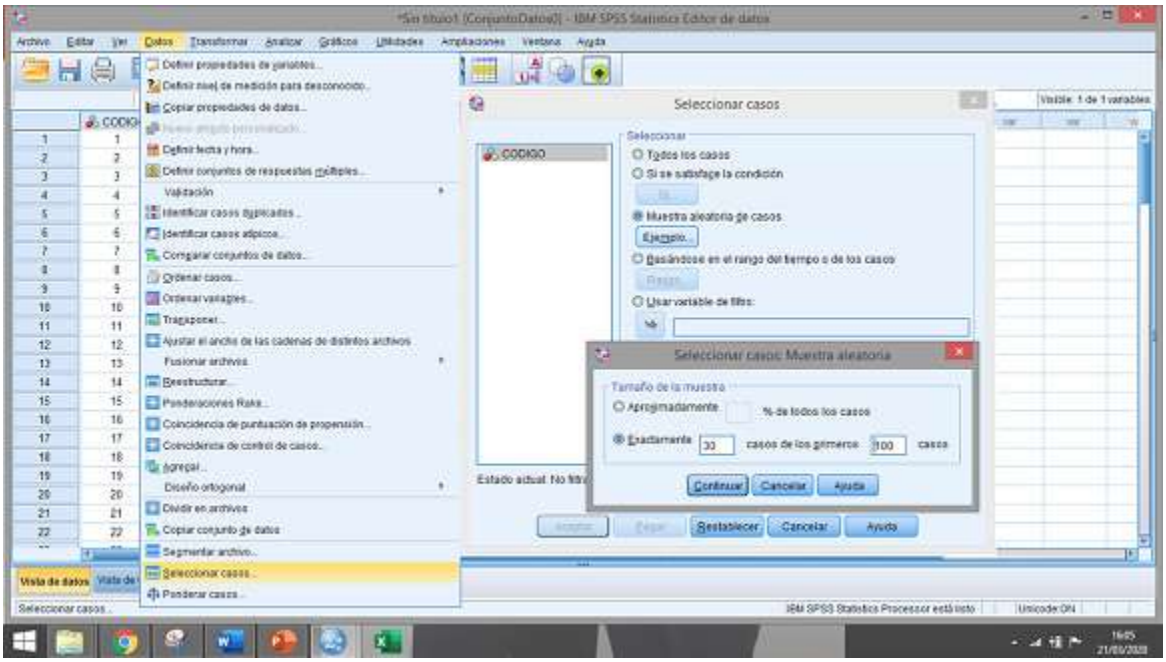


Figura 8. Muestreo en SPSS.

Una vez marcada esta última opción se debe seleccionar el botón inmediato inferior, donde se ofrecen dos opciones para elegir la muestra. La primera es seleccionando al azar un porcentaje aproximado del total de casos. En nuestro ejemplo seleccionaremos la otra posibilidad de seleccionar al azar una cantidad exacta de casos (n=30). Para ello hay que determinar cuántos van a ser los “primeros casos” a considerar

para el muestreo. Si se señala un número igual al de casos que contiene el fichero (en nuestro ejemplo son 100), implicará que la muestra se seleccionará al azar entre todos los casos. Finalmente, se selecciona copiar casos seleccionados a un nuevo conjunto de casos y aceptar.

Tras aplicar este procedimiento el programa generará una nueva base de datos de 30 fichas seleccionadas aleatoriamente (muestra1), esta base de datos está compuesta por todas las variables de investigación. En el ejemplo el programa seleccionó aleatoriamente las siguientes fichas clínicas:

1	4	5	9	10	11	12	14	18	20	21	23	24	29	31
32	33	34	43	50	60	65	66	70	72	73	79	88	90	97

Ventaja: Es el **más simple y rápido,** y además existe el software para realizarlo.

Desventaja: Se reconoce como una desventaja de este método el hecho de que no puede ser utilizado cuando el universo es grande, siendo aplicable solamente cuando la población es pequeña.

Requisito: Precisa un marco muestral o listado de todas las unidades muestrales.

1.6.1.2. Muestreo aleatorio estratificado

Un muestreo aleatorio estratificado es aquel en el que se divide la población de **N** individuos, en subpoblaciones o estratos, atendiendo a criterios que puedan ser importantes en el estudio, de tamaños respectivos.

La idea es producir grupos heterogéneos entre sí respecto de la variable de estudio, pero homogéneos dentro de cada grupo, así aseguramos la representación de cada estrato en la muestra. Las afijaciones son simple, proporcional y óptima (Díaz ***et al***., 2014).

Afijación simple: A cada estrato le corresponde igual número de elementos muestrales.

Afijación proporcional: La distribución se hace de acuerdo con el peso (tamaño) de la población en cada estrato.

Afijación óptima: Se tiene en cuenta la previsible dispersión de los resultados, de modo que se considera la proporción y la desviación típica. Tiene poca aplicación ya que no se suele conocer la desviación.

Muestreo estratificado mediante afijación proporcional

Cuando seleccionamos una característica de los individuos para definir los estratos, suele ocurrir que el tamaño de las subpoblaciones resultantes en el universo es diferente. Por ejemplo, queremos estudiar la satisfacción de atención del usuario de las unidades de salud del cantón Jipijapa y pensamos que el tipo de unidad de salud es un buen criterio para estratificar (es decir, pensamos que existen diferencias importantes en la calidad de atención según el tipo).

Definimos 3 tipos de unidades de salud: Hospital Básico Jipijapa, Centro clínico quirúrgico ambulatorio Hospital del Día IESS Jipijapa, Centro de Salud Jipijapa, estas 3 unidades no presentan igual tamaño de usuarios:

Cuadro 2. Unidades de salud del cantón Jipijapa.

Centros de salud cantón Jipijapa	Número de usuarios
Hospital básico Jipijapa	46.116
Centro clínico quirúrgico ambulatorio Hospital del Día IESS Jipijapa	43.801
Centro de Salud Jipijapa	33.233
Total	123.150

Fuente: https://www.ecuadorencifras.gob.ec/camas-y-egresos-hospitalarios/

Si usamos **mediante afijación proporcional**, la muestra deberá tener estratos que guarden las mismas proporciones observadas en la población. Si en este ejemplo queremos crear una **muestra de 380 usuarios**, los estratos tendrán que tener un tamaño como sigue:

Cuadro 3. Estratos por unidades de salud.

Centros de salud cantón Jipijapa	Número de usuarios	Proporción	Muestras
Hospital básico Jipijapa	46.116	0,37 (46.116/123.150)	142 (380*0,37)
Centro clínico quirúrgico ambulatorio Hospital del Día IESS Jipijapa	43.801	0,36 (43.801/123.150)	135 (380*0,36)
Centro de Salud Jipijapa	33.233	0,27 (33.233/123.150)	103 (380*0,27)
Total	123.150	1,00	380

Fuente: https://www.ecuadorencifras.gob.ec/camas-y-egresos-hospitalarios/

1.6.1.3. Muestreo sistemático

Este procedimiento exige, como el anterior, numerar todos los elementos de la población, pero en lugar de extraer **n números aleatorios so**lo se extrae uno. Se parte de ese número aleatorio i, que es un número elegido al azar, y los elementos que integran la muestra son los que ocupa los lugares i, i+k, i+2k, i+3k,..., i+(n-1)k, es decir, se toman los individuos de k en k, siendo k el resultado de dividir el tamaño de la población entre el tamaño de la muestra: **k = N/n**. El número i que empleamos como punto de partida será un número al azar entre 1 y k.

El riesgo este tipo de muestreo está en los casos en que se dan periodicidades en la población ya que al elegir a los miembros de la muestra con una periodicidad constante (k) podemos introducir una homogeneidad que no se da en la población. Imaginemos que estamos seleccionando una muestra sobre listas de 10 individuos en los que los 5 primeros son varones y los 5 últimos mujeres, si empleamos un muestreo aleatorio sistemático con k = **10 siempre seleccionaríamos o solo**

hombres o solo mujeres, no podría haber una representación de los dos sexos.

Por ejemplo si tenemos una población formada por 100 elementos y queremos extraer una muestra de 25 elementos, en primer lugar debemos establecer el intervalo de selección que será igual a 100/25 = **4.** A continuación, elegimos el elemento de arranque, tomando aleatoriamente un número entre el 1 y el 4, y a partir de él obtenemos los restantes elementos de la muestra: 2, 6, 10, 14,..., 98

1.6.1.4. Muestreo por conglomerados

Consiste en la identificación de conglomerados o ***clústers*** donde cada grupo presenta toda la variabilidad observada en la población, es lo opuesto al maestreo estratificado, porque los conglomerados son homogéneos entre sí, pero sus elementos son heterogéneos

En el muestreo por conglomerados la unidad muestral es un grupo de elementos de la población que forman una unidad, a la que llamamos conglomerado. Las unidades hospitalarias, las carreras universitarias, un barrio, etc., son conglomerados naturales.

El muestreo por conglomerados consiste en seleccionar aleatoriamente un cierto número de conglomerados (***el necesario para alcanzar el tamaño muestral establecido***) y en investigar después todos los elementos pertenecientes a los conglomerados elegidos (Ledesma ***et al***., 2008).

1.6.2. Muestreo no probabilístico

Cuando NO todos los elementos del universo tiene la misma probabilidad de ser parte de la muestra.

La diferencia entre muestreo no probabilístico y probabilístico es que el muestreo no probabilístico no involucra ***selección al azar*** y el muestreo probabilístico sí hace. ¿Eso significa que las muestras no probabilísti-

cas no son representativas de la población? No necesariamente. Pero sí significa que las muestras de no probabilidad no pueden depender de la lógica de la teoría de probabilidad. Al menos con una muestra probabilística, sabemos las probabilidades o la probabilidad de haber representado bien a la población. Con muestras no probabilísticas, podemos o no representar bien a la población, y a menudo nos será difícil saber qué tan bien lo hemos hecho. En general, los investigadores prefieren los métodos de muestreo probabilístico o aleatorio sobre los no probabilísticos, y los consideran más precisos y rigurosos. Sin embargo, en la investigación social aplicada puede haber circunstancias en las que no es factible, práctico o teóricamente sensato hacer un muestreo aleatorio (Dagnino, 2014).

A veces, para estudios exploratorios, el muestreo probabilístico resulta excesivamente costoso y se acude a métodos no probabilísticos, aun siendo conscientes de que no sirven para realizar generalizaciones (estimaciones inferenciales sobre la población), pues no se tiene certeza de que la muestra extraída sea representativa, ya que no todos los sujetos de la población tienen la misma probabilidad de ser elegidos.

En algunas circunstancias los métodos estadísticos y epidemiológicos permiten resolver los problemas de representatividad aun en situaciones de muestreo no probabilístico, por ejemplo, los estudios de caso-control, donde los casos no son seleccionados aleatoriamente de la población.

Entre los métodos de muestreo no probabilísticos más utilizados en investigación encontramos: muestreo por conveniencia, muestreo voluntario, muestreo de cuota y muestreo de bola de nieve.

1.6.2.1. Muestreo por conveniencia

El muestreo por conveniencia es una técnica de muestreo no probabilística donde las muestras de la población se seleccionan solo porque están convenientemente disponibles para el investigador. Estas mues-

tras se seleccionan porque son fáciles de reclutar y porque el investigador no consideró seleccionar una muestra que represente a toda la población.
Idealmente, en la investigación es bueno analizar muestras que representen a la población. Pero en algunas investigaciones, la población es demasiado grande para evaluar y considerar a toda la población.

Esta es una de las razones por las que los investigadores confían en el muestreo por conveniencia, que es la técnica de muestreo no probabilística más común, debido a su velocidad, costo-efectividad y facilidad de disponibilidad de la muestra.

Un ejemplo de muestreo por conveniencia sería utilizar a estudiantes universitarios voluntarios que sean conocidos del investigador. El investigador puede enviar la encuesta a los estudiantes y ellos en este caso actuarían como muestra.

1.6.2.2. Muestreo voluntario
Se hace un llamado a la población a participar del estudio, el informante, voluntariamente, suministra información sin ser seleccionado.

Ejemplo: Se desea conocer la opinión de los estudiantes de la carrera de Laboratorio Clínico de la Universidad Estatal del Sur de Manabí en cuanto a la calidad de la enseñanza que se les brinda. Por lo que se lanza un anuncio diciendo que cierto día habrá una reunión para aportar sus opiniones, y el que esté interesado en compartir su opinión que se acerque al lugar indicado. A dicha reunión llegaron 50 estudiantes, se les pasa una encuesta, preguntando si el personal docente está altamente capacitado para las materias que se imparten, si el plan de estudio esta actualizado o necesita actualizarse y si está satisfecho con la enseñanza brindada.

1.6.2.3. Muestreo de cuotas

En el muestreo de cuotas, se selecciona personas de forma no aleatoria de acuerdo con una cuota fija. Hay dos tipos de muestreo de cuotas: proporcional y no proporcional. En el muestreo de cuota proporcional se busca representar las principales características de la población mediante el muestreo de una cantidad proporcional de cada uno (Trochim, 2019).

Por ejemplo, si se sabe que la población tiene 40% de mujeres y 60% de hombres, y desea un tamaño de muestra total de 100, continuará tomando muestras hasta obtener esos porcentajes y luego se detendrá. Entonces, si ya tiene las 40 mujeres para su muestra, pero no los 60 hombres, continuará tomando muestras de hombres, pero incluso si las mujeres encuestadas legítimas se presentan, no se las considerará porque ya han "cumplido con su cuota".

1.6.2.4. Muestreo de bola de nieve

En el muestreo de bola de nieve, se comienza identificando a alguien que cumpla con los criterios para su inclusión en su estudio. Luego se le pide que recomienden a otros que puedan conocer y que también cumplan con los criterios. Aunque este método difícilmente conduciría a muestras representativas, hay momentos en que puede ser el mejor método disponible. El muestreo de bola de nieve es especialmente útil cuando intentas llegar a poblaciones que son inaccesibles o difíciles de encontrar (Trochim, 2019).

Se utiliza cuando la población es de difícil acceso por razones sociales (alcohólicos, drogadictos, sexoservidoras, etc.) para tener acceso se contactará con una persona del grupo que se desee estudiar y a partir de éste poco a poco se va llegando a un mayor número de individuos.

Por ejemplo, si se está estudiando a las personas sin hogar, es probable que no pueda encontrar buenas listas de personas sin hogar dentro de un área geográfica específica. Sin embargo, si va a esa

área e identifica uno o dos, es posible que sepan muy bien quiénes son las otras personas sin hogar en su vecindad y cómo puede encontrarlos.

Cuándo utilizar el muestreo no probabilístico

- Este tipo de muestreo puede ser utilizado cuando se quiere mostrar que existe un rasgo determinado en la población.
- También se puede utilizar cuando el investigador tiene como objetivo hacer un estudio cualitativo, piloto o exploratorio.
- Se puede utilizar cuando es imposible la aleatorización, y cuando la población es casi ilimitada.
- Se puede emplear cuando la investigación no tiene como objetivo generar resultados que se utilicen para hacer generalizaciones respecto de toda la población.
- También es útil cuando el investigador tiene un presupuesto, tiempo y mano de obra limitados.
- Esta técnica también se puede utilizar en un estudio inicial que será llevado a cabo nuevamente utilizando un muestreo probabilístico aleatorio.

Ventajas y limitaciones del muestreo

Ventajas

- Ahorro de tiempo.
- Resultados más confiables.
- Presentación periódica entre los eventos casuales.
- Costos manejables.
- Se puede estudiar las características de la población con mayor profundidad.
- Cuando la observación implica destrucción de los elementos.

Limitaciones

- Las investigaciones por muestreo están sujetas a un error por muestreo.
- Cuando se requiere un alto grado de desagregación.

- Cuando por la naturaleza del fenómeno se requiere investigar al total de la población.

1.7. Determinación del tamaño óptimo de una muestra

En los apartados anteriores se realizó un análisis de la forma en que seleccionará la muestra, ahora se debe conocer cuántos elementos requerimos estudiar en ella. Este conocimiento es importante, ya que, si la muestra es muy pequeña, se corre el riesgo de no detectar resultados válidos y dar por negativo un resultado por estimación inadecuada de su tamaño, como lo demostró Freiman en 1978, y si es demasiado grande, puede exponer a los sujetos de estudio a un riesgo innecesario y a desperdicio de recursos.

De acuerdo con principios matemáticos conocidos como teorema del límite central, la muestra mientras mayor sea su tamaño más se aproxima a la población, esto indicaría que mientras mayor sea la muestra mejor será el estudio, pero esto no se puede garantizar, ni afirmar. En este sentido al calcular el tamaño de la muestra se debe asegurar que la misma cumpla los objetivos, considerando aspectos importantes como la disponibilidad de tiempo, recursos humanos, recursos económicos, etc.

Recordemos que lo óptimo de una muestra depende de cuánto se aproxima su distribución a la distribución de las características de la población. Esta aproximación mejora al incrementarse el tamaño de la muestra.

Cuando las muestras están constituidas por 100 o más elementos tienden a presentar distribuciones normales y esto sirve para el propósito de hacer estadística inferencial (generalizar la muestra al universo). A lo anterior se le llama teorema del límite central (Hernández Sampieri & Mendoza Torres, 2018).

Al iniciar un proceso investigativo los investigadores siempre se plantean la misma interrogante ¿Cuántos elementos debe contener la muestra?, algunos plantean que estudiar al 10% de la población es suficiente, pero el tamaño de la muestra no puede reflejarse en función de un porcentaje considerando que, a ***menor tamaño de la población, constituye un porcentaje alto de la población***, mientras que a ***mayor tamaño de la población, el porcentaje es bajo***, a continuación unos ejemplos:

Tamaño de la población (N)	Porcentaje (%)	Tamaño de la muestra (n)
100	80%	80
500	60%	300
5.000	8%	400
50.000	0,8%	400
100.000	0,4%	400

1.7.1. Requerimientos para el cálculo del tamaño de muestra

Para calcular el tamaño de la muestra se requiere conocer previamente algunas situaciones que se enlistan a continuación:

a. El tipo de estudio.
b. La confiabilidad que se espera del estudio.
c. Estadística empleada para probar la validez de la hipótesis.

Tipo de estudio

En forma general podemos decir que se requiere calcular el tamaño de muestra en aquellos estudios que quieran conocer la frecuencia de un fenómeno (prevalencia), probar hipótesis de causalidad, los que busquen relación entre un factor de riesgo y una enfermedad, que busquen correlación entre variables, o que busquen precisar que un tratamiento sea mejor que otro. Las series de casos, por lo general no requieren del cálculo de tamaño de muestra ya que son reportes descriptivos que solamente presentan resultados en un número de casos existentes (Flores Ruiz, Miranda Novales, 2017).

La confiabilidad que se espera del estudio

Qué tanta seguridad quiero tener de que, si se repite mi estudio, los resultados que obtengan sean similares en un nivel de probabilidad aceptado (generalmente 95%, aunque esto dependerá de que tan estricto requiero ser con dicha probabilidad). Este valor está dado por $Z_{1-\beta}$ El valor Z no es otra cosa que la transformación de un valor cualquiera, independientemente de sus unidades de medida, en un valor cuya unidad de medida es la cantidad de desviaciones estándar que dicho valor se aleja de la media de la muestra estudiada.

Así, el valor Zα corresponde a la distancia de su media que tendrá el valor de probabilidad asignado a la confianza, por ejemplo, un alfa de 0,05, de acuerdo con las tablas de Z de la distribución normal, se encuentra a 1,96 desviaciones estándar del valor de la media en dicha distribución, en hipótesis de dos colas y a 1,64 desviaciones estándar en hipótesis de una cola. El siguiente cuadro presenta lo valores Z más frecuentemente utilizados.

Cuadro 4. Valor de Z según el nivel de confianza.

Nivel de confianza	α	α/2	Z α/2
0,90	0,1	0,05	1,64
0,93	0,07	0,035	1,81
0,95	0,05	0,025	1,96
0,99	0,01	0,005	2,58

Para determinar el valor de Z a distintos niveles de α/2 mediante el uso de Microsoft Excel se aplica la función.

=DISTR.NORM.ESTAND.INV(probabilidad)
=DISTR.NORM.ESTAND.INV(0,05/2), **Z= 1,96**

Estadística empleada para probar la validez de la hipótesis

El tipo de estadístico con el cual se probará la hipótesis (estadístico Z, t de Student, X, r de Pearson, etc.) depende principalmente del tipo de

estudio y del tipo de hipótesis a probar (Velasco, Martínez, Hernández, Huazano, & Nieves, 2003).

1.7.2. Fórmulas para el cálculo de la muestra en investigaciones de salud

En la investigación en salud, es muy difícil estudiar a toda la población que presenta la variable de interés, por lo que es necesario realizar un muestreo que resulte representativo de la población objetivo. El cálculo de la muestra permite responder a la pregunta del investigador de ¿cuántos individuos se deben considerar para estudiar un parámetro con un grado de confianza determinado?, o ¿cuántos individuos se deben estudiar para detectar en los resultados de los dos grupos, una diferencia que sea estadísticamente significativa?

A continuación, se presentan de una manera sencilla, las fórmulas comunes para el cálculo del tamaño de muestra y algunos ejemplos que permitan clarificar su aplicación.

a) Para una población infinita (cuando se desconoce el total de unidades de observación que la integran o la población es mayor a 10.000):

$$n = \frac{Z^2 * p^2 * q^2}{e^2}$$

Z = nivel de confianza (correspondiente con tabla de valores de Z)
p = porcentaje de la población que tiene el atributo deseado
q = porcentaje de la población que no tiene el atributo deseado = 1 - p
Nota: cuando no hay indicación de la población que posee o no el atributo, se asume 50% para p y 50% para q
e = error de estimación máximo aceptado
n = tamaño de la muestra

Ejemplo: Se desea conocer la prevalencia de obesidad en los habitantes de la parroquia urbana Parrales y Guale del cantón Jipijapa, qué tan grande debe ser una muestra si se desea tener una confianza de

al menos 93%. Realice las consideraciones necesarias para calcular n. ¿Cuál será el número adecuado de habitantes?

$$n = \frac{Z^2 * p * q}{e^2}$$

$$n = \frac{1,81^2 * 0,5 * 0,5}{0,07^2}$$

$$n = \frac{3,28 * 0,5 * 0,5}{0,0049}$$

$$n = \frac{0,82}{0,0049}$$

Z = 1,81
p = 0,5
q = 0,5
e = 0,07
n = tamaño de la muestra

n = 167 habitantes

b) Para una población finita (cuando se conoce el total de unidades de observación que la integran):

$$n = \frac{N^2 * Z^2 * p * q}{e^2 * (N - 1) + (Z^2 * p * q)}$$

Z = nivel de confianza (correspondiente con tabla de valores de Z)
p = porcentaje de la población que tiene el atributo deseado
q = porcentaje de la población que no tiene el atributo deseado = 1 - p
Nota: cuando no hay indicación de la población que posee o no el atributo, se asume 50% para p y 50% para q
e = error de estimación máximo aceptado
N = Tamaño del universo (se conoce puesto que es finito)
n = tamaño de la muestra

Ejemplo: Se tiene interés en realizar una investigación sobre la Hipertensión en adultos mayores a 65 años en la población urbana del cantón Jipijapa, según los datos del último censo de Población y Vivienda realizado por el Instituto Nacional de Estadística y Censos (INEC) en el 2010, la población > a 65 años es de 3.169 habitantes, se desea tener una confianza de al menos 95% y un margen de error del 0,05, ¿determinar el número adecuado de adultos mayores para la investigación?

$$n = \frac{N * Z^2 * p * q}{e^2 * (N - 1) + (Z^2 * p * q)}$$

$$n = \frac{3.169 * 1,96^2 * 0,5 * 0,5}{0,05^2 * (3.168) + (1,96^2 * 0,5 * 0,5)}$$

$$n = \frac{3.169 * 3,84 * 0,25}{(0,0025 * 3.168) + (3,84 * 0,25)}$$

$$n = \frac{3,044}{8,88}$$

Z = 1,96
p = 0,5
q = 0,5
e = 0,05
N = 3.169
n = tamaño de la muestra

n = 343 adultos mayores

Cabe anotar que para efectuar el cálculo de tamaño de muestra se puede también aplicar el programa Decision Analyst STATS que se puede descargar de la página web, con el subprograma tamaño de la muestra [Sample Size Determination]),el resultado es el mismo o muy similar al que proporciona dicho programa.

Al abrir el subprograma Tamaño de la muestra (Sample Size Determination) en STATS®, el programa va a pedir los siguientes datos:

- Universe size (tamaño del universo).
- Maximum Acceptable Percentage Points of Error (error máximo aceptable).
- Estimated Percentage Level (porcentaje estimado de la muestra).
- Desired Confidence Level (nivel deseado de confianza).

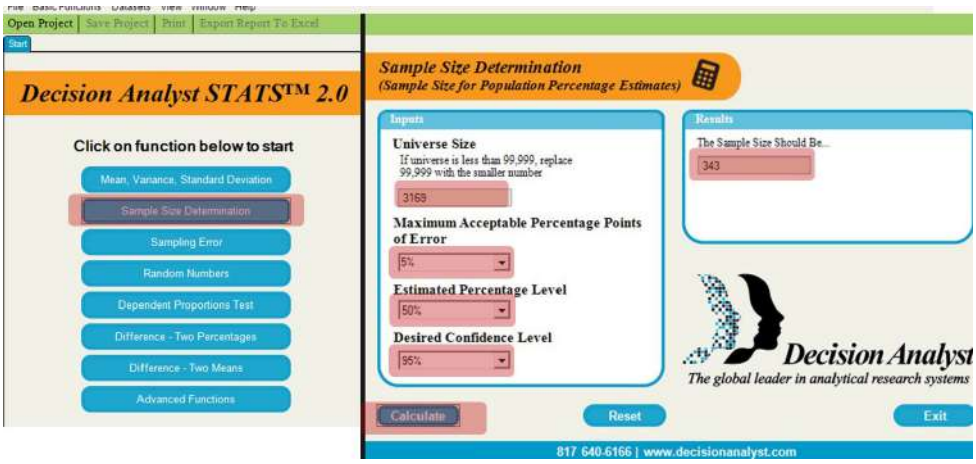


Figura 9. Programa estadístico Decision Analyst STATS.

1RA Edición

BIOESTADÍSTICA

APLICADA A INVESTIGACIONES
CIENTÍFICAS EN SALUD

CAPÍTULO II

ORGANIZACIÓN Y PRESENTACIÓN
DE DATOS ESTADÍSTICOS

MAWIL
Publicaciones Impresas
y Digitales

2.1. Organización de datos para investigación

Los datos de investigación son aquellos materiales generados o recolectados durante el transcurso de una investigación. Pueden ser hechos, observaciones o experiencias en que se basa el argumento, la teoría o la prueba.

Una base de datos es la recolección de una cantidad determinada de cúmulos de información los cuales están relacionados unos con otros, para determinar el grado de expansión de una base de datos, se debe tener conciencia de lo que se está administrando en dicha base. Los datos que están contenidos en una base de datos son los suficientes para realizar estudios estadísticos, por lo general esto se realiza con el fin de sintetizar trabajos administrativos cuando la entrada de información y de archivos de datos es constante. La organización de datos propios de un sistema debe tener un orden que facilite la rápida localización de algún dato en específico.

Los programas informáticos facilitan el manejo de esta información. Es posible adaptar un programa existente o diseñar uno nuevo, adaptado a las necesidades particulares de la investigación (hecho a la medida).

El diseño de la base de datos debe facilitar el almacenamiento sistemático de los datos. La base de datos es más que una serie de datos aislada; en ella, éstos se encuentran estructurados de forma lógica y con principios definidos. Estos datos pueden ser números o nombres organizados en filas (denominadas registros) y columnas (campos) (Díaz-Parreño ***et al***., 2014).

En este sentido, trataremos la forma de generar una base de datos en SPSS a partir de datos de estudios investigativos, creando e identificando los distintos tipos de variables y los casos recogidos. También aprenderemos a importar bases de datos de otras aplicaciones.

2.2. Datos estadísticos

Los datos estadísticos se obtienen mediante un proceso que comprende la observación y medición de conceptos, como ingresos anuales de los funcionarios del Ministerio de Salud Pública, o calificaciones de los exámenes de los estudiantes de cierto colegio, edad o estatura. Tales conceptos reciben el nombre de variables, ya que producen valores que tienden a mostrar cierto grado de variabilidad al efectuarse mediciones sucesivas.

Los datos estadísticos, en este marco, son los valores que se obtienen al llevar a cabo un estudio de tipo estadístico. Se trata del producto de la observación de aquel fenómeno que se pretende analizar. Los datos son todas aquellas unidades de información relevantes a la hora de hacer un estudio estadístico, constituyen la materia prima de trabajo de la estadística.

Además de todo lo expuesto, hay otra serie de aspectos que podemos destacar acerca de los datos estadísticos como son los siguientes:

- Tienen la particularidad de que en todo momento están sujetos a una interpretación, que es la que realiza la persona que los lleva a cabo o que los emplea. De la misma manera, también el lector de los mismos o el oyente de la presentación de aquellos podrá realizar su propia interpretación.
- Cuentan con la singularidad de que son utilizados en muchas ocasiones para persuadir o convencer. Y es que es más convincente hablar del 90% de la gente, por ejemplo, que decir “muchas personas”.

En concreto la labor de la estadística será transformar los datos que disponemos en información útil para el propósito de nuestra investigación.

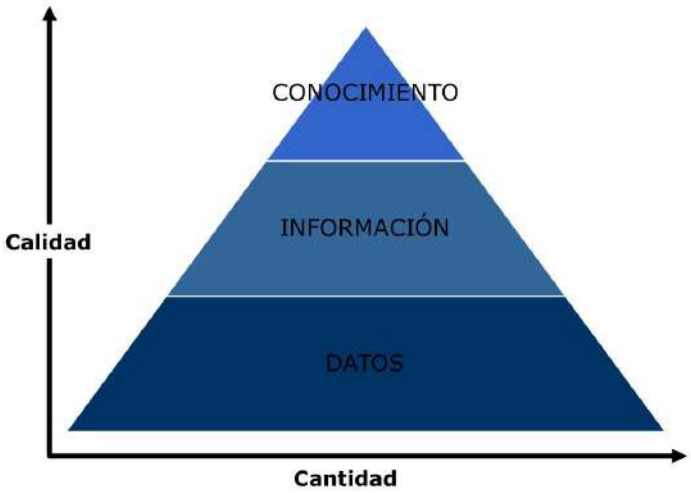


Figura 10. *Datos estadísticos y el conocimiento.*

Como se organizan los datos: variables y unidades experimentales
En estadística aplicada a ciencias de la salud, los datos disponibles se refieren a personas. Por ejemplo, en el registro de admisión de un hospital se puede tomar de cada uno de los pacientes atendidos (los pacientes serán las unidades experimentales) datos sobre su edad, sexo, motivo de ingreso, ciudad de residencia, etc.

Cada uno de estos datos de interés se recogerán como una variable distinta. La forma en la que se suelen almacenar los datos es mediante una tabla en la que la información de cada individuo se recoge en una fila, mientras que cada característica de cada variable se representa en una columna.

N°	Edad	Sexo	Nivel de glucemia (mg/dl)	Leucocitos totales (4.5 a 11.0 × 109/L)
1	39	Hombre	230	4.400
2	48	Mujer	298	3.960
3	33	Mujer	254	3.850
4	53	Mujer	119	4.070
5	60	Hombre	124	3.520
6	72	Mujer	116	5.350
7	44	Hombre	110	4.620
8	45	Mujer	136	7.260
9	56	Hombre	120	6.700
10	59	Mujer	119	3.630
11	48	Hombre	114	6.380
12	43	Mujer	160	10.340
13	36	Hombre	117	6.630
14	62	Mujer	227	5.510
15	48	Hombre	112	6.730

Figura 11. Organización de datos estadísticos.

Exactitud y precisión en la toma de datos

La calidad de la toma de decisiones estadísticas depende de la calidad de los datos y éstos dependen de la calidad de los equipos e instrumentos de medición, de la responsabilidad y profesionalismo de los investigadores y equipos de apoyo, así como de las escalas de medición y unidades de medida. La calidad de los datos se valora mediante la exactitud y la precisión de los datos.

La exactitud de los datos se refiere a qué tan cerca se encuentra el valor medido del valor real. En términos estadísticos, la exactitud está relacionada con el sesgo de una estimación. Cuanto menor es el sesgo es más exacta la estimación. Esto depende del investigador, del evaluar, del asistente o tomador de datos, quienes deben ser personas responsables y bien formadas en los procedimientos. Un evaluador de campo debe “saber lo que hace y para qué lo hace”. La exactitud de un resultado se expresa mediante el error absoluto, que es la diferencia entre el valor experimental y el valor verdadero. **La precisión** se refiere a la dispersión de un conjunto de valores obte-

nidos de mediciones repetidas. A menor dispersión mayor es la precisión. Una medida común de la variabilidad es la desviación estándar (s) de las mediciones y la precisión se puede estimar como una función de ella.

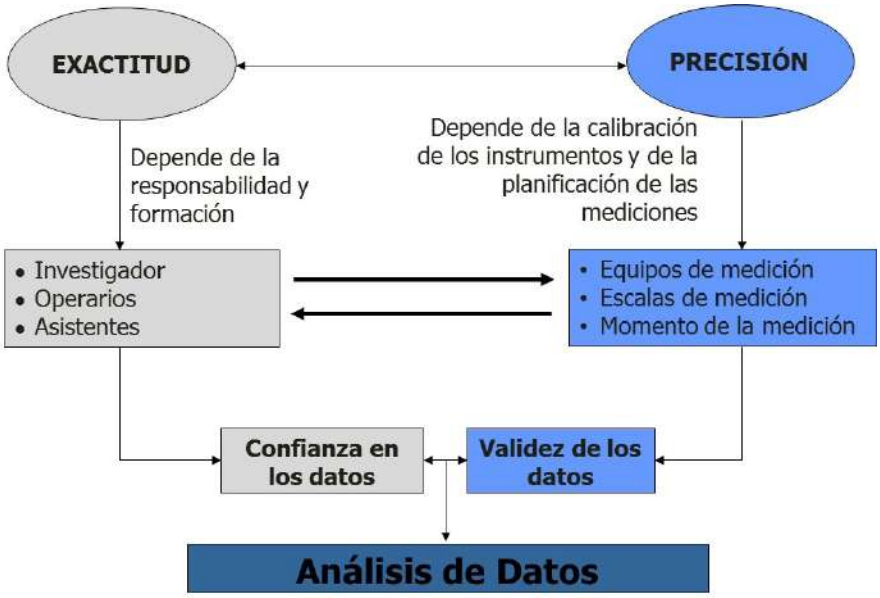


Figura 12. Exactitud y precisión de datos estadísticos.

2.3. Clasificación de variables

La información medida para cada unidad experimental, lo que llamamos variables, está sujeta a variabilidad y rodeada de incertidumbre. Si pensamos en el color de ojos de una persona, su altura, su tensión arterial, varía de un individuo a otro.

Por este hecho, nos solemos referir a las variables como **variables aleatorias**, ya que somos incapaces de predecir qué valores tomará cada individuo, al menos antes de realizar cualquier análisis estadístico.

Los datos para un estudio estadístico vienen recogidos como variables sobre unidades experimentales. Las unidades experimentales son todos aquellos individuos que albergan información sobre el objeto de interés de nuestro estudio y que por ello son incluidos en éste. En la

definición anterior entendemos el concepto de individuo de forma bastante amplia, de forma que pueden ser individuos para un estudio, bien personas o grupos de personas como, por ejemplo, unidades de salud, los trabajadores de salud, o grupos que no estén formados necesariamente por personas, como un conjunto de muestras serológicas.

Así, los datos en un estudio estadístico no son más que un conjunto de una o más variables sobre una colección de unidades experimentales.

Ejemplo: *Se desea realizar un estudio sobre hipertensión arterial en población adulta mayor.*

Las unidades experimentales serán todos aquellos adultos mayores integrantes del estudio. Las variables de nuestro estudio serán: la presión arterial de adultos mayores que es la variable de interés sobre la que queremos aprender y otras variables que deseáramos conocer si están relacionadas o no con la hipertensión como edad, sexo, consumo de calorías diarias, consumo de sal, etc.

Antes de llevar a cabo el análisis de los datos se ha de tener claro de qué tipo es cada una de las variables de que disponemos. Así, podemos clasificar las variables según el siguiente criterio:

Variables cuantitativas: Son aquellas que responden a la pregunta ¿cuánto?, y pueden ser expresadas numéricamente (es decir, siempre tomarán un valor numérico). A su vez se dividen en:

Variables continuas: Podrán tomar cualquier valor (entero o no) dentro de un rango determinado de valores.

Variables discretas: Solo podrán tomar ciertos valores concretos (habitualmente números enteros).
Variables cualitativas o categóricas: Responden a la pregunta ¿de qué tipo? Pueden tomar cualquier valor, numérico o de cualquier otro

tipo. Cada uno de los posibles valores que puede tomar estos tipos de variables se dicen categorías. Las variables cualitativas a su vez se dividen en:

Variables ordinales: Serán aquellas variables de tipo cualitativo en el que las posibles respuestas admiten una ordenación lógica.

Variables nominales: Serán aquellas variables de tipo cualitativo en el que las posibles respuestas NO admiten ningún tipo de ordenación lógica.

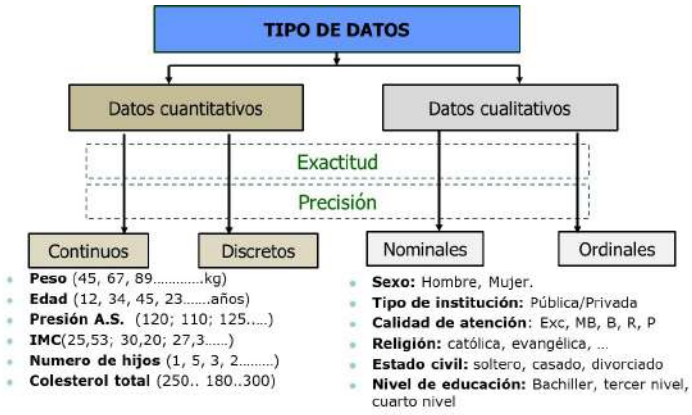


Figura 13. Tipo de datos estadísticos.

2.3.1. Medición de la variabilidad

La medición de la variación constituye el acto de registrar la información (toma de datos) de cada variable, en los individuos que conforman una muestra o población. Para realizar las mediciones se usan escalas. A cada escala de medición le corresponde un cierto conjunto de operaciones admisibles.

Las escalas de medición son: nominal, ordinal, de intervalo constante sin cero real (frecuentemente identificada como escala de intervalo) y de intervalo constante con cero real (identificada también como escala de razón o de proporción).

Escala nominal: La escala nominal permite solo clasificar los individuos u objetos de una muestra o población por sus cualidades, que no tienen una relación entre sí. Los números en la escala nominal simplemente etiquetan los datos; en realidad no son verdaderos números. Cuando los datos han sido obtenidos usando escala nominal solo caben los análisis no paramétricos.

Por ejemplo: ¿Cuáles son las 5 primeras causas de morbilidad ambulatoria de la provincia de Manabí?

a. Rinofaringitis aguda
b. Infección de vías urinarias
c. Amigdalitis aguda
d. Parasitosis intestinal
e. Diarrea y gastroenteritis de presunto origen infeccioso

https://public.tableau.com/app/profile/darwin5248/viz/Perfildemorbilidadambulatoria2016/Men?publish=yes

Escala ordinal: La medición en escala ordinal de individuos u objetos considera la relación entre categorías, estableciendo un orden o rango de mayor a menor, o viceversa.

Los datos obtenidos con escala ordinal pueden ser analizados usando las técnicas no paramétricas y, en algunos casos, apoyado en transformaciones especiales, se pueden usar las técnicas paramétricas.

Cuadro 5. Escala para medir la preocupación por el contagio de la COVID-19.

Escala	Código	Descripción
0	N	Nunca o en raras ocasiones
1	P	Poco o baja
2	MD	Moderada
3	M	Mucha
4	CS	Casi todo el tiempo

Las puntuaciones más altas indican una preocupación más frecuente por el contagio de la COVID-19.

Escala de intervalo constante sin cero real: La escala de intervalo constante sin cero real, frecuentemente identificada solo como escala de intervalo, tiene los elementos de la escala ordinal y, además, permite conocer la distancia entre dos números consecutivos de la medición, a partir de un cero arbitrario. Los datos obtenidos con escalas de intervalo pueden ser analizados usando las técnicas no paramétricas y paramétricas, con algunas restricciones.

Por ejemplo: la medición de la temperatura (grados centígrados o grados Fahrenheit (^oF = 9/5 ^oC + 32) o la medición del tiempo en horas (00HH = 24H00). Para estas mediciones se usan instrumentos de medición como son los termómetros y los relojes.

Escala de intervalo constante con cero real o de proporción: La escala de intervalo constante con cero real, también es identificada como escala de proporción o razón, tiene su origen un punto cero verdadero. Datos obtenidos con escalas de razón pueden ser analizados usando las técnicas no paramétricas y paramétricas.

Los datos son obtenidos mediante mediciones usando instrumentos (cinta métrica, balanza, tensiómetro, entre otros), de esta manera se conoce la distancia exacta entre dos números consecutivos de la medición. Estos son verdaderos números con los cuales se pueden hacer todos los cálculos y pruebas paramétricas y no paramétricas. Por ejemplo: Estatura de personas, diámetro de cráneo, peso recién nacido, miligramos (mg) / decilitro (dl).

2.4. Creación de bases de datos para investigación

Una base de datos para investigación es un conjunto de información estructurada en registros y almacenada en un soporte electrónico legible desde un ordenador. Cada registro constituye una unidad autóno-

ma de información que puede estar a su vez estructurada en diferentes campos o variables. Por ejemplo, en una base de datos de pacientes, un registro será cada uno de los pacientes. En cada registro se recogerán determinados datos, como nombre, edad, enfermedades, etc., cada una de las cuáles constituye un campo o variable. Las variables pueden ser de diferentes tipos:

Cuantitativas: Formadas por un conjunto de números. A su vez pueden ser:

Continuas: Variable que adquiere cualquier valor dentro de un intervalo especificado de valores. Por ejemplo, la presión arterial, el peso, la talla.

Discretas: Variable que presenta separaciones o interrupciones en la escala de valores que puede tomar. Por ejemplo, el número de hijos, episodios de infección urinaria.

Categóricas (cualitativas): Variable que expresa distintas cualidades, característica o modalidad. Cada modalidad que se presenta se denomina atributo o categoría, y la medición consiste en una clasificación de dichos atributos. Puede ser:

Ordinal o cuasicuantitativa: La variable toma distintos valores ordenados siguiendo un orden lógico, aunque no es necesario que el intervalo entre las mediciones sea uniforme. Por ejemplo, estratificación de los tumores, índice de gravedad de una enfermedad (leve, moderada y severa).

Nominales: Los valores de la variable no siguen un orden lógico. Pueden ser:

Dicotómicas: Cuando la variable toma dos categorías. Por ejemplo, vivo/fallecido o el sexo hombre/mujer.

Policotómicas: Cuando la variable toma más de dos categorías. Por ejemplo, los grupos sanguíneos (A, B, AB y O) o la raza.

Base de datos con SPSS

El programa SPSS «***S**tartical **P**roduct and **S**ervice **S**olutions*» es un conjunto de herramientas de tratamiento de datos para el análisis estadístico. Al igual que el resto de aplicaciones que utilizan como soporte el sistema operativo Windows el SPSS funciona mediante menús desplegables, con cuadros de diálogo que permiten hacer la mayor parte del trabajo simplemente utilizando el puntero del ratón (Díaz-Parreño ***et al***., 2014).

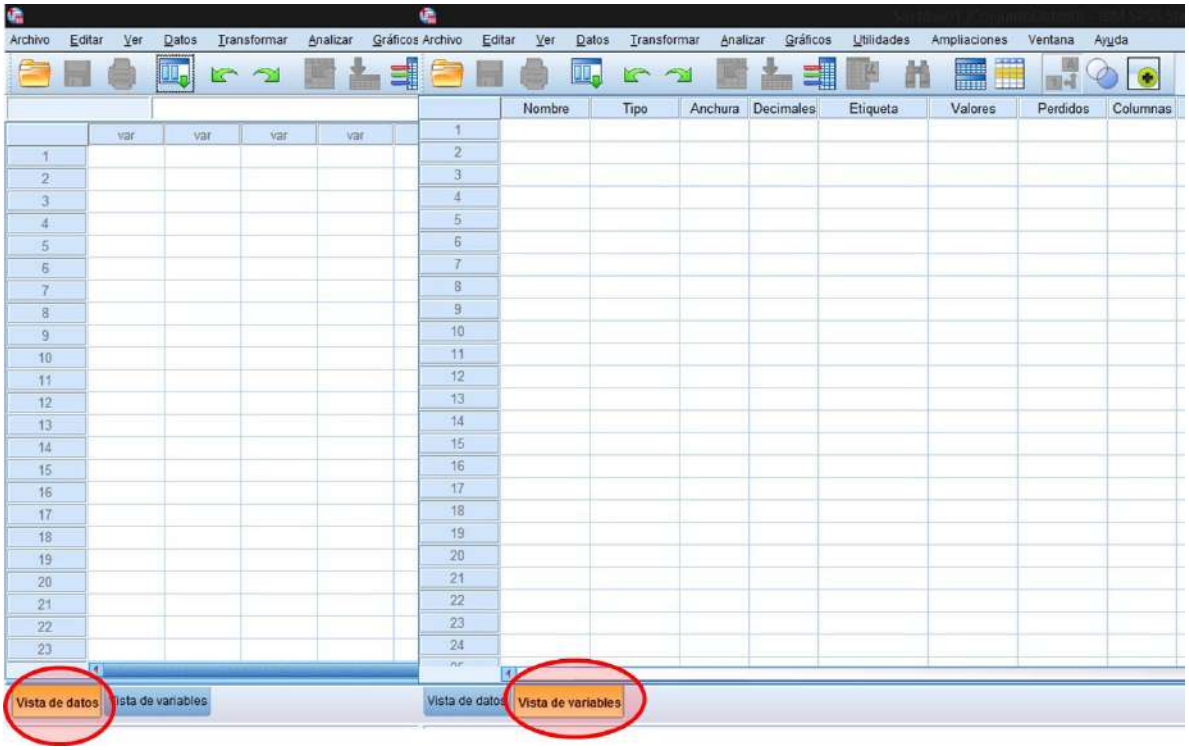


Figura 14. Vista de datos del programa SPSS.

Para cambiar de pantalla entre la vista de datos y la de variables lo único que hay que hacer es dar clic en las pestañas que se señalan en la figura 9 y que corresponden a cada vista.

En la tabla de **"Vista de datos",** las columnas contienen distribuidas cada una de las variables y en las filas cada uno de los registros. En la "Vista de variables" en las columnas aparecen distribuidas las propiedades de las variables y en las filas cada una de las variables.

Antes de introducir los registros, debemos ir a la pantalla de "Vista de variables" y definir cada una de ellas en función de sus características:

Nombre: Es el nombre que se le da a la variable, por ejemplo, edad, sexo, etc. Debe cumplir los siguientes requisitos:

- Debe ser única. No duplicados.
- No más de 64 bytes.
- Comenzar siempre por una letra y no terminar por punto.
- No utilizar espacios ni caracteres especiales (?¿, etc.)
- No utilice palabras reservadas (ALL, AND, BY, EQ, GE, GT, LE, LT, NE, NOT, OR, TO, WITH).

Tipo de variable:

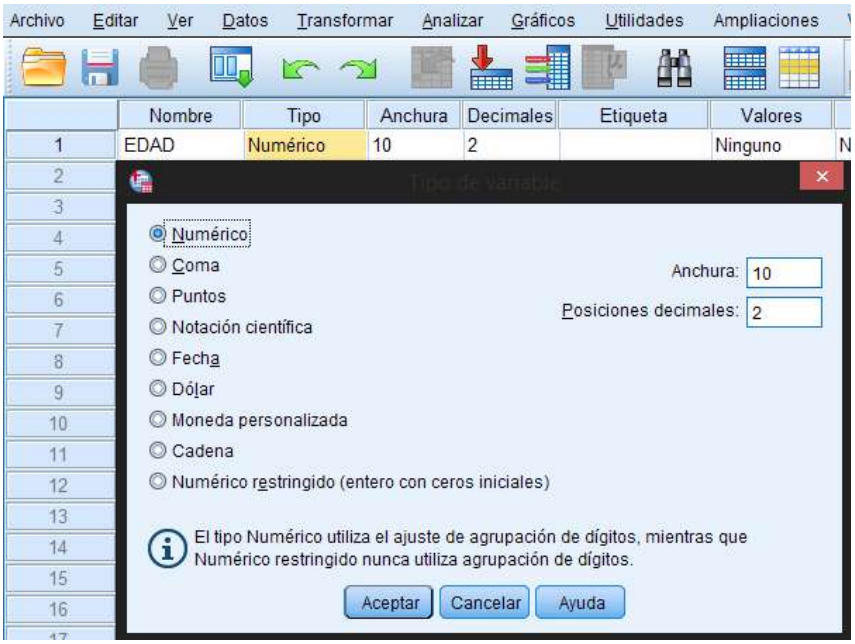


Figura 14. Tipo de variables del programa SPSS.

Numérico: Se utiliza para una variable numérica que representa magnitudes o cantidades. Asume la notación por defecto de Windows para la separación decimal (Enteros (,) Decimales) “1000,00”.

Coma y/o punto: Estos dos tipos se emplean en una variable numérica cuyos valores representan magnitudes o cantidades.

Coma: La coma se utiliza para delimitar el valor cada 3 posiciones y el punto actúa como delimitador decimal: “1,000.00”

Punto: El punto se utiliza para delimitar el valor cada 3 posiciones y la coma como delimitador decimal: “1.000,00”

Notación científica: Se utiliza en una variable numérica cuyos valores son demasiado grandes o pequeños. Así, se emplea un exponente con signo que representa una potencia en base diez. 1’000.000.00 = 1.0E+6 o 0.000001 = 1.0E(-6). SPSS nos permite representarlo de varias formas como 1000000, 1.0E6, 1.0D6, 1.0E+6, 1.0+6. La notación es útil cuando manejamos cifras extremas de lo contrario es mejor manejarlo de forma numérica

Fecha: Este tipo de variable se emplea cuando los valores de la variable representan fechas de calendario u horas de reloj; al seleccionarla aparece en el cuadro de diálogo una casilla con el listado de los diferentes formatos que el programa reconoce. Para elegir alguno de ellos basta con hacer clic sobre el formato y luego en Aceptar.

Dólar: se emplea en una variable numérica cuyos valores representan dinero en dólares. Al seleccionar este tipo de variable aparece en el cuadro de diálogo un listado de formatos monetarios. Debemos seleccionar el formato que más se acomode a los datos.

Moneda personalizada: Este tipo de variable se emplea cuando los valores de una variable representan dinero diferente al dólar (pesos,

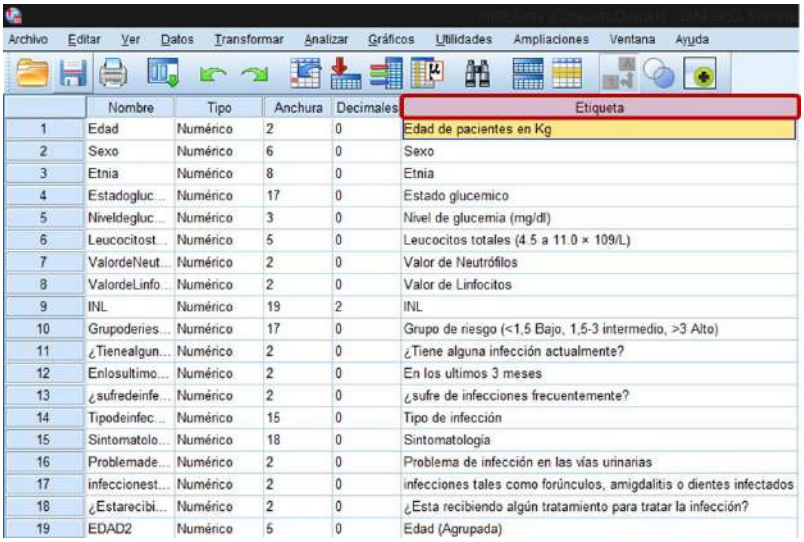
pesetas, euros, etc.); al seleccionar esta opción aparece un nuevo listado, en el cual debemos seleccionar uno de los formatos existentes.

Cadena: Se utiliza para valores no numéricos (alfanuméricos) y por lo tanto no son utilizados en los cálculos estadísticos. Pueden contener cualquier carácter y no debe exceder la longitud máxima de 255.

Anchura y decimales: Número de dígitos y decimales de la variable numérica.

Etiqueta: Significado de la variable. Se puede poner lo que se quiera hasta 255 caracteres.

El rótulo puede ser la misma pregunta o ítem en caso de que el instrumento de recolección haya sido un cuestionario, aunque también se puede utilizar una definición de lo que la variable representa. Para escribir o nombrar la etiqueta de la variable, únicamente hay que posicionarse sobre la celda correspondiente y escribir el título o rótulo que se requiera para dicha variable, por ejemplo: “Edad” o “Edad de pacientes en kg”, tal como se observa en la **figura 15.** Es muy importante cuidar que al teclear la etiqueta no se cometan errores ortográficos.



	Nombre	Tipo	Anchura	Decimales	Etiqueta
1	Edad	Numérico	2	0	Edad de pacientes en Kg
2	Sexo	Numérico	6	0	Sexo
3	Etnia	Numérico	8	0	Etnia
4	Estadogluc...	Numérico	17	0	Estado glucemico
5	Niveldegluc...	Numérico	3	0	Nivel de glucemia (mg/dl)
6	Leucocitost...	Numérico	5	0	Leucocitos totales (4.5 a 11.0 × 109/L)
7	ValordeNeut...	Numérico	2	0	Valor de Neutrófilos
8	ValordeLinfo...	Numérico	2	0	Valor de Linfocitos
9	INL	Numérico	19	2	INL
10	Grupoderies...	Numérico	17	0	Grupo de riesgo (<1,5 Bajo, 1,5-3 intermedio, >3 Alto)
11	¿Tienealgun...	Numérico	2	0	¿Tiene alguna infección actualmente?
12	Enlosultimo...	Numérico	2	0	En los ultimos 3 meses
13	¿sufredeinfe...	Numérico	2	0	¿sufre de infecciones frecuentemente?
14	Tipodeinfec...	Numérico	15	0	Tipo de infección
15	Sintomatolo...	Numérico	18	0	Sintomatología
16	Problemade...	Numérico	2	0	Problema de infección en las vías urinarias
17	infeccionest...	Numérico	2	0	infecciones tales como forúnculos, amigdalitis o dientes infectados
18	¿Estarecibi...	Numérico	2	0	¿Esta recibiendo algún tratamiento para tratar la infección?
19	EDAD2	Numérico	5	0	Edad (Agrupada)

Figura 15. Tipo de etiqueta del programa SPSS.

Valores: Valor de la variable, por ejemplo, para la variable sexo el valor "1" se puede asignar a hombres y "2" a mujeres.

Para llevar a cabo esta acción se deberá dar clic en los tres puntos suspensivos que aparecen a la derecha de la celda correspondiente y en donde se editarán los valores. Al realizar esto, automáticamente se abrirá un cuadro de diálogo en el cual se definirán los valores que corresponderán a cada categoría. Para hacerlo se tendrá que establecer un número en la caja correspondiente a valor, que tendrá el significado que el investigador defina y que se deberá establecer en el espacio correspondiente a etiqueta. Posteriormente, se tendrá que dar clic en el botón **Añadir**. Con esta acción, el valor y su significado pasarán a una nueva caja, la cual agrupará cada valor establecido y su correspondiente significado. Estos pasos se deberán repetir tantas veces como posibles categorías u opciones de valor y significado tenga la pregunta, ítem o variable. Finalmente, después de haber definido todos los valores de las categorías se deberá dar clic en el botón **Aceptar** para que el programa guarde la información vertida, en caso de no realizar este paso el SPSS no respetará los cambios realizados en el cuadro de etiquetas de valor.

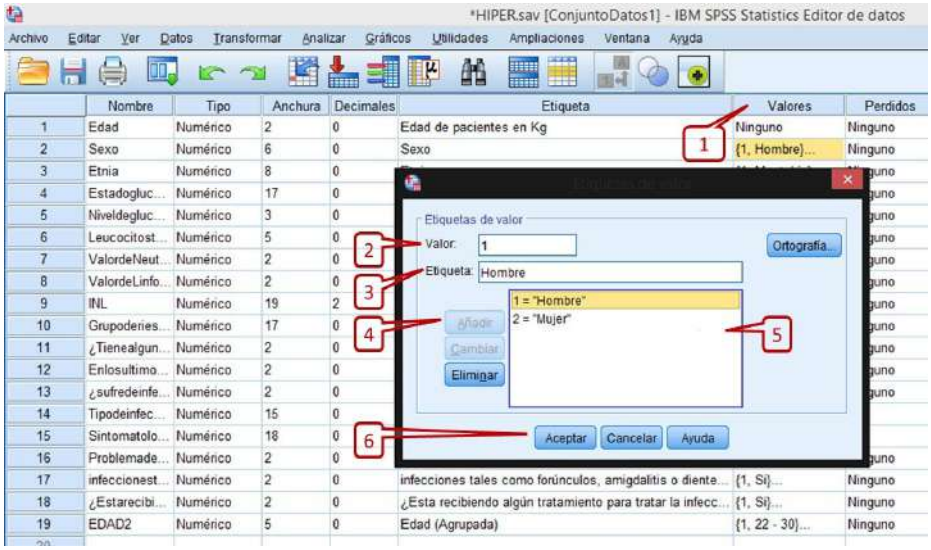


Figura 16. Configuración de valores del programa SPSS.

Perdidos: Se pueden definir los valores perdidos. Normalmente si están vacíos no se incluyen en el análisis. Al aplicar un instrumento de medición (como por ejemplo, un cuestionario en una encuesta se consulta sobre la etnia), suelen presentarse valores perdidos o no válidos, los cuales se pudieran deber a que los participantes no hubieran contestado ciertas preguntas (5 = "No responde") por desconocimiento o a que hubieran respondido de manera equivocada. En estos casos, si no se filtra la información obtenida, el análisis de los datos pudiera llegar a proporcionar resultados inexactos o confusos.

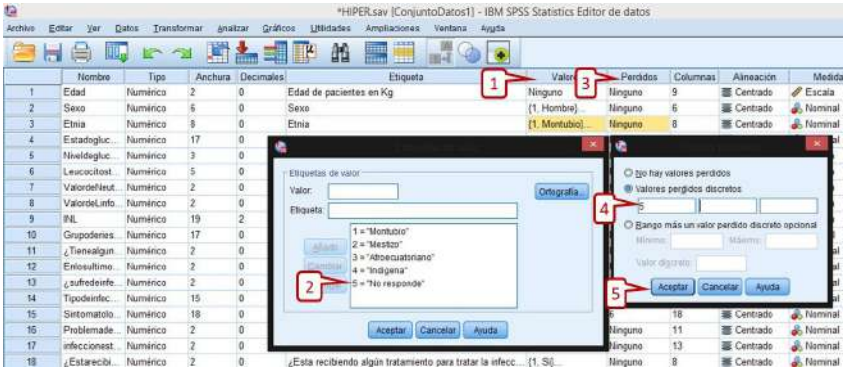


Figura 17. Configuración de valores perdidos del programa SPSS.

Columnas y alineación: Definir el ancho de columna y la alineación
Medida:

- *Escala:* Variables cuantitativas continuas. Por ejemplo, edad, altura, etc.
- *Nominal:* Variable numérica que indica una categoría de pertenencia sin orden lógico. Por ejemplo, raza, género, estado civil,…
- *Ordinal:* Variable numérica que indica una categoría de pertenencia con un orden lógico. Ejemplo: nivel de ingresos, etc.

Las variables se usan para representar los datos que se hayan recopilado, un ejemplo muy
común es el de las encuestas. En este caso, cada una de las preguntas o ítems del cuestionario equivaldría a una variable.

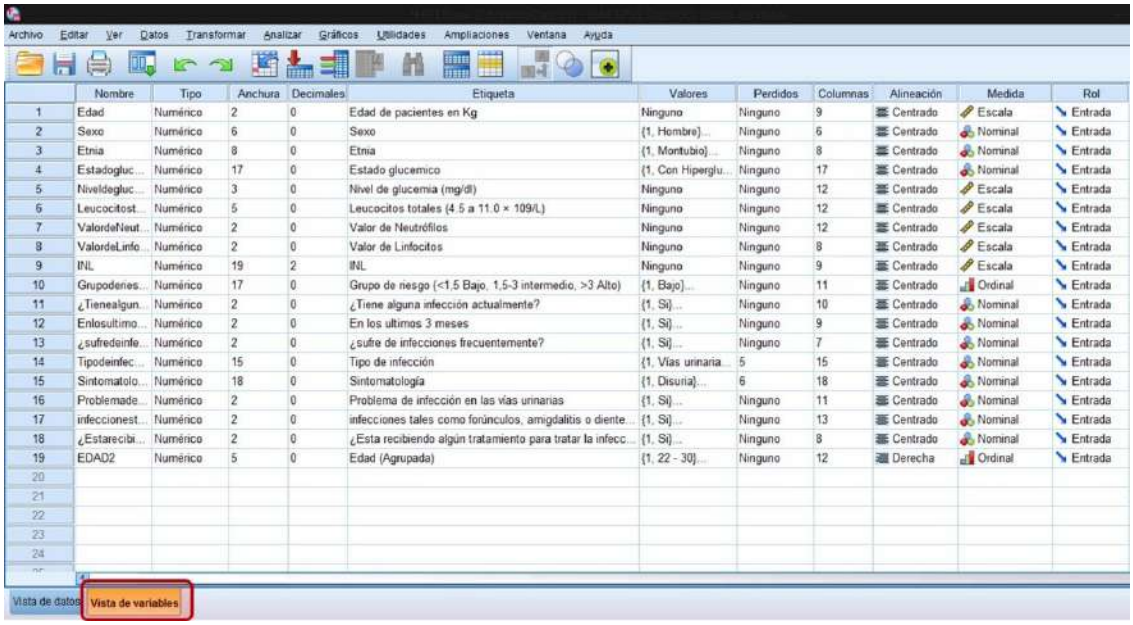
Llenado de datos en la matriz en la pestaña "Vista de los datos"

El SPSS ofrece la oportunidad de vaciar los datos obtenidos en la etapa de recolección mediante tres maneras: directo en SPSS, copiando la(s) base(s) de dato(s) de Excel o importando las hojas completas del mismo programa de Microsoft Excel.

Captura de los datos directamente del SPSS

Para capturar directamente los datos que se consiguieron en la etapa de recolección, simplemente habrá que ir llenando las casillas con los códigos que se asignaron a las opciones de respuesta, en cada variable o ítem y por cada caso.

Aunque el SPSS permite utilizar texto, lo más recomendable es utilizar códigos numéricos que permitan aprovechar al máximo las ventajas del programa, incluso en las preguntas con nivel de medición nominal.



	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
1	Edad	Numérico	2	0	Edad de pacientes en Kg	Ninguno	Ninguno	9	Centrado	Escala	Entrada
2	Sexo	Numérico	6	0	Sexo	{1, Hombre}...	Ninguno	6	Centrado	Nominal	Entrada
3	Etnia	Numérico	8	0	Etnia	{1, Montubio}...	Ninguno	8	Centrado	Nominal	Entrada
4	Estadogluc...	Numérico	17	0	Estado glucemico	{1, Con Hiperglu...	Ninguno	17	Centrado	Nominal	Entrada
5	Niveldegluc...	Numérico	3	0	Nivel de glucemia (mg/dl)	Ninguno	Ninguno	12	Centrado	Escala	Entrada
6	Leucocitost...	Numérico	5	0	Leucocitos totales (4.5 a 11.0 × 109/L)	Ninguno	Ninguno	12	Centrado	Escala	Entrada
7	ValordeNeut...	Numérico	2	0	Valor de Neutrófilos	Ninguno	Ninguno	12	Centrado	Escala	Entrada
8	ValordeLinfo...	Numérico	2	0	Valor de Linfocitos	Ninguno	Ninguno	8	Centrado	Escala	Entrada
9	INL	Numérico	19	2	INL	Ninguno	Ninguno	9	Centrado	Escala	Entrada
10	Grupoderies...	Numérico	17	0	Grupo de riesgo (<1.5 Bajo, 1.5-3 intermedio, >3 Alto)	{1, Bajo}...	Ninguno	11	Centrado	Ordinal	Entrada
11	¿Tienealgun...	Numérico	2	0	¿Tiene alguna infección actualmente?	{1, Si}...	Ninguno	10	Centrado	Nominal	Entrada
12	Enlosultimo...	Numérico	2	0	En los ultimos 3 meses	{1, Si}...	Ninguno	9	Centrado	Nominal	Entrada
13	¿sufredeinfe...	Numérico	2	0	¿sufre de infecciones frecuentemente?	{1, Si}...	Ninguno	7	Centrado	Nominal	Entrada
14	Tipodeinfec...	Numérico	15	0	Tipo de infección	{1, Vías urinaria...	5	15	Centrado	Nominal	Entrada
15	Sintomatolo...	Numérico	18	0	Sintomatología	{1, Disuria}...	6	18	Centrado	Nominal	Entrada
16	Problemade...	Numérico	2	0	Problema de infección en las vías urinarias	{1, Si}...	Ninguno	11	Centrado	Nominal	Entrada
17	infeccionest...	Numérico	2	0	infecciones tales como forúnculos, amigdalitis o diente...	{1, Si}...	Ninguno	13	Centrado	Nominal	Entrada
18	¿Estarecibi...	Numérico	2	0	¿Esta recibiendo algún tratamiento para tratar la infecc...	{1, Si}...	Ninguno	8	Centrado	Nominal	Entrada
19	EDAD2	Numérico	5	0	Edad (Agrupada)	{1, 22 - 30}...	Ninguno	12	Derecha	Ordinal	Entrada
20											
21											
22											
23											
24											

Figura 18. Vista de variables de una base de datos del programa SPSS.

Copiar desde una hoja de datos de Excel

Una opción bastante útil, sobre todo cuando varias personas participan en la investigación como encuestadores-codificadores y no se

cuenta más que con una licencia del SPSS. Cada uno de ellos vacía los datos codificados en una matriz elaborada en el programa Excel. Dicha matriz deberá ser idéntica a la generada en el SPSS y seguir la misma lógica de la vista de datos de este programa (las columnas son las variables o ítems del instrumento de recolección y las filas los casos u observaciones conseguidas).

Para llevar los datos vaciados en la matriz de Excel al SPSS, se deberán **seleccionar las celdas que nos interesan de la matriz de Excel ordenando al programa copiar las mismas** Después de copiar las celdas desde Excel, habrá que pegarlas en la Vista de datos del SPSS cuidando que al hacerlo las casillas copiadas correspondan a las celdas donde se están pegando (recordemos que la matriz de Excel debe ser idéntica a la matriz del SPSS). Para esto se sugiere **posicionarse en la primera celda del primer caso** y ahí dar la instrucción de **pegar con nombre de variables** , tal como se observa en **la figura 19.**

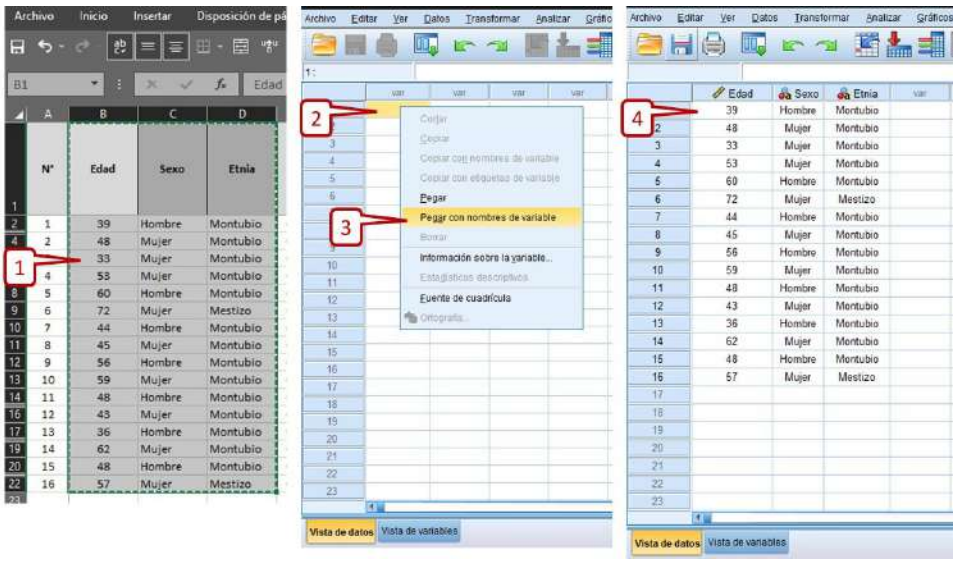


Figura 19. Copiar variables y datos desde una hoja de datos de Excel al programa SPSS.

Importar bases de datos de formato Excel a SPSS

Con el programa SPSS, podemos importar también una base de datos desde otros programas estadísticos, por ejemplo, desde Excel.

Para ello, una vez abierto el programa SPSS, hacer clic en abrir documento de datos o bien hacer clic en archivo y en abrir datos. En la ventana **"Abrir datos"**, señalar Excel (*.xls, *.xlsx, *.xlsm) en el apartado de **"Archivos de tipo"**, elegir nuestro archivo Excel y hacer clic en abrir.

Figura 20. Buscar archivo de formato Excel.

Posteriormente aparecerá una ventana, "Apertura de origen de datos de Excel" donde debemos señalar la pestaña "Leer nombre de variables de la primera fila de datos" si la primera fila de datos de nuestra hoja Excel contiene las variables. También debemos indicar dónde están situados nuestros datos, si lo están en la hoja 1, 2, etc. de Excel.

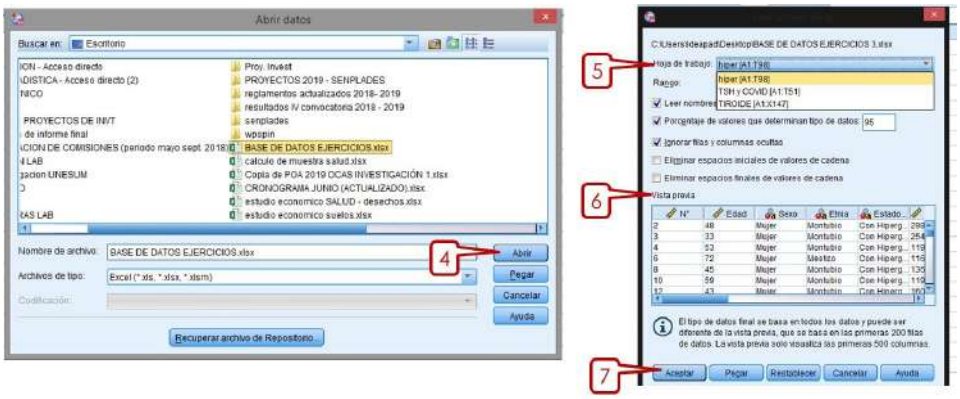


Figura 21. Abrir base de datos de formato Excel al programa SPSS.

Podemos señalar, si lo deseamos el rango de datos que queremos incluir. Finalmente hacer clic en Aceptar y debe aparecer la base de SPSS con nuestros datos de la hoja de Excel.

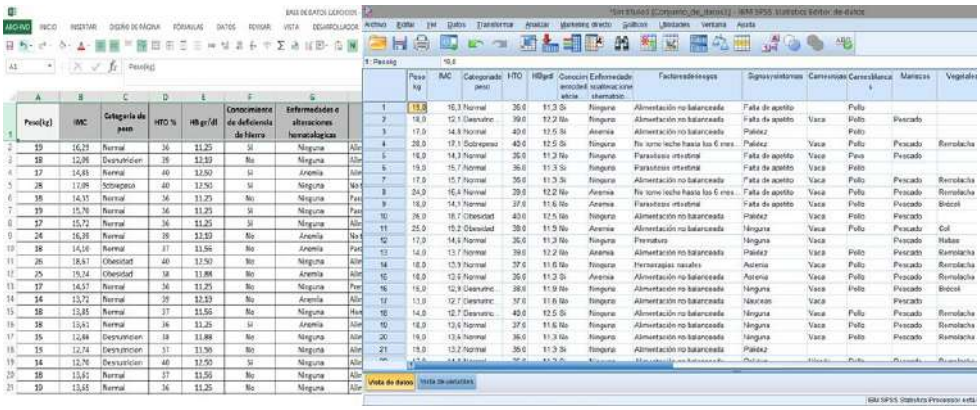


Figura 22. Importar bases de datos de formato Excel al programa SPSS.

Es importante que comprobemos siempre que las variables se hayan transmitido correctamente, pues a veces el SPSS puede cambiar el nombre o el tipo si no están correctamente definidas. Para evitar esto es muy importante que las variables en la hoja de Excel cumplan los criterios de las variables de la base de datos de SPSS. Por ejemplo, un error muy común es en la transferencia de las variables con decimales. El SPSS utiliza por defecto la "coma" para los decimales y si tenemos definido en Excel "punto", los datos se van a transferir como "cadena", es decir, como si fuera texto y no con los valores numéricos.

Vista de resultados

Como ya lo habíamos comentado, el SPSS tiene una tercera vista, denominada de resultados.

En ésta se pueden visualizar los productos de los análisis realizados con la ayuda del programa, además de otra información relevante.

La vista de resultados cuenta con un explorador que permite dirigirse rápidamente a cualquier elemento de la misma. El explorador trabaja de forma similar al de Windows, pues permite abrir directorios y subdirectorios así como la fácil navegación entre resultados contrayendo, desplegando, borrando o cambiando de lugar los titulares **(figura 23)**

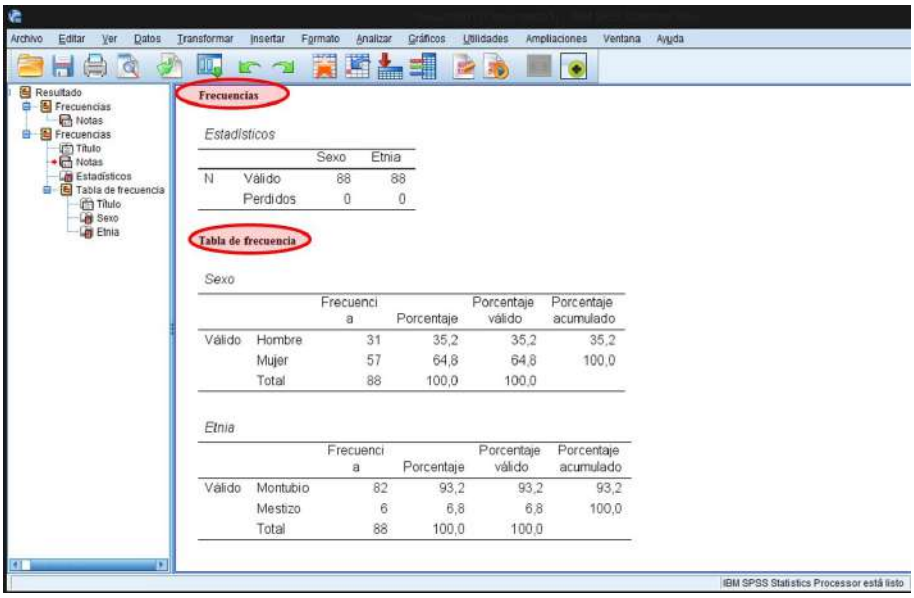


Figura 23. Entorno de vista de resultados del programa SPSS.

2.5. Modificar una base de datos

¿Cómo transformar una variable con el programa SPSS?

En determinadas ocasiones vamos a necesitar crear una nueva variable a partir de otra según nos interese. Esta nueva variable puede surgir de una operación matemática.

Con la opción “Transformar” del programa SPSS se pueden realizar diferentes acciones. Las más interesantes son “Calcular variable” y “Recodificar variable”. Con “Calcular variable” se puede transformar la variable en otra diferente a través de un cálculo matemático.

Ejemplo. Calcular el índice de masa corporal (IMC) de 12 pacientes que registran su peso (kg) y su estatura (m). El IMC es igual al peso en kg /altura $(m)^2$

Pacientes	1	2	3	4	5	6	7	8	9	10	11	12
Peso (kg)	65	43	66	89	45	78	73	88	60	65	63	70
Estatura (m)	1,56	1,78	1,45	1,6	1,62	1,7	1,65	1,66	1,55	1,67	1,6	1,78

Para ello, hacer clic en **Transformar → Calcular variable.** En la ventana **"Calcular variable",** poner en **"Variable destino"** el nombre de la nueva variable, p. ej. IMC y en expresión numérica introducir la variable que queremos transformar. Con la calculadora que aparece en la misma ventana podemos **desarrollar la fórmula**. En nuestro ejemplo, para calcular el IMC debemos **introducir las variables peso y altura** de la siguiente manera: peso / (altura*altura) y hacer clic **en aceptar.** La nueva variable aparecerá en la última columna de la "Vista de datos".

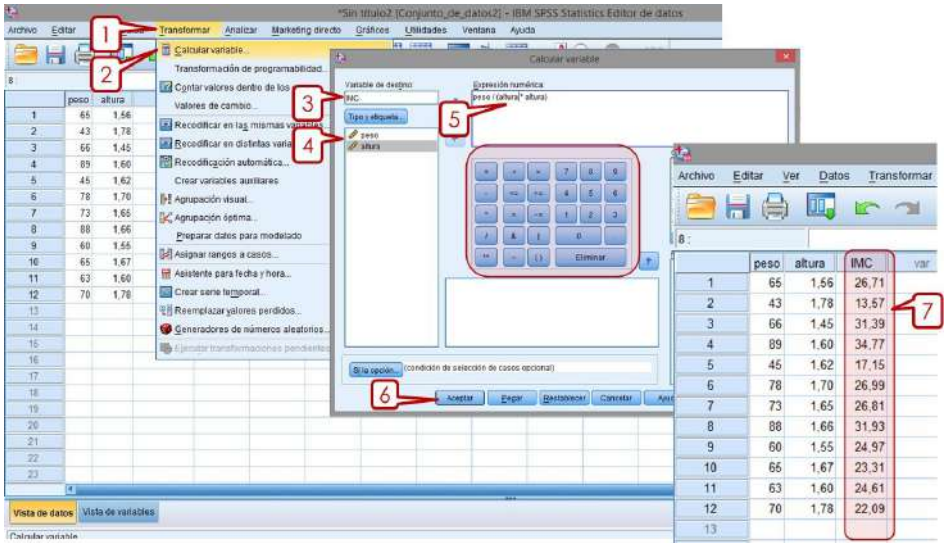


Figura 24. Calcular una variable con el programa SPSS.

2.6. Recodificación de valores de variables en una nueva variable

El diálogo recodificar en distintas variables proporciona opciones para reasignar los valores de variables existentes o para contraer los rangos de valores existentes en nuevos valores para una nueva variable.

Cuadro 6. Tabla del estado corporal según IMC.

IMC	Estado
< 18,5	Bajo peso
18,5 - 24,99	Peso normal
25,0 - 29,99	Pre-obesidad o sobrepeso
30,0 - 34,99	Obesidad clase I
35,0 - 39,99	Obesidad clase III
> 40	Obesidad clase III

Fuente: Estado corporal según la OMS (2020).

Por ejemplo, podría agrupar el IMC una nueva variable que contenga categorías de rangos del estado nutricional.

Hacer clic en Transformar → en la ventana “Recodificar en distintas variables: Valores antiguos y nuevos”, introducir en “Rango, INFERIOR, hasta valor:” el valor inferior, en este ejemplo poner 18,5 (será 18,5). En “Valor nuevo” dar un valor, en este caso se le da el valor “1”. Luego hacer clic en Añadir en el recuadro “Antiguo→Nuevo:”. Ahora, en “Rango, valor hasta añadir el valor 18,5 y en hasta 24,9 y asignar en “Valor nuevo” el valor “2” y hacer clic en Añadir en el recuadro “Antiguo→Nuevo” y así con todos los rangos (25,0 hasta 29,9), (30,0 hasta 34,9), (35,0 hasta 39,9), finalmente rango valor hasta SUPERIOR (40) y asignar en “Valor nuevo” el valor “6”. Por último, hacer clic en continuar y en la última columna de la hoja de “Vista de datos” aparecerá esta nueva variable, ESTADO NUTRICIONAL.

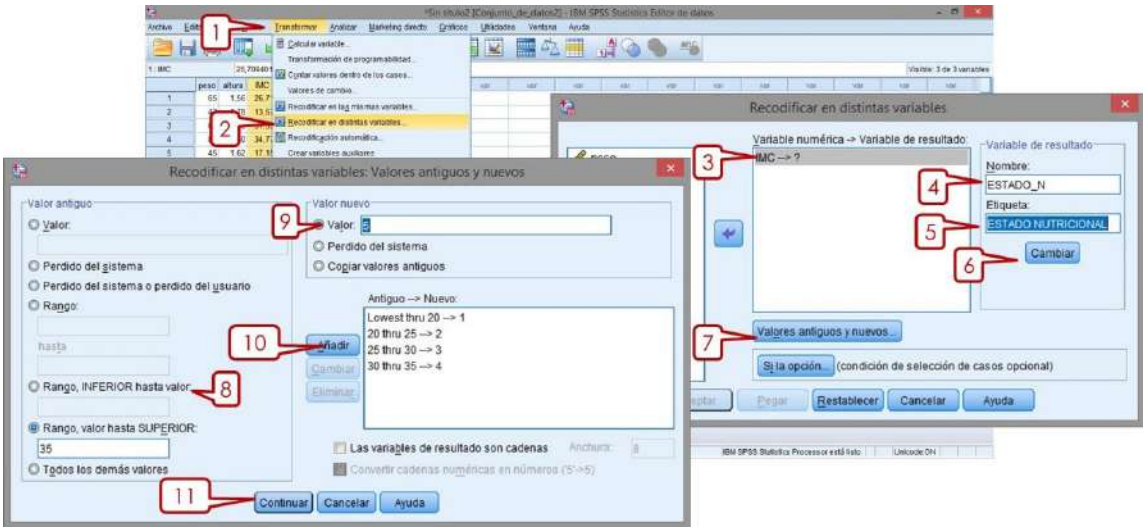


Figura 25. Proceso de recodificación en distintas variables.

2.7. Datos atípicos u outliers

Se denominan casos atípicos u ***outliers*** a aquellas observaciones con características diferentes de las demás. Este tipo de casos no pueden ser caracterizados categóricamente como benéficos o problemáticos, sino que deben ser contemplados en el contexto del análisis y debe evaluarse el tipo de información que pueden proporcionar.

Su principal problema radica en que son elementos que pueden no ser representativos de la población pudiendo distorsionar seriamente el comportamiento de los contrastes estadísticos. Por otra parte, aunque diferentes a la mayor parte de la muestra, pueden ser indicativos de las características de un segmento válido de la población y, por consiguiente, una señal de la falta de representatividad de la muestra.

Gran parte del éxito del análisis estadístico de datos subyace en la recogida de la información u obtención del conjunto de datos; no obstante, por mucho cuidado que se tenga no se estará libre de errores de muestreo y de valores anómalos (valores atípicos, discrepantes, inusitados, extraños, ***outliers***, entre otras denominaciones). Estos valores se encuentran alejados del comportamiento general del

resto del conjunto de datos y no pueden ser considerados totalmente como una manifestación del proceso bajo estudio. Los valores anómalos pueden generar resultados erróneos producto del análisis estadístico y, en consecuencia, es improbable obtener respuestas precisas que permitan caracterizar el proceso en estudio; en razón de ello, es fundamental detectar estos valores, en el conjunto de datos, ya sea para eliminarlos o para atenuar sus efectos en el análisis.

Los casos atípicos pueden clasificarse en 4 categorías:

a. *Casos atípicos que surgen de un error de procedimiento,* tales como la entrada de datos o un error de codificación. Estos casos atípicos deberían subsanarse en el filtrado de los datos, y si no se puede, deberían eliminarse del análisis o recodificarse como datos ausentes.
b. *Observación que ocurre como consecuencia de un acontecimiento extraordinario*. En este caso, el *outlier* no representa ningún segmento válido de la población y puede ser eliminado del análisis.
c. *Observaciones cuyos valores caen dentro del rango de las variables observadas pero que son únicas en la combinación de los valores de dichas variables.* Estas observaciones deberían ser retenidas en el análisis, pero estudiando qué influencia ejercen en los procesos de estimación de los modelos considerados.
d. *Datos extraordinarios para los que el investigador no tiene explicación.* En estos casos lo mejor que se puede hacer es replicar el análisis con y sin dichas observaciones con el fin de analizar su influencia sobre los resultados. Si dichas observaciones son influyentes el analista debería reportarlo en sus conclusiones y debería averiguar el porqué de dichas observaciones.

Identificación de outliers

Existe una gran controversia en la literatura en relación con la eliminación de los valores atípicos. Se ha planteado que se debe conocer

su causa y la influencia que pueden tener en los resultados de los experimentos, antes de tomar la decisión de eliminarlos o incluirlos en el análisis, porque cambian las inferencias que se obtienen y, en ocasiones, su eliminación puede conducir a la pérdida de una información importante.

¿Cómo apartar los datos atípicos?

Se debe examinar la distribución de observaciones para cada variable, seleccionando como casos atípicos aquellos casos cuyos valores caigan fuera de los rangos de la distribución. La cuestión principal consiste en el establecimiento de un umbral para la designación de caso atípico.

Hay dos herramientas eficaces para determinar datos atípicos o erróneos: el dispersograma y el análisis de los límites superior e inferior dados por la "media ± 3 desviaciones estándar". Un dato no confiable, generalmente corresponde a un valor extremo, conocido como "outliers" que de ser usado en el análisis afecta la validez de los resultados.

Por ejemplo: si en una serie de 10 datos hay valores que fluctúan entre 20 y 35, pero hay un dato de 55, éste debería considerarse, en principio "atípico".

Datos	1	2	3	4	5	6	7	8	9	10	$\overline{X}$	S	$\overline{X}$ +3S	$\overline{X}$ -3S
Con dato "atípico"	23	20	35	33	55	30	29	26	28	32	31,1	9,6	59,8	2,4
Sin dato "atípico"	23	20	35	33		30	29	26	28	32	28,4	4,8	42,9	14

En el ejemplo presentado el análisis de los límites superior e inferior dados por la "media ± 3 desviaciones estándar" ($\overline{x}$ ± 3S), los datos atípicos son aquellos casos cuyos valores caigan fuera de los rangos de la distribución 14 y 42,9.

La mayoría de los métodos univariantes más antiguos para la detección de valores atípicos se basan en el supuesto de una distribución subyacente conocida de los datos, que se supone que se distribuye de

forma idéntica e independiente. El problema con los criterios anteriores es que supone una distribución normal de los datos, algo que frecuentemente no ocurre.

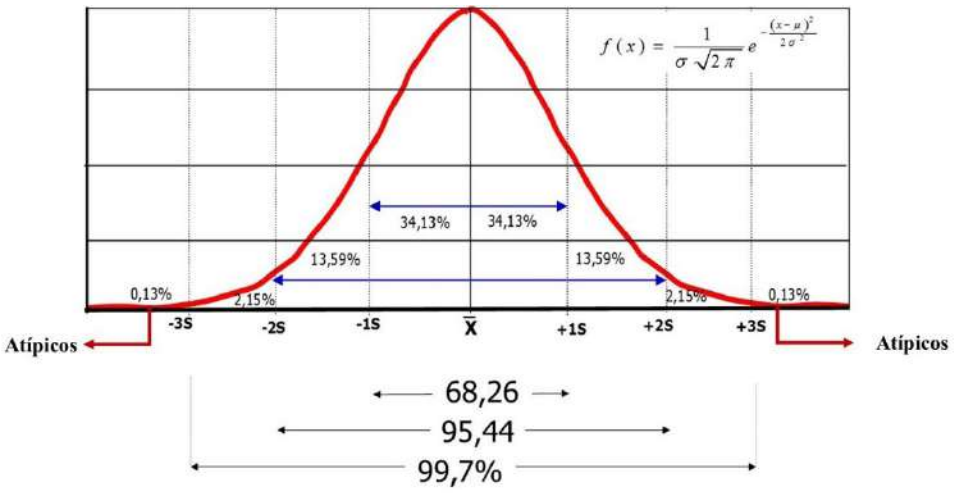


Figura 26. Valores atípicos univariantes.

1RA Edición

BIOESTADÍSTICA

APLICADA A INVESTIGACIONES
CIENTÍFICAS EN SALUD

CAPÍTULO III

ANÁLISIS DESCRIPTIVO
PARA INVESTIGACIONES

MAWIL
Publicaciones Impresas
y Digitales

3.1. ¿De qué depende el análisis estadístico de los datos?

Es muy frecuente, y mucho más cuando se lleva a cabo una investigación en salud de carácter académico-científico o investigación formativa, preguntarnos; ¿Qué procedimientos estadísticos debo utilizar en relación al análisis de datos?, ¿se debe aplicar análisis univariado, bivariado o multivariado u otro procedimiento estadístico?.

Los principales criterios se utilizan para decidir por el tipo de análisis estadístico a realizar son:

- El tipo de investigación
- El nivel de investigación
- El diseño de investigación
- Los atributos de la variable de investigación
- El objetivo de investigación
- El comportamiento de los datos

Cuadro 7. Principales criterios que se utilizan para decidir por el tipo de análisis estadístico.

Criterio	Descripción
El diseño de investigación	Diseños en investigación (según el origen): • Epidemiológicos • Experimentales • Comunitarios o ecológicos • Otros diseños
El nivel de investigación	• Exploratorio • Descriptivo • Relaciona • Explicativo • Predictivo • Aplicativo
El tipo de investigación	• Si hay o no intervención del investigador (experimental u observacional). • Si los datos son primarios o secundarios (retrospectivo o prospectivo). • El número de mediciones sobre la variable de estudio (transversal o longitudinal). • Número de variables analíticas (descriptivo o analítico).

Los atributos de la variable de investigación	Tipo de variable (cuantitativa o cualitativa), como su escala (nominal, ordinal, intervalo o razón)
El objetivo de investigación	Propósito del estudio según el nivel de investigación.
El comportamiento de los datos	Analizar, conocer y precisar la distribución de los datos, ya que su distribución solo va a ser conocida cuando se hayan obtenido los datos.

3.2. Análisis descriptivo

El análisis descriptivo tiene el propósito de describir los rasgos, atributos o caracteres fenotípicos de las poblaciones humanas, vegetales o animales; así como de las características de los objetos que son medibles mediante diferentes escalas, en determinadas condiciones de tiempo y espacio.

La estadística descriptiva proporciona los métodos para recabar información acerca de una determinada población que se desea conocer o investigar con fines específicos, entonces es importante obtener muestras adecuadas que permitan inferir el comportamiento de dicha población (Fernández & Minuesa, 2018).

Si su investigación es de nivel descriptivo, entonces requiere de análisis estadístico univariado, **no existe relación entre variables.**

La estadística univariada incluye todas las técnicas que hacen referencia a la **descripción e inferencia de una sola variable.**

Cuadro 8. Objetivos de investigación descriptiva y tipo de análisis estadístico.

Objetivos de la investigación descriptiva	Pruebas estadísticas para variable cualitativa	Pruebas estadísticas para variables cuantitativas
Describir, identificar, clasificar, cuantificar, verificar, caracterizar….	Distribución de frecuencias absolutas y relativas	Distribución de frecuencias absolutas y relativas, medidas de tendencia, central, medidas de posición, medidas de dispersión

3.3. Distribución de frecuencia

El principal objetivo de la estadística descriptiva es sintetizar conjuntos de datos mediante tablas o gráficos resumen, con el fin de poder identificar el comportamiento característico de un fenómeno y facilitar su análisis exhaustivo.

Cualquier investigación que se emprenda puede conducir a la acumulación de valores cuantitativos y cualitativos correspondientes a las diversas medidas efectuadas. Esta posibilidad, convierte a la estadística en una herramienta vital para el tratamiento de volúmenes de datos mediante tablas resúmenes conocidas como **"Tablas de frecuencia"**. Cuando los datos son agrupados, la interpretación resulta ser más sencilla.

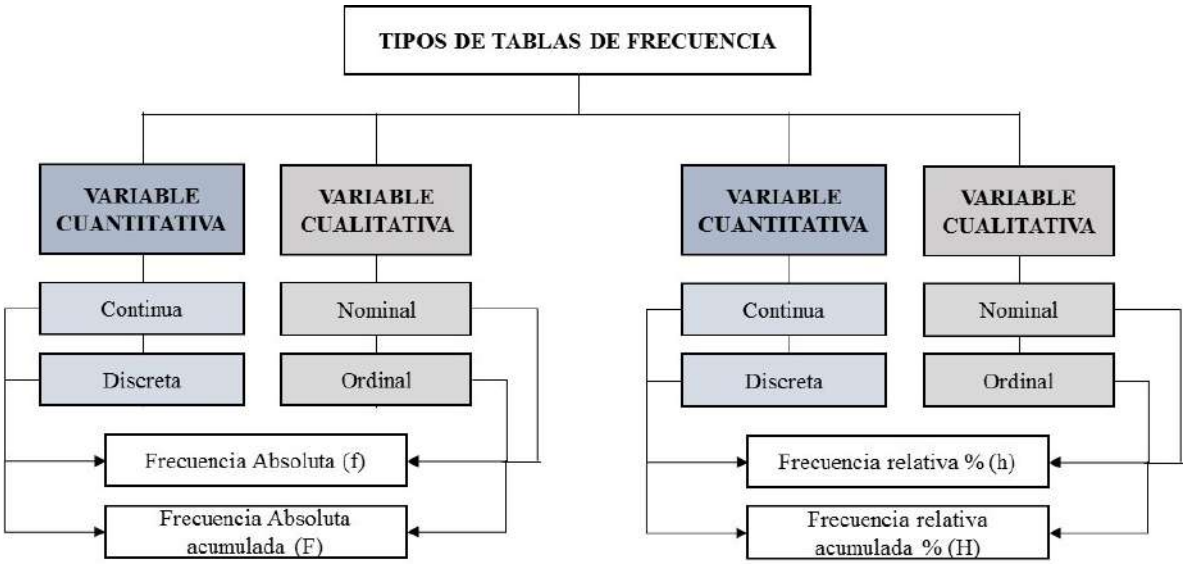


Figura 27. Tablas de frecuencia según el tipo de variable.

Una frecuencia se refiere a la **cantidad de veces que se repite un determinado valor de la variable** en una serie de datos. La distribución de frecuencias es un conjunto de datos agrupados en categorías donde se indica el número de observaciones para cada una de ellas. El análisis de frecuencias incluye la distribución de frecuencias y la tratificación de los resultados. La distribución de frecuencias es un resumen

tabulado de un conjunto de datos que muestra la frecuencia de datos de cada una de las clases que no se traslapan, con el objetivo de proporcionar una perspectiva de los datos. El análisis de frecuencias es una técnica no paramétrica.

El análisis de frecuencias puede ser de **dos tipos**: de una variable (univariada) o de dos variables (bivariada).

Análisis de frecuencias de una sola variable. Este análisis trata una sola variable, por lo que se conoce también como análisis univariado, en un sentido o en una sola dirección.

Análisis de frecuencias de dos variables. Este análisis trata simultáneamente dos variables, por lo que se conoce también como análisis en dos direcciones, bivariado, en dos sentidos, tablas cruzadas o tablas de contingencia

Resultados que se obtienen con un análisis de frecuencias
Un análisis de frecuencias permite clasificar una serie de datos y representarlos en frecuencias absolutas, frecuencia absoluta acumulada, frecuencias relativas, frecuencia relativa acumulada.

Frecuencia absoluta. Es el número de datos que corresponden a cada categoría o clase. La frecuencia absoluta se representa gráficamente en un histograma o en un polígono de frecuencias. Con las frecuencias ordenadas de mayor a menor se construye un diagrama de Pareto.

Frecuencia absoluta acumulada. Se calcula sumando las frecuencias de clase desde el primer intervalo hasta la frecuencia de clase del intervalo de interés. El valor último de las frecuencias acumuladas corresponderá al valor de la sumatoria total de los datos.

Frecuencia relativa. Se obtiene dividiendo cada una de las frecuencias de clase entre el número total de observaciones. Puede ser representada en porcentaje. La frecuencia relativa puede representarse gráficamente en un histograma o en un polígono de frecuencias. Además, se puede representar en gráficas de pastel.

Frecuencia relativa acumulada. Se calcula sumando las frecuencias de clase relativas desde el primer intervalo hasta la frecuencia de clase del intervalo de interés. El valor último de las frecuencias relativas acumuladas corresponderá al valor de cien por ciento.

3.3.1. Tablas de frecuencia para datos cualitativos (nominales y ordinales) en SPSS

El procedimiento que genera tablas de frecuencias muestran el número y el porcentaje de los casos de cada valor observado de una variable.

Elija en el menú:

1. Analizar.
2. Estadísticos descriptivos.
3. Frecuencias.
4. Seleccione las variables con las que desee hacer la tabla y muévalas hasta la lista Variable(s).
5. Pulse en Aceptar para ejecutar el procedimiento **(ver Fig. 28**).

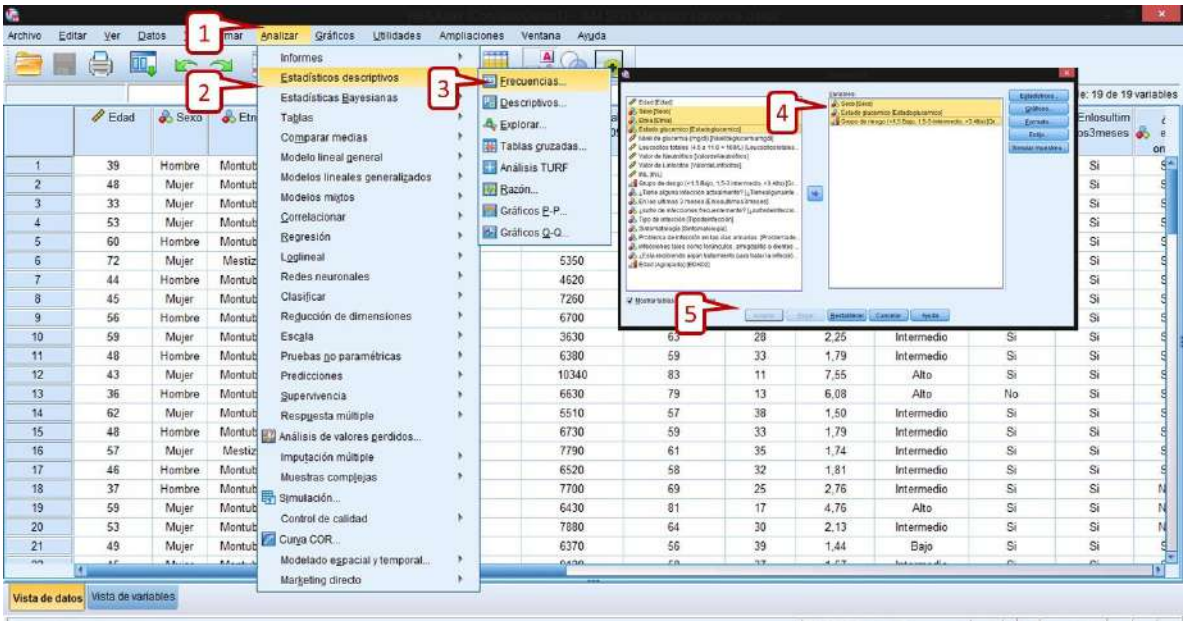


Figura 28. Procedimiento para generar tablas de frecuencia.

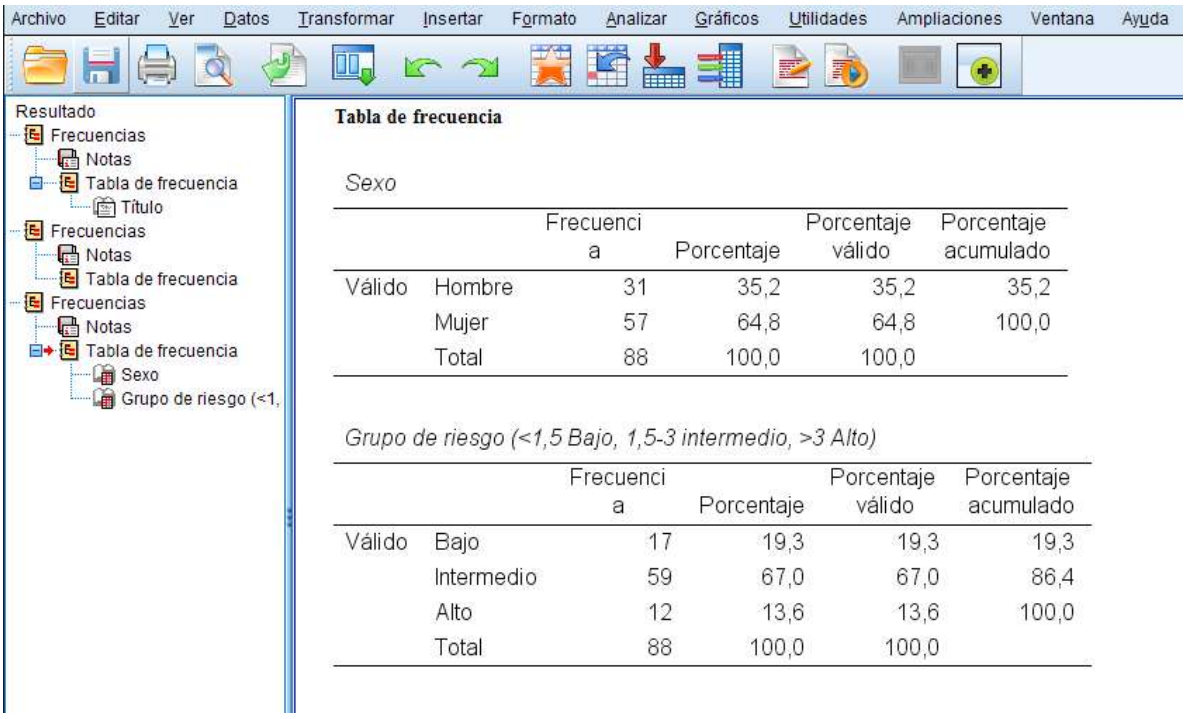


Figura 29. Vista de resultados de tablas de frecuencia.

Las tablas de frecuencias que se despliegan en la vista de resultados cuentan siempre con la siguiente información (ver Figura 29):

Válidos (datos): Esto nos señala las opciones de respuesta o categorías que se consideraron y de las que se encontraron registros, para el análisis de esa variable (en el ejemplo es el sexo, grupo de riesgo).

- **Frecuencia:** Es el número de veces que apareció determinada categoría o respuesta (en el ejemplo las categorías Hombre y Mujer).
- **Porcentaje:** Éste es el porcentaje que representa de la muestra total sobre la que se recolectaron los datos, cada una de las respuestas o categorías posibles.
- **Porcentaje válido:** Representa el porcentaje de registros que verdaderamente se consideraron en cada categoría o respuesta para llevar a cabo los análisis, eliminando los valores perdidos.
- **Porcentaje acumulado:** Su nombre describe la función que realiza, la cual es acumular los porcentajes de las respuestas posibles hasta conseguir el cien por ciento **(no se utiliza en variables nominales, p. ej. sexo)** sí se utiliza en variables ordinales.

Las tablas de frecuencias aparecen en la ventana Visor, en el ejemplo en el que se ha comparado el sexo de los pacientes: Hombre y Mujer. Las tablas de frecuencias revelan que el 35,2% de los pacientes son hombres, y que la mayoría de los pacientes son de sexo femenino (64,8%).

3.3.2. Tablas de frecuencia para datos cuantitativos (continuos o discretos) con intervalos de clase en SPSS

Este tipo de tablas suelen ser utilizadas cuando el número de resultados posibles que puede obtener una variable son tan amplios, que una tabla simple haría muy poco en resumirlos (estos datos representan un rango muy amplio).

Debido a esta cantidad de valores, será necesario agruparlos mediante intervalos (la estadística los llama “Intervalos de clases”).

Por ejemplo, en el caso de contar con una valoración del 1 al 100 (un rango equivalente a 99), una tabla de frecuencia sin intervalos de clase se encargaría de buscar cuántas veces se repite cada uno de los 99 posibles resultados en un conjunto de datos, teniendo una función contraria a la de resumir los datos.

Agrupar los valores de la variable en intervalos podría simplificar estas fuentes de datos. Por ejemplo, podríamos hablar de las frecuencias para los valores comprendidos entre 0-20, 20-40, 40-60, 60-80 y 80-100.

En el intervalo 0-20 (que de ahora en adelante le llamaremos intervalo de clase), se sumaran las frecuencias de los datos cuyos resultados estén entre 0 y 20.

Recomendaciones para su elaboración:

- Su construcción requiere, en primer lugar, la selección de los límites de los intervalos de clase.
- Para definir la cantidad de intervalos de clase (k), se puede usar:

La regla de Sturges: $k = 1 + 3{,}33*\log(n)$

- La cantidad de clases no puede ser tan pequeño (menos de 5) o tan grande (más de 15), que la verdadera naturaleza de la distribución sea imposible de visualizar.
- La amplitud de todas las clases deberá ser la misma.
- Se recomienda que sea impar y que los puntos medios tengan la misma cantidad de cifras significativas que los datos en bruto.
- Los límites de las clases deben tener una cifra significativa más que los datos en bruto.

Intervalo de clase: Intervalos empleados en las tablas de frecuencias estadísticas, capaz de contener diversas medidas de una variable. Consta de un límite inferior (Lm) y un límite superior (Ls).

Otro punto importante que el estadista debe definir, es la cantidad de intervalos de clase que empleará en la tabla de frecuencia. Esta cantidad de intervalos **no deberían ser muchos, debido a que no se cumpliría el objetivo de resumir la información, y no tan pocos intervalos, ya que se perdería mucha información**.

Número de intervalos (IC): Cantidad de intervalos con los cuales se compone una tabla de frecuencia.

Algunos autores han propuestos fórmulas que permiten ayudar en la tarea de conseguir el número ideal de intervalos.

Para determinar el número óptimo de intervalos de clase, en los cuales nuestros datos quedarán perfectamente distribuidos, aplicamos la regla de Sturges (Hernández Sampieri & Mendoza Torres, 2018):

Regla de Sturges = N.º de intervalos de clase = 1+3,33*log (n)

En donde **"n"** representa el número total de datos u observaciones que tenemos recopilados. Ejemplo en una **muestra de 88 pacientes** ¿Cuál será el número óptimo de intervalos de clase?

Utilizando la hoja de cálculo de Excel: = 1+3,33*log (n)
= 1+3,33*log (88)
= **7,47**

Evidentemente, el número de intervalos debe ser exacto; es decir, un número entero y siguiendo las recomendaciones la cantidad de intervalos debe ser un número impar. Para **una muestra de 88 pacientes** se recomienda trabajar la tabla de frecuencia con **7 intervalos de clase.**

Cada intervalo posee un número máximo de resultados que puede agrupar. A este valor lo conoceremos como el "Ancho del intervalo de clase (A)".

<table>
<tr><td colspan="2">Ancho del intervalo de clase (A): Matemáticamente se expresa:</td></tr>
<tr><td rowspan="2">Ancho del intervalo de clase =</td><td>Rango (R)</td></tr>
<tr><td>Número de Intervalos (IC)</td></tr>
<tr><td colspan="2">Rango: Matemáticamente se expresa en la diferencia del valor máximo y el valor mínimo.
Rango (R) = valor máximo-valor mínimo</td></tr>
</table>

Ejercicio aplicado en SPSS

Para realizar un estudio sobre la estatura de los estudiantes de la carrera de Laboratorio Clínico de la Universidad Estatal del Sur de Manabí, seleccionamos, mediante un proceso de muestreo aleatorio, una muestra de 35 estudiantes, obteniendo los siguientes resultados (medidos en centímetros):
176,179, 173, 150,181, 170, 158, 174, 172, 166, 194, 185, 162, 187, 198, 177, 178, 165, 154, 188, 166, 171, 175, 182, 167, 169, 172, 186, 172, 176, 168, 187,189,154,160

Solución:

1. Determinar rango: mediante el uso del análisis descriptivo se determinó el rango en 48 (ver Fig. 30).
2.

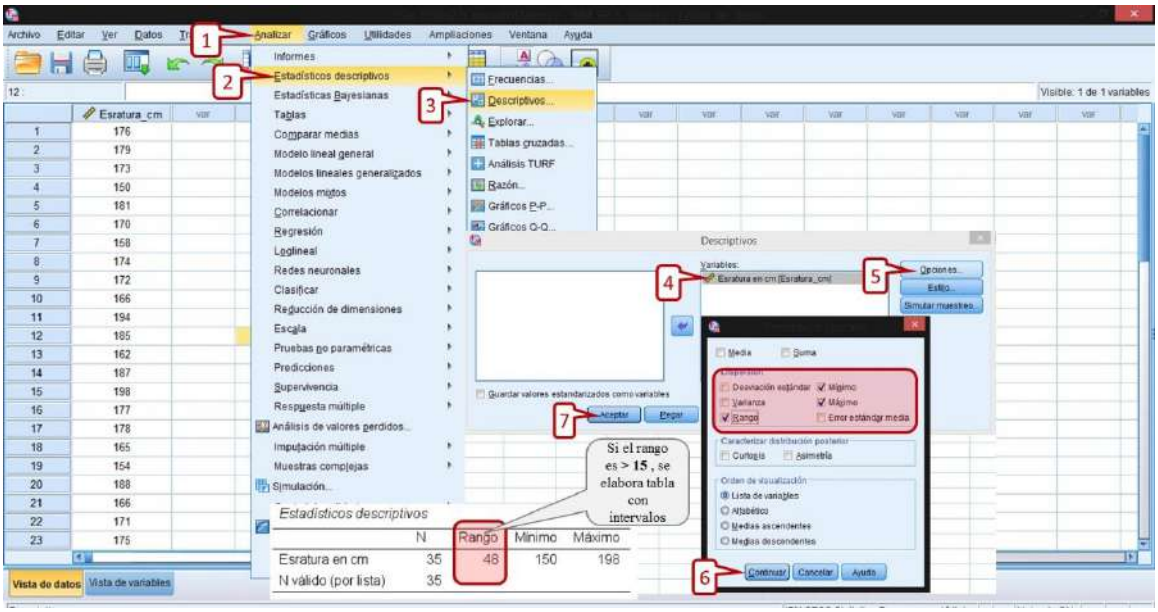


Figura 30. Cálculo del rango (R).

3. Determinar número de intervalos (IC)

Utilizando la hoja de cálculo de Excel: = 1+3,33*log(n)

= 1+3,33*log(35)

= **6,14**

Evidentemente, el número de intervalos debe ser exacto; es decir, un número entero y siguiendo las recomendaciones debe ser un número impar la cantidad de intervalos para una **muestra de 35 pacientes** se recomienda trabajar la tabla de frecuencia con **5 intervalos de clase.**

4. Determinar ancho de intervalo

Ancho del intervalo de clase =	Rango (R)
	Número de intervalos (I)

Ancho del intervalo de clase =	48
	5

Ancho del intervalo de clase = 9,6

Considerando que los datos de estatura están presentados en números enteros **el ancho del intervalo a utilizar debe ser 10**

5. Elaborar tabla de frecuencia con intervalos de clase:
Con los datos calculados, Rango (48), Número de intervalos (5) y ancho de intervalo (10), mediante el programa SPSS procedemos a elaborar la tabla de los intervalos de clase (ver Fig. 31).

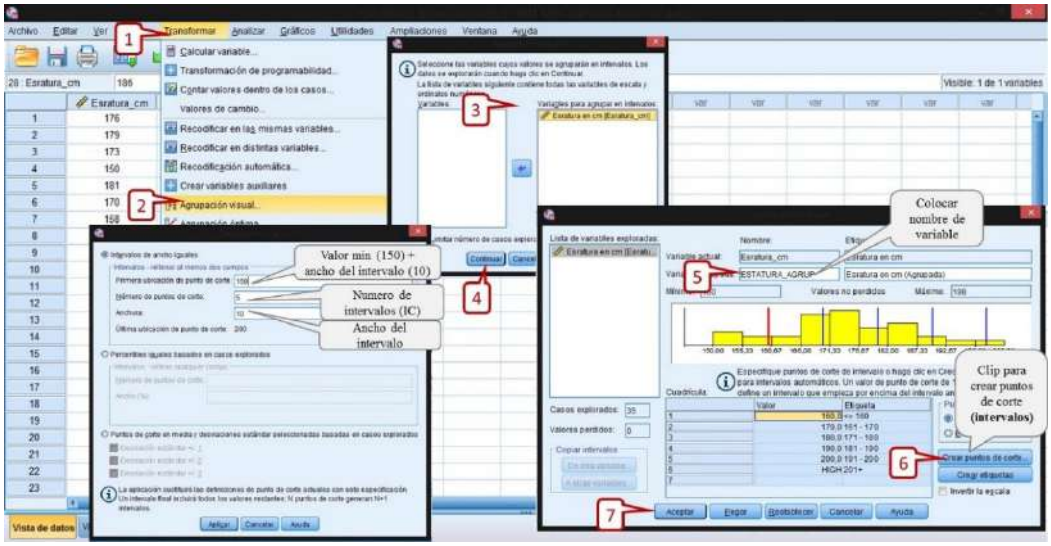
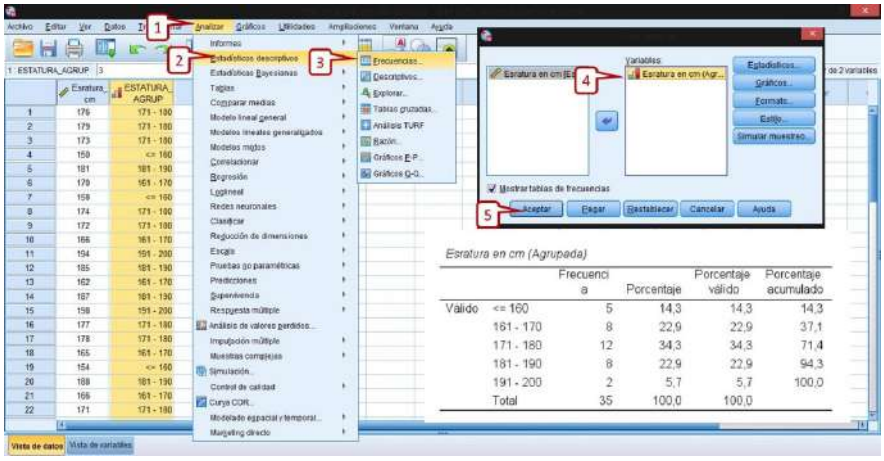


Figura 31. Transformación de una variable cuantitativa a cualitativa.

Mediante la opción agrupación visual se crea una variable cualitativa que se utiliza para la elaboración de la tabla de frecuencia con intervalos de clase (ver Fig. 32).



Estatura en cm (Agrupada)

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	<= 160	5	14,3	14,3	14,3
	161 - 170	8	22,9	22,9	37,1
	171 - 180	12	34,3	34,3	71,4
	181 - 190	8	22,9	22,9	94,3
	191 - 200	2	5,7	5,7	100,0
	Total	35	100,0	100,0	

Figura 32. Elaboración de tabla de frecuencia con intervalos de clase.

3.4. Tablas de contingencia

En investigaciones en ciencias de la salud es muy frecuente recurrir a la tabulación cruzada de los datos cuando, además de describir (análisis univariable), nos interesa comparar (análisis bivariable). Las tablas de contingencia resultan, especialmente indicadas, cuando disponemos de variables cualitativas (nominales u ordinales), suponiendo que una de ellas depende de la otra (variable independiente y dependiente). La elaboración de tablas de contingencia o tablas bivariables no se encuentra estandarizada, basta con que ésta se lea e interprete correctamente y es posible realizar con las siguientes variables:

- Implica siempre a variables cualitativas, categóricas o nominales, u ordinales con pocos valores: nominal*nominal, ordinal*ordinal; nominal*ordinal.
- También puede implicar a una variable nominal y otra de intervalo: nominal (sexo)*intervalo (edad).
- Los datos se organizan en tablas de doble entrada, distribuidos según un criterio de clasificación (variable nominal/variable ordinal). Resultado: frecuencias y porcentajes.
- Observar asociación o relación entre las categorías o valores de las variables implicadas (lectura cruzada).

Para realizar una tabla de contingencia por medio del generador de tablas, se debe ingresar una variable categórica a cada una de las dimensiones de la tabla (fila y columna). En las tablas de contingencia también se puede incluir diferentes estadísticos como el porcentaje de columna, el porcentaje de fila y el porcentaje de tabla; para lograrlo debemos activarlo en el cuadro de diálogo del generador de tablas (Díaz-Parreño ***et al***., 2014).

Tabla de contingencia en SPSS

En un estudio realizado en la población urbana del cantón Jipijapa para saber si la infección en las vías urinarias (IVU) está relacionada con el sexo, se preguntó a 40 personas. De los cuales 14 son hombres y 26 mujeres ¿Podemos concluir que existe el mismo porcentaje de

ocurrencia de IVU en ambos sexos, a continuación de exponen los siguientes resultados:

N.°	Sexo	IVU	N.°	Sexo	IVU	N.°	Sexo	IVU	N.°	Sexo	IVU
1	Hombre	No	**11**	Hombre	No	**21**	Mujer	Si	**31**	Hombre	Si
2	Mujer	Si	**12**	Mujer	Si	**22**	Mujer	Si	**32**	Mujer	Si
3	Mujer	Si	**13**	Hombre	No	**23**	Mujer	Si	**33**	Mujer	Si
4	Mujer	Si	**14**	Mujer	Si	**24**	Hombre	Si	**34**	Mujer	Si
5	Hombre	No	**15**	Hombre	No	**25**	Hombre	Si	**35**	Mujer	No
6	Mujer	Si	**16**	Mujer	Si	**26**	Mujer	Si	**36**	Hombre	Si
7	Hombre	No	**17**	Hombre	No	**27**	Mujer	Si	**37**	Mujer	No
8	Mujer	Si	**18**	Hombre	Si	**28**	Mujer	Si	**38**	Hombre	Si
9	Hombre	No	**19**	Mujer	Si	**29**	Mujer	Si	**39**	Mujer	No
10	Mujer	Si	**20**	Mujer	Si	**30**	Mujer	Si	**40**	Mujer	No

Solución

Para llevar a cabo el análisis de tablas de contingencia empezamos definiendo la tabla con las dos variables seleccionadas.

1. Acceder al cuadro de diálogo de tablas de contingencia, seleccionando Estadísticos descriptivos: Tablas cruzadas del menú principal Analizar.
2. Indicar las dos variables que van a formar la tabla de doble entrada recordando que: en las **filas situaremos a la independiente-causa** (Sexo) y en las **columnas a la variable dependiente-efecto** (IVU).
3. Una vez que ya tenemos definida la relación de las dos variables que suponemos, ***a priori***, asociadas, deberemos seleccionar en la **opción casillas los porcentajes en columnas** considerando que aquí está ubicada la variable dependiente **(Figura 33).**

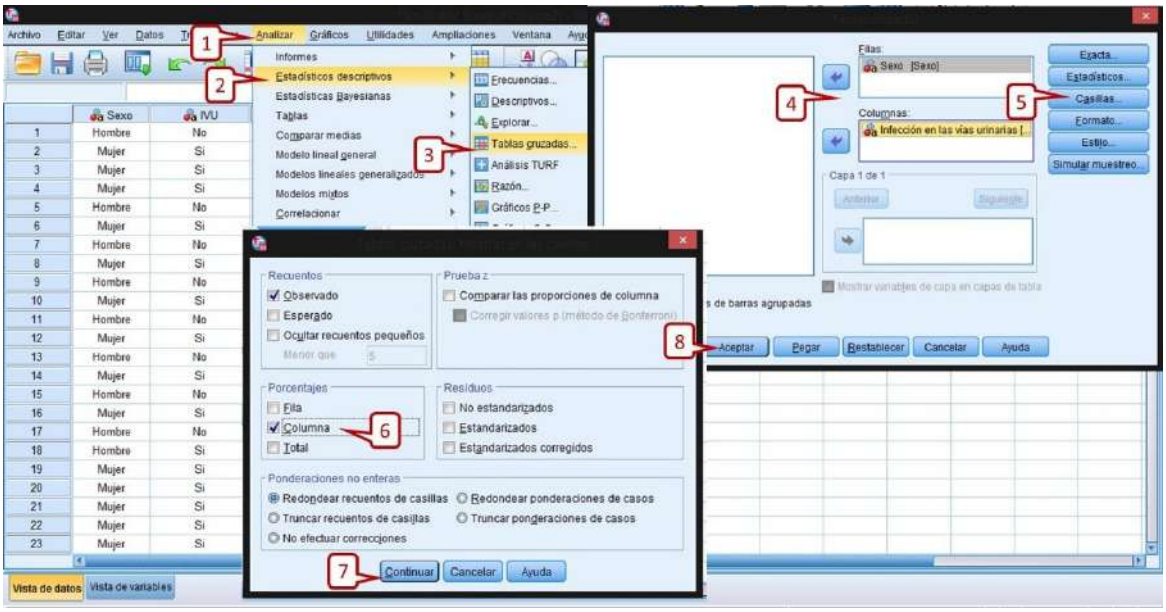


Figura 33. Elaboración de tabla de contingencia para dos variables.

Por último, y como previo paso a la salida definitiva del resultado derivado de las restricciones y peticiones a las que hemos sometido al análisis de las tablas de contingencia (cuadro de diálogo tablas cruzadas), deberemos especificar con qué formato queremos que se presente el resultado.

Como resultados del ejemplo planteado podemos concluir que existe un porcentaje superior de ocurrencia de IVU para las **mujeres (78,6%)**, en contraste con los hombres que representan el 21,4% de los casos positivos con IVU **(ver Figura 34).**

Tabla cruzada Sexo *Infección en las vías urinarias

Sexo	Infección en las vías urinarias		Total
	No	Si	
Hombre	8	6	14
	66,70%	21,40%	35,00%
Mujer	4	22	26
	33,30%	78,60%	65,00%
Total	12	28	40
	100,00%	100,00%	100,00%

Figura 34. Tabla de contingencia para dos variables.

3.5. Representación gráfica de los datos

Los graficas o diagramas son esquemas formados por líneas, figuras, áreas o mapas, volúmenes, etc., que sirven para representar datos. En la mayoría de las ocasiones los gráficos son utilizados como un soporte visual a los resultados, permitiéndonos identificar fácilmente el comportamiento de los datos, agilizando el reconocimiento de las conclusiones que el análisis nos arroja.

Los gráficos pueden ser agrupados de acuerdo a su objetivo, por lo que se incluyen las categorías (General y Específico). Estas categorías, hacen referencia a las diferentes aplicaciones que se pueden ejercer con los gráficos. Se asume que un gráfico es general, cuando se puede aplicar a diferentes procesos o análisis modificando solo su estructura básica, como, por ejemplo, el gráfico de barras, el cual puede ser empleado para describir las categorías de una variable o para comparar las categorías de múltiples variables. Por el contrario, los gráficos específicos son aquellos que solo se pueden aplicar a un proceso o análisis, a pesar de que se modifique su estructura. ejemplo

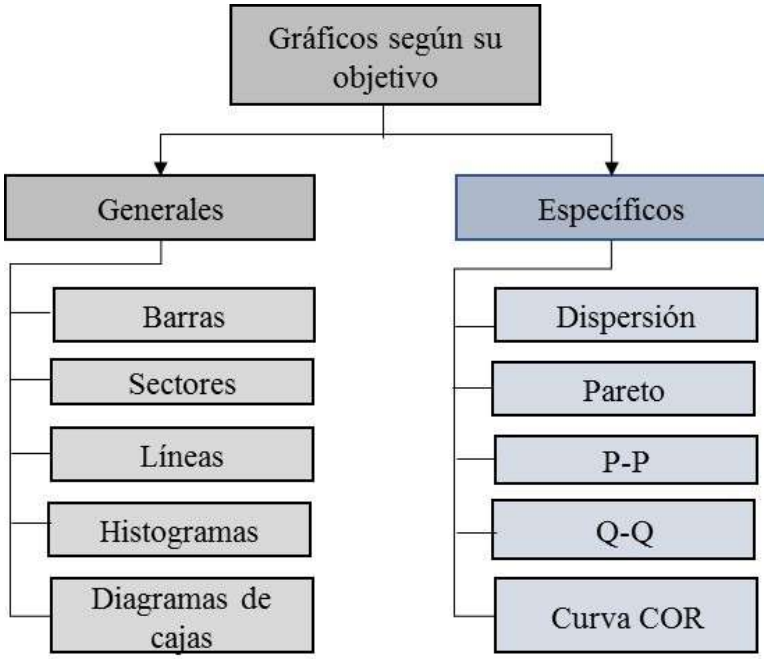


Figura 34. Tipos de gráficos según su objetivo.

Dentro de los diferentes tipos de gráficos encontramos las barras, las líneas, las áreas, los sectores, diagramas de caja, histograma, puntos, líneas verticales, dispersión, Pareto, control, P-P, Q-Q, y curva COR.

En cuanto a la representación gráfica de las variables cualitativas destacamos dos tipos de gráfico por ser los que se utilizan con mayor frecuencia.

Diagrama de sectores. El primero de ellos, el diagrama de sectores, se utiliza para visualizar de forma sencilla las frecuencias relativas de las variables. En los gráficos de sectores se divide una figura, habitualmente de forma circular, de forma que el área correspondiente a cada posible respuesta de la variable será proporcional a la frecuencia relativa de la variable. Esta representación se puede adornar de etiquetas en el interior o exterior del gráfico, además suele ser habitual incluir para cada categoría de la variable la frecuencia relativa (o si se desea absoluta de la variable).

Un diagrama de sectores se puede utilizar para todo tipo de variables, pero se usa frecuentemente para las variables cualitativas.

Los datos se representan en un círculo, de modo que el ángulo de cada sector es proporcional al porcentaje correspondiente.

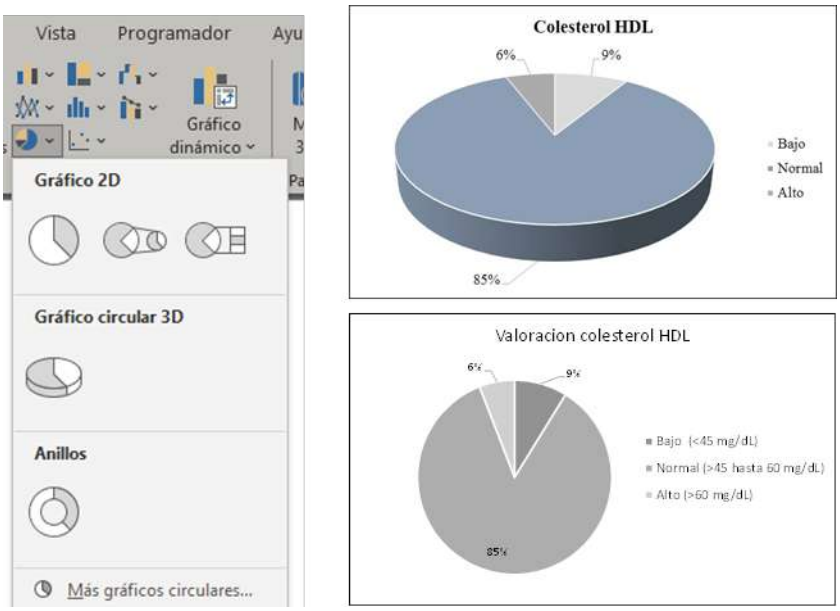


Figura 35. Tipos de gráficos de sectores.

Gráfico de barras. El segundo tipo de representaciones gráficas que vamos a contemplar son los gráficos de barras. En este tipo de gráfico se representa una barra vertical (u horizontal si se desea) para cada una de las categorías de la variable de altura proporcional a su frecuencia, bien absoluta o relativa. Al igual que los diagramas de sectores los gráficos de barras se suelen personalizar al gusto del usuario de forma que su configuración resulte lo más ilustrativa posible. Los gráficos de barras suelen ser preferibles a los diagramas de sectores ya que según se ha podido comprobar el ojo humano está particularmente entrenado para comparar longitudes y no para comparar áreas, sin embargo, dada la popularidad de estos últimos en la literatura conviene conocer su interpretación y ser conscientes de su posible uso.

Se construye de forma que la altura representa el valor de la variable y la anchura debe ser igual. Una gráfica de barras se puede usar para describir cualquier nivel de medición (nominal, ordinal, continuo o discreto).

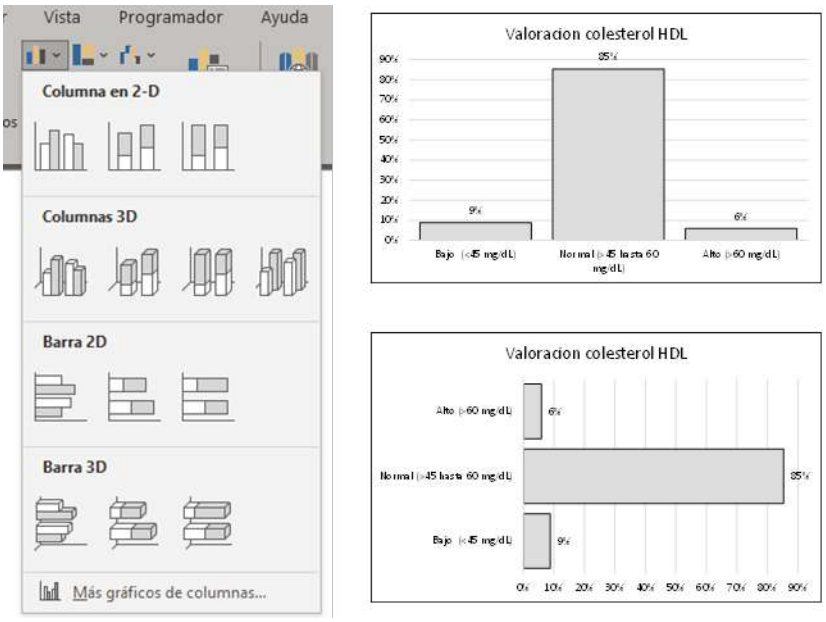


Figura 36. Tipos de gráficos de barras.

Gráfica de líneas o curvas: Para construir esta gráfica los puntos se localizan mediante las coordenadas que representan y después se unen los puntos.

Los gráficos de líneas muestran una serie como un conjunto de puntos conectados mediante una sola línea en un informe paginado. Los gráficos de líneas se usan para representar grandes cantidades de datos que tienen lugar durante un período continuado de tiempo.

Variaciones

Línea suavizada. Gráfico de líneas que usa una línea curva en lugar de una línea normal.

Línea escalonada. Gráfico de líneas que usa una línea escalonada en lugar de una línea normal. La línea escalonada conecta puntos mediante una línea que adopta la apariencia de los peldaños de una escalera.

Consideraciones sobre los datos para los gráficos de líneas:

- Para mejorar el impacto visual del gráfico de líneas predeterminado, considere la posibilidad de cambiar el ancho del borde de la serie a 3. Esto creará un gráfico de líneas mucho más oscuro.
- Si el conjunto de datos incluye valores vacíos, el gráfico de líneas agregará puntos vacíos en forma de líneas de marcador de posición para mantener la continuidad en el gráfico.
- Un gráfico de líneas requiere al menos dos puntos para dibujar una línea. Si el conjunto de datos solo tiene un punto de datos, el gráfico de líneas se mostrará como un marcador de punto de datos único.

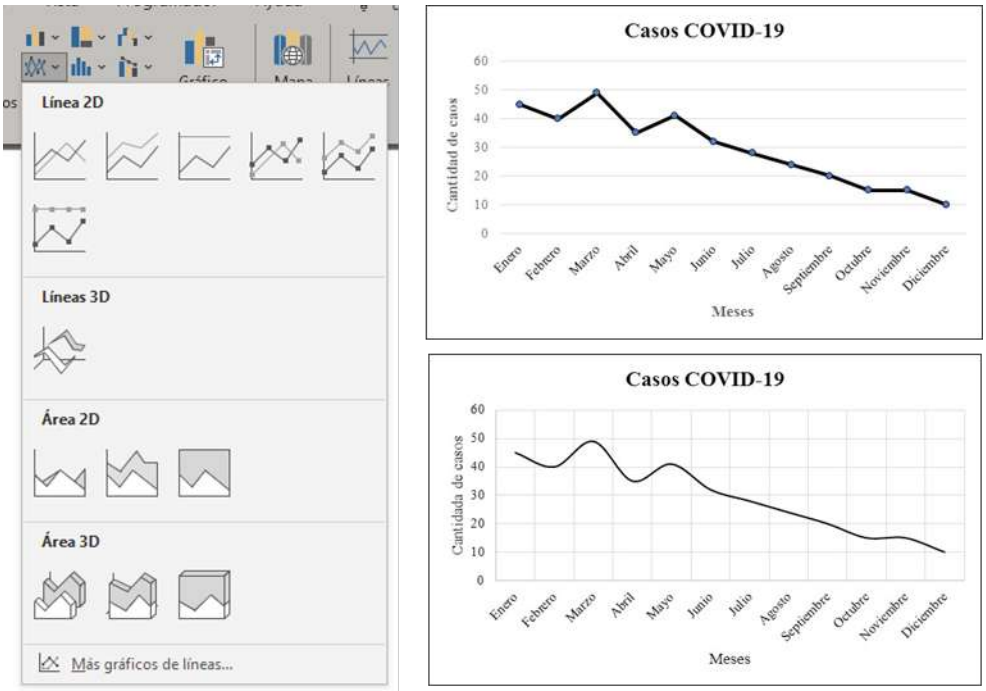


Figura 37. Tipos de gráficos de líneas.

Gráfico de dispersión o puntos: Un gráfico de dispersión muestra una serie como un conjunto de puntos en un informe paginado. Los valores se representan mediante la posición de los puntos en el gráfico. Las categorías se representan mediante distintos marcadores en el gráfico. Los gráficos de dispersión suelen usarse para comparar datos agregados de las categorías.

Consideraciones sobre los datos para los gráficos de dispersión:

- Los gráficos de dispersión se usan normalmente para mostrar y comparar valores numéricos, como datos científicos.
- Use el gráfico de dispersión cuando desee comparar grandes cantidades de puntos de datos sin tener en cuenta el tiempo. Cuantos más datos incluya en un gráfico de dispersión, mejores comparaciones podrán realizar.
- Los gráficos de dispersión son ideales para controlar la distribución de los valores y los clústeres de los puntos de datos. Es el mejor tipo de gráfico si el conjunto de datos contiene muchos puntos (p. ej., varios miles). Se debe evitar mostrar varias series en un gráfico de puntos porque visualmente puede resultar confuso.
- De forma predeterminada, los gráficos de dispersión muestran los puntos de datos como círculos. Si tiene varias series en un gráfico de dispersión, plantéese la posibilidad de cambiar la forma del marcador de cada punto por un cuadrado, un triángulo, un rombo o cualquier otra forma.

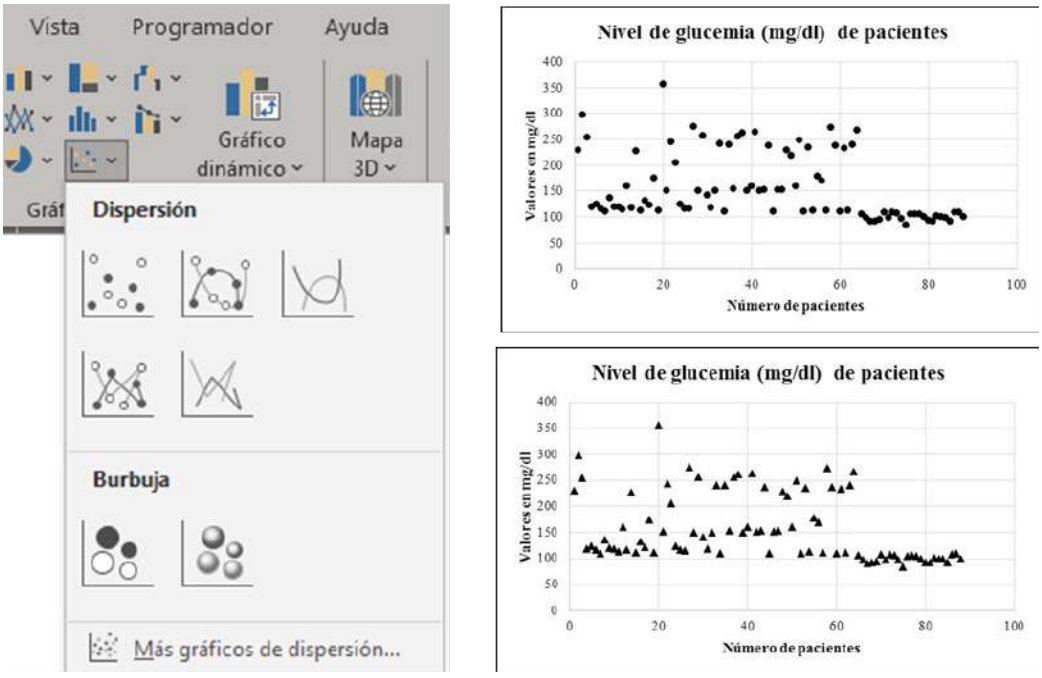


Figura 38. Tipos de gráficos de dispersión.

Histograma: Uno de los procedimientos gráficos más utilizados en el análisis de variables de escala es el histograma, a través de él se puede reunir los valores de una variable en grupos con un mismo rango o distancia denominados intervalos o clases, los cuales a su vez incorporan el recuento del número de casos dentro de cada grupo. Este recuento o frecuencia, se puede expresar en forma de porcentaje, lo que es especialmente útil para comparar conjuntos de datos de diferentes tamaños o unidades de medida. A través del histograma se puede detectar parámetros como los valores atípicos y las desviaciones de la asimetría, quienes nos pueden indicar si la variable es o no adecuada para ser analizada mediante un procedimiento que asuma una distribución normal.

SPSS cuenta con dos diferentes modalidades de gráficos; la primera corresponde a los estándar o normales y la segunda corresponde a los interactivos. Estas dos modalidades hacen referencia a las posibilidades de edición, así como la forma de obtenerlos. La principal diferencia que podemos encontrar entre estos tipos, radica en que los gráficos normales carecen de ciertas funciones interactivas como, por ejemplo, la capacidad de cambiar las variables representadas directamente sobre el gráfico, cambiar las funciones estadísticas de resumen después de ser creado o insertar elementos adicionales.

Para comprender la estructura organizacional de las diferentes opciones de gráficos con que cuenta el programa, en el cuadro de la figura 34 encontramos un diagrama comparativo de los diferentes tipos de gráficos de cada una de las modalidades. Si observamos el cuadro, notaremos que la mayoría de los tipos de gráfico se encuentran incluidos en las dos modalidades, por lo que hemos empleado el color rojo para facilitar su identificación. SPSS cuenta con 19 diferentes tipos de gráficos, algunos de los cuales iremos conociendo a través de este capítulo, haciendo una pequeña descripción de su utilidad y generando algunos ejemplos.

3.5.1. Generador de gráficos en SPSS

Además de las gráficas producidas por los anteriores procedimientos descriptivos, SPSS cuenta con un menú dedicado expresamente para la obtención de resultados gráficos. Sirvan estas notas como una breve exposición de las características generales en el manejo de los procedimientos gráficos. Una exposición más detallada de estos procedimientos requeriría una extensión que sobrepasaría los objetivos de este documento introductorio.

Seleccionando en el menú principal **"Gráficos",** se obtiene la siguiente ventana, donde se muestran los distintos gráficos que se pueden realizar.

Para proceder a realizar algún tipo de gráfico interactivo se selecciona en el **menú principal Gráficos/Generador de gráficos**… y se muestra la siguiente figura:

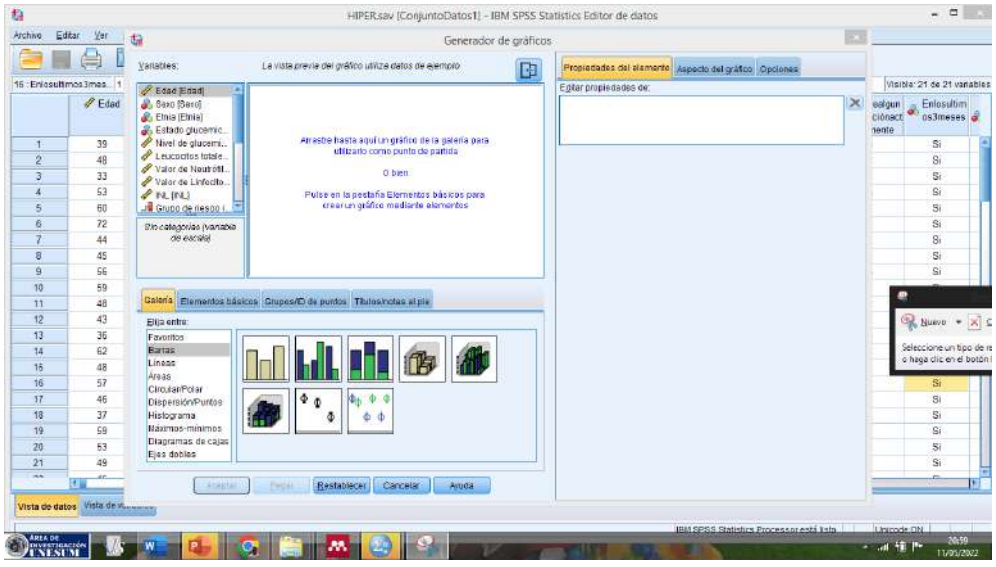


Figura 39. Menú principal Gráficos/Generador de gráfico.

Donde se selecciona el gráfico que se desee realizar (en el ejercicio guiado de esta práctica se muestra cómo se realiza un histograma interactivo).

Seleccionando en el **menú principal Gráficos/Cuadros de diálogo antiguos** se muestra la siguiente figura

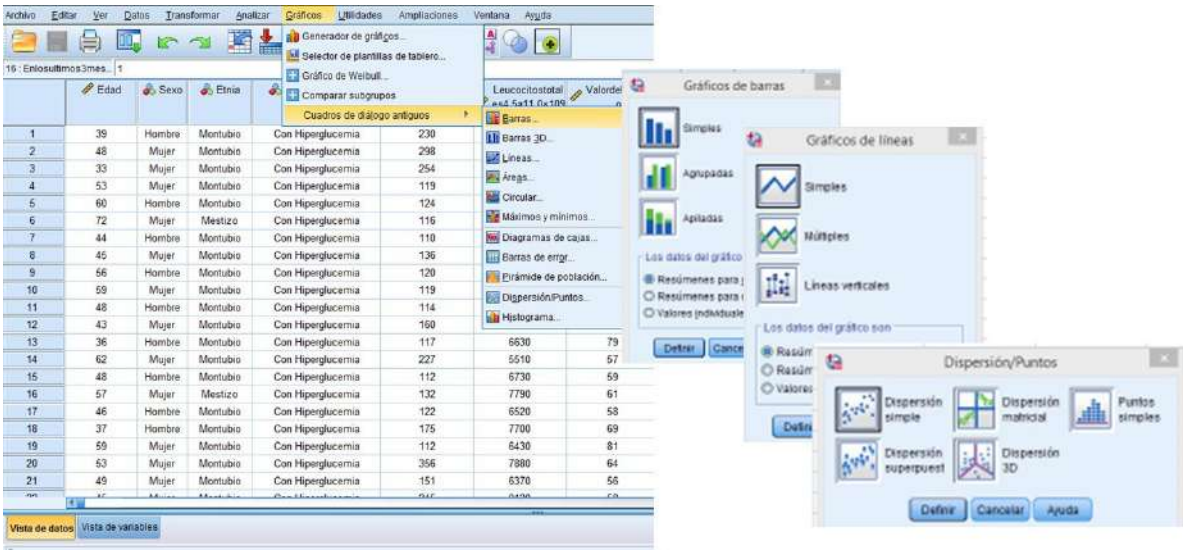


Figura 40. Menú principal Gráficos/Cuadros de diálogo antiguos.

3.6. Estadísticos descriptivos para variables cuantitativas

En cualquier análisis estadístico, la estadística descriptiva es la primera parte y más importante, pues permite conocer el comportamiento de las variables, consideradas una a una, o la posible relación existente entre ellas. El análisis descriptivo de las variables incluidas en el estudio dependerá del tipo de variables que necesitamos resumir.

Para resumir variables de tipo cuantitativo se dispone de una diversidad de pruebas estadísticas. Podemos tabular los datos en tablas de frecuencias, o bien calcular medidas de resumen específicas de este tipo de variables, que se pueden clasificar de manera general de la siguiente forma:

Medidas de centralización: Resumen la localización alrededor de la cual se distribuyen los datos (la media, moda y mediana).

Medidas de dispersión: Resumen la variabilidad que presentan los datos alrededor de alguno de los estadísticos de centralización. Estudiaremos como medidas de dispersión el rango, rango intercuartílico, varianza, error estándar y desviación estándar.

Medidas de posición (localización): Informan sobre distintas características de los datos a partir de la ordenación de los valores observados. Las medidas de orden más utilizadas son los percentiles y cuartiles.

3.6.1. Medidas de centralización

Son indicadores estadísticos que muestran hacia qué valor (o valores) se agrupan los datos.

La primera gama de indicadores corresponde a las "Medidas de tendencia central". Existen varios procedimientos para expresar matemáticamente las medidas de tendencia central, de los cuales, los más conocidos son: la media aritmética, media ponderada, la moda y la mediana.

3.5.1.2. La Media Aritmética

La media es comúnmente conocida como "promedio" y corresponde a un valor de tendencia central para una variable con medida de escala. En las variables nominales u ordinales no tiene sentido utilizar este estadístico por su naturaleza.

La media aritmética es una cifra que obtienes al sumar todos los valores observados y dividirlos por el número de valores. ¿No te resulta familiar? Claro, estás muy acostumbrado al cálculo del promedio. Se denota por los símbolos **μ** (letra griega mu) o **X** (se lee equis media), la media conserva las unidades de medida de la variable en su estado original, o sea, que la media de un grupo de edades en años se expresará asimismo en años.

Ventajas

- Es la medida de tendencia central más usada.
- El promedio es estable en el muestreo.
- Es sensible a cualquier cambio en los datos (puede ser usado como un detector de variaciones en los datos).
- Se emplea a menudo en cálculos estadísticos posteriores.
- Presenta rigor matemático.
- En la gráfica de frecuencias representa el centro de gravedad.

Desventajas

- Es sensible a los valores extremos.
- No es recomendable emplearla en distribuciones muy asimétricas.
- Si se emplean variables discretas o cuasicualitativas, la media aritmética puede no pertenecer al conjunto de valores de la variable.

3.5.1.3. La Mediana

Aquí tienes otra de las medidas de tendencia central. Al igual que la media, puedes utilizarla para describir el "centro" de un grupo de datos. No tiene un símbolo específico que la denote.

Mediana (Med.):

- Valor que divide una serie de datos en dos partes iguales.
- La cantidad de datos que queda por debajo y por arriba de la mediana son iguales.
- La definición de geométrica se refiere al punto que divide en dos partes a un segmento.

Ventajas

- Es estable a los valores extremos.
- Es recomendable para distribuciones muy asimétricas.

Desventajas

- No presenta todo el rigor matemático.
- Se emplea solo en variables cuantitativas.

3.5.1.4. La Moda

La moda una medida realmente sencilla, tanto de determinar como de interpretar. Es muy intuitiva, y consiste en el valor, clase o categoría que aparece con más frecuencia en una serie de datos; o sea, es el que más se repite. Por ejemplo, si de seis pacientes, tres tienen 20 años, y los otros tienen 18, 21 y 25, respectivamente, entonces dirías que 20 años es la moda, o edad modal.

La mayor ventaja de la moda radica en que no requiere cálculo alguno, para beneplácito de algunos que no cuentan a las matemáticas entre su círculo de amistades. Sin embargo, puede que no exista, e incluso puede no ser única. Por ejemplo, la serie 2, 5, 6, 6, 6, 7, 8, 8, 8, es una serie bimodal, pues cuenta con el seis y el ocho como modas. En el caso de que dos valores presenten la misma frecuencia, decimos que existe un conjunto de datos bimodal. Para más de dos modas hablaremos de un conjunto de datos multimodal.

Ventajas

- Es estable a los valores extremos.
- Es recomendable para el tratamiento de variables cualitativas.

Desventajas

- Pueda que no se presente.
- Puede existir más de una moda.
- En distribuciones muy asimétricas suele ser un dato muy poco representativo.
- Carece de rigor matemático.

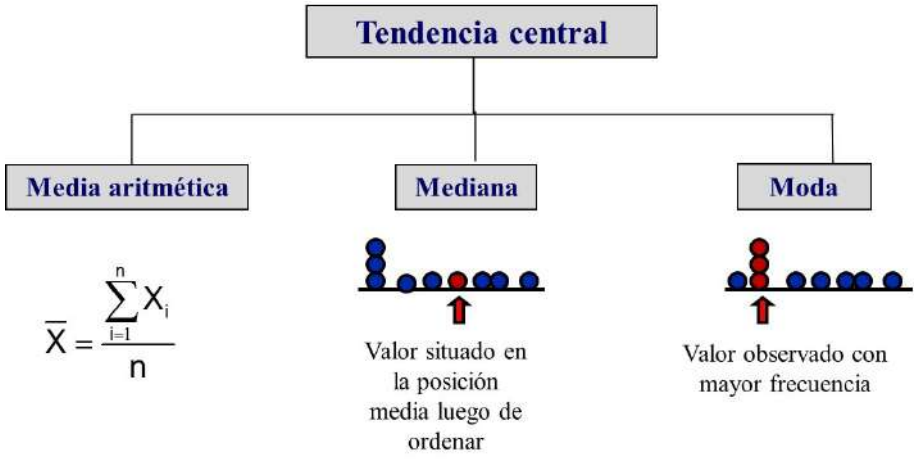


Figura 41. Medidas de centralización.

3.5.1.5. Relación entre la media, mediana y moda

Media: Es el promedio de cierto número de datos. Es como cuando sumas la calificaciones de todas tus materias y las divides entre el número de materias para ver tu promedio con los datos 7, 8, 8, 9,10 la media es (7+8+8+9+10)/5 **= 8,4**

Moda: Es el valor que más se repite en cierto número de datos, por ejemplo, si tú ves tus calificaciones y son: 7, 8, 8, 9,10 la moda es **8**, ya que es el valor que más se repite.

Mediana: Es cuando acomodas tus datos del mayor a menor y tomas el valor del centro, por ejemplo, si tienes: 7, 8, 8, 9,10 los ordenas de menor a mayor y te quedan: 7, 8, **8,** 9 y 10, entonces el valor centro es el **8** y esa es tu mediana.

Para poder establecer una relación empírica entre media, mediana y moda hay que saber diferenciar las curvas de distribución de frecuencia de nuestros datos estadísticos de la siguiente forma:

Si la **curva de distribución es simétrica o bien formada**: es decir, si las observaciones tienen un equilibrio en sus frecuencias que van subiendo al respecto a sus frecuencias hasta llegar a una máxima y

después descienden las frecuencias los valores de **media, moda y mediana son el mismo.**

Si la **curva de distribución es asimétrica sesgada a la derecha:** si la cola mayor se presenta en la parte derecha de la curva de distribución de frecuencia se dice que está sesgada a la derecha, que tiene sesgo positivo y que su relación es: **media ≥ mediana ≥ moda.**

Si la **curva de distribución es asimétrica sesgada a la izquierda:** si la cola mayor se presenta en la parte izquierda de la curva de distribución de frecuencia se dice que está sesgada a la izquierda, que tiene sesgo negativo y que su relación es: **media ≤ mediana ≤ moda.**

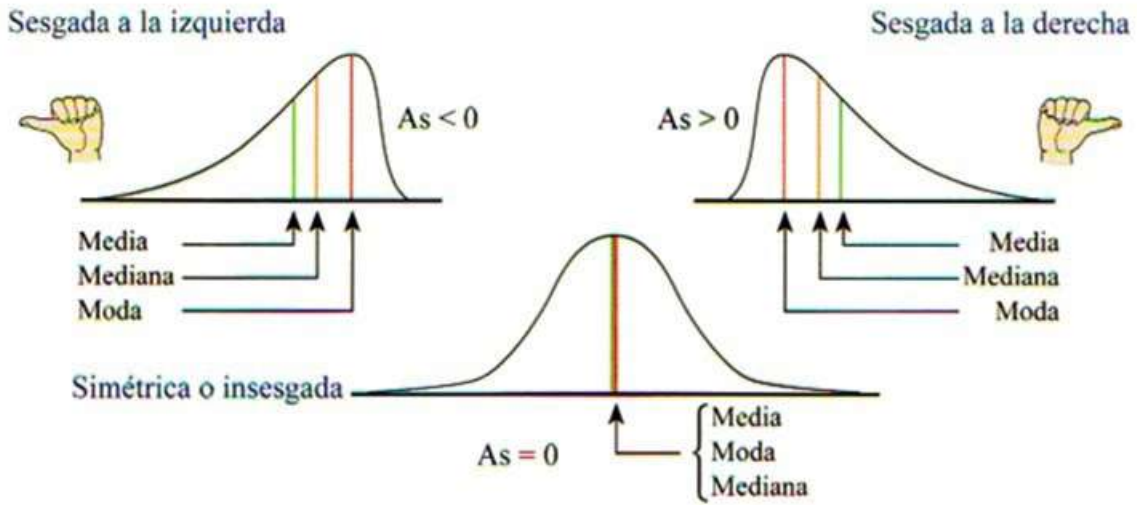


Figura 42. Relación entre la media, mediana y moda.

3.6.2. Medidas de dispersión

Nos permiten reconocer que tanto se dispersan los datos alrededor del punto central; es decir, nos indican cuánto se desvían las observaciones alrededor de su promedio aritmético (media). Son la varianza y desviación estándar o típica.

3.6.2.1. La Varianza

Es una medida de dispersión definida como la esperanza del cuadrado de la desviación de dicha variable respecto a su media. Su unidad de medida corresponde al cuadrado de la unidad de medida de la variable; por ejemplo, si la variable mide una distancia en metros, la varianza se expresa en metros al cuadrado. La varianza tiene como valor mínimo 0.

Dicha medida recibe el nombre de varianza o variancia. Se denota por los símbolos **S2** o **σ2** (letra griega sigma minúscula al cuadrado), al igual que con la media, la distinción entre ellos se hará importante en temas de estadística inferencial. De momento, usaremos el primero. Su cálculo para datos simples no agrupados se verifica según la fórmula:

$$S^2 = \frac{\sum_{i=1}^{n} (X_j - \overline{X})^2}{n-1} \qquad \sigma^2 = \frac{\Sigma(X-\mu)^2}{N}$$

Datos de muestra **Datos de población**

Este promedio es calculado elevando cada una de las diferencias al cuadrado (con el fin de eliminar los signos negativos), y calculando su promedio o media; es decir, sumado todos los cuadrados de las diferencias de cada valor respecto a la media y dividiendo este resultado por el número de observaciones que se tengan.

Esta medida logra describir adecuadamente la dispersión del conjunto de datos, pero tiene un inconveniente: su resultado se expresa en unidades cuadradas, algo harto engorrosas y difíciles de entender en la mayoría de las situaciones prácticas, y por demás disonante en relación con la medida de tendencia central utilizada. Sería algo así como años cuadrados, o pesos cuadrados (¿?).

A fin de eliminar este aparente escollo, puedes hallar la raíz cuadrada positiva del número obtenido, con lo que tendrás de vuelta a las unidades originales, obteniendo así una medida denominada **desviación**

típica o estándar, y es la medida de variación más ampliamente utilizada en el mundo de las estadísticas. Su símbolo es S (por ser la raíz cuadrada de la varianza), aunque se utiliza también DS (desviación standard) o SD (standard deviation). Tiene, además, la ventaja de que hasta las calculadoras de bolsillo —las científicas, claro está— la calculan, y casi la totalidad de los paquetes estadísticos existentes en el mercado del software.

3.6.2.2. Desviación estándar

Habíamos visto que la varianza transforma todas las distancias a valores positivos elevándolas al cuadrado, con el inconveniente de elevar consigo las unidades de los datos originales.

La desviación estándar nos da como resultado un valor numérico que representa el promedio de diferencia que hay entre los datos y la media. Para calcular la desviación estándar basta con hallar la raíz cuadrada de la varianza.

La **S** representa la desviación estándar de una **muestra**, mientras que **σ** la desviación para todos los datos de una **población**. Ampliando las fórmulas tenemos:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}}$$

Población

$$S = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n}}$$

Muestra

3.6.2.3. El error estándar

El error estándar es el término utilizado para referirse a una estimación de la desviación estándar, derivado de una muestra especial utilizada para calcular la estimación en las estadísticas. Es la más común, error estándar es un proceso de estimación de la desviación estándar de la distribución de muestreo asociada con el método de estimación.

Cada estadística tiene un error estándar asociado. Una medida de la precisión de la estadística puede deducir que el error estándar de 0 representa que la estadística tiene ningún error aleatorio y el más grande representa menos preciso de las estadísticas. Error estándar no es constantemente informados y no siempre fáciles de calcular.

Lo que a menudo no se logra apreciar totalmente es que las estadísticas también se comportan de una manera aleatoria, similar a la de las mediciones individuales, y esto se mide con el error estándar. Cuando se informa la media de una muestra, no se informa el promedio "verdadero" sino una estimación. La estadística muestral puede resultar levemente superior o inferior al valor verdadero desconocido. El error estándar de la media mide la diferencia que puede existir entre la media verdadera y la estadística que se informa. En términos más generales, podemos hablar del "error estándar de la estimación" cada vez que se informa una cantidad estadística estimada. Cuando se calcula un dato estadístico único, es posible calcular el error estándar de la estimación. En general, cuanto mayor sea el tamaño de la muestra, menor será el error estándar de una cantidad estimada.

El error estándar de la muestra representa la desviación estándar de las medias y permite conocer, exactamente, el probable campo de localización de la media de la población μ. Cuando los datos provienen de un censo, el error estándar de la media es igual a cero (Gutiérrez, 2017).

En un muestreo siempre existen errores en la estimación de la media que es un estimador de la media de la población.

Fórmula:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Datos de muestra **Datos de población**

A partir del error estándar se construye el intervalo de confianza de la medida correspondiente.

1) Para una media de colesterol total de 197, una desviación estándar de 38,43. El error estándar de la media estimado en la muestra del ejemplo es 3,05. Se calcula dividiendo la desviación estándar por la raíz cuadrada del tamaño muestral 38,43/√159=3,05
2) Calculado a partir de él, el intervalo de confianza al 95% para la media va desde 191,03 a 202,97

- **Límite inferior =** media - 1,96 veces el error estándar = 197 – 1,96 * 3,05 = 191,03 (límite inferior)
- **Límite superior =** media + 1,96 veces el error estándar = 197 + 1,96 * 3,05 = 202,97 (límite superior)

Este es uno de los métodos estadísticos que exige normalidad de la población. Quiere decir que podemos afirmar, con una confianza del 95%, que la media poblacional está incluida en dicho intervalo.

3.6.2.4. Coeficiente de variación %

En estadística, cuando se desea hacer referencia a la relación entre el tamaño de la media y la variabilidad de la variable, se utiliza el coeficiente de variación (suele representarse por las siglas “CV”).

Su fórmula expresa la desviación estándar como porcentaje de la media aritmética, mostrando una interpretación relativa del grado de variabilidad, independiente de la escala de la variable, a diferencia de la desviación típica o estándar. Por otro lado, presenta problemas ya que,

a diferencia de la desviación típica, este coeficiente es fuertemente sensible ante cambios de origen en la variable. Por ello es importante que todos los valores sean positivos y su media dé, por tanto, un valor positivo. A mayor valor del coeficiente de variación mayor heterogeneidad de los valores de la variable; y a menor CV, mayor homogeneidad en los valores de la variable.

Por ejemplo, si el **CV es menor o igual al 30%,** significa que la media aritmética es representativa del conjunto de datos, por ende, el conjunto de datos es **"homogéneo".** Por el contrario, si el **CV supera al 30%,** el promedio no será representativo del conjunto de datos (por lo que resultará **"heterogéneo"**).

El coeficiente de variación permite comparar la dispersión entre dos poblaciones distintas e, incluso, comparar la variación producto de dos variables diferentes (que pueden provenir de una misma población).

Coeficiente de variación (CV): Equivale a la razón entre la media aritmética y la desviación típica o estándar.

Ante casos como el descrito con anterioridad, es imprescindible contar con una medida de variabilidad relativa, como el **coeficiente de variación (CV)**, que expresa a la desviación típica como porcentaje de la media, y su cálculo se realiza mediante:

$$CV = \left| \frac{S}{X} \right| *100$$

Observa que, por tener la desviación estándar y la media las mismas unidades de medida, quedan canceladas dichas unidades, de ahí que el coeficiente de variación no tenga unidades propias, lo que facilita la comparación.

En el ejemplo siguiente, si comparas las desviaciones estándares de los dos grupos, pudieras creer que ambos tienen igual dispersión:

Grupo 1: X: 60 cm; S:4 cm = 4/60= 6,7%
Grupo 2: X: 170 cm; S:4 cm = 4/170= 2,4%

Al contrastarlos, ves algo bien diferente, pues en realidad **el grupo 1 tiene casi tres veces más dispersión que el grupo 2**.

3.6.3. Medidas de posición

Medidas de orden o posición (localización) informan sobre distintas características de los datos a partir de la ordenación de los valores observados. Las medidas de orden que estudiaremos son los percentiles y cuartiles.

Las medidas de posición se suelen dividir en dos grandes grupos: la de **tendencia no central y las centrales**. Las medidas de posición no centrales son **los cuantiles**. Estos realizan una serie de divisiones iguales en la distribución ordenada de los datos. De esta forma, reflejan los valores superiores, medios e inferiores.

Los más habituales son:

- **El cuartil:** Es uno de los más utilizados y divide la distribución en cuatro partes iguales. Así, existen tres cuartiles. Los valores inferiores de la distribución se sitúan por debajo del primero (Q1). La mitad o mediana son los valores menores iguales al cuartil dos (Q2) y los superiores son representados por el cuartil tres (Q3). La diferencia entre el tercer cuartil y el primero (Q3-Q1) se conoce como rango intercuartílico (RIC).

Se representa gráficamente como la anchura de las cajas en los llamados diagramas de cajas. El Q2 representa a la mediana.

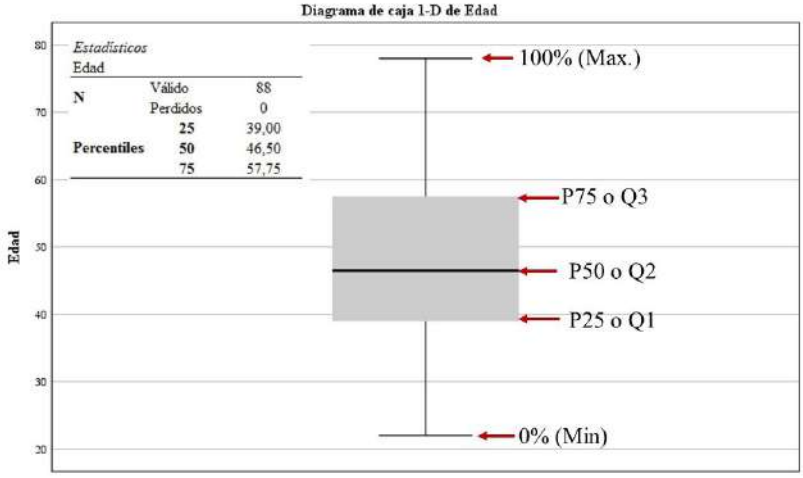


Figura 43. Representación de cuartiles en diagrama de cajas y bigotes.

- **El quintil:** En este caso, divide la distribución en cinco partes. Por tanto, hay cuatro quintiles. Además, no existe ningún valor que divida la distribución en dos partes iguales. Es menos frecuente que el anterior.
- **El decil**: Estamos ante un cuantil que divide los datos en diez partes iguales. Existen nueve deciles, de D1 a D9. El D5 se corresponde con la mediana. Por su lado, los valores superiores e inferiores (equivalentes a los diferentes cuartiles) se sitúan en puntos intermedios entre éstos.
- **El percentil:** Por último, este cuantil divide la distribución en cien partes. Hay 99 percentiles. Tiene, a su vez, una equivalencia con los deciles y cuartiles.

3.7. Medidas de frecuencia de una enfermedad

Para poder cuantificar el impacto que realmente tiene una enfermedad, en epidemiología se suele trabajar con diferentes tipos de fracciones:

a. **Proporción (probability):** Es un cociente donde el numerador está incluido en el denominador. No tiene dimensión, sus valores van de 0 a 1 y suele expresarse en porcentajes. Por ejemplo, si en una población de 10.000 habitantes se diagnostican 100

casos de tuberculosis pulmonar, la proporción de tuberculosis pulmonar en esa población será de 100/10.000 = 0,01 (1%).

b. **Razón (ratio):** Se obtiene de dividir dos cantidades donde el numerador no está contenido en el denominador. Puede tener o no dimensiones. En el ejemplo anterior, la razón entre la población con infección por tuberculosis pulmonar y la no infectada es de 100/9.900 = 0,01 Cuando se calcula la probabilidad de que ocurra un evento entre la probabilidad que no ocurra se le denomina *Odds*. En el ejemplo anterior, la Odds es de 0,01 e indica que por cada 1/0,01 = 100 pacientes que no tienen tuberculosis pulmonar hay 1 que sí la tiene. En los siguientes capítulos se abordará factores de riesgo.
c. **Tasa (rate):** Es similar a la proporción, con la diferencia de que las tasas llevan incorporado el concepto de tiempo. Hace referencia al ritmo con que aparecen nuevos eventos en un grupo de individuos a medida que transcurre el tiempo. El numerador lo constituye la frecuencia absoluta de casos del problema a estudiar y el denominador lo forma la suma de los periodos individuales de riesgo a los que han estado expuestos los sujetos susceptibles de la población en estudio. Tiene dimensiones (inversa del tiempo o tiempo^{-}1) y un rango de 0 a infinito. Por ejemplo, si de 100 pacientes con infección por tuberculosis pulmonar se controlan 80 en un año, tendríamos 80/100*1 = 80 por 100 pacientes/año se controlan al tomar la medicación (Gutiérrez, 2017).

En epidemiología las medidas de frecuencia de enfermedad más comúnmente utilizadas son la prevalencia e incidencia.

a. Prevalencia: Hace referencia a la proporción de individuos de una población que presentan una enfermedad en un momento determinado.

$$P = \frac{\text{N.° de casos con la enfermedad en un momento dado}}{\text{Total de la población en ese momento}}$$

La prevalencia no tiene dimensión y sus valores oscilan entre 0 y 1.

Por ejemplo si en una población de 10.000 habitantes existen 100 pacientes con infección por tuberculosis, la prevalencia de la enfermedad en esa población será: P = 100/10.000 = 0,01 = 1%

La prevalencia de período se calcula como la proporción de personas que han presentado la enfermedad en alg**ún momento a lo largo de un perí**odo de tiempo determinado, por ejemplo, la prevalencia de tuberculosis en los últimos 5 años.

Incidencia: Hace referencia al número de casos nuevos de una enfermedad que se desarrollan en una población en un tiempo determinado. Son de especial interés para identificar factores de riesgo o factores pronósticos. Son la incidencia acumulada y la tasa de incidencia.

a. **Incidencia acumulada (IA):** Es la proporción de individuos sanos que desarrollan la enfermedad en un periodo de tiempo concreto.

$$IA = \frac{\text{N.° de casos nuevos de enfermedad durante el seguimiento}}{\text{Total de la población en riesgo al inicio del seguimiento}}$$

Por ejemplo: Durante 5 años se siguió a una población de 600 varones con prácticas de riesgo para adquirir la infección por el VIH, al final del seguimiento 120 pacientes adquirieron la infección. La incidencia acumulada sería: **IA =** 120 / 600= 0,20 **= 20%** en 5 años.

La IA asume que la población entera a riesgo al principio del estudio ha sido seguida durante todo el período de tiempo, pero esto no siempre sucede puesto que hay sujetos que se pierden durante el seguimiento, por ello es más correcto utilizar la IA actuarial que tiene en cuenta a los sujetos perdidos:

IA actuarial=	N.° de eventos		
	Sujetos a riesgo	-	Sujetos perdidos
			2

IA actuarial=	120		
	600	-	50
			2
IA actuarial=	120		
	600	-	25
IA actuarial=	120		
	575		
IA actuarial=	20,9%		

Aquí, el denominador es el "número efectivo" de personas en riesgo, y asume que la media de abandonos ocurre en la mitad de tiempo de seguimiento.

b. Tasa de incidencia (TI). También denominada densidad de incidencia (DI). Es el cociente entre el número de casos nuevos de una enfermedad ocurridos durante el periodo de seguimiento y la suma de todos los tiempos individuales de observación.

TI=	N.° de casos nuevos durante el tiempo de seguimiento
	Suma de los tiempos individuales de observación

Su valor no puede ser inferior a cero, pero no tiene límite superior. Es muy útil para estudiar las hipótesis etiológicas de las enfermedades con periodos de latencia largos.

Por ejemplo, supongamos que tenemos 6 pacientes con tuberculosis y los seguimos durante 48 semanas y queremos calcular la TI de curación tras inicio del tratamiento tuberculostático y observamos que el paciente 1 se ha curado a las 24 semanas, el 2 no se ha curado, el 3 lo ha hecho a las 16 semanas, el 4 a las 32 semanas, el 5 a las 24 semanas y el 6 no se ha curado. La incidencia de curaciones sería 4/6= 0,66, es decir, en 48 semanas se han curado el 66%.
En cambio la **TI** sería 4/(24+48+16+32+24+48) **= 4/192** = 0,02 pacientes por semana. Es decir, la velocidad de curación de la infección **sería del 2%** de control/semana.

3.8. Análisis descriptivo con SPSS

LA primera tarea después de contar con los datos que se recolectaron para la investigación, es describirlos en cada una de las variables del estudio, y hacerlo mediante el SPSS resulta bastante sencillo.
En SPSS se puede hacer de varias formas.

1. Con la opción de frecuencias se requiere seguir la siguiente ruta:

- Analizar
- Estadísticos descriptivos
- Frecuencias

En la ventana de frecuencias introducir la/s variable/s que se quiere analizar. Puede ser cuantitativa o categórica, por ejemplo, la edad, sexo, peso, altura, IMC, etc.

En **"Estadísticos",** señalar las pestañas que nos interese, por ejemplo cuartiles, percentiles, medidas de tendencia central (media, mediana, moda y suma), de dispersión (desviación típica, varianza, etc.) o de distribución (asimetría y curtosis).

En **"Gráficos",** podemos señalar el tipo de gráfico, por ejemplo de barras o sectores para una variable cualitativa (ejemplo, el sexo o la raza) y el histograma para una variable cuantitativa continuar (por ejemplo la edad, talla, peso, IMC, etc.). En este último podemos indicar que necesitamos que aparezca la curva de normalidad. Además, podemos indicar si queremos que aparezcan las frecuencias o porcentajes.

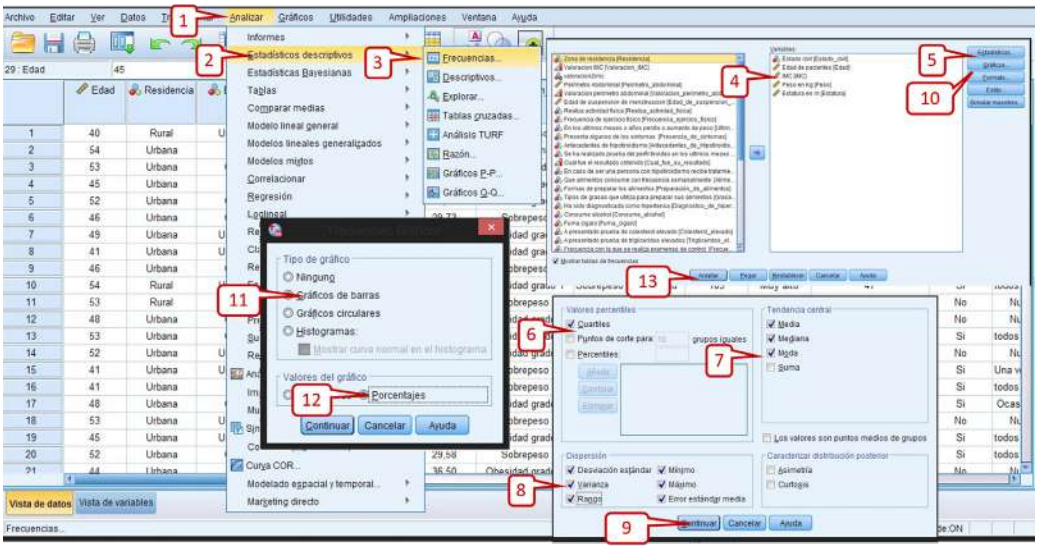


Figura 44. Estadística descriptiva con opción frecuencia.

Estadísticos					
		Edad de pacientes	IMC	Peso en kg	Estatura en m
N	Válido	159	159	159	159
	Perdidos	0	0	0	0
Media		47,69	31,37	77,42	1,57
Error estándar de la media		0,395	0,337	0,898	0,004
Mediana		47,00	31,07	80,00	1,57
Moda		53	34,17	80,00	1,56
Desv. Desviación		4,98	4,25	11,32	0,05
Varianza		24,76	18,04	128,23	0,002
Rango		15	19,68	49,00	0,20
Mínimo		40	22,06	50,00	1,45

Máximo		55	41,74	99,00	1,65
	25	44,0	28,4	70,0	1,5
Percentiles	**50**	47,0	31,1	80,0	1,6
	75	53,0	34,3	85,0	1,6

En esta tabla se representan los estadísticos que hemos elegido. En este caso, la edad media de esta muestra de 159 pacientes es de 47,69 ± 4,98 años, el pacientes de mayor edad es 55 años y podemos decir que el 75% (P75) tienen 53 años o menos.

Edad de pacientes				
Edad en años	**Frecuencia**	**Porcentaje**	**Porcentaje válido**	**Porcentaje acumulado**
40	9	5,7	5,7	5,7
41	19	11,9	11,9	17,6
42	3	1,9	1,9	19,5
43	3	1,9	1,9	21,4
44	18	11,3	11,3	32,7
45	13	8,2	8,2	40,9
46	10	6,3	6,3	47,2
47	6	3,8	3,8	50,9
48	9	5,7	5,7	56,6
49	10	6,3	6,3	62,9
50	3	1,9	1,9	64,8
52	10	6,3	6,3	71,1
53	20	12,6	12,6	83,6
54	13	8,2	8,2	91,8
55	13	8,2	8,2	**100,0**
Total	**159**	**100,0**	**100,0**	

En esta tabla se representan todos los valores de la variable, la frecuencia y el porcentaje con que aparece. Por ejemplo, en el recuadro, 3 (porcentaje) pacientes (12,6%) tenían 53 años y el 83,6% tiene 53 años y menos (porcentaje acumulado).

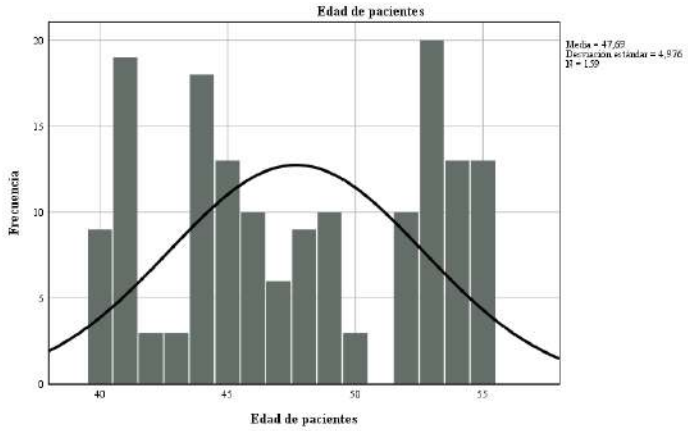


Figura 45. Histograma de la edad con curva de normalidad.

2. Con la opción de Descriptivos: Hacer clic en:

- Analizar
- Estadísticos descriptivos
- Descriptivos

En la ventana de Descriptivos introducir la variable que queremos analizar, por ejemplo, la edad. En opciones elegir el estadístico que nos interese (aquí tenemos menos opciones que con el apartado anterior), por ejemplo media, desviación típica, etc.

3. Con la opción de Explorar: Esta opción permite hacer estratificaciones, por ejemplo distribuir la edad por el sexo (hombre/mujer). Hacer clic en:

- Analizar
- Estadísticos descriptivos
- Explorar

En la ventana de **"Explorar"** añadir en lista de dependientes la variable cuantitativa continua que se quiere analizar (por ejemplo la edad) y en lista de factores la variable por la que se quiere estratificar (por ejemplo por el sexo).

En **“Estadísticos”** señalar por ejemplo Descriptivos (son la media, mediana, Desv. típica, máximo, mínimo, rango, amplitud intercuartílico, asimetría y curtosis) y percentiles (son el P5, P10, P25, P75, P90 y P95). En **“Gráficos”** se puede elegir diagramas de caja (factores juntos o dependientes juntos), descriptivos (tallo y hojas e histograma) y dispersión por nivel con prueba de Levane. Además se puede obtener los **gráficos con pruebas de normalidad**.

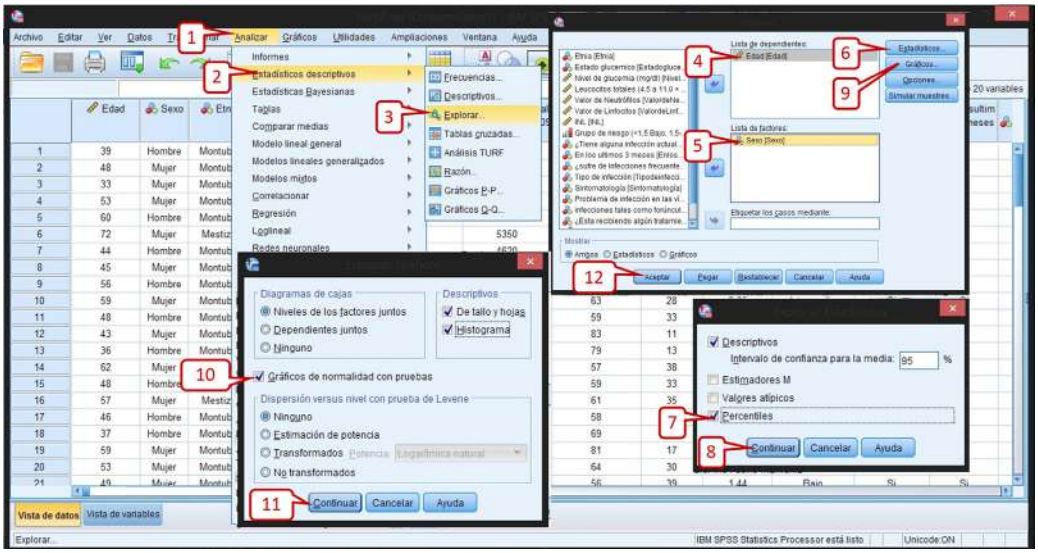


Figura 46. Estadística descriptiva con opción explorar.

En la hoja de resultados obtenemos una tabla con el número de casos válidos por cada factor, cada uno de los estadísticos de la variable distribuidos por cada factor (sexo) y los percentiles por cada factor.

Descriptivos					
	Sexo			Estadístico	Desv. Error
Edad	Hombre	Media		47,16	2,21
		95% de intervalo de confianza para la media	Límite inferior	42,65	
			Límite superior	51,68	
		Media recortada al 5%		46,64	
		Mediana		45,00	
		Varianza		151,473	
		Desv. Desviación		12,307	
		Mínimo		25	
		Máximo		78	
		Rango		53	
		Rango intercuartil		17	
	Mujer	Media		47,61	1,65
		95% de intervalo de confianza para la media	Límite inferior	44,31	
			Límite superior	50,91	
		Media recortada al 5%		47,68	
		Mediana		48,00	
		Varianza		154,706	
		Desv. Desviación		12,438	
		Mínimo		22	
		Máximo		72	
		Rango		50	
		Rango intercuartil		19	

Tabla en la que se representan cada uno de los estadísticos de la variable por cada factor (sexo). En este ejemplo se representan las edades medias, mediana, Desv. típica, etc., de cada hombre y mujer.

Percentiles									
Sexo			Percentiles						
			5	10	25	50	75	90	95
Promedio ponderado	Edad	Hombre	29,2	36,0	39,0	45,0	56,0	67,00	75,00
		Mujer	25,7	28,0	39,5	48,0	58,0	62,00	67,50
Bisagras de Tukey	Edad	Hombre			39,0	45,0	55,0		
		Mujer			41,0	48,0	58,0		

Tabla en la que se representan los percentiles de la variable distribuidos por cada factor, en este caso los percentiles de la edad por varones y mujeres.

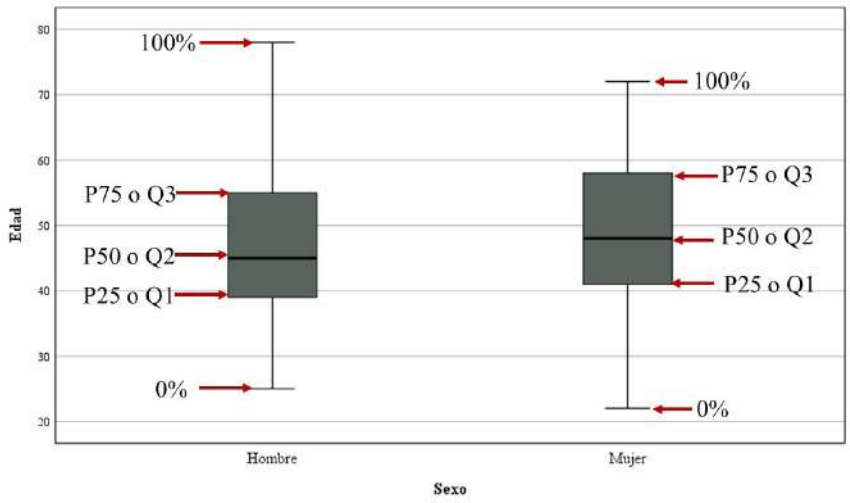


Figura 47. Gráfico de cajas por cada factor. Se representa las cajas y bigotes de la edad distribuido por el sexo.

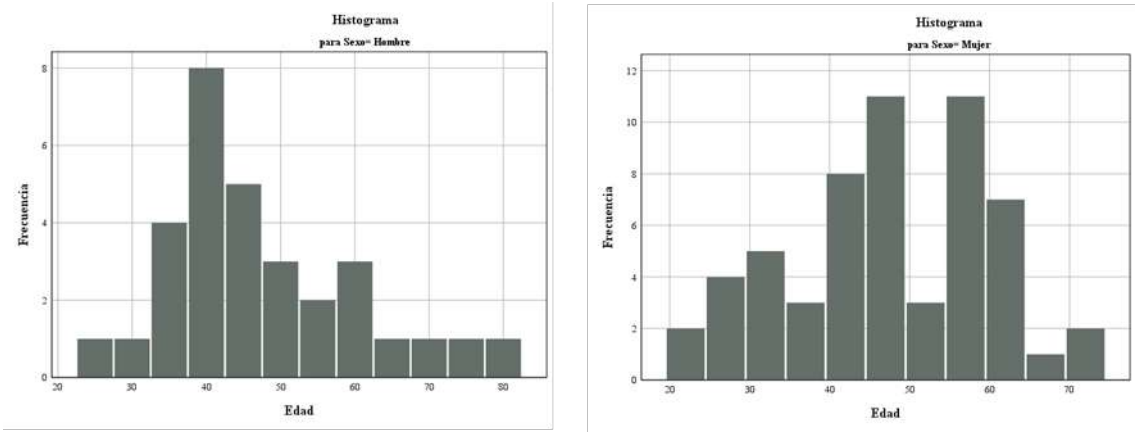


Figura 48. Histograma de la edad distribuida por sexo (varón y mujer).

Edad Diagrama de tallo y hojas de Sexo= Hombre

Frecuencia	Stem &	Hoja
1,00	2 .	5
10,00	3 .	2667789999
10,00	4 .	0024556788
5,00	5 .	14689
3,00	6 .	038
2,00	7 .	38

Edad Diagrama de tallo y hojas de Sexo= Mujer

Frecuencia	Stem &	Hoja
6,00	2 .	236788
8,00	3 .	22334668
19,00	4 .	1123334455567788889
14,00	5 .	33355567788999
8,00	6 .	00122247
2,00	7 .	22

Gráfico de tallo y hojas. En el primer ejemplo hay 2 hombres que tienen 73 y 78 años; en el segundo ejemplo 2 mujeres que tienen 72 años.

Pruebas de normalidad

Sexo		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Estadístico	gl	Sig.	Estadístico	gl	Sig.
Edad	Hombre	0,150	31	0,072	0,936	31	0,065
	Mujer	0,092	57	,200*	0,975	57	0,282

*. Esto es un límite inferior de la significación verdadera.

a. Corrección de significación de Lilliefors

Test estadísticos que podemos utilizar para este propósito. El test de Kolmogorov- Smirnov es el más extendido en la práctica (>50 datos). La otra prueba que se utiliza para determinar la normalidad es la de

Shapiro-Wilk, que es una de las más potentes, sobre todo en poblaciones pequeñas (<50 datos). Los datos nos muestran una **significación >0,05** aceptamos H0 (hipótesis nula) **los datos tienen distribución normal,** tanto para la edad de mujeres y hombres.

3.9. Análisis de normalidad de datos en SPSS

La distribución normal (en ocasiones llamada distribución gaussiana) es la distribución continua que se utiliza más comúnmente en estadística, es un modelo que aproxima el valor de una variable aleatoria a una situación ideal, dependiendo de la media y la desviación estándar.

- Muchos fenómenos que podemos medir tanto en las ciencias exactas como las sociales de asemejan en su frecuencia a esta distribución.
- La distribución normal tiene ciertas propiedades matemáticas que nos permiten predecir qué proporción de la población (estadística) caerá dentro de cierto rango si la variable tiene distribución normal.
- Varios tests de significancia de diferencia entre conjuntos de datos presumen que los datos del conjunto tienen una distribución normal.

Propiedades de la distribución normal:

1. Tiene una única moda, que coincide con su media y su mediana.
2. Es simétrica con respecto a su media. Según esto, para este tipo de variables existe una probabilidad de un 50% de observar un dato mayor que la media, y un 50% de observar un dato menor.
3. La distancia entre la línea trazada en la media y el punto de inflexión de la curva es igual a una desviación típica. Cuanto mayor sea, más aplanada será la curva de la densidad.
4. El área bajo la curva comprendida entre los valores situados aproximadamente a dos desviaciones estándar de la media es igual a 0,95. En concreto, existe un 95% de posibilidades de observar un valor comprendido en el intervalo.
5. La forma de la campana de Gauss depende de los parámetros

media y desviación estándar.

6. La media indica la posición de la campana, de modo que para diferentes valores de media la gráfica es desplazada a lo largo del eje horizontal. Por otra parte, la desviación estándar determina el grado de apuntamiento de la curva. Cuanto mayor sea el valor de la desviación típica, más se dispersarán los datos en torno a la media y la curva será más plana. Un valor pequeño de este parámetro indica, por tanto, una gran probabilidad de obtener datos cercanos al valor medio de la distribución (Del Campo & Matamoros, 2019).
7. La proporción de mediciones situada entre la media y las desviaciones es una constante en la que:

a. La media ± 1 * desviación estándar = cubre el 68,26% de los casos
b. La media ± 2 * desviación estándar = cubre el 95,44% de los casos
c. La media ± 3 * desviación estándar = cubre el 99,7% de los casos

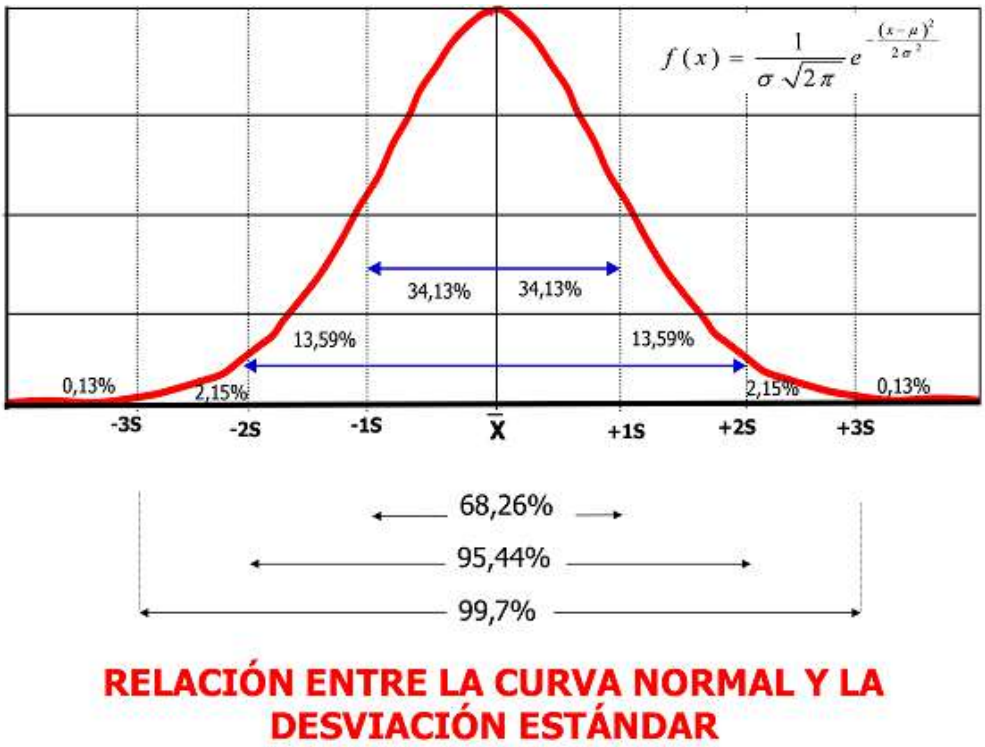


Figura 49. Relación entre la curva normal y la desviación estándar.

Gráfico de normalidad

Los gráficos de probabilidad normal permiten comprobar si un conjunto de datos puede considerarse o no procedente de una distribución normal. En un mismo gráfico se enfrentan los datos que han sido observados frente a los datos teóricos que se obtendrían de una distribución normal. Si la distribución de la variable coincide con la distribución gaussiana, los puntos tenderán a concentrarse en torno a una línea recta, teniendo en cuenta que siempre habrá una mayor variabilidad en los extremos.

Los gráficos Q-Q se obtienen de modo análogo, esta vez representando los cuantiles respecto a los cuantiles de la distribución normal. Además de permitir valorar la desviación de la normalidad, los gráficos de probabilidad permiten conocer la causa de esa desviación.

Una curva en forma de "U" o con alguna curvatura, significa que la distribución es asimétrica con respecto a la gaussiana, mientras que un gráfico en forma de "S" significará que la distribución tiene colas mayores o menores que la normal, esto es, que existen pocas o demasiadas observaciones en las colas de la distribución.

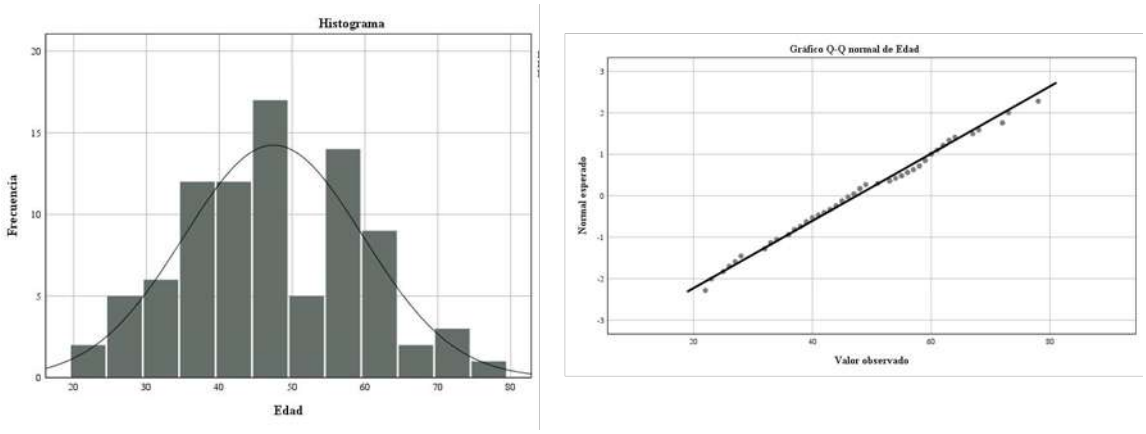


Figura 50. Gráfico Q-Q e histograma con curva de normalidad de una variable con distribución normal.

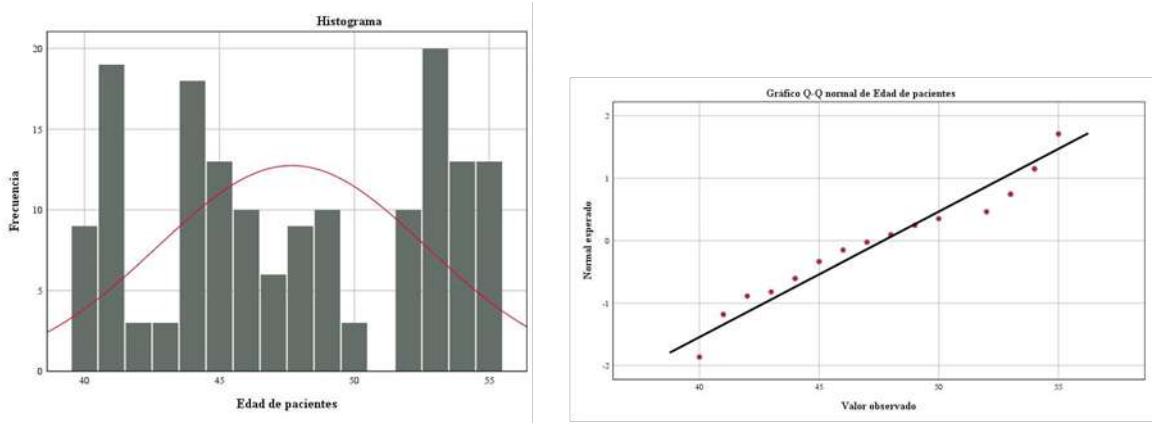


Figura 51. Gráfico Q-Q e histograma con curva de normalidad de una variable sin una distribución normal.

Parece lógico que cada uno de estos métodos se complemente con procedimientos de análisis que cuantifiquen de un modo más exacto las desviaciones de la distribución normal. Existen distintos test estadísticos que podemos utilizar para este propósito. El test de Kolmogorov-Smirnov es el más extendido en la práctica. Se basa en la idea de comparar la función de distribución acumulada de los datos observados con la de una distribución normal, midiendo la máxima distancia entre ambas curvas. Como en cualquier test de hipótesis, la hipótesis nula se rechaza cuando el valor del estadístico supera un cierto valor crítico que se obtiene de una tabla de probabilidad. Cuando se dispone de un número suficiente de datos, cualquier test será capaz de detectar diferencias pequeñas aun cuando éstas no sean relevantes para la mayor parte de los propósitos (Del Campo & Matamoros, 2019).

El test de Kolmogorov-Smirnov, en este sentido, otorga un peso menor a las observaciones extremas y por la tanto es menos sensible a las desviaciones que normalmente se producen en estos tramos. La otra prueba que se utiliza para determinar la normalidad es la de Shapiro-Wilk, que es una de las más potentes, sobre todo en poblaciones pequeñas (<50).

En general, utilizaremos la prueba de **Kolmogorov-Smirnov si hay más de 50 unidades** de análisis o la de **Shapiro-Wilk si hay menos de 50 unidades** de análisis.

Prueba de Kolmogorov-Smirnov con SPSS

Definir en la hoja de recogida de datos la variable cuantitativa continua que se quiere analizar. En la hoja de recogida de datos, hacer clic en:

- Analizar
- Pruebas no paramétricas
- Cuadro de diálogos antiguos
- K-S de 1 muestra.

En la ventana de prueba de Kolmogorov-Smirnov para una muestra introducir la variable prueba (por ejemplo, la edad) en la casilla de lista contrastar variables y hacer clic en aceptar.

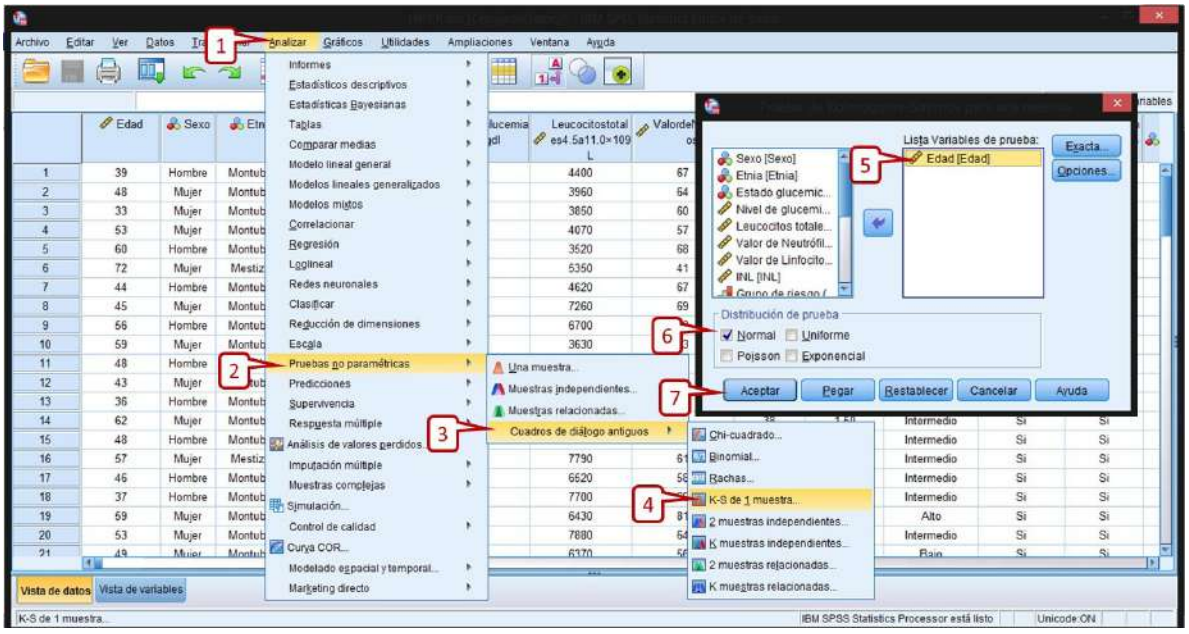


Figura 52. Prueba de Kolmogorov-Smirnov para una muestra en SPSS.

En la hoja de resultados obtenemos una tabla en la que se representan el número (N), la media, desviación típica, las diferencias más extremas, el valor de Z de Kolmogorov-Smirnov y la significación es-

tadística. En esta prueba se considera la H (hipótesis nula) como que la distribución es normal. Por lo tanto, si **la significación estadística es ≥ 0,05 no podremos rechazar la hipótesis nula** y tendremos que decir que **la variable tiene una distribución NORMAL.**

Prueba de Kolmogorov-Smirnov para una muestra		
		Edad
N		**88**
Parámetros normales[a,b]	Media	47,45
	Desv. Desviación	12,32
Estadístico de prueba		0,085
Sig. asintótica(bilateral)		,164[c]

a. La distribución de prueba es normal.

b. Se calcula a partir de datos.

c. Corrección de significación de Lilliefors.

Prueba de normalidad mediante la opción "explorar" con SPSS

También podemos averiguar la normalidad de una variable distribuida por un factor, por Ejemplo, la edad distribuida por el sexo (entre los hombres y mujeres). Para ello, hacer clic en:

- Analizar
- Estadísticos descriptivos
- Explorar

En la ventana de Explorar introducir en lista de dependientes la variable a analizar (por ejemplo la edad) y en lista de factores las variables por las que se quiere distribuir la variable dependiente (por ejemplo el sexo). En gráficos señalar la pestaña "Gráficos de normalidad con pruebas" y hacer clic en Continuar.

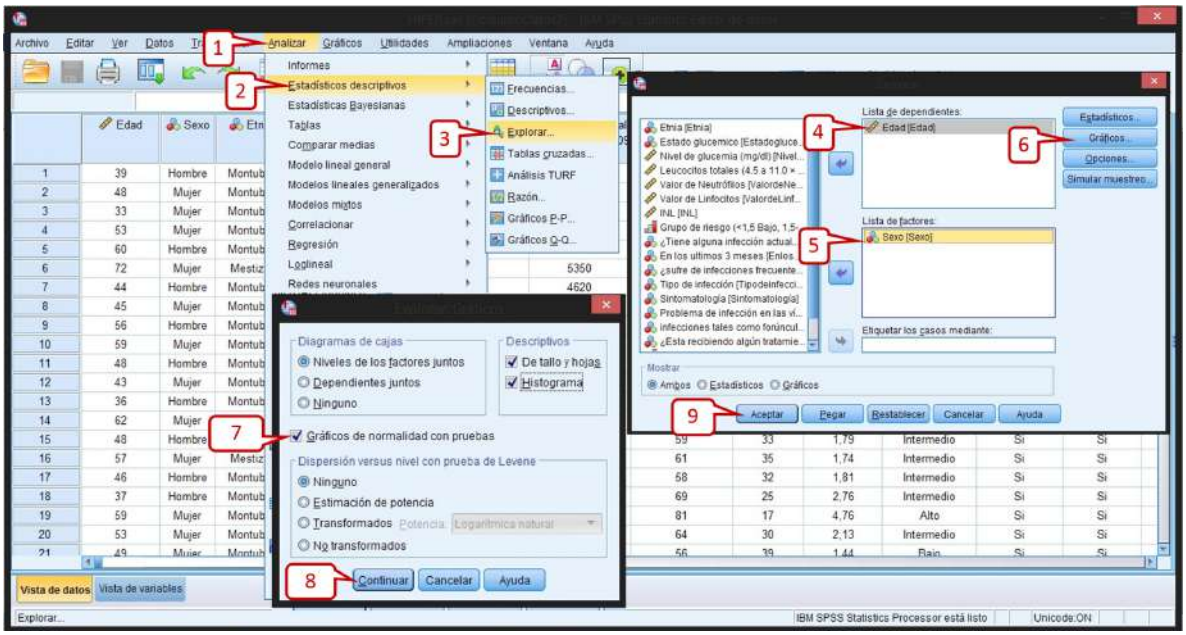


Figura 53. Prueba de normalidad con opción explorar en SPSS.

En la ventana de resultados aparece una tabla con las "Pruebas de normalidad". La prueba de **Kolmogorov-Smirnov se debe utilizar > 50 unidades** de análisis. Se considera que **tiene una distribución NORMAL si la Sig. es ≥ 0,05**. En cambio, la prueba de **Shapiro-Wilk se debe utilizar < 50 unidades** de análisis. Al igual que la anterior se considera que **tiene una distribución NORMAL si la Sig. es ≥ 0,05.**

Pruebas de normalidad							
Sexo		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Estadístico	gl	Sig.	Estadístico	gl	Sig.
Edad	Hombre	0,150	31	0,072	0,936	31	0,065
	Mujer	0,092	57	,200*	0,975	57	0,282
*. Esto es un límite inferior de la significación verdadera.							
a. Corrección de significación de Lilliefors							

Tabla con las pruebas de normalidad de la edad distribuida por sexo. En este ejemplo **la edad tiene una distribución normal** tanto en hombres como en mujeres, para las **mujeres hemos utilizado la prueba de Kolmogorov-Smirnov** al tener una **n > 50** (p=0,20) y **para los hombres la de Shapiro-Wilk** al ser una **n < 50** (p=0,065).

1RA Edición

BIOESTADÍSTICA

APLICADA A INVESTIGACIONES
CIENTÍFICAS EN SALUD

CAPÍTULO IV

ANÁLISIS INFERENCIAL
PARA INVESTIGACIONES

MAWIL
Publicaciones Impresas
y Digitales

4.1. Inferencia estadística

En la mayoría de las investigaciones realizadas en el ámbito de salud, se realizan estudios comparativos entre dos o más muestras confrontando, la mayoría de ellos, el efecto producido por terapias o tratamientos. En este último caso, algunos estudios enmascaran el placebo de fármaco activo, aunque esta práctica ha sido objeto de amplio debate por plantear dudas sobre su ética. La finalidad de estas investigaciones es contestar preguntas tales como: ¿es igual el tratamiento A al tratamiento B? ¿Cuál es la efectividad del tratamiento? En estos casos es cuando necesitamos evaluar si las diferencias que se obtienen a partir de una muestra, se deben a factores distintos al azar y están directamente relacionadas con la administración de un tratamiento u otro.

Para conocer en qué se basan este tipo de estudios, deberemos introducir conceptos como las pruebas de hipótesis y los errores asociados a ellas. Además, veremos que esta probabilidad puede ser calculada a partir de pruebas estadísticas paramétricas, en las que se supone la normalidad de los datos, y las no paramétricas, usadas en condiciones no idóneas de normalidad.

La inferencia se define como el conjunto de métodos estadísticos que permiten deducir cómo se distribuye la población e inferir las relaciones entre variables a partir de la información que proporciona la muestra recogida. Por tanto, los **objetivos fundamentales** de la inferencia estadística son **la estimación y el contraste de hipótesis**.

Para que un método de inferencia estadística proporcione buenos resultados debe basarse en una técnica matemática (estadística) adecuada al problema planteado, además la muestra seleccionada debe ser representativa de la población y de tamaño suficiente.

El análisis de datos a nivel inferencial busca obtener conclusiones sólidas y más profundas que una simple descripción de la información, basados en el trabajo con muestras y su posterior generalización. Se

obtiene información sobre una población de datos mediante el estudio de una muestra de los mismos. Las muestras deben reflejar el verdadero comportamiento de las poblaciones.

La estadística inferencial estima cómo serían los resultados de la población objetivo si fuéramos capaces de estudiar a todos sus individuos. Para ello, extrae conclusiones a partir de los resultados obtenidos en la muestra, por lo que existirá una probabilidad de error (Monterrey & Gómez-Restrepo, 2007).

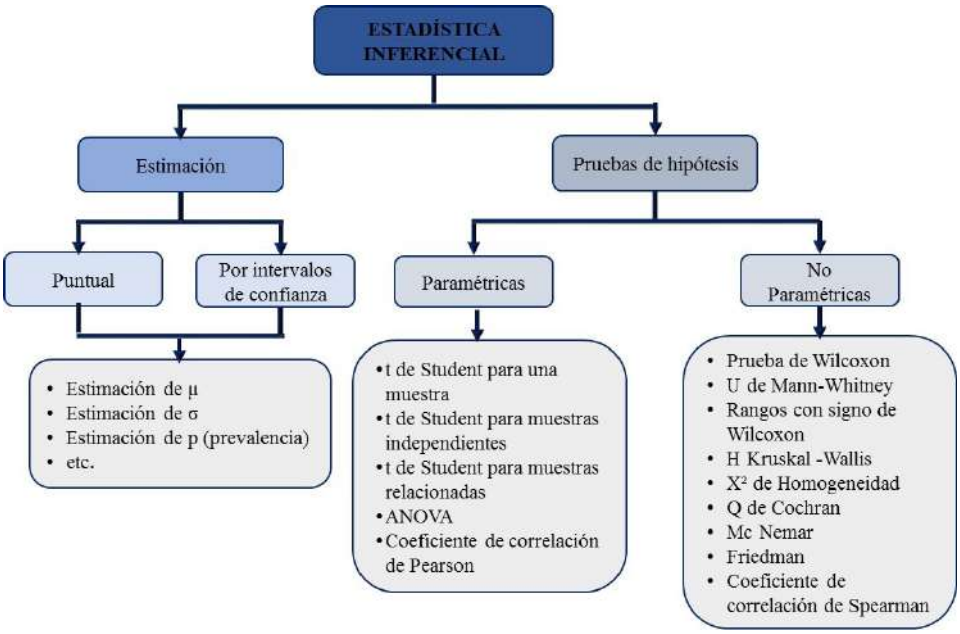


Figura 54. Principales de pruebas de estadística inferencial.

4.2. Bases para la elección de una prueba estadística

El primer aspecto para la selección de una prueba estadística es el diseño de investigación. En los capítulos anteriores se describen los diferentes diseños de investigación, se describe que cuando solamente existe un grupo y el objetivo de la investigación únicamente es especificar una o más características de dicha población, el tipo de estudio se denomina descriptivo, por lo que, como su nombre lo indica, solo es necesario emplear estadística descriptiva.

Por su parte, los estudios correlacionales pretenden responder a preguntas de investigación como las siguientes: ¿la obesidad en adultos mayores de 60 años está vinculada a un mayor riesgo de padecer diabetes? (Hernández Sampieri & Mendoza Torres, 2018).

En ocasiones solo se analiza la relación entre dos conceptos o variables, pero con frecuencia se ubican en el estudio vinculaciones entre tres, cuatro o más variables. Los estudios correlacionales, al evaluar el grado de asociación entre las variables, primero miden cada una de ellas (presuntamente relacionadas) y las describen, y después cuantifican y analizan la vinculación.

La utilidad principal de los estudios correlaciónales es saber cómo se puede comportar un concepto o una variable al conocer el comportamiento de otras variables vinculadas. Las correlaciones pueden ser positivas (directamente proporcionales) o negativas (inversamente proporcionales). Si es positiva, significa que los casos que muestren altos valores en una variable tenderán también a manifestar valores elevados en la otra variable. Si es negativa, implica que casos con valores elevados en una variable tenderán a mostrar valores bajos en la otra variable (Hernández Sampieri & Mendoza Torres, 2018).

Número de mediciones

El segundo aspecto por considerar en la selección de una prueba estadística es el número de mediciones de las variables de resultado. Los investigadores pueden realizar una **sola medición en un momento dado** (*estudios transversales*) o analizar de diferentes formas los cambios de una variable a lo largo de un período (*estudios longitudinales*).

Por ejemplo, la respuesta al tratamiento para hipertensión arterial después de 4 meses en pacientes adultos mayores se puede efectuar con un análisis del valor inicial previo al tratamiento y el valor final obtenido a los 3 meses, o bien, evaluar el cambio mensual durante los 3 meses.

Escala de medición de las variables

El tercer aspecto trascendente, cuando se planea un análisis estadístico, es la escala de medición de las variables, la cual ya ha sido explicada en otros capítulos de este documento. En resumen, es necesario definir la naturaleza de cada uno de los datos o las mediciones que se realizan durante el desarrollo de una investigación; en general se pueden dividir en *cualitativos o cuantitativos*. A su vez, las variables **cualitativas** se clasifican en *nominales y ordinales*; las **nominales** agrupan las características similares entre sí en las que no hay diferencia entre una y otra, tales como el sexo (hombre/mujer) o el estado civil (soltero/casado/unión libre). Por su parte, las variables cualitativas **ordinales** ya tienen cierta dimensión, como el estadio o gravedad de una enfermedad (leve/moderada/grave).

Por su parte, las variables cuantitativas pueden ser de 2 tipos: *cuantitativas continuas y cuantitativas discretas.* Las cuantitativas **continuas:** kilogramos de peso corporal, estatura en centímetros, mililitros de orina y edad de una persona. Las **discretas** con variables numéricas que de alguna manera no se pueden dividir (número de hijos, número de embarazos, etcétera).

Objetivos de la investigación

Si el estudio es de nivel correlacional, entonces todo el análisis estadístico es bivariado, relaciona a todas las variables respecto de una variable dependiente, algunos ejemplos:

Cuadro 9. Relación de objetivos de investigación correlacional y pruebas estadísticas.

Objetivo	Variable cualitativa	Variable cuantitativa	
		Con distribución normal	Sin distribución normal
Comparar	Chi-cuadrado	t de Student para muestras independientes	U de Mann-Whitney
Asociar, comparar	McNemar	t de Student para muestras dependientes	Rangos de Wilcoxon
Correlacionar	Coeficiente de correlación de Spearman	Coeficiente de correlación de Pearson	Coeficiente de correlación de Spearman

Elección de la prueba estadística

En el cuadro 9 se resume la manera de selección de las pruebas estadísticas, tomando en cuenta el objetivo, número de grupos y la escala de medición de las variables. Como parte del análisis global de los datos nunca debe omitirse la inclusión del análisis descriptivo de los datos, es decir, es necesario que los investigadores resuman cada una de las variables estudiadas en medidas de tendencia central y de dispersión, tomando en cuenta la escala de medición de las variables y su distribución (Gómez-Gómez ***et al***., 2013).

Cuadro 10. Pruebas estadísticas más utilizadas de acuerdo al tipo de variable.

TIPO DE ESTUDIO	VARIABLE FIJA	VARIABLE ALEATORIA				
		PRUEBAS NO PARAMÉTRICAS			NUMÉRICA	
		NOMINAL DICOTÓMICA	NOMINAL POLITÓMICA	ORDINAL	Con distribución normal	Sin distribución normal
					Paramétrica	No paramétrica
Estudio transversal	1 grupo	X^2 Bondad de ajuste binomial	X^2 Bondad de ajuste	X^2 Bondad de ajuste	t de Student para una muestra	Prueba de Wilcoxon
Muestras independientes	2 grupos	X^2 de homogeneidad	X^2 de homogeneidad	U de Mann-Whitney	t de Student para muestras independientes	U de Mann-Whitney
	Más de 2 grupos	X^2 de homogeneidad	X^2 de homogeneidad	H Kruskal-Wallis	ANOVA de un factor	H Kruskal -Wallis
Estudio longitudinal	2 medidas	McNemar	Q de Cochran	Rangos con signo de Wilcoxon	t de Student para muestras relacionadas	Rangos con signo de Wilcoxon
Muestras relacionadas	Más de 2 medidas	Q de Cochran	Q de Cochran	Friedman	ANOVA para medias repetidas	H Kruskal-Wallis
	Correlación entre dos variables			Coeficiente de correlación de Spearman	Coeficiente de correlación de Pearson	Coeficiente de correlación de Spearman

4.3. Prueba o contraste de hipótesis

La estadística inferencial recoge bajo el título genérico de prueba o contraste de hipótesis. Implica, en cualquier investigación, la existencia de dos teorías o hipótesis implícitas, que denominaremos hipótesis nula e hipótesis alternativa, que de alguna manera reflejarán esa idea *a priori* que tenemos y que pretendemos contrastar con la "realidad".

Los contrastes de hipótesis son capaces de responder a preguntas concretas que nos podemos formular sobre los parámetros poblacionales de interés, por ejemplo: ¿La cantidad media diaria de sal ingerida por hipertensos es mayor que la que ingieren las personas con presión arterial normal?, ¿la temperatura corporal de los pacientes que han sufrido cierta infección bacteriana es superior a los 36 grados centígrados?, ¿la proporción de personas diabéticas con problemas de vista es superior a la de la población general? Resulta evidente que un mecanismo capaz de dar respuesta a cuestiones como las anteriores sería una herramienta muy valiosa, en consecuencia, los contrastes o tests

de hipótesis son una de las utilidades más valoradas y extendidas en la realización de estudios estadísticos.

De la misma manera aparecen, implícitamente, diferentes tipos de errores que podemos cometer durante el procedimiento. No podemos olvidar que, habitualmente, el estudio y las conclusiones que obtengamos para una población cualquiera, se habrá apoyado exclusivamente en el análisis de solo una parte de ésta.

4.3.1. Elaboración de las hipótesis nula y alternativa

Muy a menudo, en la práctica, se tienen que tomar decisiones sobre poblaciones, partiendo de la información muestral de las mismas. Tales decisiones se llaman decisiones estadísticas. Por ejemplo, se puede querer decidir a partir de los datos del muestreo, si un suero nuevo es realmente efectivo para la cura de una enfermedad, si los niños de diferentes comunidades tienen la misma altura, si un sistema educacional es mejor que otro, etc.

Cualquier investigación implica la existencia de **dos hipótesis** o **afirmaciones** acerca de las poblaciones que se estudian. Tales afirmaciones que pueden ser o no ciertas se llaman hipótesis estadísticas (Dagnino, 2014).

Hipótesis nula

La hipótesis nula **(H0)** se refiere siempre a un valor especificado del parámetro de población, no a una estadística de muestra. El planteamiento de la hipótesis nula siempre contiene un signo de igualdad con respecto al valor especificado del parámetro.

En una prueba de hipótesis, debemos establecer el valor supuesto o hipotético del parámetro de la población antes de comenzar a tomar la muestra. La suposición que deseamos probar se conoce como hipótesis nula y se simboliza **H_0 "H sub-cero".**

Supongamos que deseamos probar la hipótesis de que la media de la población es **igual a 200.** En símbolos se escribe como sigue y se lee "la hipótesis nula es que la media de población es igual a 200
H_0: μ= 200

¿Por qué se llama "hipótesis nula"?

El término hipótesis nula surge de las primeras aplicaciones agrícolas y médicas de la estadística.
Con el fin de probar la efectividad de una nueva medicina, la hipótesis que se probaba era que **no hubo efecto**, es decir, **no hubo diferencia entre las muestras tratadas y las no tratadas**.

Ejemplo: La hipótesis nula es que la media de la población es igual a 200.

$$H_0 : \mu = 200$$

Formas
Si H_0: μ = 200
Si H_0: μ ≤ 200
Si H_0: μ ≥ 200

La condición "*igual*" siempre se considera en la hipótesis nula

Hipótesis alternativa
La **hipótesis alternativa H_1 "H sub-uno"** es cualquier hipótesis que difiera de la hipótesis nula. El planteamiento de la hipótesis alternativa nunca contiene un signo de igualdad con respecto al valor especificado del parámetro.

Formas

Si H_0: μ = 200	H_1 : μ ≠ 200
Si H_0: μ ≤ 200	H_1 : μ > 200
Si H_0: μ ≥ 200	H_1 : μ < 200

Tipos de errores

Una hipótesis **estadística** es una asunción relativa a una o varias poblaciones, que puede ser cierta o no. Las hipótesis estadísticas se pueden contrastar con la información extraída de las muestras y tanto si se aceptan como si se rechazan se puede cometer un error.

La hipótesis formulada con intención de rechazarla se llama hipótesis nula y se representa por H_0, **rechazar H_0** implica **aceptar una hipótesis alternativa (H_1).**

En la muestra seleccionada nos puede conducir a aceptar la H_0 o rechazarla. Si aceptamos la H_0 y en la población es cierta, no cometeremos ningún error y a esto se denomina nivel de confianza (1-alfa). En cambio, si rechazamos la H_0 y es cierta, cometeremos un error denominado tipo I o alfa. Si aceptamos la H_0 y en la población es falsa cometeremos un error denominado tipo II o beta. En cambio, si aceptamos la H_1 y en la población es cierta, no cometeremos ningún error y esto se le conoce como potencia (1-beta) e indica la capacidad de un test estadístico para encontrar realmente asociaciones.

En general, se formulan hipótesis donde se intenta rechazar la H_0 y se busca minimizar al máximo el error tipo I o alfa. En la mayoría de las ocasiones suele ser inferior a 0,05 o 0,01 ($p<0,05$ o $p<0,01$). Es decir, si rechazamos la H_0 la probabilidad de cometer un error en la población sería inferior al 5% o 1%.

Un error tipo I se presenta si la **hipótesis nula es rechazada** cuando de hecho es verdadera y **debía ser aceptada**.

Un **error tipo II** se presenta si la **hipótesis nula es aceptada** cuando de hecho es falsa y **debía ser rechazada**.

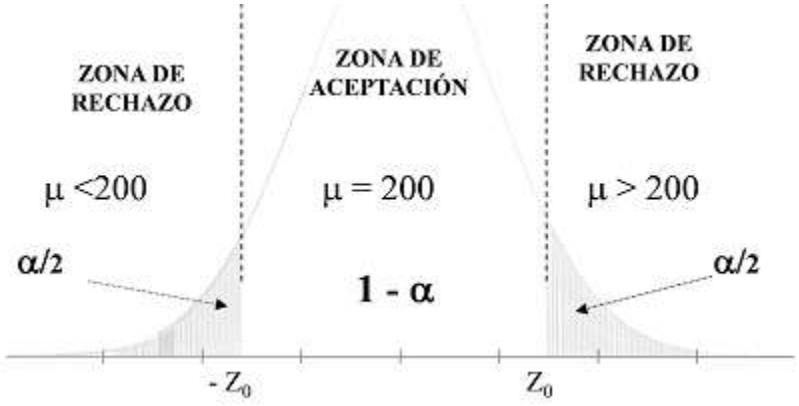
Tipos de prueba

a) **Prueba bilateral o de dos extremos (dos colas):** la hipótesis nula se formula con la igualdad

Ejemplo

$H_0 : \mu = 200$

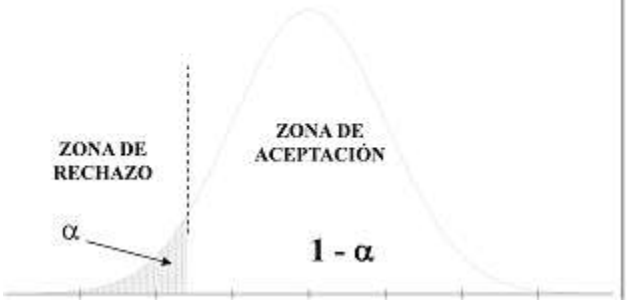
$H_1 : \mu \neq 200$



b) Pruebas **unilateral o de un extremo (una cola)**: la hipótesis nula se formula con ≥ o ≤

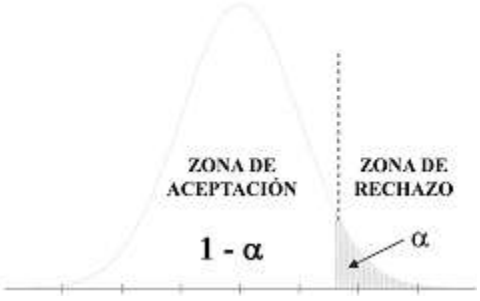
$H_0 : \mu \geq 200$

$H_1 : \mu < 200$



$H_0 : \mu \leq 200$

$H_1 : \mu > 200$



4.3.2 Elementos fundamentales en contrastes de hipótesis

Las hipótesis

En cualquier contraste de hipótesis tendremos 2 alternativas complementarias en las que se especificarán distintos valores de un parámetro poblacional y a la vista de los datos habremos de optar por una de ellas. Por ejemplo, si deseamos conocer si el valor de un parámetro ¿puede ser igual a 25 o, por el contrario, es inadmisible a la vista de los datos que disponemos, nuestras hipótesis serán:

$$\mu = 25 \text{ y } \mu \neq 25$$

Estas 2 hipótesis que hemos señalado no jugarán el mismo papel dentro de cualquier contraste de hipótesis, por tanto, cada una de ellas recibirá un nombre específico:

Hipótesis nula, a la que habitualmente nos referimos como H_0
Hipótesis alternativa, a la que habitualmente nos referimos como H_1

A la hipótesis nula siempre se le concederá el beneficio de la duda e intentaremos encontrar en nuestra muestra evidencias en contra de ella. Así, al terminar el contraste habremos de optar por aceptar H_0 (si no tenemos evidencia suficiente en su contra) o rechazarla (si los datos hacen que la descartemos). Se podría hacer un símil entre el papel de la hipótesis nula en un contraste de hipótesis y el acusado de un juicio: ambos tienen presunción de inocencia y si los datos no aportan evidencias suficientes. En consecuencia, si en un contraste de hipótesis rechazamos la hipótesis nula será porque disponemos de evidencias suficientes en su contra, es decir, estamos razonablemente seguros de que dicha hipótesis es falsa.

Por el contrario, si aceptamos H_0 será porque no hemos encontrado evidencias suficientes en su contra, pero esto no implica que estemos más o menos seguros de que realmente dicha hipótesis sea cierta, podría darse el caso de que H_0 fuera falsa, pero que los datos no aportan evidencia suficiente como para que lleguemos a dicha conclusión.

Mecánica de las pruebas de hipótesis

Una vez hemos descrito los elementos fundamentales de los contrastes de hipótesis estamos en condiciones de describir la mecánica habitual para llevar a cabo este proceso.

Dividimos este proceso en las siguientes fases:

1) Plantear las hipótesis:

H_0 : $\mu = 25$

H_1 : $\mu \neq 25$

2) Seleccionar el nivel de significación: $\alpha = 0,05$ o $\alpha = 0,01$

3) Elegir la prueba estadística:

Los supuestos son:

- la población está normalmente distribuida;
- la muestra ha sido seleccionada al azar.

4) Determinación de los criterios de decisión

5) Aceptación/rechazo de la hipótesis nula

4.4. Pruebas paramétricas y no paramétricas

Siguiendo la secuencia del cuadro 10 se debe conocer que las diferentes pruebas estadísticas se dividen en 2 grandes conjuntos: las paramétricas y las no paramétricas. Una vez que se definió con claridad los 3 aspectos señalados se deberá establecer a cuál de estos 2 conjuntos pertenece la prueba. Tomando en cuenta la escala de medición de las variables, al conjunto de pruebas estadísticas paramétricas les corresponde las cuantitativas continuas y discretas (con distribución normal); mientras que para las variables cualitativas (ya sean nominales u ordinales) y las cuantitativas (sin distribución normal) se incluyen las pruebas estadísticas no paramétricas.

Además, un requisito indispensable para seleccionar una prueba paramétrica es la distribución de los datos; en este sentido, solamente se debe utilizar este tipo de prueba cuando los datos muestran una distribución normal (es decir, semejante a una curva de Gauss). Se debe recordar que para determinar el tipo de distribución existen diferentes pruebas estadísticas, tales como Kolmogorov-Smirnov, Shapiro-Wilk (ver capítulo III).

4.4.1. Pruebas paramétricas

La estadística paramétrica es una rama de la estadística inferencial que comprende los procedimientos estadísticos y de decisión que están basados en distribuciones conocidas. Éstas son determinadas usando un número finito de parámetros. Esto es, por ejemplo, si conocemos que la altura de las personas sigue una distribución normal, pero desconocemos cuál es la media y la desviación de dicha normal. Cuando desconocemos totalmente qué distribución siguen nuestros datos entonces deberemos aplicar primero un test no paramétrico, que nos ayude a conocer primero la distribución.

La mayoría de procedimientos paramétricos requiere conocer la forma de distribución para las mediciones resultantes de la población estudiada. Para la inferencia paramétrica es requerida como mínimo una escala de intervalo, esto quiere decir que nuestros datos deben tener un orden y una numeración del intervalo. Sin embargo, datos categorizados en niños, jóvenes, adultos y ancianos no pueden ser interpretados mediante la estadística paramétrica ya que no se puede hallar un parámetro numérico (como por ejemplo la media de edad) cuando los datos no son numéricos (Monterrey & Gómez-Restrepo, 2007).

La estadística paramétrica, como parte de la inferencia estadística pretende:

- Estimar los parámetros de una población en base a una muestra.
- Conocer el modelo de distribución de la población, presenta variables cuantitativas continuas (medibles).
- Mientras más grande sea la muestra más exacta será la estimación, mientras más pequeña, más distorsionada será la media de las muestras.

Condiciones que deben cumplir las pruebas paramétricas
Una prueba paramétrica debe cumplir con los siguientes elementos:

Normalidad: El análisis y observaciones que se obtienen de las muestras deben considerarse normales. Para esto se deben realizar pruebas de bondad de ajuste donde se describe que tan adaptadas se encuentran las observaciones y cómo discrepan de los valores esperados.

Homocedasticidad: Los grupos deben presentar variables uniformes, es decir, que sean homogéneas.

Errores: Los errores que se presenten deben ser independientes. Esto solo sucede cuando los sujetos son asignados de forma aleatoria y se distribuyen de forma normal dentro del grupo.

Tipos de pruebas paramétricas más utilizadas:

- Prueba t de Student (una sola muestra).
- Prueba t de Student para datos relacionados (muestras dependientes).
- Prueba t de Student para datos no relacionados (muestras independientes).
- Prueba F (análisis de varianza o ANOVA).
- Coeficiente de correlación de Pearson.

Ventajas y desventajas de las pruebas paramétricas

Algunas de las **ventajas de las pruebas paramétricas son:**

- Son más eficientes.
- Son perceptibles a las características de la información obtenida.
- Los errores son muy poco probables.
- Los cálculos probabilísticos son muy exactos.

Las **desventajas de las pruebas paramétricas son:**

- Los cálculos son difíciles de realizar.
- Los datos que se pueden observar son limitados.

Las pruebas paramétricas son una herramienta útil para múltiples situaciones, cálculo e interpretaciones. Gracias a que se utilizan comúnmente, es posible observar los resultados obtenidos a través de un análisis. Son un método muy poderoso si se cumplen las condiciones de su aplicación. Sin embargo, los investigadores deben tener en cuenta que, si las variables que están estudiando no siguen una ley normal, no pueden elegirse.

4.4.1.1. Prueba t de Student para una sola muestra

La distribución de probabilidad de la t de Student **permite estimar el valor de la media poblacional** de **una variable aleatoria** que sigue una distribución normal **cuando el parámetro se extrae de una muestra** pequeña y se desconoce la varianza poblacional.

La condición más importante para usar una prueba t-Student, es que los datos con los que se trabajan deben provenir de una distribución normal.

La prueba T para una muestra se utiliza cuando queremos averiguar si las medias de una variable son superiores o inferiores a un valor fijo.

En un proyecto que pretende estudiar si **la obesidad en adultos mayores de 60 años está vinculada a un mayor riesgo de padecer diabetes.**

Para valorar la eficacia del tratamiento se ha recogido **una muestra de 16 adultos mayores**, **con sobrepeso y obesidad,** obteniendo los siguientes resultados del examen de glucemia en ayunas (en mg/dl):

112	107	113	90	112	90	96	114
96	102	112	110	108	111	103	125

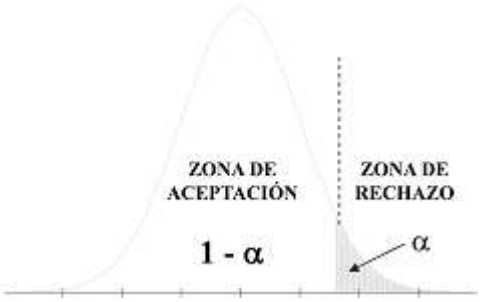
a) ¿**Pueden concluir** estos investigadores que los resultados del examen de glucemia en ayunas aplicado a los adultos mayores es **significativamente mayor a 100 mg/dl?** Calcula el p-valor del contraste. Utiliza un nivel de **significación = 0,05.**

Contraste en el programa SPSS

La significación estadística que nos da el programa SPSS es la de una prueba bilateral, por lo que tenemos que calcular la "p" en función de si la prueba es unilateral hacia la izquierda o hacia la derecha.

En las pruebas unilaterales los valores pueden ser superiores o inferiores a un punto. Por ejemplo, podemos demostrar que el nivel de glucemia en ayunas es **significativamente mayor a 100 mg/dl.**

En este caso la hipótesis nula es H_0 = 100 mg/dl. La prueba es unilateral hacia la derecha, por lo tanto, la significación estadística sería p/2.



1) Planteamiento de hipótesis:

H_0: El nivel de glucemia en ayunas es menor o igual a 100 mg/dl

H_1: El nivel de glucemia en ayunas es significativamente >100 mg/dl

2) Nivel de significación: α = 0,05

3) Prueba estadística:

Prueba de t-Student para una muestra

Los supuestos son:

- La población está normalmente distribuida.
- La muestra ha sido seleccionada al azar.

4) Determinación de los criterios de decisión

Para realizarlo con SPSS, hacer clic en:

- Analizar
- Comparar medias
- Prueba T para una muestra

En la ventana "Prueba T para una muestra" introducir la variable que queremos analizar, ej. **nivel de glucemia** y en el valor de prueba, el valor, ej. **100 mg/dl.**

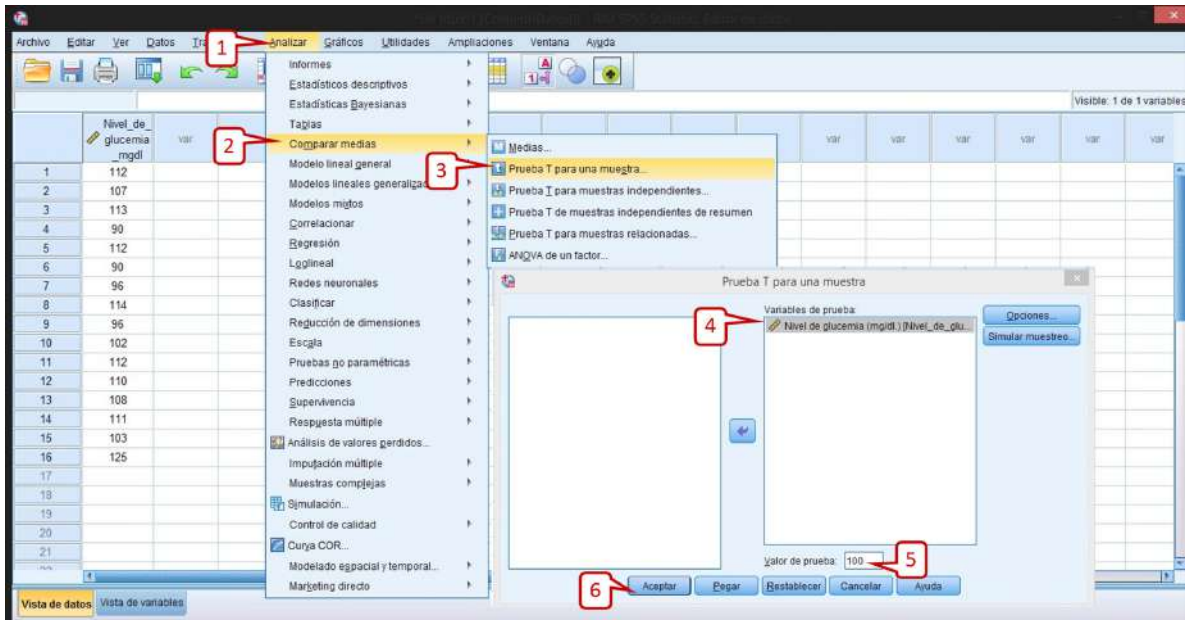


Figura 55. Prueba t Student para una sola muestra.

Presentación de hoja resultados

Estadísticas para una muestra				
	N	**Media**	**Desv. Desviación**	**Desv. Error promedio**
Nivel de glucemia (mg/dl.)	16	106,31	9,534	2,383

Prueba para una muestra						
	Valor de prueba = 100					
	t	gl	Sig. (bilateral)	Diferencia de medias	95% de intervalo de confianza de la diferencia	
					Inferior	Superior
Nivel de glucemia (mg/dl.)	2,648	15	0,018	6,313	1,23	11,39

En la ventana de resultados aparece una tabla con el valor de la t de Student, los grados de libertad y la significación **como si la prueba fuera bilateral**.

Si queremos **realizar una prueba unilateral** hacia la izquierda (<) o derecha (>), debemos dividir entre 2 el valor de la p (p/2). Por ejemplo, en la prueba bilateral el p = 0,018, el valor de la **p real** en una prueba unilateral sería **p/2 = 0,018/2** = 0,0091.

En el ejemplo propuesto, queremos demostrar que **$H_{1=}$ el nivel de glucemia en ayunas es significativamente >100 mg/dl,** la prueba es unilateral hacia la derecha, en este caso la p real sería (0,018/2), **p = 0,0091**, por lo que **se rechaza la hipótesis nula,** en consecuencia, **se acepta H_1:** El nivel glucemia en ayunas es significativamente >100 mg/dl.

5) Aceptación/rechazo de la hipótesis nula

Se **RECHAZA** la hipótesis nula "El nivel de glucemia en ayunas es menor o igual a 100 mg/dl" y se **ACEPTA la hipótesis alternativa (H_1)** "El nivel de glucemia en ayunas es significativamente mayor a 100 mg/dl" a un nivel de significación de α = 0,05. La prueba resultó **significativa**.

4.4.1.2. Prueba t-Student para dos muestras independientes

La prueba T para muestras independientes permite **comparar las medias de dos grupos** de casos. Lo ideal es que para esta prueba los sujetos se asignen aleatoriamente a dos grupos, de forma que cualquier diferencia en la respuesta sea debida al factor en estudio.

Por ejemplo, un estudio pretende comparar el nivel de calcio en plasma sanguíneo en hombres y mujeres, obteniéndose que el nivel medio para los hombres es 3,6 mmol/l, mientras que para las mujeres el nivel medio es 2,9 mmol/l. ¿Es significativa la diferencia obtenida en el nivel de calcio entre hombres y mujeres (= 0,05)?

En el caso de muestras independientes **las varianzas de las dos poblaciones se suponen iguales, aunque desconocidas**. Según el caso puede ser conveniente realizar previamente el **prueba de Levene para igualdad de varianzas.**

Un ejemplo nos ayudará a fijar las ideas.

Ejemplo: La tabla siguiente contiene valores del índice neutrófilo/linfocito (INL) en relación al estado glucémico de 28 pacientes con hiperglucemia (grupo 1) y sin hiperglucemia (grupo 2).

Estado glucémico	**Índice neutrófilo/linfocito (INL)**													
Grupo 1	2,79	2,06	3,76	3,58	3,58	2,80	2,58	2,76	3,50	2,25	2,79	2,55	3,08	3,50
Grupo 2	2,39	1,78	1,51	2,03	1,80	1,24	2,32	3,00	2,24	3,80	2,06	3,08	1,74	2,40

a) Se puede concluir que los **pacientes con hiperglucemia** presentan un **mayor INL que los pacientes sin hiperglucemia**. Calcula el p-valor del contraste. Utiliza un nivel de **significación = 0,05.**

Contraste en el programa SPSS

La comparación de medias entre dos grupos independientes se realiza con la prueba t de Student, para ello es necesario que la variable cuan-

titativa continua tenga una distribución normal en cada grupo. Aunque se puede asumir que ésta tendrá una distribución normal si el tamaño de la muestra es superior a 30, debe comprobarse con las pruebas de normalidad.

1) Planteamiento de hipótesis:
H_0: Los pacientes con hiperglucemia presentan un INL **igual** que los pacientes sin hiperglucemia
H_1: Los pacientes con hiperglucemia presentan un **mayor** INL que los pacientes sin hiperglucemia

2) Nivel de significación: $\alpha = 0,05$

3) Prueba estadística:
Prueba de t-Student para muestras independientes
4) Determinación de los criterios de decisión
Primero, definir en la base de datos las variables que vamos a utilizar. En el siguiente ejemplo vamos a comprobar si hay diferencias estadísticamente significativas entre los pacientes con y sin hiperglucemia en cuanto a la variable índice neutrófilo/linfocito (INL).

Segundo, debemos comprobar que la variable cuantitativa tiene una distribución normal en cada grupo. Para ello recurrimos a la realización de las pruebas de normalidad (ver capítulo III) En este ejemplo, **índice neutrófilo/linfocito (INL)** tiene una distribución normal en cada grupo según el test de Shapiro-Wilk (<50 datos), en caso de no tener distribución normal se debe optar por la prueba U de Mann-Whitney que es la versión no paramétrica de la habitual prueba t de Student aplicada a dos muestras independientes.

Pruebas de normalidad				
Estado glucémico		Shapiro-Wilk		
		Estadístico	gl	Sig.
INL	Con hiperglucemia	0,927	14	0,28*
	Sin hiperglucemia	0,943	14	0,45*
*>0,05 Distribución normal				

Tercero, en la hoja de vista de datos ejecutamos el análisis. Para ello hacer clic en Analizar → en Comparar medias → finalmente en Prueba T para muestras independientes. Nos aparece una ventana (Prueba T para muestras independientes) donde debemos introducir en la casilla "variables para contrastar", las variables cuantitativas continuas (en este caso, índice neutrófilo/linfocito (INL)) y en "variable de agrupación" la variable categórica dicotómica (en este ejemplo, con y sin hiperglucemia) definiendo los grupos y asignando el valor que se le ha dado previamente (ej. "1" con hiperglucemia y "2" sin hiperglucemia) y hacer clic en Aceptar.

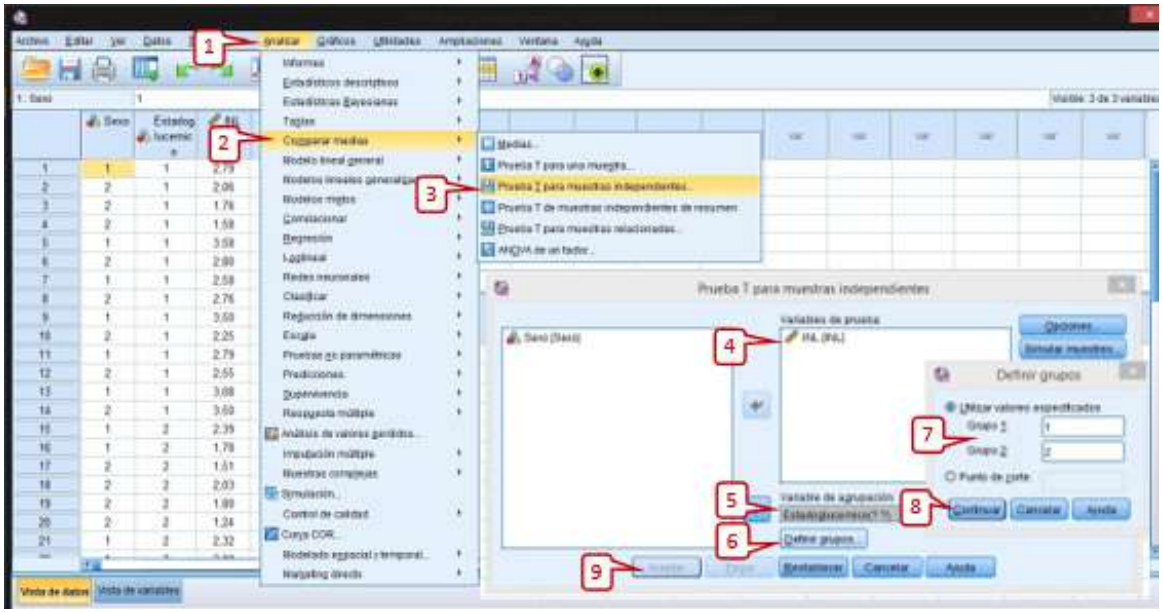


Figura 56. Prueba t Student para dos muestras independientes.

Presentación de resultados

Cuarto: en la hoja de resultados aparece:

Tabla estadísticos de grupo con la n, media, desviación típica y error típico de la media de la/s variable/s cuantitativas continuas por cada grupo donde se evidencia los pacientes con hiperglucemia media 2,97±0,54 superior a los pacientes sin hiperglucemia 2,24±0,68

Estadísticas de grupo

	Estado glucémico	n	Media	Desv. Desviación	Desv. Error promedio
INL	Con hiperglucemia	14	2,97	0,54	0,14
	Sin hiperglucemia	14	2,24	0,68	0,18

Prueba de muestras independientes

		Prueba de Levene de igualdad de varianzas		prueba t para la igualdad de medias						
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Diferencia de error estándar	95% de intervalo de confianza	
									Inferior	Superior
INL	Se asumen varianzas iguales	0,14	0,71	3,14	26	0,004*	0,728	0,232	0,251	1,205
	No se asumen varianzas iguales			3,14	24,6	0,004	0,728	0,232	0,250	1,206

Esta tabla nos informa de la significación estadística de la prueba de Levene para igualdad de varianzas y la prueba T para igualdad de medias para el **índice neutrófilo/linfocito (INL)** en pacientes con y sin hiperglucemia. Si nos fijamos en la **prueba de Levene para igualdad de varianzas**, vemos que la **p > 0,05** por lo que se acepta la H para igualdad de varianzas y **se puede decir que las varianzas son iguales** por lo que se elige la p superior en la prueba T para igualdad de medias. Así, podemos decir que hay diferencias en relación al

índice neutrófilo/linfocito (INL**)** en pacientes con y sin hiperglucemia p = 0,004; < 0,05.

También se representa el valor de la prueba de t de Student (t) y los grados de libertad (gl), la diferencia de medias, el error típico de la diferencia y el intervalo de confianza al 95% para la diferencia.

5) Aceptación/rechazo de la hipótesis nula
Se **RECHAZA** la hipótesis nula "Los pacientes con hiperglucemia presentan un INL **igual** que los pacientes sin hiperglucemia" y se **ACEPTA** la hipótesis alternativa (H_1) **"Los pacientes con hiperglucemia presentan un mayor INL que los pacientes sin hiperglucemia"** a un nivel de significación de $\alpha = 0,05$. La prueba resultó **significativa**.

4.4.1.3. Prueba t-Student para dos muestras dependientes o relacionadas
La prueba t de Student para muestras relacionadas permite **comparar las medias de dos series** de mediciones realizadas **sobre las mismas unidades estadísticas**. El procedimiento calcula las **diferencias entre las dos mediciones** y contrasta **si la media difiere de 0**.

Es una prueba paramétrica de comparación de dos muestras relacionadas, que debe cumplir las siguientes características:

- Asignación aleatoria de los grupos.
- Homocedasticidad (homogeneidad de las varianzas de la variable dependiente de los grupos).
- Distribución normal de la variable dependiente en los dos grupos.
- Su función es comparar dos mediciones de puntuaciones (medias aritméticas) y determinar que la diferencia no se deba al azar (que las diferencia sea estadísticamente significativa).

En un estudio sobre la hipertensión sanguínea, se toma la tensión a todos los pacientes al comienzo del estudio, se les aplica un tratamiento

y se les toma la tensión otra vez. De esta manera, a **cada sujeto le corresponden dos medidas**, normalmente denominadas **medidas antes y después**. Un diseño alternativo para el que se utiliza esta prueba consiste en un estudio de pares relacionados o un estudio de control de casos en el que cada registro en el archivo de datos contiene la respuesta del paciente y de su sujeto de control correspondiente (Dagnino, 2014).

Ejemplo: Queremos contrastar el efecto de una nueva dieta que prometen revolucionaria, y para ello sometemos a esta dieta a **30 personas durante 7 días,** obteniendo los siguientes resultados sobre el peso antes y después de esta dieta:

N°	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Inicio (kg)	66	50	81	85	85	79	83	97	73	82	70	87	83	84	71
Final (kg)	56	64	57	59	50	56	69	66	53	63	67	50	68	59	53

a) Se puede concluir que **hay efecto de la nueva dieta** revolucionaria en relación a la **disminución del peso de los pacientes**. Calcula el p-valor del contraste. Utiliza un nivel de **significación = 0,05.**
Contraste en el programa SPSS

La idea básica es simple. Si el tratamiento no tuvo efecto, la **diferencia promedio** entre las mediciones **es igual a 0 y se cumple la hipótesis nula**. Por otro lado, si el tratamiento tuvo un efecto (intencionado o no), la diferencia **promedio no es 0 y se rechaza la hipótesis nula**. Si los datos de la variable de contraste no son cuantitativos, pero están ordenados, o bien **no están distribuidos normalmente, utilice la prueba de Wilcoxon de los rangos con signo**.

1) Planteamiento de hipótesis:
H_0: No hay diferencias en entre la medición del peso de inicio y la medición final de peso de pacientes.
H_1: Hay diferencias en entre la medición del peso de inicio y la medición final de peso de pacientes.

2) Nivel de significación: $\alpha = 0,05$

3) Prueba estadística:
Prueba de t-Student para muestras dependientes

4) Determinación de los criterios de decisión

Primero: definir en la base de datos las variables que vamos a utilizar. En el siguiente ejemplo vamos a comprobar si hay diferencias estadísticamente significativas entre los pesos antes y después.

Segundo: debemos comprobar que la variable cuantitativa tiene una distribución normal en cada grupo. Para ello recurrimos a la realización de las pruebas de normalidad (ver capítulo III) En este ejemplo, **el peso antes y después** tiene una distribución normal en cada grupo según el test de Shapiro-Wilk (<50 datos), en **caso de tener no distribución normal** se debe optar por la prueba **de Wilcoxon de los rangos con signo,** que es la versión no paramétrica de la habitual prueba t de Student aplicada a dos muestras dependientes o relacionadas.

Pruebas de normalidad			
	Shapiro-Wilk		
	Estadístico	gl	Sig.
Inicio (kg)	0,907	15	0,12*
Final (kg)	0,934	15	0,31*
*>0,05 Distribución normal			

Tercero: en la hoja de vista de datos ejecutamos el análisis. Para obtener una prueba T para muestras relacionadas elija en los menús: Analizar > Comparar medias > Prueba T para muestras relacionadas, seleccione uno o más pares de variables, si lo desea, puede pulsar en opciones para controlar el tratamiento de los datos perdidos y el nivel del intervalo de confianza, donde debemos introducir en la casilla "variables emparejadas", las variables cuantitativas continuas (en este caso, peso inicio y final) y hacer clic en Aceptar.

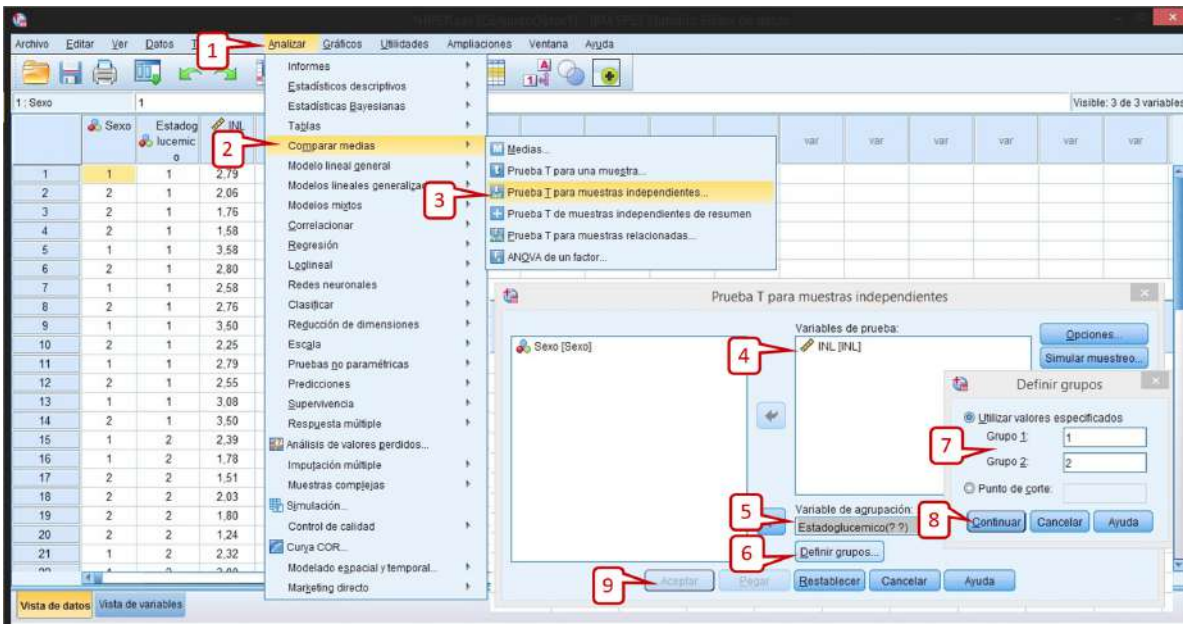


Figura 57. Prueba t-Student para dos muestras dependientes o relacionadas.

Presentación de resultados

Cuarto: en la hoja de resultados aparece:

Tabla estadísticos de grupo con la n, media, desviación típica y error típico de la media de la/s variable/s cuantitativas continúas por cada grupo, donde se evidencia el peso de los pacientes antes de la dieta con una media 78,40 ± 11,07 superior al peso final medio de 59,33 ± 6,49 kg.

Estadísticas de muestras emparejadas

		Media	N	Desv. Desviación	Desv. Error promedio
Par 1	Inicio (kg)	78,40	15	11,07	2,858
	Final (kg)	59,33	15	6,49	1,675

Prueba de muestras emparejadas									
Diferencias emparejadas									
		Media	Desv. Desviación	Desv. Error promedio	95% de intervalo de confianza		t	gl	Sig. (bilateral)
					Inferior	Superior			
Par 1	Inicio (kg) - Final (kg)	19,07	12,87	3,32	11,94	26,19	5,74	14	0,000

El software muestra un valor de p de 0,000 para la prueba bilateral. Esto indica que la posibilidad de encontrar una diferencia de medias de muestra de 19,07 con una diferencia de medias subyacente de cero, es de unas 0 de cada 100. Tenemos confianza en nuestra decisión de rechazar la hipótesis nula (<0,05). Así, podemos decir que **hay diferencias** en relación al peso inicial y final en pacientes que siguieron la dieta revolucionaria propuesta, p = 0,000; < 0,05

También se representa el valor de la prueba de t de Student (t) y los grados de libertad (gl), la diferencia de medias, el error típico de la diferencia y el intervalo de confianza al 95% para la diferencia.

5) Aceptación/rechazo de la hipótesis nula

Se **RECHAZA** la hipótesis nula "No hay diferencias en entre la medición del peso de inicio y la medición final de peso de pacientes" y se **ACEPTA la hipótesis alternativa (H_1)** "Hay diferencias en entre la medición del peso de inicio y la medición final de peso de pacientes" a un nivel de significación de $\alpha = 0,05$. La prueba resultó **significativa**.

4.4.1.4. Análisis de varianza (ANOVA)

En este capítulo introducimos el análisis de la varianza, cuyo objetivo es **comparar dos o más medias simultáneamente**. Si, por ejemplo, queremos valorar la hemoglobina (Hb) de los niños en edad escolar de tres unidades educativas diferentes (UEl, UE2 y UE3) querríamos comparar 3 medias simultáneamente: $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ y $\boldsymbol{\mu}_3$, detrás de la comparación de estas tres medias podemos ver el estudio de la relación de dos va-

riables, una cuantitativa (hemoglobina (Hb)) y otra cualitativa nominal (unidades educativas). A la variable cualitativa (UE) se le suele llamar factor y en este caso se trata de un factor con tres categorías (UEl, UE2 y UE3). Para llevar a cabo esta comparación plantearíamos un modelo de análisis de la varianza de una vía (de un factor). Si tuviéramos más de un factor para estudiar, por ejemplo, queremos comparar diferentes unidades educativas y diferenciar entre hombres y mujeres, tendríamos dos factores en estudio: unidades educativas (UEl, UE2 y UE3) y sexo (H y M), y el modelo sería de análisis de la varianza con más de un factor.

Para aplicar un análisis de varianza, **es obligatorio** que los datos de la variable respuesta sean cuantitativos, aleatorios, que la distribución de **las poblaciones comparadas sea normal y que exista homogeneidad en el valor de varianzas**, existen diferentes pruebas para llevar a cabo este contraste, pero las más extendidas en uso son el **test de Barlett y el test de Levene.** Si al realizar alguna de estas dos pruebas obtenemos un p-**valor superior al nivel de significatividad (normalmente = 0,05)** no podremos rechazar la hipótesis de homogeneidad de varianzas, y por tanto podemos asumirla como cierta. En caso de no cumplirse estas condiciones habría que recurrir a la transformación de datos y en caso de no hacerse esta transformación no se podrá aplicar el análisis de varianza, **si no tienen una distribución normal** se debe utilizar la prueba alternativa no paramétrica de **Kruskal-Wallis.**

Los datos deberán clasificarse ya sea en columnas o en renglones, en donde cada columna o renglón identificará un grupo o tratamiento a comparar. **El número de observaciones por grupo puede ser el mismo o no.** Esto es, no es obligatorio que las muestras sean del mismo tamaño para ser comparadas, aunque si es deseable porque facilita los cálculos.

Para plantear el par de hipótesis, nula (H_0) y alternativa (H_1), se parte de la suposición de que los diversos tratamientos no conducen a re-

sultados diferentes y entonces la hipótesis nula establecería que todos los tratamientos aplicados funcionan igual, en promedio. Mientras que la hipótesis alternativa establecería posibles diferencias, en promedio, parciales o totales.

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_n$,
H_1: Al menos un par de medias es diferente

Ejemplo: Supongamos que estamos interesados en comprobar si existen diferencias significativas en el nivel medio de hemoglobina (Hb) en niños de 3 unidades académicas del cantón Jipijapa. Con el fin de realizar la comparación correspondiente se toman 45 niños en edad escolar y se reparten al azar entre las 3 unidades académicas del cantón Jipijapa (15 en cada grupo). A continuación mostramos los datos recolectados correspondiente de los valores de hemoglobina (Hb) expresados en (gr/dl) para llevar a cabo esta comparación.

Unidad educativa	Hemoglobina HB (gr/dl)														
EU 1	11,25	12,19	11,56	11,25	11,95	12,44	11,85	11,15	11,25	12,19	11,56	12,12	11,88	11,25	12,19
EU 2	12,81	11,88	11,25	12,19	12,02	12,81	12,50	11,54	11,72	11,35	11,56	11,96	11,44	11,25	12,33
EU 3	11,56	11,94	12,50	12,50	11,88	12,25	12,94	11,88	11,58	12,68	12,50	11,56	12,75	13,60	11,56

a) Se puede concluir que **niveles medios de hemoglobina HB (gr/dl)** en los niños de las distintas unidades educativas **son diferentes o al menos una de ellas lo es.** Calcula el p-valor del contraste. Utiliza un nivel de **significación = 0,05.**

Contraste en el programa SPSS
1) Planteamiento de hipótesis:
H_0: Nivel medio de Hb igual en las tres unidades educativas
H_1: Nivel medio de Hb no es igual en las tres unidades educativas

2) Nivel de significación: $\alpha = 0,05$

3) Prueba estadística:

Prueba de F análisis de varianza (ANOVA)

4) Determinación de los criterios de decisión

Primero: en la hoja de vista de variables debemos definir las variables cuantitativas continuas y la variable de grupo asignando un número diferente a cada clase del grupo. Así, en el siguiente ejemplo, se ha utilizado una base de datos de niños en edad escolar agrupados por unidades educativas y se le ha asignado los siguientes números a la variable "Unidades Educativas": 1 = UE 1, 2 = UE 2, 3 = UE 1. Se pretende evaluar si existen diferencias entre cada uno de los grupos en cuanto a los niveles medios de hemoglobina HB (gr/dl).

Segundo: debemos comprobar que la variable cuantitativa tiene una distribución normal en cada grupo. Para ello recurrimos a la realización de las pruebas de normalidad (ver capítulo III). En este ejemplo, niveles medios de hemoglobina HB (gr/dl) tiene una distribución normal en cada grupo según el test de Shapiro-Wilk (<50 datos), en caso de tener no distribución normal se debe optar por la prueba de Kruskal-Wallis que es la versión no paramétrica de la habitual prueba F (ANOVA).

Pruebas de normalidad

Unidad Educativa		Shapiro-Wilk		
		Estadístico	gl	Sig.
HB gr/dl	UE 1	0,895	15	0,08*
	UE 2	0,928	15	0,256*
	UE 3	0,914	15	0,16*

*>0,05 Distribución normal

Tercero: ejecutar el análisis. Hacer clic en analizar → en comparar medias → en ANOVA de un factor. En la ventana de ANOVA de un factor, introducir en la casilla lista de dependientes la/s variable/s cuantitativa/s continua/s que se quiere comparar (en este caso niveles medios de hemoglobina HB (gr/dl)) y en la casilla "Factor", la variable con los grupos definidos previamente (Unidades Educativas).

En “Opciones” señalar las pestañas descriptivas, prueba de homogeneidad de varianzas y gráfico de medias.

En **“Post Hoc”** señalar una de las pruebas de asumiendo varianzas iguales, habitualmente prueba de diferencia honestamente significativa de **Tukey** (Tukey’s HSD test por sus siglas en inglés) que sirven para hacer comparaciones entre cada uno de los grupos una vez que la prueba de la ANOVA ha sido significativa (Monterrey & Gómez-Restrepo, 2007).

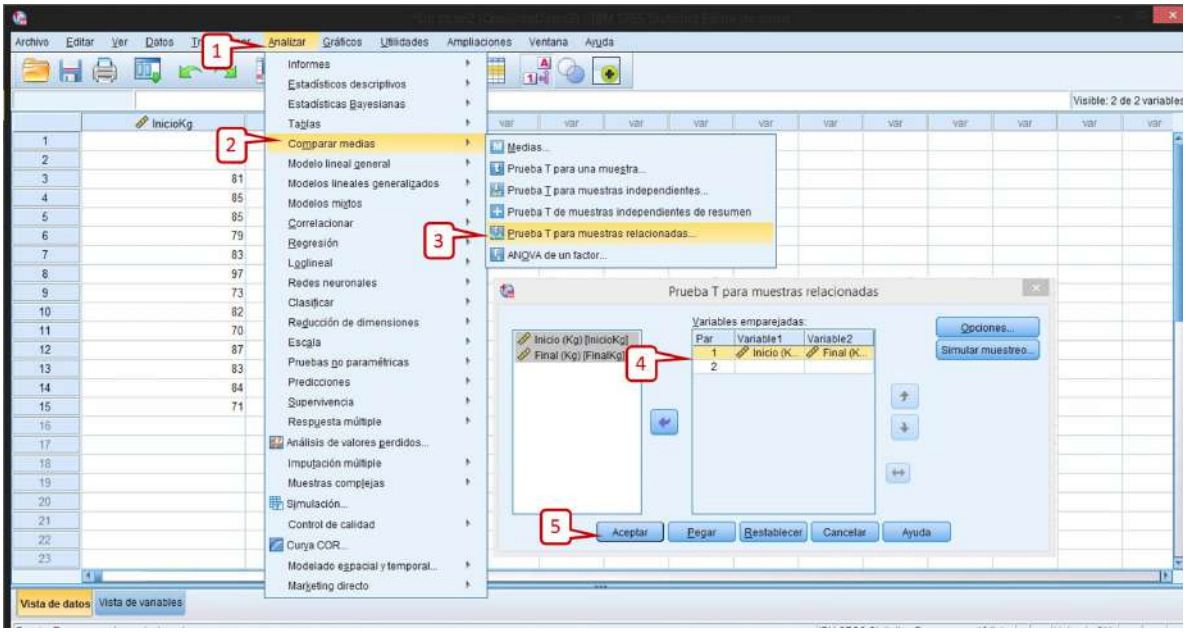


Figura 58. Análisis de varianza (ANOVA).

Presentación de resultados

Cuarto: en la hoja de resultados aparece:

1. Tabla con la estadística descriptiva de las variables cuantitativas continúas por cada grupo.

Descriptivos HB gr/dl								
	N	Media	Desv. Desviación	Desv. Error	95% del intervalo de confianza para la media		Mínimo	Máximo
					Límite inferior	Límite superior		
UE 1	15	11,74	0,44	0,11	11,50	11,98	11,15	12,44
UE 2	15	11,91	0,53	0,14	11,61	12,20	11,25	12,81
UE 3	15	12,25	0,61	0,16	11,91	12,58	11,56	13,60
Total	45	11,96	0,56	0,08	11,80	12,13	11,15	13,60

Tabla donde se representan la media, desviación típica, error típico, intervalos de confianza al 95% y valores máximo y mínimo de los niveles de colesterol distribuidos por cada uno de los grupos.

2. Tabla “Prueba de homogeneidad de varianzas” donde aparece el estadístico de Levene.

Si la Sig. ≥ 0,05 no se rechaza la hipótesis nula (H = las varianzas son homogéneas) y por lo tanto, podemos decir que las varianzas son homogéneas y continuar con el análisis.

Prueba de homogeneidad de varianzas					
		Estadístico de Levene	gl1	gl2	Sig.
HB gr/dl	Se basa en la media	0,703	2	42	0,51*
	Se basa en la mediana	0,715	2	42	0,49*
	Se basa en la mediana y con gl ajustado	0,715	2	39,3	0,50*
	Se basa en la media recortada	0,733	2	42	0,49*
*>0,05 varianzas homogéneas					

3. Tabla "ANOVA". Debemos fijarnos para comprobar si hay diferencias entre los grupos. La H0 = no hay diferencias y la H = hay diferencias intragrupo. Si la $p \geq 0,05$ el análisis lo damos por finalizado y tenemos que decir que no hay suficiente evidencia para encontrar diferencias entre las medias de los grupos. Por el contrario, si la $p < 0,05$ podemos rechazar la H0 y decir que al menos un grupo es diferente y debemos continuar realizando el análisis para averiguar cuál o cuáles son diferentes.

ANOVA HB gr/dl					
Suma de cuadrados		**gl**	**Media cuadrática**	**F**	**Sig.**
Entre grupos	1,997	2	0,999	3,55	0,037*
Dentro de grupos	11,798	42	0,281		
Total	13,795	44			
* < 0,05 Diferencias significativas para prueba de F					

Tabla con la prueba de ANOVA como **$p < 0,05$, se rechaza la H_0** y podemos decir que **al menos un grupo es diferente** y continuar con el análisis.

4. En la tabla de **"Pruebas post-Hoc"** aparecen las comparaciones múltiples entre cada uno de los grupos con el **valor del estadístico y la significación estadística**. También aparece el **gráfico de medias** de la variable cuantitativa continua por cada grupo.

HSD Tukey[a]			
Unidad Educativa	**N**	**Subconjunto para alfa = 0.05**	
		1	**2**
UE 1	15	11,74	
UE 2	15	11,91	11,91
UE 3	15		12,25
Sig.		0,661	0,200
Se visualizan las medias para los grupos en los subconjuntos homogéneos.			
a. Utiliza el tamaño de la muestra de la media armónica = 15,000.			

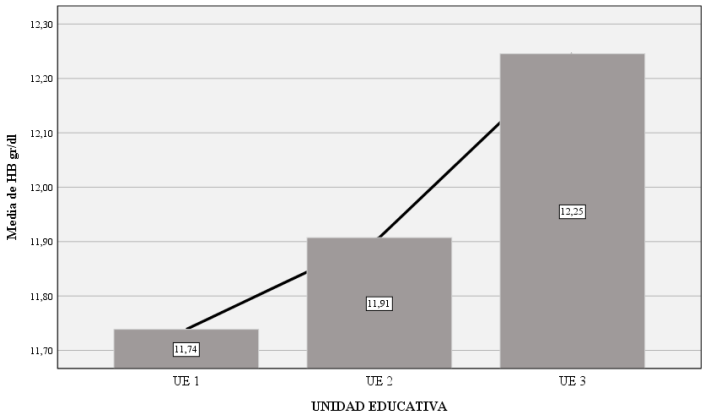


Figura 59. Gráfico de comparaciones de medias.

Tabla de comparaciones múltiples diferencia honestamente significativa de Tukey. En este ejemplo se puede decir que los niños de la Unidad Educativa 1 tienen niveles medios de hemoglobina HB (11,74 gr/dl)) significativamente inferiores con los niños de la Unidad Educativa 3 que tienen los niveles medios de hemoglobina HB (12,25 gr/dl)) significativamente superiores al resto de grupos, pero todos los grupos **se ubican en el rango normal de 11,8 a 14,6 gr/dl** para la edad de 6 a 11 años.

5) Aceptación/rechazo de la hipótesis nula
Se **RECHAZA** la hipótesis nula "Nivel medio de Hb igual en las tres unidades educativas" y se **ACEPTA la hipótesis alternativa (H_1)** "Nivel medio de Hb no es igual en las tres unidades educativas" a un nivel de significación de $\alpha = 0,05$. La prueba resultó **significativa**.

4.4.1.5. Correlación de variables
El concepto de relación o correlación se refiere **al grado de variación conjunta existente entre dos o más variables**. En este apartado nos vamos a centrar en el estudio de un tipo particular de relación llamada lineal y nos vamos limitar a considerar únicamente dos variables (simple). Una relación lineal positiva entre dos variables X_i e Y_i indica que los valores de las dos variables varían de forma parecida: los sujetos que puntúan alto en X_i tienden a puntuar alto en Y_i y los que puntúan bajo en X_i tienden a puntuar bajo en Y_i. Una relación lineal negativa significa que los valores de las dos variables varían justamente al revés: los sujetos que puntúan alto en X_i tienden a puntuar bajo en Y_i y los que puntúan bajo en X_i tienden a puntuar alto en Y_i.

La forma más directa e intuitiva de formarnos una primera impresión sobre el tipo de relación existente entre dos variables es a través de un diagrama de dispersión. Un diagrama de dispersión es un gráfico en el que una de las variables (X_i) se coloca en el eje de abscisas, la otra (Y_i) en el de ordenadas y los pares (X_i , Y_i) se representan como una nube de puntos. La forma de la nube de puntos nos informa sobre el tipo de relación existente entre las variables.

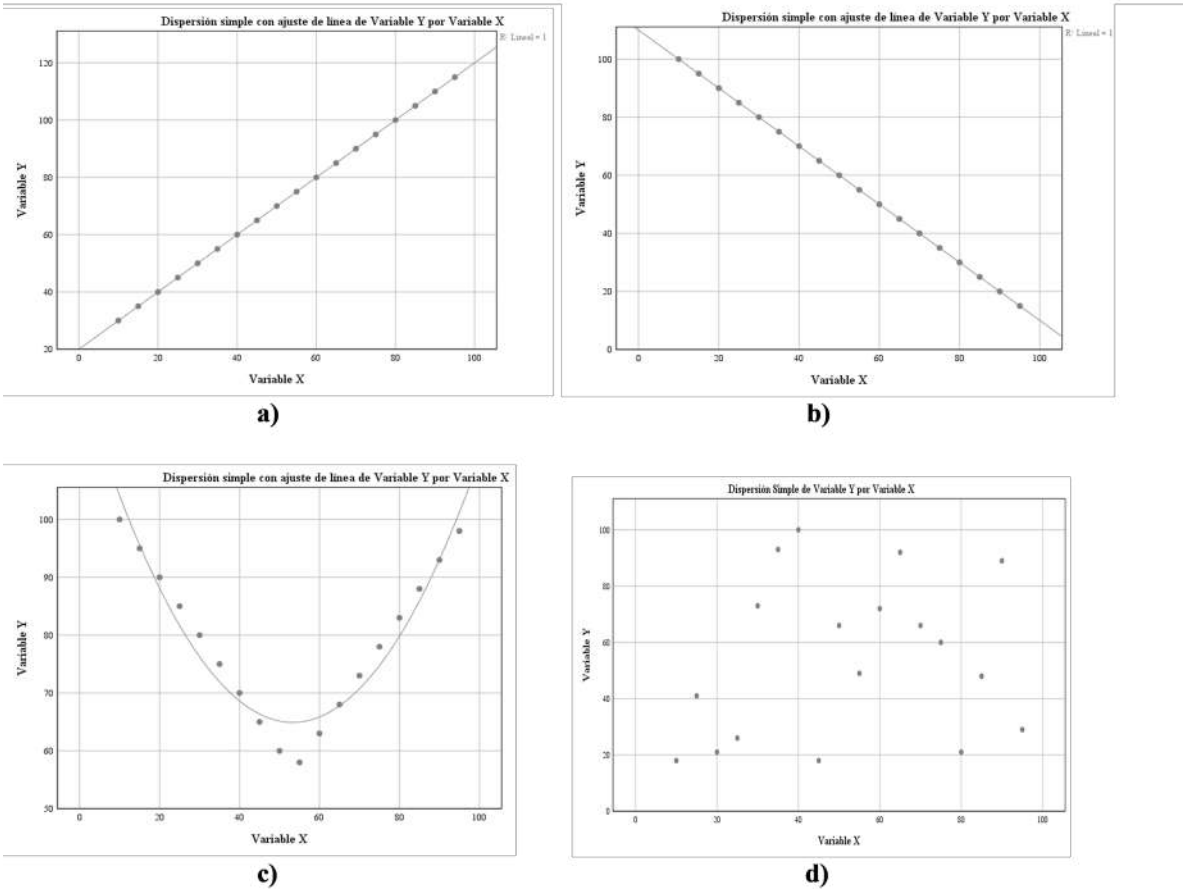


Figura 60. Diagramas de dispersión en función del tipo de relación.

La figura **a)** muestra una situación en la que cuanto mayores son las puntuaciones en una de las variables, mayores son también las puntuaciones en la otra; cuando ocurre esto, los puntos se sitúan en una línea recta ascendente y hablamos de relación lineal positiva. La figura **b)** representa una situación en la que cuanto mayores son las puntuaciones en una de las variables, menores son las puntuaciones en la otra; en este caso, los puntos se sitúan en una línea recta descendente y hablamos de relación lineal negativa. En la situación representada en la figura **c)** también existe una pauta de variación clara, pero no es lineal: los puntos no dibujan una línea recta. Y en la figura **d)** no parece existir ninguna pauta de variación clara, lo cual queda reflejado en

una nube de puntos dispersa, muy lejos de lo que podría ser una línea recta.

Cuando se desea establecer la relación de 2 variables cuantitativas continuas con distribución normal (por ejemplo, peso en kg con el nivel de colesterol LDL) se utilizará el coeficiente de correlación de Pearson (r de Pearson). Sin embargo, **cuando alguna de las 2 variables por correlacionar no sigue una distribución normal**, la prueba que corresponde es el coeficiente de **correlación de Spearman (rho de Spearman).** Esta última es la que se aplica si se trata de **analizar variables ordinales** (ejemplo, el grado de desnutrición y su correlación con el grado de anemia).

Pearson. El coeficiente de correlación de Pearson (1896) es, quizá, el mejor coeficiente y el más utilizado para estudiar el grado de relación lineal existente entre dos variables cuantitativas. Se suele representar por **r** y se obtiene tipificando el promedio de los productos de las puntuaciones diferenciales de cada caso (desviaciones de la media) en las dos variables correlacionadas:

El coeficiente de correlación de Pearson toma valores entre -1 y 1: un valor de **1 indica relación lineal perfecta positiva**; un valor de **-1 indica relación lineal perfecta negativa** (en ambos casos los puntos se encuentran dispuestos en una línea recta); un **valor de 0 indica relación lineal nula**. El coeficiente r es una medida simétrica: la correlación entre X_i e Y_i es la misma que entre Y_i y X_i.

Es importante señalar que un coeficiente de correlación alto no implica causalidad. Dos variables pueden estar linealmente relacionadas (incluso muy relacionadas) sin que una sea causa de la otra.

Tau-b de Kendall. Este coeficiente de correlación es apropiado para estudiar la **relación entre variables ordinales**. Toma valores entre -1 y 1, y se interpreta exactamente igual que el coeficiente de correlación de Pearson.

La utilización de este coeficiente tiene sentido si las variables no alcanzan el nivel de medida de intervalo y/o no podemos suponer que la distribución poblacional conjunta de las variables sea normal.

Spearman. El coeficiente de correlación rho de Spearman (1904) es el coeficiente de correlación de Pearson, pero aplicado después de transformar las puntuaciones originales en rangos. Toma valores entre -1 y 1, y se interpreta exactamente igual que el coeficiente de correlación de Pearson.

Al igual que ocurre con el coeficiente **Tau-b de Kendall, el de Spearman puede utilizarse** como una **alternativa al de Pearson** cuando las **variables estudiadas son ordinales y/o se incumple el supuesto de normalidad.** Los valores de intervalo de confianza están disponibles para Pearson y Spearman.

Ejemplo: Se desea investigar si el **peso en kg** tiene **influencia lineal sobre el colesterol LDL** de un grupo de mujeres que acuden a un centro de salud para su control, utilizar un nivel de significación de α = 0,05.

Peso (kg)	66	67	72	82	90	93	65	77	81	85	79	83	97	73	82
LDL (mg/dl)	93	131	143	121	178	189	97	123	112	125	118	120	180	144	125

Determinar, en función del test de independencia sobre el coeficiente de correlación lineal, si la asociación entre las dos variables es significativa.

Primero: en la hoja de vista de variables debemos definir las variables cuantitativas continuas en nuestro caso la variable peso en kg y el colesterol LDL

Segundo: debemos comprobar que la variable cuantitativa tiene una distribución normal en la variable dependiente (Y). Para ello recurrimos a la realización de la prueba de normalidad (ver capítulo III). En este

ejemplo, **peso en kg** y colesterol LDL tiene una distribución normal según el test de Shapiro-Wilk (<50 datos), en **caso de no tener distribución normal** se debe optar por el coeficiente de **Spearman** que es la versión no paramétrica de la habitual del coeficiente de correlación de Pearson.

Prueba de normalidad			
	Shapiro-Wilk		
	Estadístico	gl	Sig.
Peso (kg)	0,963	15	0,74*
LDL (mg/dl)	0,895	15	0,08*
*>0,05 Distribución normal			

Tercero: ejecutar el análisis. Hacer clic en Analizar → en Correlaciones → Bivariadas
Este ejemplo muestra cómo obtener los coeficientes de correlación y los estadísticos del procedimiento Correlaciones bivariadas.

En el **cuadro de diálogo Correlaciones bivariadas** seleccionar las variables , peso en kg y colesterol LDL y trasladarlas a la lista Variables.

Marcar la **opción Pearson**, Tau-b de Kendall y Spearman del recuadro Coeficientes de correlación.
Pulsar el **botón Opciones**... para acceder al cuadro de diálogo Correlaciones bivariadas:

Opciones y, en el recuadro **Estadísticos**, marcar las opciones **Medias y desviaciones estándar** y Productos cruzados y covarianzas.

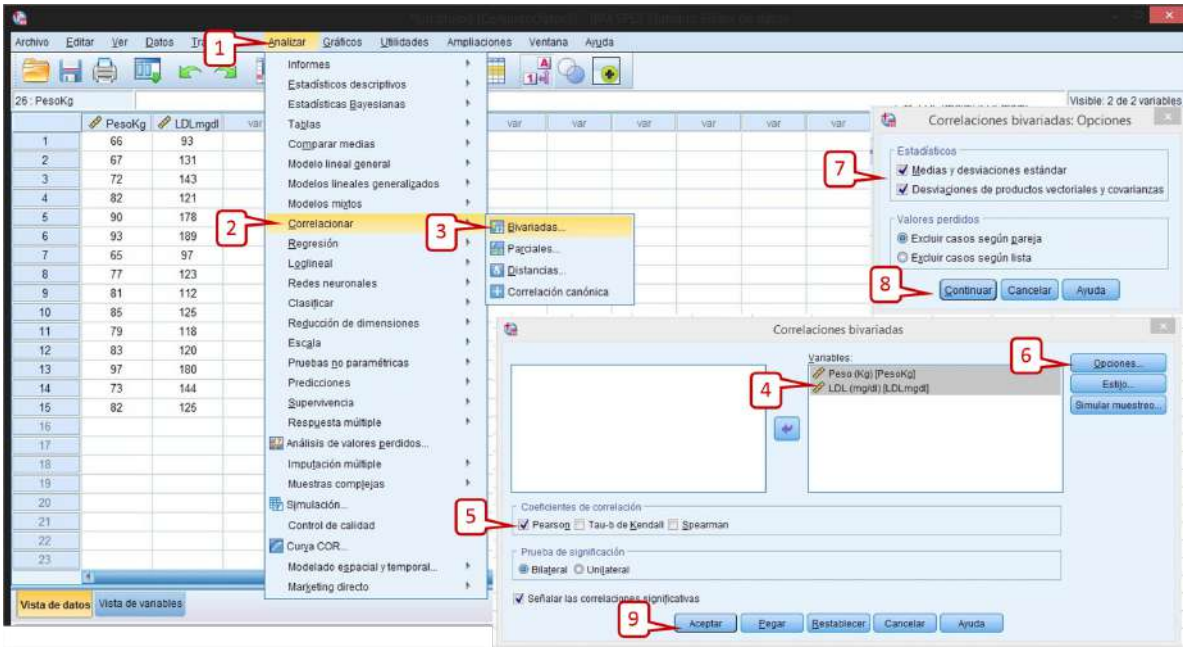


Figura 61. Coeficiente de correlación de Pearson.

Presentación de resultados

Cuarto: en la hoja de resultados aparece:

1. Tabla con la estadística descriptiva de las variables cuantitativas continuas.

Estadísticos descriptivos			
	Media	**Desv. Desviación**	**N**
Peso (kg)	79,47	9,67	15
LDL (mg/dl)	133,27	28,91	15

Tabla donde se presentan Estadísticos descriptivos la media, desviación estándar de cada variable

2. La siguiente tabla ofrece la información referida al coeficiente de **correlación de Pearson**. Cada celda contiene valores referidos al cruce entre cada dos variables: 1) el valor del coeficiente de correlación de Pearson; 2) el nivel crítico bilateral que corresponde a ese coeficiente

(Sig. bilateral; el nivel crítico unilateral puede obtenerse dividiendo por 2 el bilateral); 3) la suma de cuadrados (para el cruce de una variable consigo misma) y la suma de productos cruzados (para el cruce de dos variables distintas); 4) la covarianza; y 5) el número de casos válidos (N) sobre el que se han efectuado los cálculos.

El nivel crítico permite decidir sobre la hipótesis nula de independencia lineal (o lo que es lo mismo, sobre la hipótesis de que el coeficiente de correlación vale cero en la población). Rechazaremos la hipótesis nula de independencia (y concluiremos que existe relación lineal significativa) cuando el nivel crítico sea menor que el nivel de significación establecido (0,05). Así, basándonos en los niveles críticos podemos afirmar que las variables Peso (kg) y colesterol LDL (mg/dl) se correlacionan significativamente (Sig. = 0,002), con un coeficiente de correlación de 0,729 valorado como Buena correlación.

Correlaciones

		Peso (kg)	LDL (mg/dl)
Peso (kg)	Correlación de Pearson	1	0,729**
	Sig. (bilateral)		0,002
	Suma de cuadrados y productos vectoriales	1309,73	2853,13
	Covarianza	93,55	203,79
	N	15	15
LDL (mg/dl)	Correlación de Pearson	0,729**	1
	Sig. (bilateral)	0,002	
	Suma de cuadrados y productos vectoriales	2853,13	11696,93
	Covarianza	203,79	835,49
	N	15	15

**. La correlación es significativa en el nivel 0,01 (bilateral).

Adicionalmente, podemos generar un diagramas de dispersión. Estos diagramas son útiles para representar datos multivariantes. Pueden ayudar a determinar posibles relaciones entre variables de escala. Un diagrama de dispersión simple utiliza un sistema de coordenadas 2-D

para representar dos variables. Un diagrama de dispersión 3-D utiliza un sistema de coordenadas 3-D para representar tres variables.

Creación de un diagrama de dispersión simple

En la sección Tipos de gráfico del generador de gráficos

- Para el tipo de gráfico, haga clic en **Dispersión / Punto**.
- Haga clic en la imagen que dice **Matriz de diagrama de dispersión**.
- En el cuadro **Variables** en la parte superior izquierda, mantenga presionada la tecla Ctrl y haga clic en los nombres de variables Eje X. (peso) Y. (colesterol LDL) arrástrelos al cuadro en la parte inferior del gráfico
- Por último, haga **clic** en **Aceptar** .

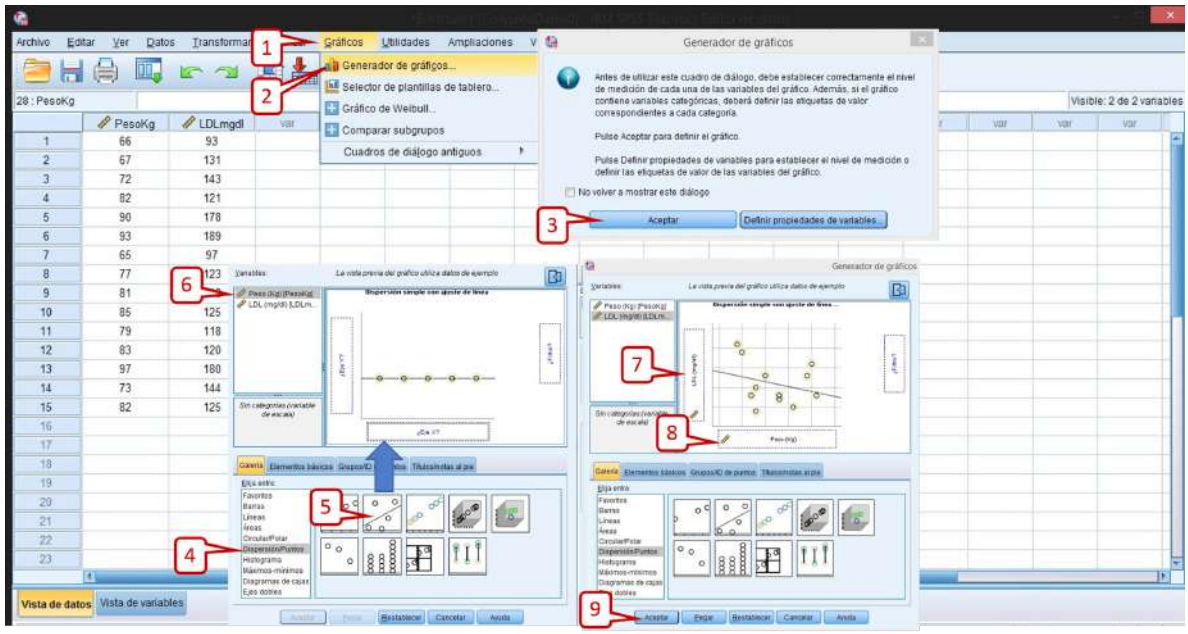


Figura 62. Creación de un diagrama de dispersión simple.

La siguiente matriz de gráficos de dispersión aparecerá automáticamente:

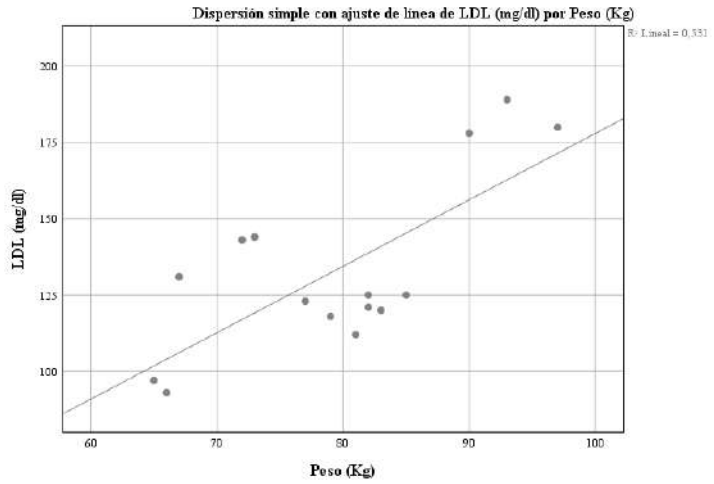


Figura 63. Gráfico de dispersión.

4.4.2. Estadística no paramétrica

La estadística no paramétrica es una rama de la inferencia estadística **cuyos cálculos y procedimientos están fundamentados en distribuciones desconocidas.** En otras palabras, la estadística no paramétrica intenta averiguar la naturaleza de una variable aleatoria. Las pruebas no paramétricas engloban una serie de pruebas estadísticas que tienen como denominador común la ausencia de asunciones acerca de la ley de probabilidad que sigue la población de la que ha sido extraída la muestra. Por esta razón es común referirse a ellas como pruebas de distribución libre.

Las pruebas no paramétricas reúnen las siguientes características: 1) se puede utilizar estas pruebas, aunque se desconozca los parámetros de la población en estudio; 2) nos permiten analizar datos en escala nominal u ordinal; 3) se pueden usar cuando dos series de observaciones provienen de distintas poblaciones; 4) son la única alternativa cuando el tamaño de muestra es pequeño; y 5) son útiles a un nivel de significancia previamente especificado.

Algunos autores utilizan el término ***no paramétricos*** para referirse únicamente a los contrastes que no plantean hipótesis sobre parámetros y que se limitan a analizar las propiedades nominales u ordinales de los datos, y añaden el término ***de distribución libre*** para referirse a los contrastes que no necesitan establecer supuestos (o establecen supuestos poco exigentes, como simetría o continuidad) sobre las poblaciones originales de las que se extraen las muestras.

4.4.2.1. Análisis no paramétrico para variables cualitativas

Existen pruebas específicas para la comparación de grupos cuando la escala de medición de las variables es cualitativa, ejemplo 2, grupos independientes (sexo), se puede comparar el porcentaje de pacientes que alcanzan el nivel establecido de colesterol total normal utilizando la prueba de chi-cuadrado, pero si las muestras son relacionadas deberá emplearse la prueba de McNemar. En caso de comparar 3 o más grupos independientes también se utiliza chi-cuadrado; en caso de muestras relacionadas, Q de Cochran.

Un punto a destacar cuando se emplean las pruebas de comparación de proporciones es que se deben cumplir ciertas condiciones: cuando el número de datos sea menor a 30 se aplicará la corrección de Yates, mientras que la prueba exacta de Fisher debe ser utilizada en lugar de X^2 cuando se comparan 2 grupos independientes si en algunas de las casillas de la tabla de contingencia se encuentra algún valor menor de 5.

4.4.2.1.1. Prueba chi-cuadrado de bondad de ajuste para una muestra

La prueba ***chi-cuadrado*** para una muestra permite averiguar si la distribución empírica de una variable categórica se ajusta o no (se parece o no) a una determinada distribución teórica (uniforme, binomial, multinomial, etc.). Esta hipótesis de ajuste, o mejor, de bondad de ajuste, se pone a prueba utilizando un estadístico originalmente propuesto por Pearson (1900; ver también Cochran, 1952) para comparar las frecuencias observadas o empíricas con las esperadas o teóricas de cada

categoría, es decir, un estadístico diseñado para comparar las frecuencias de hecho obtenidas en una muestra concreta (frecuencias observadas: ***n¡)*** con las frecuencias que deberíamos encontrar si la variable realmente siguiera la distribución teórica propuesta en la hipótesis nula (frecuencias esperadas (f_t), se utiliza para variables nominal politómica y ordinal.

$$\chi^2 = \sum \frac{(fo - ft)^2}{ft}$$

Las frecuencias esperadas (f_t) se obtienen multiplicando la probabilidad teórica de cada categoría (la que corresponde a cada categoría de acuerdo con la hipótesis nula) por el número de casos válidos: Si no existen casillas vacías y el número de frecuencias esperadas menores de 5 no superan el 20% del total de frecuencias esperadas (Cochran, 1952), el estadístico X^2 se distribuye según el modelo de probabilidad *chi-cuadrado* con *k-1* grados de libertad (donde *k* se refiere al número de categorías de la variable cuyo ajuste se está intentando evaluar).

En las investigaciones **podemos realizar suposiciones** sobre el valor de algún parámetro estadístico. **Estas proposiciones se deben contrastar con la realidad** (mediante el muestreo de datos) **para tomar una decisión entre aceptar o rechazar la suposición.**

Estos **supuestos se denominan hipótesis** y el **procedimiento para decidir si se aceptan o se rechazan se llama prueba de hipótesis o de significación.**

En ocasiones estaremos interesados en comparar los resultados obtenidos al realizar un experimento multinomial con los resultados esperados (teóricos). Para ello, recurriremos a la distribución chi-cuadrado, la cual nos permitirá realizar un contraste sobre la bondad del ajuste.

Para obtener la prueba chi-cuadrado de bondad de ajuste para una muestra mediante SPSS

Ejemplo: Se realiza un estudio a un grupo de 40 pacientes hipertensos sobre la frecuencia de realización de alguna actividad física, con intención de averiguar si las frecuencias de las categorías de esa variable se ajustan a una distribución uniforme.

0	3	4	3	3	2	2	2
0	2	2	2	2	3	1	4
1	3	3	0	2	1	4	4
4	2	3	0	2	2	3	2
2	2	4	3	0	3	1	4

Mediante la siguiente escala de Likert

Nunca	0
Casi nunca	1
Ocasionalmente	2
Casi todos los días	3
Todos los días	4

1) Planteamiento de hipótesis:

H_0: Las frecuencias de realización de alguna actividad física se ajustan a una distribución uniforme.

H_1: Las frecuencias de realización de alguna actividad física no se ajustan a una distribución uniforme.

2) Nivel de significación: $\alpha = 0{,}05$

3) Prueba estadística:

Prueba chi-cuadrado para una muestra

4) Determinación de los criterios de decisión

Primero: en el menú seleccionar la opción:

- Analizar

- Pruebas no paramétricas
- Cuadro de diálogos antiguos
- Chi-cuadrado
- **clic** para acceder al cuadro de diálogo *Prueba de chi-cuadrado.*

La lista de variables del archivo de datos ofrece un listado de todas las variables con formato numérico. Para contrastar la hipótesis de bondad de ajuste referida a una variable categórica:

Trasladar esa variable **actividad física** a la lista **Contrastar variables.** Si se selecciona más de una variable, el SPSS ofrece tantos contrastes como variables.

Todas las categorías iguales. Las frecuencias esperadas se obtienen dividiendo el número total de casos válidos entre el número de categorías de la variable. Equivale a efectuar el ajuste a una distribución uniforme.

Segundo: El botón **Opciones** permite obtener algunos estadísticos descriptivos y decidir qué tratamiento se desea dar a los valores perdidos.
Estadísticos. Las opciones de este recuadro permiten obtener algunos estadísticos descriptivos:
Descriptivos. Ofrece el número de casos válidos, la media, la desviación típica, el valor mínimo y el valor máximo. |
Cuartiles. Ofrece los centiles 25, 50 y 75.

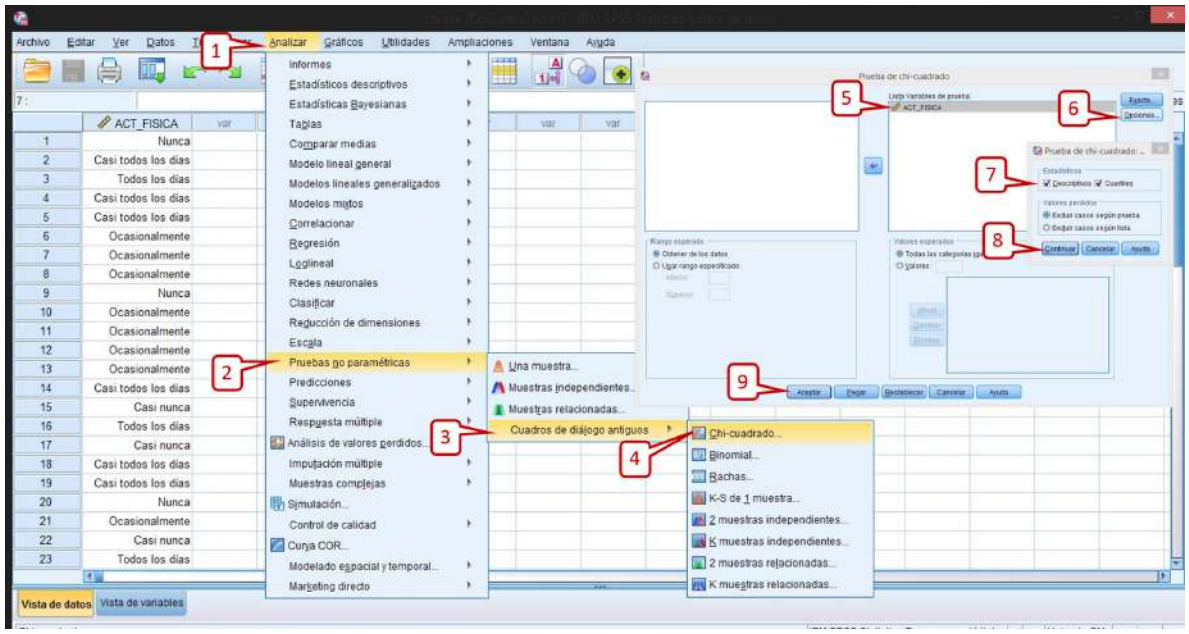


Figura 64. Cuadro de diálogo prueba chi-cuadrado.

Presentación de resultados

Cuarto: en la hoja de resultados aparece:
Aceptando estas elecciones, el *Visor* ofrece los resultados que muestran a continuación:

La tabla siguiente recoge la información descriptiva solicitada al marcar las opciones **Descriptivos y cuartiles:** tenemos una muestra de 40 casos, con un promedio de frecuencia de actividad física 2,25 (ocasionalmente). Los cuartiles indican, por ejemplo, que la mitad de los sujetos realiza ejercicios ocasionalmente (2) y que el 25% de los sujetos realiza ejercicios nunca (0) y casi nuca (1).

Estadísticos descriptivos								
Actividad Física	N	Media	Desv. Desviación	Mínimo	Máximo	Percentiles		
						25	50 (Mediana)	75
	40	2,25	1,235	0	4	2,00	2,00	3,00

La segunda tabla de resultados (contiene las frecuencias observadas y las esperadas, así como las diferencias entre ambas (residual).

Frecuencia de actividad física			
	N observado	N esperada	Residuo
Nunca	5	8,0	-3,0
Casi nunca	4	8,0	-4,0
Ocasionalmente	14	8,0	6,0
Casi todos los días	10	8,0	2,0
Todos los días	7	8,0	-1,0
Total	40		

Estadísticos de prueba	
	Actividad física
Chi-cuadrado	8,250[a]
gl	4
Sig. asintótica	,083

5) Aceptación/rechazo de la hipótesis nula

La información presentada permite tomar una decisión sobre la hipótesis de bondad de ajuste: el valor del estadístico *chi-cuadrado* (8,250), sus grados de libertad 4 (*gl* = número de categorías menos uno) y su nivel crítico (*Sig.* = 0,083). Puesto que el **nivel crítico es mayor que 0,05,** podemos ACEPTAR la hipótesis nula de bondad de ajuste y concluir que **la variable actividad física se ajusta a una distribución uniforme.**

4.4.2.1.2. Prueba chi-cuadrado de independencia

La prueba chi-cuadrado es una de las más conocidas y utilizadas para analizar variables nominales o cualitativas, es decir, para determinar la

existencia o no de independencia entre dos variables. Que dos variables sean independientes significa que no tienen relación, y que por lo tanto una no depende de la otra, ni viceversa.

Así, con el estudio de la independencia, se origina también un método para verificar si las frecuencias observadas en cada categoría son compatibles con la independencia entre ambas variables.

Para realizar este contraste se disponen los datos en una tabla de frecuencias. Para cada valor o intervalo de valores se indica la frecuencia absoluta observada o empírica. A continuación, y suponiendo que la hipótesis nula es cierta, se calculan para cada valor o intervalo de valores la frecuencia absoluta que cabría esperar o frecuencia esperada. El estadístico de prueba se basa en las diferencias entre la O y E y se define como:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Para realizar un contraste chi-cuadrado: la secuencia de una prueba no paramétrica de comparación de proporciones para dos y más de dos muestras independientes, debe cumplir las siguientes características:

- Nivel nominal de la variable dependiente.
- Su función es comparar dos o más de dos distribuciones de proporciones y determinar que la diferencia no se deba al azar (que las diferencia sea estadísticamente significativa).

Parte de la distribución de frecuencias de dos variables cruzadas, representadas en las llamadas tablas cruzadas.

- Se pueden comparar 2 tipos de distribuciones de frecuencias o proporciones:
- Cuando las dos variables tienen cada una dos valores (2 x 2).
- Cuando alguna o las dos variables tiene más de dos valores.

Para comprobar la asociación entre dos variables categóricas dicotómicas (por ejemplo, el sexo (hombre/mujer) con parásitos (sí/no)) se utilizan tablas de 2 x 2 o de contingencia y se evalúa mediante los estadísticos **chi-cuadrado, de Pearson o la prueba la exacta de Fisher.** En general **se utilizará la prueba de chi-cuadrado** si los **valores esperados en cada celda son iguales o mayores de 5**. Se permite que haya una celda (el 25%, 1 de las 4 celdas) con una frecuencia esperada inferior a 5. Cuando los **valores esperados son menores a 5** en más del 25% de las celdas **se utilizará la prueba exacta de Fisher**.

Si se quiere evaluar la **asociación entre 2 variables categóricas politómicas**, por ejemplo, rangos de edad (p. ej. 3-5; 6-8; 9-11; 12-14) y parasitados (sí o no) se utilizará la chi-cuadrado de Pearson, si se cumple la misma premisa, que la frecuencia esperada sea mayor o igual a 5 al menos el 75% de las celdas, si no es así, en este caso no se puede utilizar la prueba exacta de Fisher y habrá que reagrupar para aumentar las frecuencias esperadas, por ejemplo, unir los rangos de edad: 3-7; 8-11 etc.

Tabla cruzada Sexo *Parasitados

			Parasitados		Total
			No	Si	
Sexo	Mujer	f	114	72	186
		%	50,7%	62,6%	54,7%
	Hombre	f	111	43	154
		%	49,3%	37,4%	45,3%
Total		f	225	115	340
		%	100,0%	100,0%	100,0%

Ejemplo de tabla de contingencia de 2 x 2. Se representan el número y la frecuencia esperada de los pacientes con y sin parásitos entre los hombres y mujeres. En cada casilla **la frecuencia esperada es superior a 5**, por lo que se podría utilizar el estadístico chi-cuadrado de Pearson.

Prueba chi-cuadrado de independencia mediante SPSS

Ejemplo: se tiene interés en comprobar si existe una asociación entre la presencia de parásitos y el sexo, para ello se le ha asignado el valor "1" a la presencia de parásitos y el "0" a la ausencia de parásitos. También se ha asignado el valor "1" a los hombres y el "0" a las mujeres.

1) Planteamiento de hipótesis:

H_0: No existe asociación de la presencia de parásitos con el sexo de pacientes.

H_1: Existe asociación de la presencia de parásitos con el sexo de pacientes.

2) Nivel de significación: $\alpha = 0{,}05$

3) Prueba estadística:

Prueba chi-cuadrado para de independencia.

4) Determinación de los criterios de decisión

Primero: ejecutar el análisis con SPSS hacer clic en Analizar → Estadísticos → Tablas de contingencia. En la ventana de "Tablas de contingencia" introduciremos en filas variables independientes y columnas cada una de las variables categóricas aleatorias que se quiere contrastar.

En casillas indicaremos si queremos que nos muestre la frecuencia esperada y los **porcentajes en columnas** o filas. En Estadísticos señalaremos la pestaña de **"Chi cuadrado"**.

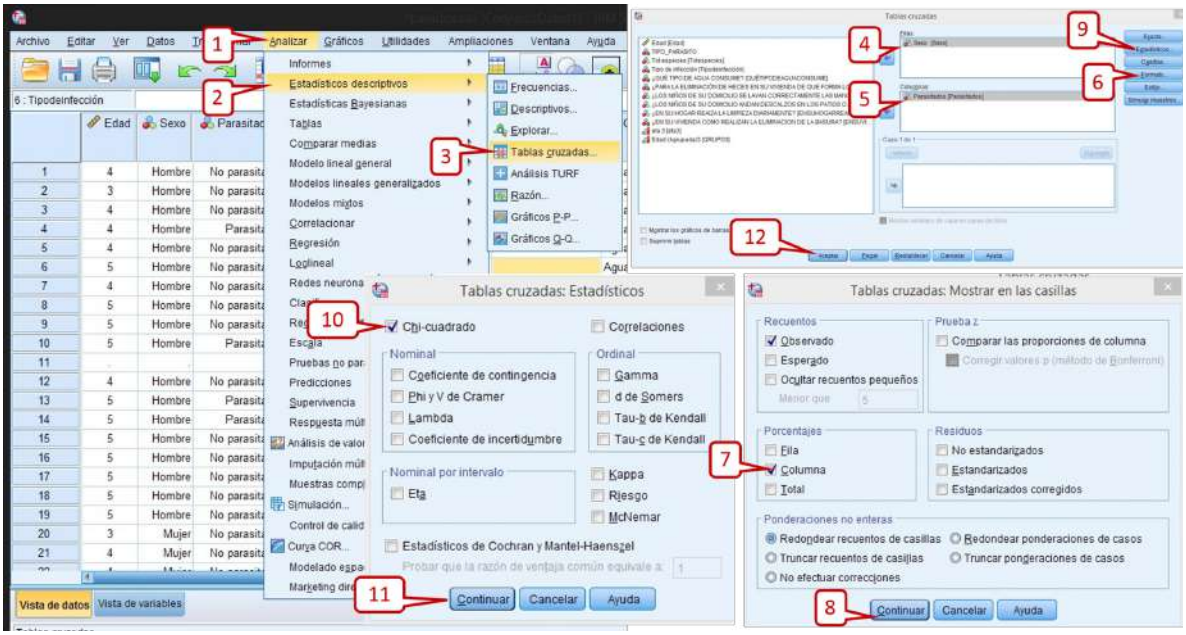


Figura 65. Cuadro de diálogo Prueba chi-cuadrado de independencia.

Presentación de resultados

Segundo: en la hoja de resultados obtenemos una tabla de contingencia como en el ejemplo de arriba y una tabla con las pruebas de chi-cuadrado con el valor, los grados de libertad y la significación estadística.

Pruebas de chi-cuadrado

	Valor	df	Significación asintótica (bilateral)	Significación exacta (bilateral)	Significación exacta (unilateral)
Chi-cuadrado de Pearson	4,380[a]	1	,036		
Corrección de continuidad[b]	3,911	1	,048		
Razón de verosimilitud	4,416	1	,036		
Prueba exacta de Fisher				,039	,024
Asociación lineal por lineal	4,367	1	,037		
N de casos válidos	340				

a. 0 casillas (0,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 52,09.

b. Sólo se ha calculado para una tabla 2 x 2.

5) Aceptación/rechazo de la hipótesis nula
Como más del 75% de las casillas tienen una frecuencia esperada superior a 5 (ninguna casilla tiene frecuencia esperada inferior a 5), podemos utilizar la significación estadística de chi-cuadrado de Pearson. En este ejemplo la significación es inferior a 0,05 (**p = 0,036**), por lo que podemos **RECHAZAR** la hipótesis nula y decir, en este caso, que los **resultados positivos de parásitos son más frecuentes en hombres.**

4.4.2.1.3. Prueba de McNemar (proporciones relacionadas)
Una variante de los **diseños longitudinales** recién estudiados consiste en medir una misma variable dicotómica (positivo-negativo, deprimido-no deprimido, etc.) en dos momentos temporales diferentes. Esta situación es propia de diseños **antes-después** y resulta especialmente útil para medir el cambio. Se procede de la siguiente manera: se toma una medida de una variable dicotómica, se aplica un tratamiento (o simplemente se deja pasar el tiempo) y se vuelve a tomar una medida de la misma variable a los mismos sujetos.

Estos diseños permiten contrastar la hipótesis nula de igualdad de proporciones antes-después, es decir, la hipótesis de que la proporción de éxitos es la misma en la medida antes y en la medida después (la categoría éxito se refiere a una cualquiera de las dos categorías de la variable dicotómica estudiada).

Una prueba no paramétrica McNemar de comparación de proporciones para dos muestras relacionadas, debe cumplir las siguientes características:

- Los datos se ajustan a la distribución de chi-cuadrado.
- Nivel nominal de la variable dependiente.

Su función es comparar el cambio en la distribución de proporciones entre dos mediciones de una variable dicotómica y determinar que la diferencia no se deba al azar (que las diferencia sea estadísticamente significativa).

$$\chi^2 = \frac{(b-c)^2}{b+c}$$

Un ejemplo clásico de evaluación de respuestas clínicas, es cuando un sujeto es evaluado antes de someterse a un tratamiento y luego después de terminado el tratamiento, es decir, el paciente es su propio control. Si la variable respuesta está medida en escala de intervalo **o de razón, las metodologías de análisis son bastante conocidas: el test t-Student para datos pareados o el test de Wilcoxon para datos pareados. Sin embargo, si la respuesta es binaria (con mejoría versus sin mejoría) la situación se complejiza pues la transición de los estados puede ser de la siguiente forma: Sin mejoría (0)**. Con mejoría (1).

Para obtener la prueba McNemar mediante SPSS

Ejemplo: Frente a la COVID-19 aparecen sentimientos de tristeza, hasta llegar incluso a la depresión, para analizar se realizó un estudio en mediante la observación por un periodo de tiempo a 30 personas con tendencia a la depresión, para ello se aplicó un test antes y deputes de sufrir ellos o familiares afectación del COVID-19. Los datos se exponen a continuación:

0 = No depresión; 1 = Depresión

N.°	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Antes	0	1	1	0	1	0	0	1	1	0	1	1	1	1	0
Después	1	1	1	1	0	1	1	1	1	0	1	0	1	0	0
N.°	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Antes	0	1	1	0	0	1	1	0	0	0	0	0	0	0	0
Después	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1

1) Planteamiento de hipótesis:

H_0 : No se presentan cambios en los pacientes depresivos.

H_1 : Se presentan cambios en los pacientes depresivos.

2) Nivel de significación: $\alpha = 0,05$

3) Prueba estadística:
Prueba McNemar

4) Determinación de los criterios de decisión
Este ejemplo muestra cómo obtener e interpretar el estadístico de McNemar del procedimiento Tablas de contingencia.

Primero: ejecutar el análisis con SPSS hacer clic en Analizar → Estadísticos → Tablas
de contingencia. En la ventana de "Tablas de contingencia" introduciremos en filas (antes) y columnas (después).

Pulsar el **botón Estadísticos...** para acceder al subcuadro de diálogo Tablas de contingencia: Estadísticos y marcar la **opción McNemar.** En casillas indicaremos si queremos que nos muestre los **porcentajes total**.

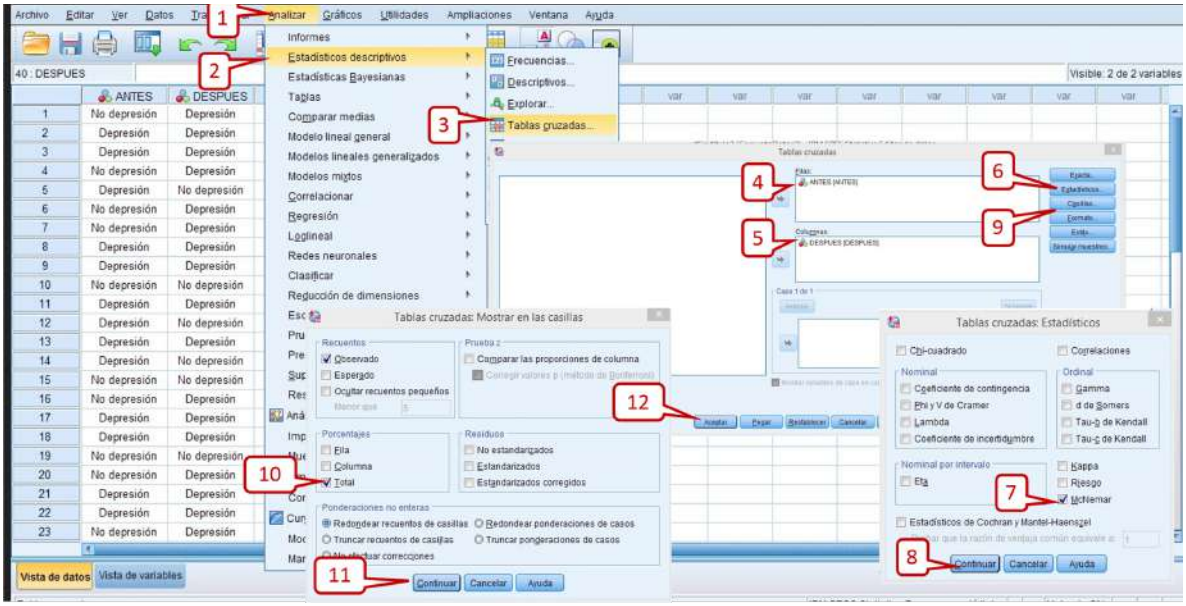


Figura 66. Prueba de McNemar del procedimiento tablas de contingencia.

Presentación de resultados

Segundo: en la tabla se muestra el nivel crítico asociado al número de cambios observados (significación exacta bilateral) y el número de casos válidos. El hecho de que la tabla no muestre el valor del estadístico de McNemar significa que el nivel crítico se ha calculado utilizando la distribución binomial (la cual permite obtener la probabilidad exacta en lugar de la aproximada). El nivel crítico indica el grado de compatibilidad existente entre los datos muestrales y la hipótesis nula de igualdad de proporciones antes-después.

Tabla cruzada antes*después

			Después		
			No depresión	Depresión	Total
Antes	No depresión	f	4	13	17
		%	13,3%	43,3%	56,7%
	Depresión	f	3	10	13
		%	10,0%	33,3%	43,3%
	Total	f	7	23	30
		%	23,3%	76,7%	100,0%

Pruebas de chi-cuadrado

	Valor	Significación exacta (bilateral)
Prueba de McNemar		,021[a]
N.º de casos válidos	30	

a. Distribución binomial utilizada.

5) Aceptación/rechazo de la hipótesis nula

Puesto que el nivel crítico es menor que 0,05, (p = 0,021) podemos **RECHAZAR la hipótesis nula** y concluir que la proporción de pacientes que tenían depresión (13/30 = 0,43) ha cambiado significativamente, ha aumentado el número de pacientes con depresión (23/30 = 0,77).

4.4.2.2. Análisis no paramétrico para variables cuantitativas

Cuando la distribución de **datos cuantitativos no sigue una distribución normal** hay diferentes pruebas estadísticas con las que se comparan las medianas. La prueba de **Wilcoxon se utiliza para comparar un grupo antes y después**, **es decir, muestras relacionadas.** Para la comparación de **grupos independientes se debe emplear U de Mann-Withney**. En el caso de 3 o más grupos independientes se debe utilizar la prueba de Kruskal-Wallis (la cual es equivalente al ANOVA de una vía). La prueba Friedman es la que se recomienda cuando se comparan 3 o más muestras relacionadas (equivalente a ANOVA de 2 vías).

4.4.2.2.1. Prueba U de Mann-Withney

La prueba U de Mann-Whitney es una **prueba no paramétrica alternativa a la prueba t de muestras independientes** (una prueba de hipótesis estadística utilizada para determinar si una media poblacional desconocida es diferente de un valor específico).

La prueba U de Mann-Whitney resulta útil si tenemos dos muestras independientes y **queremos si hay una diferencia en la magnitud de la variable** que estamos estudiando, pero no podemos usar la prueba de t independiente porque los datos no cumplen con alguno de los requisitos.

Importancia de la prueba U de Mann-Whitney

A diferencia de la prueba t de muestras independientes, la prueba U de Mann-Whitney permite sacar diferentes conclusiones sobre los datos en función de las suposiciones que se hagan sobre la distribución de los mismos.

Estas conclusiones pueden ir desde simplemente **afirmar si las dos poblaciones difieren hasta determinar** si hay diferencias en las medianas entre los grupos. Estas diferentes conclusiones dependen de la forma de las distribuciones de los datos.

¿Cómo funciona la prueba U de Mann-Whitney?

La prueba U de Mann-Whitney realiza una comparación estadística de la media y determina si existe una diferencia en la variable dependiente para dos grupos independientes.

La variable dependiente debe medirse a nivel continuo a nivel ordinal

Algunos ejemplos de variables continuas son valores de glucosa, colesterol, triglicéridos, la inteligencia (medida mediante la puntuación del coeficiente intelectual), el peso (medido en kg), el perímetro abdominal en cm, etc.

Ejemplos de variables ordinales son los ítems de la escala de Likert (una escala que va desde 0 a 5 u otra escala), entre otras formas de clasificar categorías (por ejemplo, una escala de 5 puntos que explique cuánto le ha gustado la atención en este centro de salud, desde "No mucho" hasta "Sí, mucho").

El procedimiento Prueba U de Mann-Whitney utiliza el rango de cada caso para comprobar si los grupos se han extraído de la misma población. La prueba de Mann-Whitney contrasta si dos poblaciones muestreadas son equivalentes en su posición. Las observaciones de ambos grupos se combinan y clasifican, asignándose el rango de promedio en caso de producirse empates. El número de empates debe ser pequeño en relación con el número total de observaciones. Si la posición de las poblaciones es idéntica, los rangos deberían mezclarse aleatoriamente entre las dos muestras. La prueba calcula el número de veces que una puntuación del grupo 1 precede a una puntuación del grupo 2 y el número de veces que una puntuación del grupo 2 precede a una puntuación del grupo 1. Fórmula

$$U_1 = n_1 n_2 + \frac{n_1 (n_1 + 1)}{2} - \Sigma R_1$$

$$U_2 = n_1 n_2 + \frac{n_2 (n_2 + 1)}{2} - \Sigma R_2$$

Donde:
U1 y U2 = valores estadísticos de U Mann-Whitney.
n1 = tamaño de la muestra del grupo 1
n2 = tamaño de la muestra del grupo 2
R1 = sumatoria de los rangos del grupo 1
R2 = sumatoria de los rangos del grupo 2

A manera de ejemplo vamos a suponer que tenemos datos diagnósticos de 10 mujeres y 10 hombres. Todos fueron diagnosticados con diabetes y tenemos la edad a la cual se les descubrió la enfermedad. Queremos saber si hay diferencia en la edad entre hombres y mujeres. Mediante la prueba U de Mann-Whitney podría descubrir si la edad que les diagnosticaron la enfermedad es igual en ambos sexos o difieren.

Para obtener la prueba U de Mann-Whitney mediante SPSS
Ejemplo:
La tabla siguiente contiene valores nivel de glucemia (mg/dl) de 30 pacientes de ambos sexo (mujeres 11) y (19 hombres) con sospecha de hiperglucemia.

Sexo	Nivel de glucemia (mg/dl)																		
Hombre	230	124	110	120	114	117	112	122	175	124	116								
Mujer	298	254	119	116	136	119	160	227	132	112	356	151	245	205	115	275	150	257	142

a) Se puede concluir que el nivel de glucemia (mg/dl) difiere en relación al sexo. Calcular el p-valor del contraste. Utilizar un nivel de **significación = 0,05.**

1) Planteamiento de hipótesis:
H_0 : El nivel de glucemia **no es** significativamente diferente entre hombres y mujeres.
H_1 : El nivel de glucemia **es** significativamente diferente entre hombres y mujeres.

2) Nivel de significación: $\alpha = 0,05$

3) Prueba estadística:
U de Mann Whitney para muestras independiente que no tienen una distribución normal

4) Determinación de los criterios de decisión
La prueba de la U de Mann Whitney es una prueba no paramétrica que se utiliza para evaluar la asociación o independencia de variables cuantitativas y una variable categórica dicotómica cuando no tienen una distribución normal o la muestra es muy pequeña (habitualmente n = < 30).

Esta prueba compara los rangos de la variable cuantitativa en los diferentes grupos establecidos por la variable categórica.

Primero: definir en la hoja de vista de variables, las variables que vamos a analizar. En este ejemplo vamos a comprobar si existen diferencias entre los valores de glucemia (mg/dl) entre los pacientes de ambos sexos.

Segundo: comprobar que la variable cuantitativa no tiene una distribución normal.
Tercero: ejecutar el análisis con SPSS. Para ello, en la vista de datos, hacer clic en **Analizar** → **Pruebas no paramétricas** → **Cuadros de diálogos antiguos** → en 2 muestras independientes.

En la ventana para dos muestras independientes, introducir en lista de variables a contrastar, las variables cuantitativas continuas que en nuestra **muestra no tienen una distribución normal,** en este ejemplo **valores de glucemia (mg/dl)**. En variable de agrupación, la **variable categórica, el sexo y definir los grupos** (ej. “1” para hombres y “2” para mujeres). Hacer clic en la pestaña de U de Mann-Whitney y posteriormente en aceptar.

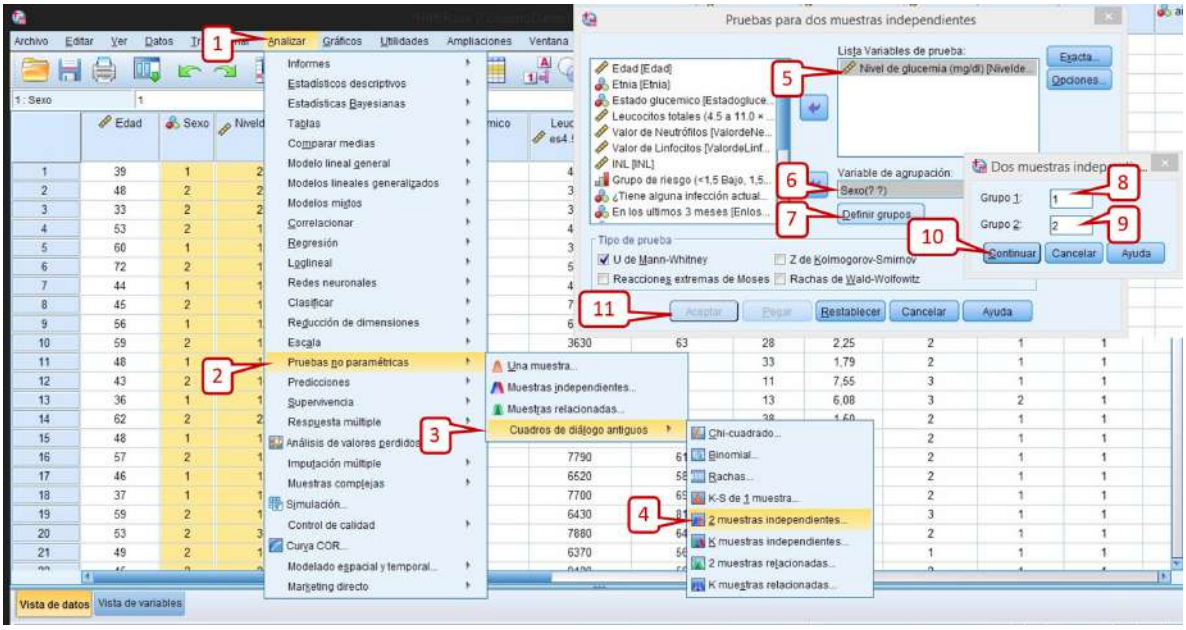


Figura 67. Prueba de U de Mann-Whitney, mediante pruebas no paramétricas.

Presentación de resultados

Primero: En este ejemplo, valores de glucemia (mg/dl) **NO** tiene una distribución normal (p = 0,000 hombres y p = 0,017 mujeres) según el test de Shapiro-Wilk (<50 datos).

Pruebas de normalidad				
	Sexo	Shapiro-Wilk		
		Estadístico	gl	Sig.
Nivel de glucemia (mg/dl)	Hombre	0,625	11	0,000*
	Mujer	0,874	19	0,017*
* < 0,05 No tienen distribución normal				

Segundo: En la ventana de resultado se presentan descriptivos de valores de glucemia (mg/dl) en la que se representa la media, desviación estándar, valor máximo y mínimo.

Estadísticos descriptivos					
	N	Media	Desv. Desviación	Mínimo	Máximo
Nivel de glucemia (mg/dl)	30	167,77	67,93	110	356

En la ventana de resultados aparece una tabla con los rangos promedios y suma de rangos por cada grupo y una segunda tabla con los estadísticos de contraste con el valor de la U de Mann-Whitney, la Z y la significación estadística. Si p<0,05, se rechaza la H y podemos decir que hay diferencias estadísticamente significativas entre ambos grupos.

Rangos				
	Sexo	N	Rango promedio	Suma de rangos
Nivel de glucemia (mg/dl)	Hombre	11	10,64	117,00
	Mujer	19	18,32	348,00
Total		30		

Estadísticos de prueba[a]	
	Nivel de glucemia (mg/dl)
U de Mann-Whitney	51,000
W de Wilcoxon	117,000
Z	-2,303
Sig. asintótica(bilateral)	,021
Significación exacta [2*(sig. unilateral)]	,021[b]
a. Variable de agrupación: Sexo.	
b. No corregido para empates.	

5) Aceptación/rechazo de la hipótesis nula

Puesto que el nivel crítico es menor que 0,05, (p=0,021) podemos **RECHAZAR la hipótesis nula** y concluir que el nivel de glucemia es significativamente diferente entre hombres y mujeres, con mayor nivel para las mujeres.

4.4.2.2.2. La prueba de Wilcoxon

El test no paramétrico prueba de los rangos con signo de Wilcoxon, también conocido como Wilcoxon signed-rank test, permite comparar poblaciones cuando sus distribuciones (normalmente interpretadas a partir de las muestras) no satisfacen las condiciones necesarias para otros test paramétricos. Es una **alternativa al t-test de muestras dependientes cuando las muestras no siguen una distribución normal** (muestran asimetría o colas) o cuando tienen un tamaño demasiado reducido para poder determinar si realmente proceden de poblaciones normales.

A la hora de elegir entre t-test o Wilcoxon, es importante tener en cuenta que, el problema de las muestras pequeñas, no se soluciona con ninguno de los dos. Si el tamaño de las muestras es pequeño, también lo es la calidad de la inferencia que se puede hacer con ellas. Ahora bien, existen dos situaciones en las que, ***a priori***, se puede recomendar utilizar un Wilcoxon antes que un t-test

Si el tamaño de las muestras no permite determinar con seguridad si las poblaciones de las que proceden se distribuyen de forma normal, y no se dispone de información que pueda orientar sobre la naturaleza de las poblaciones de origen (estudios anteriores, que sea un tipo de variable que se sabe que se distribuye casi siempre de forma normal…), entonces la más apropiada es la prueba de Wilcoxon ya que no requiere asumir la normalidad de las poblaciones.

El test de Wilcoxon presenta las siguientes características:

De modo general, el test de Wilcoxon compara si las diferencias entre pares de datos siguen una distribución simétrica en torno a un valor. Si dos muestras proceden de la misma población, es de esperar que las diferencias entre cada par de observaciones se distribuyan de forma simétrica en torno al cero.

Tienen menos poder estadístico (menor probabilidad de rechazar la hipótesis nula cuando realmente es falsa) ya que ignoran valores extremos. En el caso de los t-test, al trabajar con medias, sí los tienen en cuenta. Esto a su vez, hace que el test de Wilcoxon sea una prueba más robusta que el t-test.

Condiciones para la prueba de Wilcoxon

- Los datos tienen que ser dependientes.
- Los datos tienen que ser continuos u ordinales, se deben poder ordenar de menor a mayor o viceversa.
- A pesar de considerarse el equivalente no paramétrico del t-test, el test de Wilcoxon trabaja con medianas, no con medias.

Fórmula

$$Z = \frac{S_+ - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}$$

Obtención de S+:

- Calcular las diferencias en valor absoluto entre las dos puntuaciones de cada pareja.
- Asignar rangos a las diferencias (no incluir las diferencias nulas).
- Sumar los rangos correspondientes a las diferencias positivas (S+) y los correspondientes a las diferencias negativas (S-).

Para obtener la prueba Wilcoxon mediante SPSS

Ejemplo: Se desea estudiar la efectividad de cierta dieta y para ello se toma una muestra aleatoria de 24 mujeres adultas en el grupo de edad de 40-60 años. Se establece el colesterol total antes de iniciar la prueba y al mes de encontrarse realizando el tratamiento. Los resultados se muestran a continuación: ¿Existe una diferencia significativa al 0,05?

N°	CT antes	CT después	N°	CT antes	CT después	N°	CT antes	CT después
1	185	183	**9**	214	226	**17**	195	160
2	258	199	**10**	183	209	**18**	185	180
3	167	160	**11**	198	178	**19**	181	158
4	266	170	**12**	189	227	**20**	180	170
5	244	237	**13**	356	215	**21**	192	192
6	209	230	**14**	151	214	**22**	182	211
7	228	238	**15**	177	155	**23**	182	195
8	377	313	**16**	184	150	**24**	187	178

1) Planteamiento de hipótesis:

H_0 : No hay diferencias entre el colesterol total antes y después del tratamiento

H_1 : Hay diferencias entre el colesterol total antes y después del tratamiento

2) Nivel de significación: $\alpha = 0{,}05$

3) Prueba estadística:

Prueba de Wilcoxon para muestras dependientes que no tienen una distribución normal

4) Determinación de los criterios de decisión

Primero: definir en la hoja de vista de variables, las variables que vamos a analizar. En este ejemplo vamos a comprobar si existen diferencias entre los valores de colesterol total antes y después del tratamiento.

Segundo: comprobar que la variable cuantitativa no tiene una distribución normal.

Tercero: ejecutar el análisis con SPSS. Para ello, en la vista de datos, hacer clic en **Analizar → Pruebas no paramétricas → Cuadros de diálogos antiguos** → en 2 muestras independientes.

Cuarto: si la variable no presenta una distribución normal tenemos que recurrir a una prueba no paramétrica como la prueba de Wilcoxon. Hacer clic en **Analizar** → **Pruebas no paramétricas** → **Cuadro de diálogos antiguos** → en 2 muestras relacionadas.

En la ventana de pruebas para 2 muestras relacionadas introducir la variable colesterol total antes y después del tratamiento y señalar la pestaña de Wilcoxon.

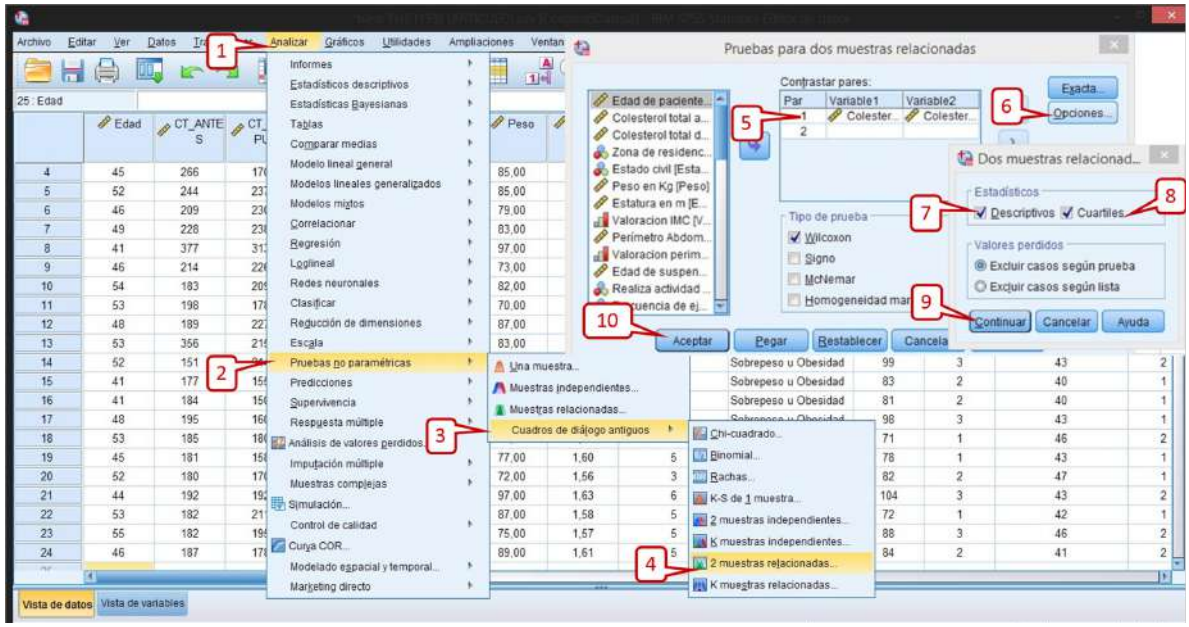


Figura 68. Test de Wilcoxon, mediante pruebas no paramétricas.

Presentación de resultados

La prueba de Wilcoxon realiza una comparación de rangos. En el ejemplo, la variable colesterol total antes y después del tratamiento que no hubiera tenido una distribución normal, obtendríamos el siguiente resultado:

Primero: En este ejemplo, valores de glucemia (mg/dl) **NO** tiene una distribución normal (p = 0,000 antes y p = 0,025 después) según el test de Shapiro-Wilk (<50 datos).

Pruebas de normalidad			
	Shapiro-Wilk		
	Estadístico	gl	Sig.
Colesterol total antes	0,731	24	0,000
Colesterol total después	0,903	24	0,025
*. Esto es un límite inferior de la significación verdadera.			
a. Corrección de significación de Lilliefors			

Segundo: En la ventana de resultado se presentan algunos estadísticos descriptivos para las dos variables seleccionadas: el número de casos válidos en ambas variables, la media, la desviación típica, el valor más pequeño, el más grande y los cuartiles.

Estadísticos descriptivos								
	N	Media	Desv. Desviación	Mínimo	Máximo	Percentiles 25	50 (Mediana)	75
Colesterol total antes	24	211,25	55,23	151	377	182,0	188,0	224,50
Colesterol total después	24	197,83	36,97	150	313	170,0	193,5	223,25

La tabla siguiente contiene información relacionada con la prueba de Wilcoxon. El número de la muestra, media y suma de los rangos negativos y de los rangos positivos. Las notas a pie de tabla permiten conocer el significado de los rangos positivos (8 incrementaron su colesterol total) y negativos (15 pacientes redujeron su colesterol total). También ofrece el número de empates (1 paciente mantuvo igual el colesterol total) y el número total de sujetos.

Rangos				
		N	**Rango promedio**	**Suma de rangos**
Colesterol total después - Colesterol total antes	Rangos negativos	15[a]	11,63	174,50
	Rangos positivos	8[b]	12,69	101,50
	Empates	1[c]		
	Total	24		
a. Colesterol total después < Colesterol total antes				
b. Colesterol total después > Colesterol total antes				
c. Colesterol total después = Colesterol total antes				

La siguiente tabla muestra el estadístico de Wilcoxon (Z) y su nivel crítico bilateral (Sig. asintót. bilateral).

Estadísticos de prueba[a]	
	Colesterol total después - Colesterol total antes
Z	-1,110[b]
Sig. asintótica(bilateral)	,267
a. Prueba de rangos con signo de Wilcoxon	
b. Se basa en rangos positivos.	

5) Aceptación/rechazo de la hipótesis nula

Considerando que el nivel crítico es mayor que 0,05 (p = 0,267) podemos **ACEPTAR la hipótesis nula** de igualdad de promedios y concluir que las variables comparadas (colesterol total inicial y final) **NO difieren significativamente**.

4.4.2.2.3. Prueba de Kruskal-Wallis

El test de Kruskal-Wallis, también conocido como test H, es la **alternativa no paramétrica al test ANOVA de una vía** para datos no pareados. Se trata de una extensión del test de Mann-Whitney para más de dos grupos. Es por lo tanto un test que emplea rangos para contrastar la hipótesis de que k muestras han sido obtenidas de una misma población.

A **diferencia del ANOVA en el que se comparan medias**, el test de Kruskal-Wallis contrasta si las diferentes muestras están equidistribuidas y que, por lo tanto, pertenecen a una misma distribución (población). Bajo ciertas simplificaciones puede considerarse que **el test de Kruskal-Wallis compara las medianas**.

La prueba H de Kruskal-Wallis es una prueba no paramétrica basada en el rango que puede utilizarse para corroborar **si existen diferencias relevantes a nivel estadístico entre dos o más grupos** de una variable independiente en una variable **dependiente ordinal o continua.**

La prueba determina si las medianas de dos o más grupos son diferentes. De esta forma, calcula un estadístico de prueba y lo compara con un punto de corte de la distribución. El estadístico de prueba utilizado se denomina estadístico H. Las hipótesis de la prueba son:

H0: las medianas de la población son iguales.
H1: las medianas de la población no son iguales.

Supóngase que se dispone de k grupos cada uno con n observaciones. Si se ordenan todas las observaciones de menor a mayor y se le asigna a cada una de ellas su rango, cuando se obtenga la suma de rangos para cada uno de los grupos (Ri) es de esperar que, si se cumple la hipótesis nula, todos los grupos tengan un valor similar. Partiendo de esta idea se calcula el estadístico H como: **Fórmula**

$$H = \left(\frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} \right) - 3(n+1)$$

Condiciones para la prueba de Kruskal-Wallis

- No es necesario que las muestras que se comparan provengan de una distribución normal.
- Homocedasticidad: dado que la hipótesis nula asume que todos los grupos pertenecen a una misma población y que por lo tanto

tienen las mismas medianas, es requisito necesario que todos los grupos tengan la misma varianza. Se puede comprobar con representaciones gráficas o con los test de Levenne o Barttlet.
- Misma distribución para todos los grupos: la distribución de los grupos no tiene que ser normal, pero ha de ser igual en todos (por ejemplo, que todos muestren asimetría hacia la derecha) (García *et al*., 2018).

Al ser no paramétrica, la prueba no asume que los datos provienen de una distribución particular. La prueba de **Kruskal-Wallis te dirá si hay una diferencia significativa** entre los grupos. Sin embargo, **no te dirá qué grupos son diferentes**.

Solo es apropiado utilizar una prueba H de Kruskal-Wallis si tus datos "pasan" por cuatro supuestos que son necesarios para que una prueba H de Kruskal-Wallis pueda arrojar un resultado válido:

Supuesto N.º 1: Es necesario medir a nivel **ordinal o continuo su variable dependiente**.

Supuesto N.º 2: Dos o más de dos grupos categóricos e independientes conforman su variable independiente. La prueba H de Kruskal-Wallis **se utiliza cuando se tienen tres o más grupos categóricos independientes**, pero puede utilizarse solo para dos grupos.

Supuesto N.º 3: Es necesario que haya independencia de las observaciones, es decir, no se presente ninguna relación entre las observaciones de los grupos o entre los grupos.

Para obtener la prueba H de Kruskal-Wallis mediante SPSS
Ejemplo: La facultad de Ciencias de la Salud de una prestigiosa universidad está interesada en evaluar un programa de reducción de peso que considera aplicar con los empleados de una universidad. La muestra la conforman 24 empleados obesos que son asignados al azar

a tres condiciones, con ocho empleados a cada una. Los sujetos en la condición 1 se someten a una dieta que reduce su ingesta calórica diaria. Los sujetos en la condición 2 reciben la misma dieta restringida, pero, además, se les pide caminar un kilómetro cada día. La condición 3 es la condición control, en la cual se pide a los sujetos que mantengan su consumo y los hábitos de ejercicio acostumbrados. Después de 3 meses de tratamientos se evidencia la pérdida o aumento de peso, un numero positivo indica pérdida de peso y un número negativo indica aumento de peso. ¿Cuál es la conclusión final del programa? Utilice $\alpha = 0,05$.

Tratamientos	Pérdida de libras de peso						
Dieta baja en calorías	1	1	0	0	9	1	8
Dieta baja en calorías + rutina de ejercicios	2	34	1	2	36	10	13
Grupo control	6	-2	-1	0	16	0	1

1) Planteamiento de hipótesis:

H_0 : No hay diferencias de libras perdidas entre los tratamientos.

H_1 : Al menos un tratamiento es diferente.

2) Nivel de significación: $\alpha = 0,05$

3) Prueba estadística:

La prueba H de Kruskal-Wallis cuando la variable continua no tiene distribución normal.

4) Determinación de los criterios de decisión

Primero: definir en la hoja de vista de variables, las variables que vamos a analizar. En este ejemplo vamos a comprobar si existen diferencias entre los tratamientos (dieta, dieta + ejercicio y grupo control).

Segundo: vamos a comprobar si la pérdida de peso en libras es diferente en cada grupo (tratamientos). Con las pruebas de normalidad de Shapiro-Wilk (<50 datos) hemos comprobado que la variable pérdida de libras no tiene una distribución normal.

Tercero: ejecutar el análisis. Para la comparación de la media de más de 2 grupos cuando las variables cuantitativas continuas no tienen una distribución normal se debe utilizar la prueba de Kruskal Wallis. Para realizarlo con SPSS, hacer clic en **Analizar → Pruebas no paramétricas → Cuadro de diálogos antiguos →** en **K muestras independientes**. En la ventana "Prueba para muestras independientes" introducir en la casilla "Lista contrastar variables", las variables cuantitativas continuas (en este caso la variable libras perdidas) y en "Variables de agrupación", la variable de grupo (en este caso, tratamientos) y definir los grupos.

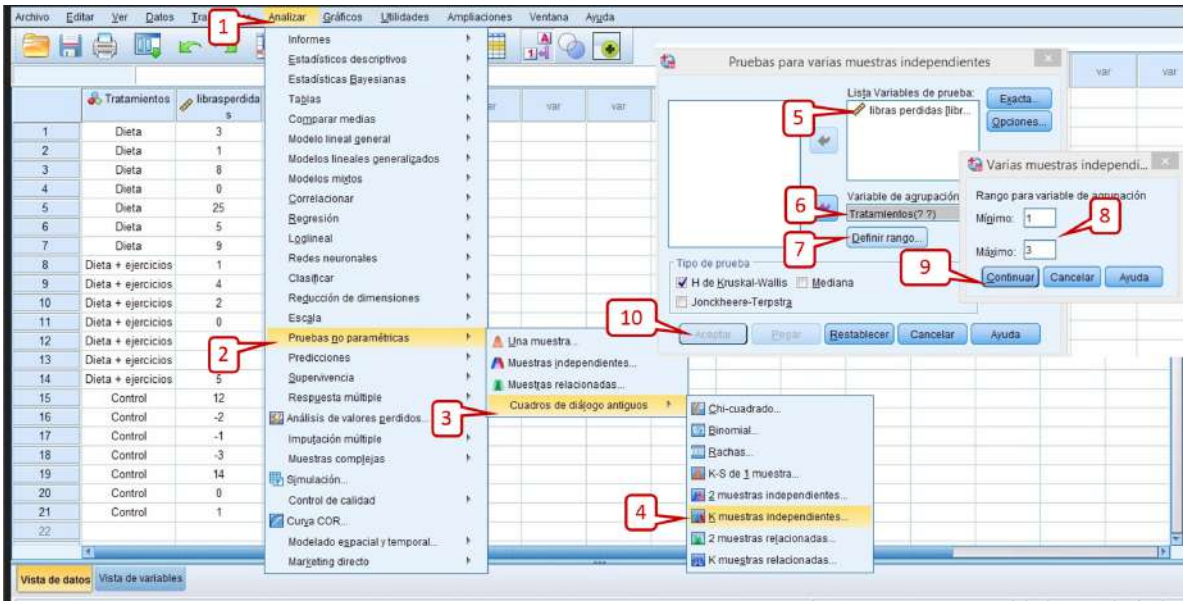


Figura 69. Test de Kruskal-Wallis, mediante pruebas no paramétricas.

Presentación de resultados

Primero: En este ejemplo, valores de glucemia (mg/dl) **NO** tiene una distribución normal (p = 0,000), según el test de Shapiro-Wilk (<50 datos).

Pruebas de normalidad				
Dietas		Shapiro-Wilk		
		Estadístico	gl	Sig.
libras perdidas	Dieta	0,710	7	0,005
	Dieta + ejercicios	0,805	7	0,046
	Control	0,754	7	0,014
*. Esto es un límite inferior de la significación verdadera.				
a. Corrección de significación de Lilliefors.				

Segundo: la tabla "Estadísticos de contraste". La prueba de Kruskal-Wallis utiliza los rangos para la comparación de los grupos. En la tabla "Estadísticos de contraste" aparece el valor de chi-cuadrado y la significación estadística. En este caso, si $p < 0,05$, podemos rechazar la H_0 al menos un tratamiento es diferente.

Rangos			
	Dietas	N	Rango promedio
libras perdidas	Dieta	7	9,57
	Dieta + ejercicios	7	15,71
	Control	7	7,71
	Total	21	

Estadísticos de prueba[a,b]	
	libras perdidas
H de Kruskal-Wallis	6,506
gl	2
Sig. asintótica	,039
a. Prueba de Kruskal-Wallis	
b. Variable de agrupación: Dietas	

Tabla con la prueba de Kruskal-Wallis. Como la significación es inferior a 0,05 ($p = 0,039$) podemos decir que **al menos un tratamiento es diferente del resto.**

Quinto: para averiguar qué grupo es diferente:

- Analizar
- Test no paramétricos
- Muestras independientes.

En la ventana de test no paramétricos, señalar en objetivo "Comparación automática distribuciones entre grupos". En Campos, introducir en "Campos de prueba" la variable **cuantitativa continua, libras perdidas** y en "Grupos", en este caso **tratamientos**, posteriormente hacer clic en Ejecutar.

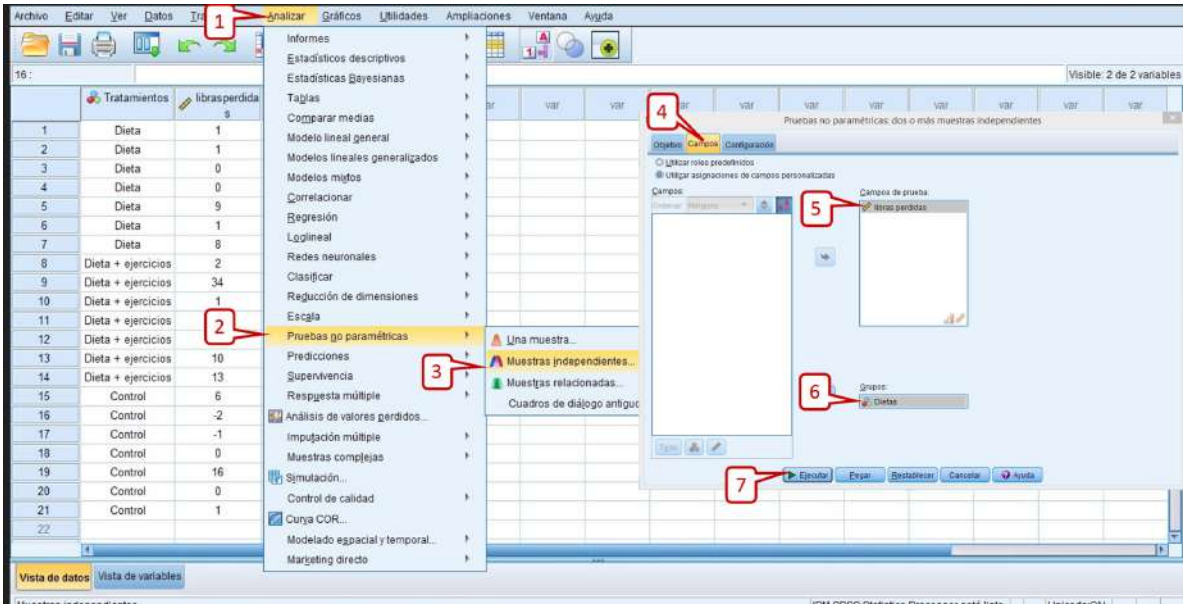


Figura 70. Test de Kruskal-Wallis, comparación de grupos.

Resumen de prueba de hipótesis

	Hipótesis nula	Prueba	Sig.	Decisión
1	La distribución de libras perdidas es la misma entre las categorías de Dietas.	Prueba de Kruskal-Wallis para muestras independientes	,039	Rechazar la hipótesis nula.

Se muestran significaciones asintóticas. El nivel de significación es de ,0

La primera columna representa la hipótesis nula; la segunda, el test estadístico utilizado (en este caso Kruskal-Wallis); la tercera, la significación estadística y, la cuarta, la decisión que debemos tomar, es decir,

rechazar la hipótesis nula y por lo tanto decir que al menos un grupo es diferente.

Ahora para saber qué grupos son diferentes entre sí, debemos **hacer doble clic sobre esta tabla y aparecerá el "visor de modelos".**

Resumen de prueba de hipótesis

	Hipótesis nula	Prueba	Sig.	Decisión
1	La distribución de libras perdidas es la misma entre las categorías de Dietas.	Prueba de Kruskal-Wallis para muestras independientes	,039	Rechazar la hipótesis nula.

Se muestran significaciones asintóticas. El nivel de significación es de ,05

Obtendremos la siguiente figura donde se representa una gráfica de cajas por cada variable, la N total, el valor de la prueba estadística, los grados de libertad y la significación estadística.

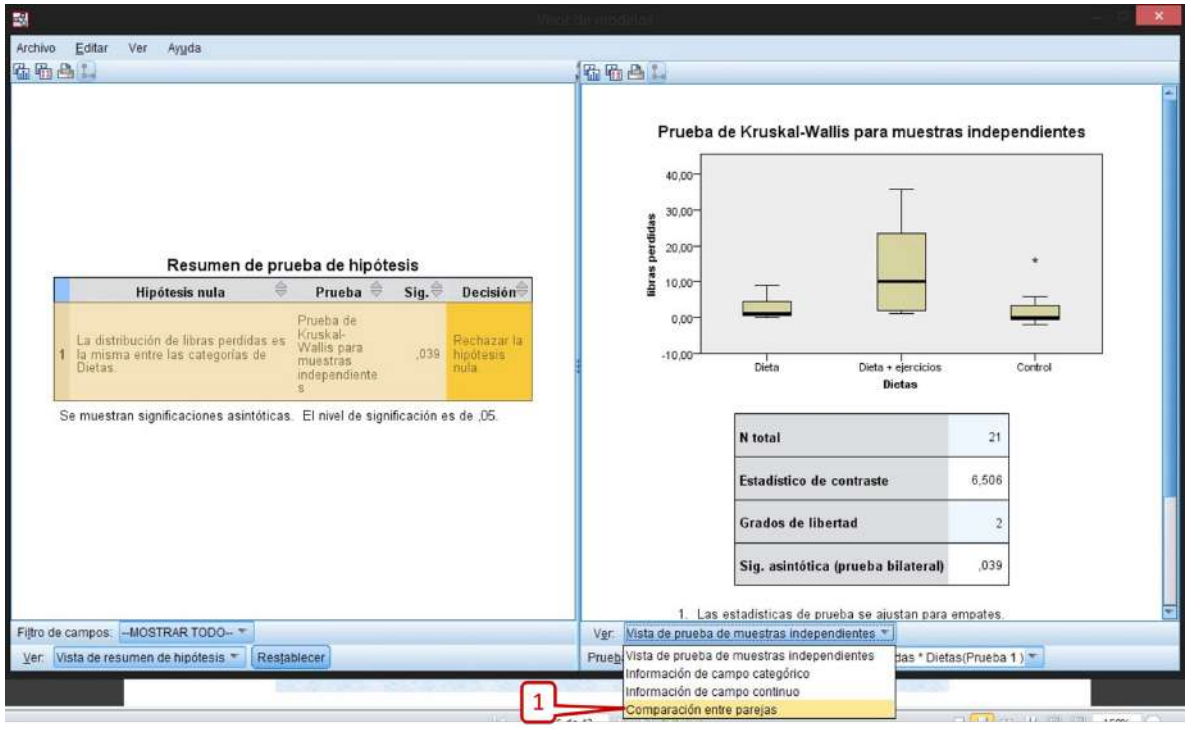


Figura 71. Test de Kruskal-Wallis, comparación entre parejas.

En la columna de la derecha hacer clic en “Ver” y posteriormente en “Comparación entre parejas” y obtendremos la siguiente imagen:

Cada nodo muestra el rango promedio de muestras de Dietas.

Muestra 1-Muestra 2	Estadístico de contraste	Error Error	Desv. Estadístico de contraste	Sig.	Sig. ajust.
Control-Dieta	1,857	3,283	,566	,572	1,000
Control-Dieta + ejercicios	8,000	3,283	2,437	,015	,044
Dieta-Dieta + ejercicios	-6,143	3,283	-1,871	,061	,184

Cada fila prueba la hipótesis nula de que las distribuciones de la Muestra 1 y la Muestra 2 son las mismas.
Se muestran las significaciones asintóticas (pruebas bilaterales). El nivel de significación es ,05.
Los valores de significación se han ajustado mediante la corrección de Bonferroni para varias pruebas.

Figura 72. Test de Kruskal-Wallis, resultados comparación entre parejas.

En esta figura aparece cada una de las comparaciones de los grupos entre sí con el valor de la prueba estadística y la significación estadística, en la que se evidencia que el control – dieta y Dieta – Dieta + ejercicios (no difieren entre sí), pero Control-Dieta + ejercicio presentan diferencias estadísticas entre sí.

5) Aceptación/rechazo de la hipótesis nula

Considerando que el nivel crítico es menor que 0,05 (p = 0,039) podemos **RECHAZAR la hipótesis nula** de igualdad y concluir que la distribución de libras perdidas difiere, al menos una es diferente, el tratamiento de Dieta + ejercicio presentó la mayor pérdida de libras de peso.

4.5. Estimación

En inferencia estadística se llama **estimación** al conjunto de técnicas que permiten dar **un valor aproximado de un parámetro de una población** a partir de los datos proporcionados por una muestra.

Por ejemplo, una estimación de la media de una determinada característica de una población de tamaño N podría ser la media de esa misma característica para una muestra de tamaño (n). La estimación se divide en tres grandes bloques, cada uno de los cuales tiene distintos métodos que se usan en función de las características y propósitos del estudio: estimación puntual, estimación por intervalos y estimación bayesiana.

Vamos a ver **dos tipos de estimaciones: puntual y por intervalo.** La segunda es la más natural. La primera, la estimación puntual, es la más sencilla, y por ese motivo vamos a comenzar por ella. Ocurre, además, que la estimación por intervalo surge, poco más o menos, de construir un intervalo de posibles valores alrededor de la estimación puntual.

Estimación puntual: Se busca un estimador, que con base en los datos muestrales dé origen a un valor puntual que utilizamos como estimación del parámetro

Estimación por intervalos: Se determina un intervalo aleatorio que, de forma probable, contiene el verdadero valor del parámetro. Este intervalo recibe el nombre de intervalo de confianza.

4.5.1. Estimación puntual

Un estadístico es un valor que se obtiene de la muestra y que representa o estima a su parámetro poblacional. Por tanto, es un instrumento mediante el cual podremos estimar parámetros, utilizando lo que llamamos inferencia o estadística inferencial. La estimación puntual hace referencia al cálculo de valores que apuntan hacia el verdadero valor poblacional, como, por ejemplo: estimación de una media o de una prevalencia. De esta forma, un buen estimador debe ser:

- **Insesgado:** Que el valor del parámetro coincida con el valor promedio del estimador. Esta propiedad la tienen la mayoría de los estimadores usados en la práctica.
- **Consistente:** Que el valor de la muestra se acerque al valor del

parámetro al aumentar el tamaño de la muestra.
- **Suficiente**: Que el estimador use toda la información que la muestra contiene respecto al parámetro de interés.
- **Eficiente:** Que el estimador tenga menor variabilidad que otro posible.

Estimar puede tener dos significados interesantes. Significa querer e inferir. Desde luego, el primer significado es más trascendente. Pero no tiene ningún peso en la estadística, el segundo significado es el importante aquí. Una estimación estadística es un proceso mediante el que establecemos qué valor debe tener un parámetro según deducciones que realizamos a partir de estadísticos. En otras palabras, estimar es establecer conclusiones sobre características poblacionales a partir de resultados muestrales.

Una estimación puntual consiste en establecer un valor concreto (es decir, un punto) para el parámetro. El valor que escogemos para decir "el parámetro que nos preocupa vale X" es el que suministra un estadístico concreto. Como ese estadístico sirve para hacer esa estimación, en lugar de estadístico suele llamársele estimador. Así, por ejemplo, utilizamos el estadístico "media aritmética de la muestra" como estimador del parámetro "media aritmética de la población". Esto significa: si quieres conocer cuál es el valor de la media en la población, estimaremos que es exactamente el mismo que en la muestra que hemos manejado.

Por ejemplo, la media muestral es un buen estimador de la media poblacional, porque su valor apunta al verdadero valor promedio en la población. Otros estimadores puntuales son, la proporción muestral para estimar proporciones poblacionales y la desviación estándar en la muestra para estimar la poblacional. En estos ejemplos, la estimación puntual permanece igual, pero como asumimos cierto error por el hecho de elegir una muestra y no otra, debemos acotar el error que cometemos y ello se realiza mediante el intervalo de confianza, pues-

to que la estimación puntual es insuficiente. El intervalo de confianza se puede definir como el intervalo de longitud mínima tal que contiene el verdadero valor del parámetro poblacional con una probabilidad igual a I-α. A efectos prácticos, esto significa que si seleccionamos 100 muestras distintas de una misma población y calculamos el intervalo de confianza del 95%, el estimador obtenido en 95 de estas muestras estará contenido en dicho intervalo.

4.5.2. Intervalos de confianza

El proceso de inferencia es aquel mediante el cual se pretende estimar el valor de un parámetro a partir del valor de un estadístico. Esta estimación puede ser puntual o bien por intervalo. La mejor estimación puntual de un parámetro es simplemente el valor del estadístico correspondiente, pero es poco informativa porque la probabilidad de no dar con el valor correcto es muy elevada, es por eso que se acostumbra a dar una estimación por intervalo, en el que se espera encontrar el valor del parámetro con una elevada probabilidad. Esta estimación recibe el nombre de estimación mediante intervalos de confianza.

Actualmente, las pruebas de hipótesis reciben algunas críticas por varios motivos. En primer lugar, se desconoce la magnitud de la diferencia que se observa y, por tanto, no se puede definir la relevancia clínica. En segundo lugar, damos como significativo un resultado con una p = 0,045 y sin embargo aceptamos la hipótesis nula con una p = 0,05. Y finalmente, con un tamaño de muestra elevado, cualquier resultado puede cobrar significación estadística (Flores Ruiz y Miranda Novales, 2017).

Un intervalo de confianza es un recorrido de valores, basados en una muestra tomada de una población, en el que cabe esperar que se encuentre el verdadero valor de un parámetro poblacional con cierto grado de confianza. En otras palabras, se puede tener gran confianza en que el intervalo resultante abarca el valor verdadero, pues dicho intervalo se ha obtenido por un método que casi siempre acierta.

Un intervalo de confianza posee la ventaja de que se puede calcular para cualquier valor.

Si se desea determinar si es verdadera la diferencia observada entre dos grupos, se calcula el intervalo de confianza de 95% de la diferencia entre sus respectivas medias. Si el intervalo abarca el valor cero, no se puede descartar que no haya una diferencia; si no lo abarca, la probabilidad de que se esté observando una diferencia que en realidad no existe se considera remota.

La **precisión de los resultados** guarda relación con el tamaño muestral y con la variabilidad de los datos, de tal manera que cuanto más grande la muestra, más se acercarán los resultados al **verdadero valor poblacional y más estrecho será el intervalo de confianza**. Asimismo, mientras más grande sea la desviación estándar de los datos, menos precisos serán los resultados y más amplio el intervalo de confianza.

Pasos para calcular el intervalo de confianza para tus datos.

1. **Escribe el fenómeno que te gustaría examinar. Supongamos la siguiente situación:** peso promedio de niños en edad escolar del cantón Jipijapa*.* **¿Cuál es el peso de los niños dentro de un intervalo de confianza?**
2. Selecciona una muestra de la población escogida.
3. Calcula el promedio y la desviación estándar de tu muestra.
4. **Elegir el nivel de confianza**. **Los niveles de confianza usa**dos con mayor frecuencia son 90%, 95% y 99%. El más utilizado 95%.
5. **Calcular error estándar = $\sigma/\sqrt{(n$**
6. **Expresa tu intervalo de confianza. Usar esta fórmula práctica para encontrar el intervalo de confianza: $\bar{x} \pm Z_{a/2} * \sigma/\sqrt{(n)}$.**

4.5.2.1. Intervalos de confianza y contrastes para la media en SPSS

Vamos a realizar a continuación algunos contrastes y a calcular intervalos de confianza, estos se llevan a cabo simultáneamente en ambos procedimientos.

Ejemplo: Se obtuvieron, a partir de una muestra de 25 hombres adultos físicamente con obesidad, los siguientes valores de triglicéridos

193,5	132,0	223,0	97,0	58,3
184,9	148,0	147,0	398,0	233,0
138,6	180,0	204,0	83,0	140,8
148,0	104,0	182,2	186,0	207,9
195,3	160,0	148,5	198,0	159,0

A partir de estos datos, calcula el intervalo de confianza al 95 por ciento para el nivel medio de esta variable de la población.

Primero: Identificar la variable en estudio, en este ejemplo los valores de triglicéridos (mg/dL) de 25 hombres adultos con obesidad.

Segundo: en la vista "Editor de datos" se seleccionar**á**:

- Analizar
- Comparar medias
- Prueba T para una muestra

En el cuadro Variables para contrastar seleccionamos la variable que contiene los valores de Triglicéridos (mg/dL). El nivel de confianza utilizado por defecto es del 95%, **para cambiarlo**, hay que seleccionar el **botón opciones.**

El Valor de prueba corresponde a μ en el contraste de hipótesis. Por defecto toma el valor cero, y pulsamos aceptar.

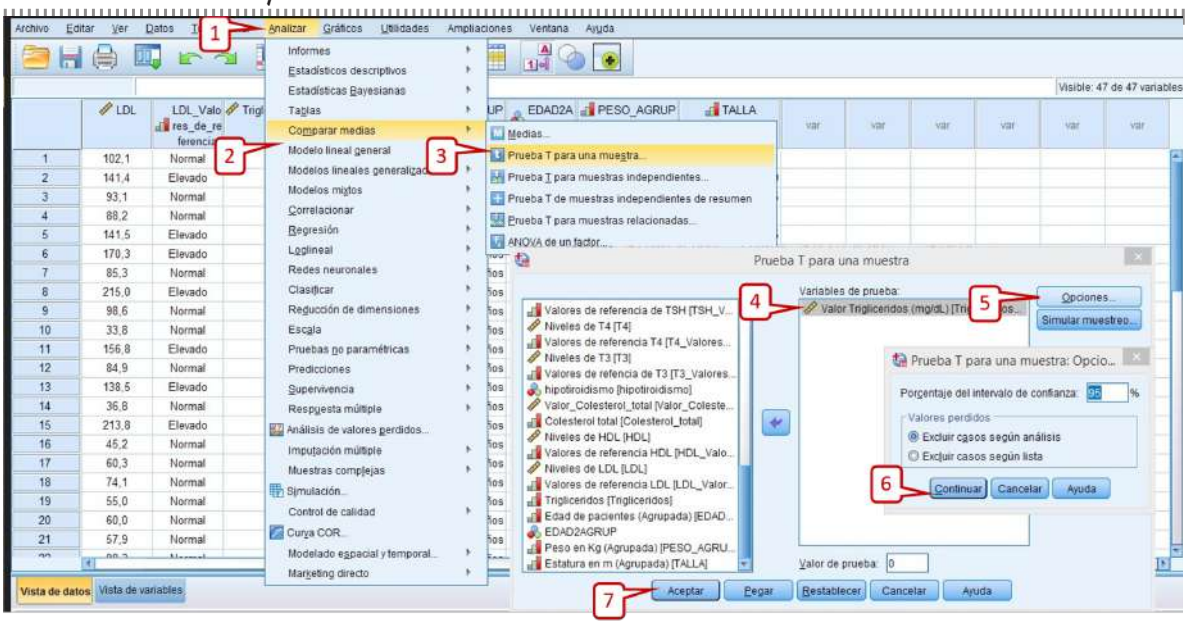


Figura 73. Intervalos de confianza para una media.

Cuarto: como resultado en general, la primera tabla muestra los descriptivos n = 25; una media de 170 ± 64,3. En este caso, el intervalo de confianza **IC =** (143,46 a 196,56), **es decir el verdadero valor poblacional** de triglicéridos de hombres adultos se ubica entre 143,46 a 196,56 (mg/dL).

Estadísticas para una muestra

	N	Media	Desv. Desviación	Desv. Error promedio
Valor triglicéridos (mg/dL)	25	170,0	64,31	12,86

Prueba para una muestra

	Valor de prueba = 0					
	t	gl	Sig. (bilateral)	Diferencia de medias	95% de intervalo de confianza de la diferencia	
					Inferior	Superior
Valor Triglicéridos (mg/dL)	13,222	24	,000	170,0	143,46	196,56

4.5.2.2. Estimación por intervalo de confianza de una proporción IC (p)
Cuando trabajamos con variables cualitativas (nominales u ordinales) no es posible calcular la media ni la desviación estándar, sino solo considerar la proporción de casos que hay en una categoría que elegimos. La proporción es la frecuencia relativa de la categoría que se elige, la cantidad de casos en esa categoría dividida por el tamaño de la muestra. Cuando se trata de variables con solo dos categorías (dicotómicas) puede elegirse cualquiera de ellas. Por ejemplo, si es el resultado de un examen de laboratorio como positivo o negativo, podemos interesarnos por la proporción de cualquiera de ellas, ya que la otra es el complemento (lo que le falta para llegar a uno). Si una es 0,60 (60%), la otra no puede sino ser 0,40 (40%). Es diferente si la variable tiene más de dos categorías, por ejemplo, si se trata de la determinación de los parásitos más prevalentes en niños de edad escolar. Allí es usual que haya más de un parásito presente, por lo que, conocer la proporción de uno de ellos no nos dice mucho sobre cada uno de los otros: si hay cinco tipos de parásitos y uno se lleva el 40%, solo sabemos que el 60% restante se reparte entre los otros cuatro, pero no sabemos cuánto le corresponde a cada uno. A estos casos los trataremos como si fueran dicotómicos: una categoría será el parásito de mayor interés y la otra categoría estará formada por todos los demás parásitos identificados. Lo que hacemos con este procedimiento es simplemente llamar la atención sobre una categoría y confrontarla con el resto indiscriminado.

Para obtener el intervalo de confianza para proporciones, deberemos aplicar la siguiente fórmula:

$$\boldsymbol{IC(p)} = p \pm Z\alpha / 2 * \sqrt{\frac{p*q}{n}}$$

Cálculo de intervalos de confianza de una proporción
Imaginemos que deseamos estimar la prevalencia de parasitosis en niños de 5 a 11 años de la población urbana del cantón Jipijapa. Para ello, se investiga una muestra de 88 niños y se les realizó el examen de

heces directo con SSF al 0,85% y coloración temporal de lugol, obteniendo los siguientes resultados:

1= (parasitados), 2 = (no parasitados)

1	2	1	2	2	2	2	1	2	2	1
2	1	1	1	2	2	1	2	2	2	2
2	2	2	2	2	2	1	2	2	1	2
1	2	1	1	2	2	2	2	1	2	2
2	2	1	2	1	2	2	1	2	1	2
1	1	1	2	2	2	2	1	2	2	1
1	2	2	2	2	2	2	1	2	2	2
2	1	1	2	2	1	2	1	1	1	2

Estimar un intervalo de confianza a un nivel del 95% para estimar la proporción de niños que padecen de parásitos en la población urbana del cantón Jipijapa.

Primero: Determinar la prevalencia de parasitados mediante un análisis de frecuencia.

Segundo: En la vista “Editor de datos”: se selecciona:

- Analizar
- Estadísticos descriptivos
- Frecuencias
- Seleccionar la variable de prueba
- Aceptar

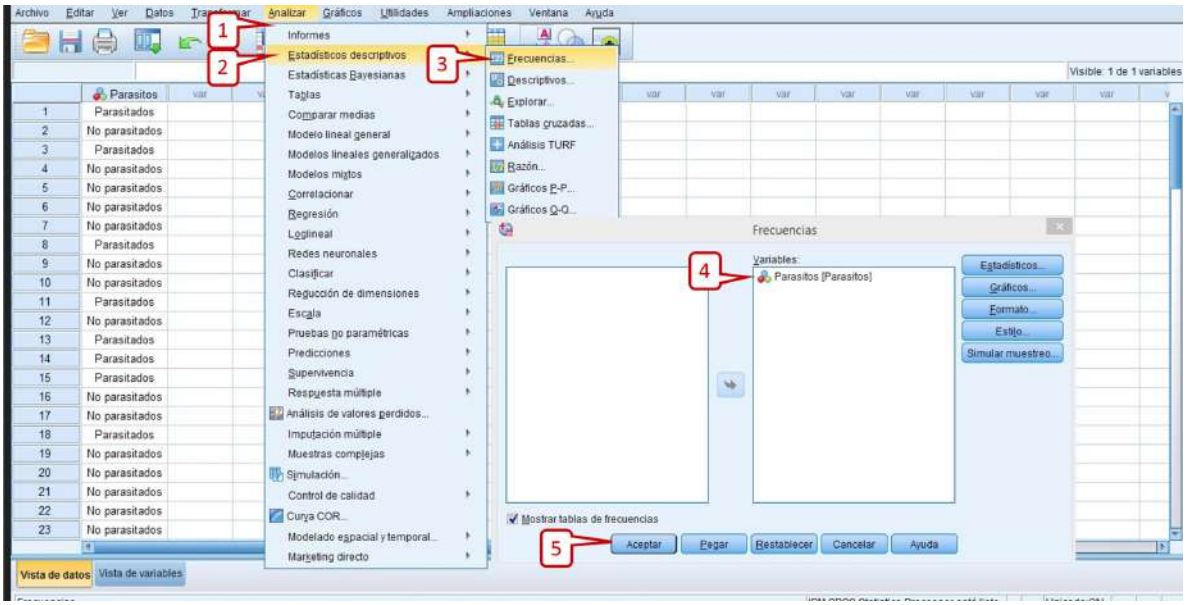


Figura 74. Determinación de frecuencias.

Estado del paciente		
	Frecuencia	**Porcentaje**
No parasitados	57	64,8
Parasitados	31	35,2
Total	88	100,0

La proporción de pacientes parasitados es **p =** 0,352 (35,2%) por lo que la proporción de pacientes no parasitados es **q =** 0,648 (64,8%).

Si decimos que la prevalencia de parasitosis es del **35,2%, estamos incurriendo en el error de no proporcionar el intervalo de confianza correspondiente** a esa proporción y que nos indicará el rango de valores reales que puede adoptar nuestra prevalencia. Una prevalencia del 30% puede obtenerse de una muestra de 30 casos o de un estudio poblacional con miles de pacientes. Cuanto mayor sea el tamaño muestral, mayor será la precisión de nuestra estimación y esto se refleja con el intervalo de confianza, que será menor. Para obtener el intervalo de confianza para proporciones, deberemos aplicar la siguiente fórmula:

$$IC(p) = p \pm Z\alpha / 2 * \sqrt{\frac{p * q}{n}}$$

Donde **n** es el tamaño muestral, **p corresponde a la proporción** obtenida en nuestra muestra, como hemos dicho anteriormente **q corresponde a l-p** y, finalmente, $\mathbf{Z_{\alpha/2}}$ es el valor crítico de la distribución normal que deja una probabilidad 1-**α** bajo la curva y corresponde al valor **1,96; en el caso de definir un nivel de confianza del 95%** se puede obtener mediante las función de Microsoft Excel (=DISTR.NORM.ESTAND.INV(0,05/2)). De esta forma, el cálculo del intervalo de confianza del 95% se obtiene como:

$$IC(p) = p \pm Z\alpha/2 * \sqrt{\frac{p*q}{n}}$$

LI: $IC(p) = 0,352 - 1,96 * \sqrt{\frac{0,352*0,648}{88}} = 0,352 - (1,96 * \sqrt{0,0026}) = 0,352 - 0,0999$

IC: LI= 0,25 = **25%**

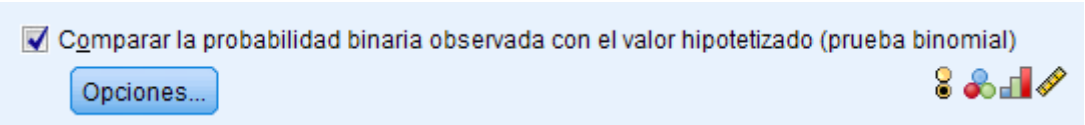
LS: $IC(p) = 0,352 - 1,96 * \sqrt{\frac{0,352*0,648}{88}} = 0,352 - (1,96 * \sqrt{0,0026}) = 0,352 + 0,0999$

IC: **LS=** 0,45 = **45%**

Por lo tanto, a partir de los datos obtenidos en la muestra podemos afirmar que **la prevalencia de la parasitosis en niños de 5 a 11 años** de la población urbana del cantón Jipijapa está entre el **25% y el 45%** con una confianza del 95%.

Estimación por intervalo de una proporción en SPSS

Primero: ejecutar el análisis para la se debe comparar la probabilidad binaria observada con el valor hipotetizado (prueba binomial).



Para realizarlo con SPSS, hacer clic en **Analizar** → en **Pruebas no paramétricas** → **Una muestra**. En la ventana "Objetivo" seleccionar personalizar análisis, en la ventana "Campos" seleccionar la variable de

prueba, en la ventana “Configuración” seleccionar comparar la probabilidad binaria observada con el valor hipotetizado (prueba binomial), en el botón “Opciones” seleccionar la prueba Clopper-Pearson (exacto) y aceptar, finalmente, ejecutar.

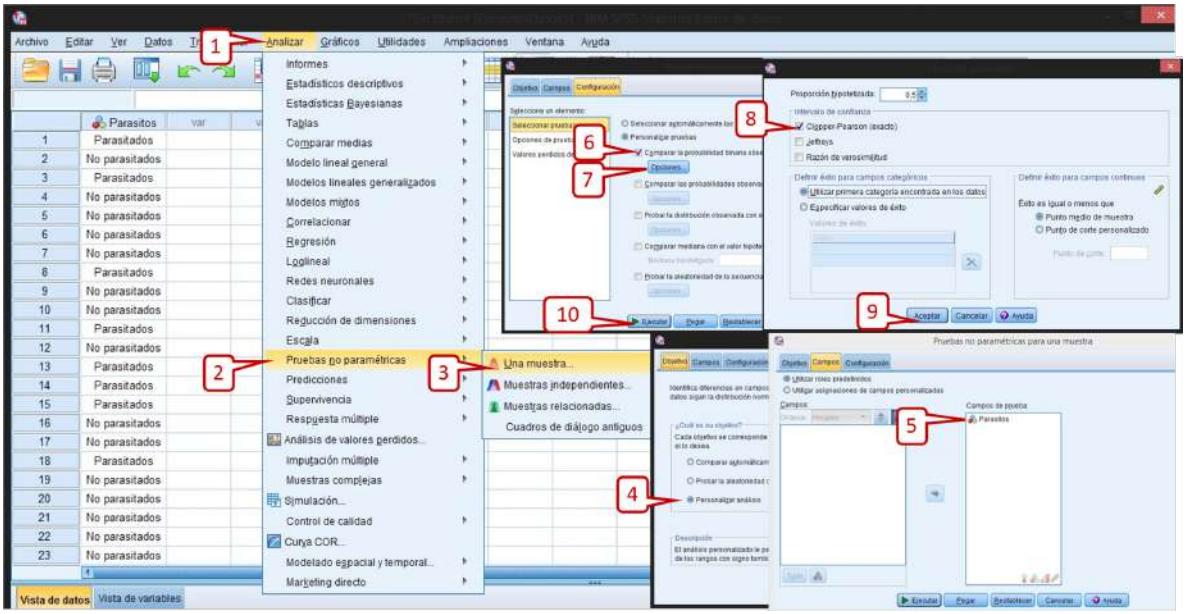


Figura 75. Probabilidad binaria observada con el valor hipotetizado (prueba binomial).

Presentación de resultados

La primera columna representa la hipótesis nula, la segunda el test estadístico utilizado (en este caso prueba binomial para una muestra), la tercera la significación estadística 0,008 y la cuarta la decisión que debemos tomar, es decir rechazar la hipótesis nula.

Resumen de prueba de hipótesis

	Hipótesis nula	Prueba	Sig.	Decisión
1	Las categorías definidas por Estado del paciente <=2 y >2 se producen con las probabilidades 0,5 y 0,5.	Prueba binomial para una muestra	,008	Rechazar la hipótesis nula.

Se muestran significaciones asintóticas. El nivel de significación es de ,05.

Ahora, para saber los **intervalos de confianza de una proporción**, debemos **hacer doble clic sobre esta tabla y aparecerá el “visor”.**

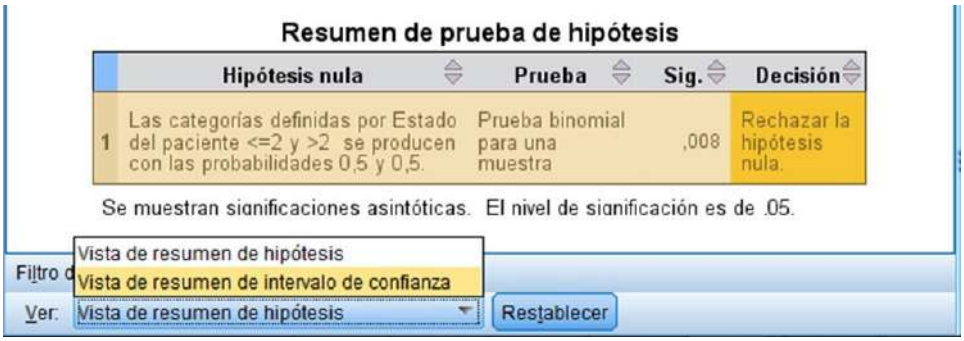


Figura 76. Vista de resumen de intervalos de confianza.

En la columna de la derecha hacer clic en "Ver" y seleccionamos vista de resumen de intervalos de confianza. y obtendremos la siguiente imagen:

Resumen de intervalo de confianza

Tipo de intervalo de confianza	Parámetro	Estimación	Intervalo de confianza 95%	
			Inferior	Superior
Tasa de éxito binomial para una muestra (Clopper-Pearson)	Probabilidad (Parasitos=Parasitados).	,352	,253	,451

Figura 77. Intervalos de confianza para una proporción.

En esta figura aparece el resumen de intervalos de confianza en la primera columna la tasa de éxito binomial para una muestra (Clopper-Pearson), en la segunda el parámetro analizado (parasitados), en la tercera la estimación (prevalencia de parasitados 0,35 = 35%), en la cuarta y quinta los intervalos de confianza, **límite Inferior 0,25 (25%)** y límite **superior 0,45 (45%)**

4.6. Índices de riesgo

Las frecuencias de una tabla de contingencia pueden obtenerse utilizando dos estrategias básicas de recogida de datos. En la estrategia habitual, que es la que hemos supuesto al aplicar todas las medidas

de asociación estudiadas hasta aquí, los datos representan un corte temporal **transversal:** se recogen en el mismo o aproximadamente el mismo punto temporal.

Si, en lugar de esto, medimos una o más variables en una muestra de sujetos y hacemos seguimiento a esos sujetos para volver a tomar una medida de esas mismas variables o de otras diferentes, nos encontramos en una situación **longitudinal:** las medidas se toman en diferentes puntos temporales. Los **índices de riesgo** que estudiaremos en este apartado resultan especialmente útiles para diseños longitudinales en los que medimos dos variables ***dicotómicas.***

El seguimiento de los estudios longitudinales puede hacerse hacia adelante o hacia atrás. En los diseños longitudinales hacia adelante, llamados diseños prospectivos o de cohortes, los sujetos son clasificados en dos grupos con arreglo a la presencia o ausencia de algún factor desencadenante (por ejemplo, el sedentarismo y se les hace seguimiento durante un espacio de tiempo hasta determinar la proporción de sujetos de cada grupo en los que se da un determinado desenlace objeto de estudio (por ejemplo, problemas hipertensivos) y se hace seguimiento hacia atrás intentando encontrar información sobre la proporción en la que se encuentra presente en cada muestra un determinado factor desencadenante.

Lógicamente, cada diseño de recogida de datos permite dar respuesta a diferentes preguntas y requiere la utilización de unos estadísticos particulares.

En los diseños de **cohortes,** en los que se establecen dos grupos de sujetos a partir de la presencia o ausencia de una condición que se considera desencadenante y se hace seguimiento hacia adelante para determinar qué proporción de sujetos de cada grupo alcanza un determinado desenlace, la medida de interés suele **ser el riesgo relativo***:* el grado en que la proporción de desenlaces es más alta en un grupo que en el otro.

Consideremos los datos de la tabla referidos a un estudio sobre la relación entre el sedentarismo*, y* la presencia de ***problemas hipertensivos*** en una muestra de 49 sujetos. (Si = 1), (No = 2).

N.°	Sedentarismo	Problemas hipertensivos	N.°	Sedentarismo	Problemas hipertensivos	N.°	Sedentarismo	Problemas hipertensivos
1	2	1	17	1	1	33	1	1
2	2	2	18	2	2	34	2	2
3	1	1	19	1	2	35	1	1
4	1	1	20	2	2	36	1	1
5	1	1	21	1	1	37	2	1
6	2	2	22	2	2	38	1	2
7	2	1	23	1	1	39	2	1
8	2	2	24	1	2	40	1	2
9	1	1	25	1	1	41	1	1
10	2	2	26	1	1	42	1	1
11	1	1	27	1	1	43	1	2
12	2	1	28	1	1	44	1	1
13	1	1	29	2	2	45	1	2
14	1	1	30	1	1	46	1	1
15	1	1	31	2	1	47	1	1
16	1	2	32	2	2	48	2	2
						49	1	1

Sedentarismo	Problemas hipertensivos		Total
	Sí	No	
Sí	25	7	32
No	6	11	17
Total	31	18	49

Entre los sedentarios, la proporción de casos con problemas hipertensivos vale 25/32 = 0,78 o 78%. Entre los no sedentarios, esa proporción vale 6/17 = 0,35 o 35%. El riesgo relativo se obtiene dividiendo ambas proporciones: 0,78/0,35 = 2,214. Este índice de riesgo **(2,214)** informa sobre el número de veces que es más probable encontrar problemas hipertensivos en sujetos sedentarios que en sujetos no sedentarios. Un índice de riesgo de 1 indica que los grupos considerados no difieren en la proporción de desenlaces.

En los diseños de caso-control, tras formar dos grupos de sujetos a partir de alguna condición de interés, se va hacia atrás buscando la presencia de algún factor desencadenante. El mismo estudio sobre sedentarismo y problemas hipertensivos podría diseñarse seleccionando dos grupos de sujetos diferenciados por la presencia de problemas hipertensivos y buscando en la historia clínica la presencia o no de la falta de ejercicios. Puesto que el tamaño de los grupos se fija a partir de la presencia o ausencia de un determinado desenlace, no tiene sentido calcular un índice de riesgo basado en las proporciones de desenlaces (incidencias), pues el número de sedentarios y no sedentarios no ha sido previamente establecido, sino que es producto del muestreo. Pero podemos calcular la **ratio sedentarios /no- sedentarios** tanto en el grupo de sujetos con problemas hipertensivos como en el grupo de sujetos sin problemas, y utilizar el cociente entre ambas *ratios* como una estimación del riesgo relativo.

Basándonos en los datos de la tabla de contingencia anterior, la ratio sedentarios/no- sedentarios en el grupo de sujetos con problemas hipertensivos vale: 25/31 **= 4,166**; y en el grupo de sujetos sin problemas: 7/11 = **0,636.** El índice de riesgo en un diseño caso-control se obtiene dividiendo ambas ratios: 4,166/0,636 = **6,55.** Este valor **se interpreta de la misma manera que el índice de riesgo relativo** (pues es una estimación del mismo), pero también admite esta otra interpretación: entre **los sujetos con problemas hipertensivos, es 6,55 veces más probable encontrar sedentarios que no sedentarios** Un índice de riesgo de 1 indica que la probabilidad de encontrarnos con el factor desencadenante es la misma en las dos cohortes estudiadas.

Determinación de riesgo en SPSS

Consideremos los datos de la tabla referidos a un estudio sobre la relación entre el sedentarismo*, y* la presencia de *problemas hipertensivos* en una muestra de 49 sujetos. (Si=1), (No=2).

N.°	Sedentarismo	Problemas hipertensivos	N.°	Sedentarismo	Problemas hipertensivos	N.°	Sedentarismo	Problemas hipertensivos
1	2	1	17	1	1	33	1	1
2	2	2	18	2	2	34	2	2
3	1	1	19	1	2	35	1	1
4	1	1	20	2	2	36	1	1
5	1	1	21	1	1	37	2	1
6	2	2	22	2	2	38	1	2
7	2	1	23	1	1	39	2	1
8	2	2	24	1	2	40	1	2
9	1	1	25	1	1	41	1	1
10	2	2	26	1	1	42	1	1
11	1	1	27	1	1	43	1	2
12	2	1	28	1	1	44	1	1
13	1	1	29	2	2	45	1	2
14	1	1	30	1	1	46	1	1
15	1	1	31	2	1	47	1	1
16	1	2	32	2	2	48	2	2
						49	1	1

Este ejemplo explica cómo obtener e interpretar los índices de riesgo del procedimiento.

Primero: ejecutar el análisis con SPSS hacer clic en **Analizar → Estadísticos → Tablas de contingencia**. En la ventana de "Tablas de contingencia" introduciremos en filas variables independientes y columnas cada una de las variables categóricas aleatorias que se quiere contrastar, seleccionar las variables sedentarismo y problemas hipertensivos como variables fila y columna, respectivamente.

En la opción **"Estadísticos"** seleccionar **riesgo** y aceptar.

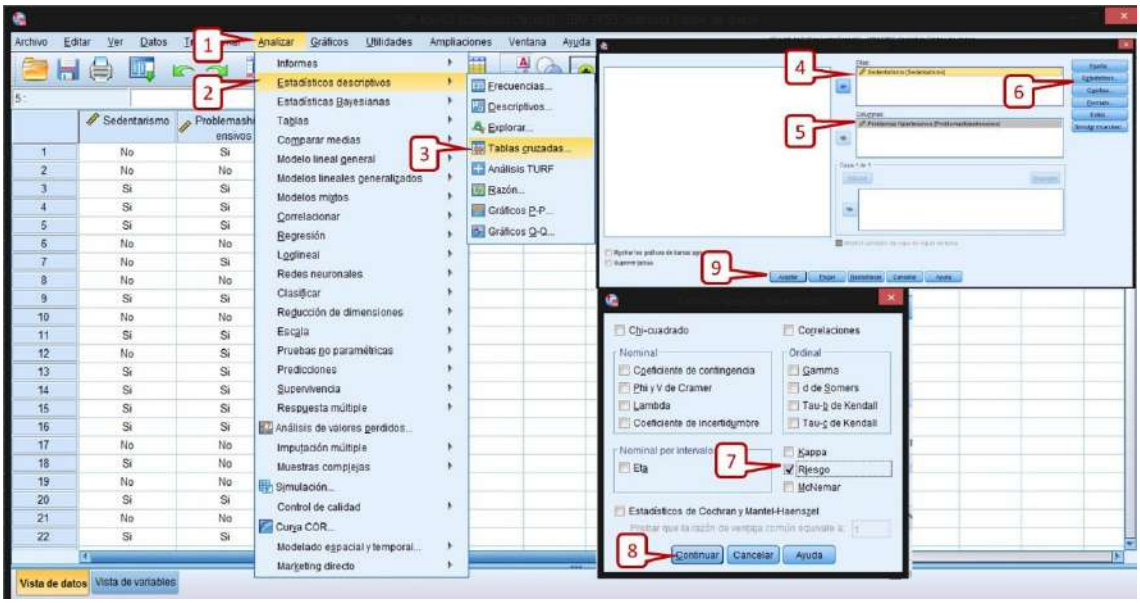


Figura 79. Estimación de riesgo.

Resultados

Aceptando estas elecciones, el visor ofrece los resultados que se muestra a continuación:

Tabla cruzada Sedentarismo*Problemas hipertensivos

Recuento

		Problemas hipertensivos		Total
		Si	No	
Sedentarismo	Si	25	7	32
	No	6	11	17
Total	31	18	49	

Estimación de riesgo

		Intervalo de confianza de 95%	
	Valor	Inferior	Superior
Razón de ventajas para sedentarismo (Si / No)	6,548	1,783	24,043
Para cohorte problemas hipertensivos = Sí	2,214	1,134	4,323
Para cohorte problemas hipertensivos = No	0,338	,161	,711
N de casos válidos	49		

La primera fila de la tabla indica que el riesgo estimado se refiere al de sedentarismo *(sí/no)* en un **diseño de caso-control** (Odds ratio). Su valor (6,55) significa que, entre los sujetos con **problemas hipertensivos, la probabilidad (el riesgo) de encontrar sedentarios es 6 veces mayor que la de encontrar no sedentarios**. La razón de ventajas también puede interpretarse como una estimación del riesgo relativo (particularmente si la proporción de desenlaces es pequeña): el riesgo de padecer problemas hipertensivos es **4 veces más entre sedentarios que entre no sedentarios**.

Los límites del intervalo de confianza calculado al 95 por ciento indican que el riesgo obtenido es mayor que 1: concluiremos que el riesgo es significativamente mayor que 1 cuando, como en el ejemplo, el valor 1 no se encuentre entre los límites obtenidos.

Las dos filas siguientes ofrecen dos índices de riesgo para un **diseño de cohortes** (dos índices porque el desenlace que interesa evaluar puede encontrarse en cualquiera de las dos categorías de la variable). Si el desenlace que interesa estudiar es los problemas hipertensivos, la probabilidad o riesgo de encontrar tal desenlace entre los sedentarios es **2,21 veces mayor que la de encontrarlo entre los no sedentarios**: por cada sujeto ***con*** problema hipertensivos entre los no sedentarios, podemos esperar encontrar 2,21 sujetos ***con*** problema hipertensivos entre los sedentarios. Si el desenlace que interesa estudiar es la ***ausencia*** de problema hipertensivos*,* la probabilidad o riesgo de encontrar tal desenlace entre los sedentarios es menor que entre los no sedentarios: por cada sujeto ***sin*** problema hipertensivos entre los no sedentarios, podemos esperar encontrar 0,338 sujetos sin problema hipertensivos entre los pacientes sedentarios.

1RA Edición

BIOESTADÍSTICA

APLICADA A INVESTIGACIONES
CIENTÍFICAS EN SALUD

BIBLIOGRAFÍA

BIBLIOGRAFÍA

MAWIL
Publicaciones Impresas
y Digitales

Dagnino, J. (2014). Elección de una prueba de hipótesis. Rev Chil Anest.

Del Campo, N. M. S., & Matamoros, L. Z. (2019). ¿Por qué emplear el análisis estadístico implicativo en los estudios de causalidad en salud? Revista Cubana de Informática Médica.

Díaz-Parreño, S. A., Rebollo, J. M. C., Limón, A. R., & Alberca, A. S. (2014). Bioestadística aplicada con SPSS. CEU Ediciones.

Díaz, I., García, C., León, M., Ruiz, F., Torres, F., & Lizama, P. (2014). Asociación entre variables (Pearson y Spearman en SPSS).

Fernández, M., & Minuesa, J. (2018). Estadística básica para ciencias de la salud. Universidad de Extremadura.

Flores Ruiz E, Miranda Novales MG, V. K. (2017). El protocolo de investigación VI: cómo elegir la prueba estadística adecuada. Estadística inferencial. Rev Alerg Mex., 364.

García, J. A. G., Alvarenga, J. C. L., Ponce, F. J., Tapia, Y. R., Pérez, L. L., & Bernal, A. R. (2018). Metodología de la investigación, bioestadística y bioinformática en ciencias médicas y de la salud. McGraw-Hill.

Gómez-Gómez, M., Danglot-Banck, C., & Vega-Franco, L. (2013). Cómo seleccionar una prueba estadística (Primera de dos partes). Revista Mexicana de Pediatría, 80, 30-34.

González, S. H. (2005). Historia de la estadística. La ciencia y el hombre. https://www.uv.mx/cienciahombre/revistae/vol18num2/articulos/historia/

Gutiérrez, A. F. (2017). Medición en epidemiología: prevalencia, incidencia, riesgo, medidas de impacto. Revista Alergia México.

Hernández Sampieri, R. (2014). Metodología de la Investigación (6.a ed.). McGraw-Hill / Interamericana Editores, S.A.

Hernández Sampieri, R., & Mendoza Torres, C. P. (2018). Metodología de la Investigación: Las rutas cuantitativa cualitativa y mixta. Mc-Graw-Hill Interamericana Editores, S.A. (1.a ed.). http://repositorio.uasb.edu.bo:8080/bitstream/54000/1292/1/Hernández- Metodología de la investigación.pdf

Ledesma, R., Macbeth, G., & Cortada de Kohan, N. (2008). Tamaño del efecto: revisión teórica y aplicaciones con el sistema estadístico ViSta. Revista Latinoamericana de Psicología.

Molina, G., & Rodríguez, M. (2019). Estadísticos de asociación entre variables. Estadística descriptiva en psicología.

Monterrey, P., & Gómez-Restrepo, C. (2007). Aplicación de las pruebas de hipótesis en la investigación en salud: ¿estamos en lo correcto? Universitas Médica, 48.

Trochim, W. (2019). Base de conocimiento de métodos de investigación.

1RA Edición

BIOESTADÍSTICA

APLICADA A INVESTIGACIONES CIENTÍFICAS EN SALUD

Publicado en Ecuador
Marzo del 2022

Edición realizada desde el mes de noviembre del 2021 hasta febrero del año 2022, en los talleres Editoriales de MAWIL publicaciones impresas y digitales de la ciudad de Quito

Quito – Ecuador

Tiraje 150, Ejemplares, A5, 4 colores; Offset MBO
Tipografía: Helvetica LT Std; Bebas Neue; Times New Roman; en tipo fuente.

BIOESTADÍSTICA

APLICADA A INVESTIGACIONES CIENTÍFICAS EN SALUD

Autores Investigadores

Ing. Kleber Dionicio Orellana Suarez.
Lcdo. José Clímaco Cañarte Vélez.

ISBN: 978-9942-602-23-7

BIOESTADÍSTICA APLICADA A INVESTIGACIONES CIENTÍFICAS EN SALUD