SINAPSIS EDITORIAL

Fundamentos de ESTADÍSTICA:

desde Conceptos Básicos hasta Análisis Avanzados

Colección: *Eugenio Espejo*

SINAPSIS EDITORIAL

ISBN: 978-9942-638-09-0

9 789942 638090

Este texto ha sido sometido a un proceso de evaluación por pares externos con base en la normativa editorial de Sinapsis.

Editorial Sinapsis
112 pag: 210x297,0 mm.
Título: Fundamentos de Estadística: desde Conceptos Básicos hasta Análisis Avanzados
Edición en español 2024
ISBN digital: 978-9942-638-09-0
Publicación en formato: PDF

AUTORES

ANA MANUELA PALMA AVELLÁN
Ingeniera Industrial
Diploma Superior en Educación
Universitaria por Competencias
Diploma Superior en Investigación
Magíster en Docencia Universitaria e
Investigación Educativa
Doctora en Administración de Empresas
Docente Titular e Investigadora
Universidad Laica Eloy Alfaro
de Manabí, Manta, Ecuador.
https://orcid.org/0000-0003-3071-1594
palmaavellan@gmail.com

LUIS EDUARDO ZAMBRANO PALMA
Empresa SINECUANON
Guayaquil, Ecuador
Investigador Independiente
C.I. 0916918014
lzambrano_85@outlook.com
teléfono 0995599149
Ingeniero Civil
https://orcid.org/0000-0002-7886-7347

CARLOS ORLANDO VÉLEZ CHÁVEZ
Ingeniero Comercial
Ingeniero en Contabilidad y Auditoría
Magíster en Administración de Empresas
Especialización en Dirección de Proyectos
Docente Investigador
Universidad Laica Eloy Alfaro de Manabí
Manta, Ecuador
C.I. 1305887364
https://orcid.org/0000-0003-0371-4773
carlos.velez@uleam.edu.ec

ALEJANDRA VIVIANA PALMA SANCHEZ
Médico Cirujano
Investigadora independiente
Posgrado Neumóloga

Hospital IESS Norte
Guayaquil
C.I: 1311588386
alejapalmasanchez08@live.com
https://orcid.org/0000-0002-3113-7865

ANDREA PATRICIA VERA PALMA

Ingeniera Comercial con mención especial
en Administración Agroindustrial y Agropecuaria
obtenidos en la Universidad Politécnica Agropecuaria
de Manabí Manuel Félix López.
Master en Administración de Empresas
Mención Gestión Pequeñas y Medianas Empresas
Agro productivas.
Portoviejo
C.I: 1314649151
andreapatriciavp@gmail.com
https://orcid.org/0000-0002-0020-8923

HAIDEE MONCERRATE PALMA AVELLÁN
Licenciada en Educación de Párvulos
Academia Naval Almirante Illingworth
Guayaquil
C.I: 1305309625
haideemoncerrat@hotmail.com,
hpalma@anai.edu.ec
https://orcid.org/0000-0002-0109-3320

CARLOS MANUEL ZAMBRANO PALMA
MÉDICO
Teoton S.A.
Investigador Independiente
Guayaquil
C.I.: 0916917990
k_rlosmzp@hotmail.com
https://orcid.org/0009-0008-0429-4188

Prólogo

Bienvenidos a un viaje fascinante hacia el mundo de la estadística. En estas páginas, se encuentra un recurso invaluable que iluminará el camino para comprender y utilizar esta poderosa herramienta matemática en su máxima expresión.

La estadística es mucho más que números y fórmulas. Es el lenguaje que nos permite interpretar el mundo que nos rodea, tomar decisiones informadas y revelar patrones ocultos en la información. En un mundo cada vez más inundado de datos, la habilidad de comprender, analizar y sacar conclusiones significativas a partir de ellos se ha vuelto esencial en prácticamente todas las disciplinas y campos de estudio.

Este libro, ***"Fundamentos de Estadística: Desde Conceptos Básicos hasta Análisis Avanzados",*** es una guía completa que lo llevará desde los principios fundamentales de la estadística hasta técnicas avanzadas de análisis de datos. Es una obra diseñada tanto para aquellos que están dando sus primeros pasos en el mundo de la estadística como para quienes buscan mejorar y expandir su conocimiento en el tema.

A medida que avance por las páginas de este libro, encontrará una estructura cuidadosamente organizada que lo guiará desde los conceptos más simples hasta los más complejos. Comenzamos por definir la naturaleza misma de la estadística, desglosando conceptos clave como población y muestra, parámetros y estadísticas, y la importancia de las distribuciones muestrales y poblacionales.

A lo largo de este viaje, descubrirá cómo describir datos utilizando tablas y gráficos, calcular medidas de tendencia central y entender la variabilidad en sus datos. También exploraremos distribuciones de probabilidad, estimación, y cómo realizar pruebas de hipótesis.

Para hacer que estos conceptos cobren vida, utilizaremos ejemplos del mundo real y ejercicios prácticos que le permitirán aplicar lo que ha aprendido. En cada paso del camino, nuestro objetivo es brindarle las herramientas necesarias para que pueda tomar decisiones fundamentadas y comunicar sus hallazgos de manera efectiva.

Este libro no solo está dirigido a estudiantes y académicos, sino también a profesionales de diversas disciplinas: desde científicos de datos hasta investigadores, médicos, economistas y profesionales de negocios. La estadística es una habilidad universalmente valiosa y relevante, y este libro le proporcionará los conocimientos necesarios para destacar en su campo.

A medida que se adentre en este viaje, no olvide que la estadística es una aventura intelectual emocionante y poderosa. Al comprender y aplicar sus principios, se embarcará en un viaje que le permitirá desvelar los secretos de los datos y dar forma a un mundo más informado y fundamentado en la evidencia.

Sin más preámbulos, le damos la bienvenida a este viaje educativo y enriquecedor a través de los fundamentos de la estadística. Estamos seguros de que encontrará este libro tanto informativo como inspirador mientras explora el fascinante mundo de la estadística. ¡Disfrute del viaje!

Autores

Introducción

La estadística es una rama de las matemáticas que se encarga de recolectar, analizar e interpretar datos para obtener información útil y relevante. Es una herramienta fundamental en la toma de decisiones en muchos campos, desde la medicina hasta los negocios, la política y la investigación científica.

Este libro sobre estadística básica está diseñado para proporcionar a los lectores una comprensión sólida de los conceptos fundamentales de la estadística y cómo aplicarlos en situaciones reales. Desde el análisis de datos simples hasta la inferencia estadística avanzada, este libro cubre todo lo que necesita saber para comenzar a utilizar la estadística como una herramienta efectiva en su vida diaria.

En las primeras secciones del libro, se cubren los conceptos básicos de la estadística, incluyendo la definición de población y muestra, variables y tipos de datos, y medidas de tendencia central y dispersión. También se discuten técnicas para representar datos visualmente, como gráficos y diagramas de dispersión.

A medida que avanzamos en el libro, se exploran temas más avanzados, como la probabilidad y la distribución de probabilidad. Se discute la importancia de la aleatoriedad en la recolección de datos y cómo se puede utilizar para hacer inferencias sobre una población más grande.

También se aborda la inferencia estadística, que es el proceso de hacer conclusiones sobre una población basándose en una muestra de datos. Se discuten los diferentes tipos de pruebas estadísticas y cómo se pueden utilizar para evaluar hipótesis y tomar decisiones informadas.

Además, el libro incluye ejemplos prácticos y ejercicios para ayudar a los lectores a aplicar los conceptos y técnicas discutidos en situaciones reales. Se cubren temas como la encuesta de opinión pública, el análisis de datos financieros y la evaluación de la efectividad de los tratamientos médicos.

En resumen, este libro sobre estadística básica es una guía completa para aquellos que desean aprender los fundamentos de la estadística y cómo aplicarlos en su vida diaria. Ya sea que esté interesado en la investigación científica, los negocios, la política o cualquier otro campo, este libro le proporcionará las herramientas necesarias para tomar decisiones informadas y basadas en datos.

Tabla de contenido

Prólogo 4
Introducción 6
1. La naturaleza de las estadísticas 11
1.1. ¿Qué es la estadística? 11
1.2. Población y muestra 12
1.3. Estadísticas descriptivas e inferenciales 15
1.4. Parámetros y estadísticas 18
1.5. Análisis estadístico de datos 19
2. Variables y organización de los datos 21
2.1. Variables 21
2.1.1. Escalas 22
2.2. Organización de los datos 23
3. Descripción de datos mediante tablas y gráficos 25
3.1. Variable cualitativa 25
3.2. Variable cuantitativa 29
3.3. Distribuciones muestrales y poblacionales 35
4. Medidas de tendencia central 39
4.1. Media Aritmética (Promedio): 39
4.2. La moda 39
4.3. La mediana 41
4.4. La media 43
4.5. ¿Qué medida elegir? 44
5. Medidas de variación 44
5.1. Rango 45
5.2. Rango intercuartílico 46
5.2.1. Resumen de cinco números y boxplots 48
5.3. Desviación típica o desviación estándar 51
5.3.1. Regla empírica para distribuciones simétricas 53
5.4. Desviación Media en Datos Agrupados: 54
5.5. Varianza en Datos Agrupados: 55
5.6. Desviación Típica en Datos Agrupados: 55

5.7. Estadísticas de la muestra y parámetros de la población 55
5.8. Varianza 56
5.9. Desviación estándar 57
6. Distribuciones de probabilidad 57
6.1. Distribuciones de probabilidad 57
6.2. Media y desviación típica de una variable aleatoria 62
6.3. Distribución normal 63
7. Distribuciones muestrales 66
7.1. Distribuciones muestrales 66
7.2. Distribuciones muestrales de las medias muestrales 67
8. Estimación 72
8.1. Estimación puntual 72
8.1.1. Estimadores puntuales de la media y la desviación típica de la población .. 73
8.2. Intervalo de confianza 74
8.2.1. Intervalo de confianza para μ cuando se conoce σ 74
8.2.2. Intervalo de confianza de muestra grande para μ 76
8.2.3. Intervalo de confianza de muestra pequeña para μ 76
9. Comprobación de hipótesis 78
9.1. Hipótesis 78
9.2. Prueba de significación para una media poblacional μ 81
9.2.1. Prueba de significación para μ cuando se conoce σ 81
9.2.2. Prueba de significación para μ en muestras grandes 83
9.2.3. Prueba de significación para μ en muestras pequeñas 84
10. Resumen de datos bivariantes 85
10.1. Variables cualitativas 85
10.2. Variable cualitativa y variable cuantitativa 91
10.3. Variables cuantitativas 97
11. Diagrama de dispersión y coeficiente de correlación 98
11.1. Gráfico de dispersión 98
11.2. Coeficiente de correlación 100
12. Covarianza: 106
13. Regresión Lineal: 107

14. Regresión Múltiple: 107
15. Prueba de Chi-Cuadrada: 107
16. Referencias 108

1. La naturaleza de las estadísticas

1.1. ¿Qué es la estadística?

La estadística, como disciplina amplia y esencial en la investigación científica y en diversos campos de aplicación, merece una definición más profunda y precisa. Según Silva et al. (2022), la estadística es la ciencia que se encarga de recopilar, organizar, analizar, interpretar y presentar datos. A través de sus métodos, permite obtener información, extraer conclusiones y tomar decisiones basadas en la evidencia proporcionada por los datos (Zuur et al., 2010).

Es importante destacar que la estadística va mucho más allá de la simple tabulación de números y la creación de gráficos. Su alcance se extiende desde la planificación detallada que precede a la recopilación de datos hasta la evaluación de la incertidumbre en las conclusiones extraídas. En este sentido, la estadística desempeña un papel fundamental en la investigación científica y ofrece respuestas a preguntas cruciales, como las que se plantean en la planificación de un estudio de investigación:

- **Diseño:** Implica la planificación y ejecución de investigaciones, determinando qué datos recopilar y en qué cantidad.
- **Descripción:** En este paso, se resumen y exploran los datos recopilados para obtener una visión más clara de los fenómenos en estudio.
- **Inferencia:** Se emplean métodos estadísticos para hacer predicciones y generalizaciones basadas en los datos, permitiendo así llegar a conclusiones respaldadas por la evidencia.

Además, la estadística es la herramienta principal para abordar la incertidumbre en la toma de decisiones. Se aplica exitosamente en diversos campos, como la evaluación de la eficacia de tratamientos médicos, el análisis de la reacción de los consumidores a la publicidad, el estudio de actitudes y comportamientos sociales, así como en cuestiones económicas y técnicas.

1.2. Población y muestra

En el contexto de la estadística, los conceptos de población y muestra desempeñan un papel fundamental en la investigación científica. Como señala Livi-Bacci (2007), la población se define como el conjunto completo de personas individuales, objetos o elementos en los cuales un investigador está principalmente interesado durante el desarrollo de su estudio. Es importante destacar que, en muchas ocasiones, obtener mediciones para todos los elementos de la población puede ser costoso o logísticamente difícil.

Por lo tanto, se recurre a la noción de muestra, que representa un subconjunto de individuos extraídos de la población en cuestión. En palabras de Pimienta (2000), la muestra consiste en un grupo seleccionado de individuos de la población, y es a través del análisis de esta muestra que se obtienen conclusiones sobre la población en su conjunto.

La **población** es el conjunto de todos los individuos o elementos considerados en un estudio estadístico (Martínez, 2019).

La **muestra** es la parte de la población de la que se recoge información (Etikan, 2016).

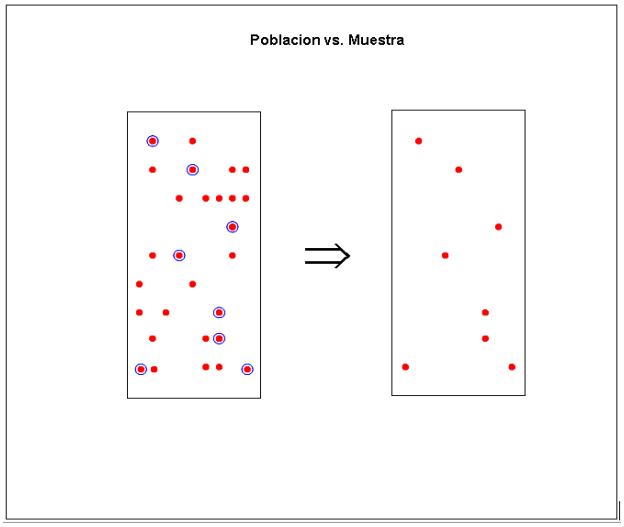


Figura 1: Población y Muestra

En el contexto de la investigación estadística, la población y la muestra son conceptos fundamentales que desempeñan un papel crucial en la formulación de estudios y en la obtención de conclusiones significativas. Como lo destaca Levy & Lemeshow (2008), la población se define como el conjunto completo de todos los individuos o elementos considerados en un estudio estadístico. Esto implica que la población abarca la totalidad de los elementos o individuos que conforman el universo de interés en una investigación dada.

Sin embargo, en la práctica, es común que un estudio estadístico se enfoque en un conjunto específico de características, relativamente limitado, de los individuos o elementos de la población. En este sentido, es fundamental reconocer que no se pretende medir todas las propiedades de los elementos de la población. Esta observación conduce a la apreciación de la importancia de un conjunto selecto de mediciones, lo que a su vez nos brinda definiciones alternativas tanto para la población como para la muestra.

La muestra, según Bechtold & Patterson (2005), se configura como la porción representativa de la población a partir de la cual se obtienen datos. En otras palabras, la muestra consiste en una selección cuidadosa de individuos o elementos de la población que se eligen para proporcionar información relevante sobre las características de interés.

La elección de una muestra adecuada es crucial para garantizar que los resultados del estudio sean representativos de la población en su conjunto. La estadística ofrece una serie de técnicas y métodos para diseñar muestras que sean representativas y que permitan realizar inferencias precisas acerca de la población a partir de la información recopilada.

Una población (estadística) es el conjunto de mediciones (o registro de algún rasgo cualitativo) correspondiente a toda la colección de unidades sobre las que se quieren hacer inferencias (Johnson & Bhattacharyya, 1992).

Una muestra de una población estadística es el conjunto de mediciones que se recogen realmente en el curso de una investigación.

Cuando la población y la muestra se definen como lo hacen Johnson y Bhattacharyya, resulta útil definir la fuente de cada medición como unidad de muestreo o simplemente una unidad.

La población siempre representa el objetivo de una investigación. Aprendemos sobre la población mediante el muestreo de la colección. Puede haber muchas poblaciones diferentes, los siguientes ejemplos demuestran posibles discrepancias sobre las poblaciones.

Ejemplo 1 (Población finita). En muchos casos, la población objeto de estudio puede enumerarse físicamente. Por ejemplo:

-Los estudiantes de la Universidad de Los Ángeles,

-Los libros de una biblioteca.

La noción de población no se limita únicamente a conjuntos de individuos o elementos tangibles. En muchas ocasiones, puede adquirir un carácter más abstracto, surgiendo directamente del fenómeno que está siendo objeto de estudio. Un ejemplo ilustrativo de esta perspectiva se encuentra en el contexto de una fábrica de bombillas. Tal como señala Keoleian & Menerey (1994), si esta fábrica mantiene la continuidad en el uso de los mismos equipos, materias primas y métodos de producción en el futuro, las bombillas que serán producidas en esa fábrica constituyen una población hipotética.

En otras palabras, esta población hipotética representa una proyección teórica de futuras bombillas que serán producidas por la fábrica. No obstante, para evaluar la calidad de estas bombillas futuras, se recurre a una muestra de bombillas tomada

de la línea de producción actual. Esta muestra se elige estratégicamente para que sea representativa de las condiciones y procesos de producción existentes en la fábrica en ese momento.

La capacidad de utilizar una muestra de bombillas actuales para hacer inferencias sobre la calidad de las bombillas futuras es un ejemplo de cómo la estadística puede desempeñar un papel fundamental en la toma de decisiones y la evaluación de la calidad en la industria. Esta perspectiva resalta la versatilidad de los conceptos de población y muestra en diferentes contextos y su relevancia en la obtención de conclusiones significativas en la investigación y la toma de decisiones.

1.3. Estadísticas descriptivas e inferenciales

Existen dos tipos principales de estadística. La rama de la estadística dedicada al resumen y la descripción de datos se denomina estadística descriptiva y la rama de la estadística que se ocupa de utilizar datos de muestra para hacer una inferencia sobre una población de datos se denomina estadística inferencial.

En el ámbito de la estadística, se distinguen dos ramas principales que desempeñan roles fundamentales en el análisis y la interpretación de datos. Como señala Ross (2018), la primera de estas ramas es la estadística descriptiva, que se centra en el resumen y la descripción de datos. La estadística descriptiva es esencial para organizar y sintetizar la información recopilada, proporcionando una visión clara y concisa de las características fundamentales de un conjunto de datos.

En contraste, la segunda rama es la estadística inferencial, cuyo propósito principal es utilizar datos de muestra para realizar inferencias sobre una población de datos más amplia. Esta rama se ocupa de evaluar la fiabilidad de las conclusiones que se pueden extraer acerca de la población a partir de la información obtenida de una muestra representativa. Los métodos de la estadística inferencial, como la estimación puntual, la estimación por intervalos y la comprobación de hipótesis, se basan en principios sólidos de la teoría de la probabilidad (Crawford et al., 2010).

La estadística descriptiva comprende diversas técnicas que incluyen la construcción de gráficos, tablas y cuadros, así como el cálculo de medidas descriptivas clave. Estas medidas incluyen, entre otras, la media, la desviación estándar, la varianza y los percentiles. Es importante destacar que gran parte del enfoque de un curso de estadística se dedica a la estadística descriptiva, ya que proporciona una base sólida para comprender y comunicar la información contenida en los datos.

La **estadística descriptiva** consiste en métodos para organizar y resumir la información (Willemse, 2009).

Por otro lado, la estadística inferencial se basa en la teoría de la probabilidad y se utiliza para realizar generalizaciones y tomar decisiones sobre poblaciones más amplias en función de los resultados observados en una muestra específica. Estas técnicas permiten a los investigadores sacar conclusiones fundamentadas sobre fenómenos más allá de los datos observados en la muestra.

La **estadística inferencial** consiste en métodos para extraer y medir la fiabilidad de las conclusiones sobre la población a partir de la información obtenida de una muestra de la población (Weiss, 2016).

La estadística descriptiva incluye la construcción de gráficos, tablas y cuadros, y el cálculo de diversas medidas descriptivas, como medias, medidas de variación y percentiles. De hecho, la mayor parte de este curso trata sobre estadística descriptiva.

La estadística inferencial incluye métodos como la estimación puntual, la estimación por intervalos y la comprobación de hipótesis, todos ellos basados en la teoría de la probabilidad.

Ejemplo. Supongamos que se lanza un dado. El dado se lanza 100 veces y los resultados forman los datos de la muestra. La estadística descriptiva se utiliza para agrupar los datos de la muestra en la siguiente tabla

Resultado de la tirada	Frecuencias en los datos de la muestra
1	10
2	20
3	18
4	16
5	11
6	25

La estadística inferencial desempeña un papel fundamental en la evaluación de situaciones y la toma de decisiones basadas en evidencia. Un ejemplo ilustrativo de su aplicabilidad se presenta en la verificación de la equidad de un dado en un juego de azar. La estadística inferencial permite determinar si el dado utilizado en el juego es justo o si existe algún sesgo en sus resultados.

Es importante reconocer que la estadística descriptiva y la estadística inferencial están estrechamente interrelacionadas y a menudo se utilizan de manera complementaria. Como señala Marshall & Jonker (2010), en la mayoría de los casos, es necesario aplicar métodos de estadística descriptiva para organizar y resumir la información recopilada de una muestra antes de llevar a cabo un análisis más exhaustivo mediante métodos de estadística inferencial. El análisis descriptivo inicial de la muestra proporciona una visión general de las características clave de los datos, lo que a su vez orienta la elección del método inferencial más apropiado para el análisis subsiguiente.

En ocasiones, es posible recopilar datos de toda la población de interés. En estos casos, se puede llevar a cabo un estudio descriptivo tanto de la población completa como de cualquier muestra obtenida de ella. Sin embargo, la esencia de la estadística inferencial radica en la extrapolación de las conclusiones de una muestra a toda la población. Como señala Zieffler et al. (2008), es cuando se realiza una inferencia sobre la población basada en la información obtenida de la muestra que el estudio se convierte en inferencial.

1.4. Parámetros y estadísticas

En el ámbito de la estadística, la distinción entre parámetros y estadísticas es fundamental para la formulación y el análisis de problemas de investigación. Como afirma Preacher & Kelley (2011), en la mayoría de los casos, las características de la población en estudio pueden expresarse y resumirse mediante parámetros numéricos. Estos parámetros, como la media, la varianza o la proporción, representan medidas que caracterizan la población en su conjunto.

Sin embargo, es importante destacar que los valores de estos parámetros poblacionales generalmente son desconocidos en la práctica. En su lugar, los investigadores recopilan datos de una muestra representativa de la población, y estos datos son utilizados para estimar y hacer inferencias sobre los parámetros poblacionales. Los valores numéricos que resumen las características de la muestra se denominan estadísticos.

Un estadístico, como la media muestral o la desviación estándar muestral, es una medida descriptiva de la muestra que se utiliza como una estimación del parámetro poblacional correspondiente. La relación entre parámetros y estadísticas es esencial para la estadística inferencial, ya que permite hacer generalizaciones y tomar decisiones basadas en evidencia recopilada de una muestra.

> Un parámetro es un resumen numérico desconocido de la población. Un estadístico es un resumen numérico conocido de la muestra que puede utilizarse para hacer inferencias sobre parámetros (Cobb & Moore, 1997).

En el proceso de investigación estadística, la inferencia acerca de un parámetro específico, que generalmente es desconocido, se apoya en el análisis de un estadístico derivado de una muestra. Como señala O'Fallon (2013), los estadísticos se utilizan como herramientas para realizar inferencias sobre los parámetros de la población. En esencia, esto implica que los valores numéricos que se calculan a partir de una muestra se convierten en representaciones estimadas de las características de la población en su conjunto.

Es importante destacar que, en la mayoría de los estudios de investigación, el objetivo principal radica en obtener información precisa sobre los parámetros poblacionales, en lugar de centrarse en los estadísticos calculados para una muestra particular. La muestra y los estadísticos que la describen adquieren relevancia en la medida en que proporcionan información crucial sobre los parámetros poblacionales, que a menudo son el enfoque central de la investigación.

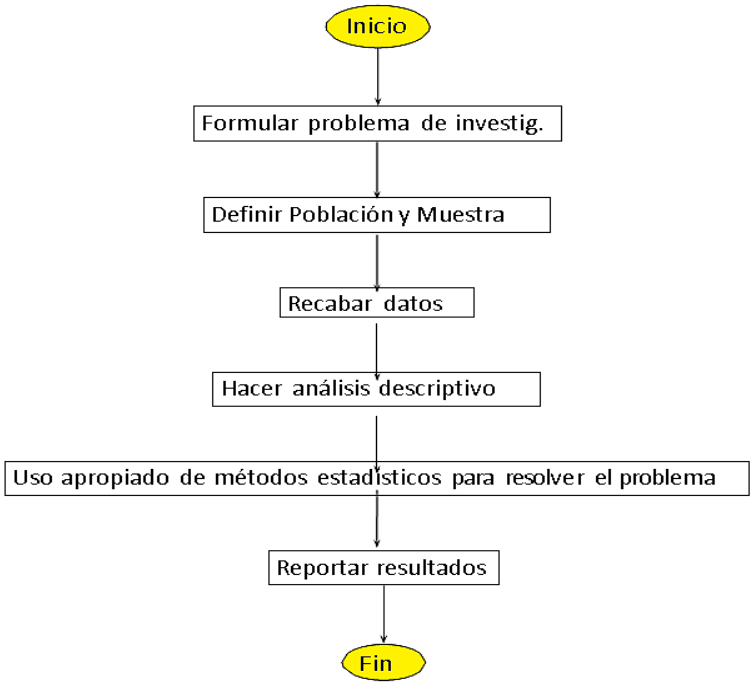
Un ejemplo ilustrativo de esta distinción entre parámetros y estadísticas se encuentra en la investigación destinada a determinar el porcentaje de personas de entre 18 y 30 años que asisten al cine al menos una vez al mes. En este caso:

- **Parámetro:** La proporción (p) de personas de 18 a 30 años que asisten al cine al menos una vez al mes representa el valor que se desea estimar y que generalmente es desconocido a priori.
- **Estadístico:** La proporción muestral ($\hat{p}$) se calcula a partir de una muestra específica de personas de 18 a 30 años y se utiliza como una estimación de la proporción poblacional p.

Este ejemplo ilustra cómo los estadísticos derivados de una muestra pueden proporcionar una estimación útil de un parámetro poblacional, permitiendo a los investigadores realizar inferencias significativas sobre la población en su conjunto. La relación entre parámetros y estadísticas es esencial en la estadística inferencial y en la toma de decisiones basadas en evidencia (Byrne, 2007).

1.5. Análisis estadístico de datos

El objetivo de la estadística es comprender los datos. Cualquier análisis de datos debe incluir los siguientes pasos:



La estadística desempeña un papel crucial en la toma de decisiones y la generación de conocimiento en una amplia variedad de disciplinas. Uno de los principales objetivos de la estadística es realizar inferencias significativas sobre la población en base al análisis de la información contenida en los datos muestrales. Esto incluye la evaluación del grado de incertidumbre asociado a estas inferencias, aspecto esencial en cualquier investigación estadística.

Como destaca Hernán et al. (2019), la capacidad de hacer inferencias sólidas a partir de datos de muestra es fundamental en la investigación y en la toma de decisiones en áreas como la medicina, la economía, la ciencia política y muchas otras. La estadística proporciona un marco metodológico riguroso que permite a los investigadores evaluar la validez de las conclusiones basadas en evidencia empírica y cuantitativa.

En última instancia, la estadística se erige como una disciplina fundamental para comprender y dar sentido a los datos, proporcionando herramientas y técnicas que permiten explorar, analizar y tomar decisiones fundamentadas en un mundo cada vez más inundado de información (Wild & Pfannkuch, 1999). Esta capacidad de

generar conocimiento a partir de datos es esencial en la era de la información en la que vivimos.

2. Variables y organización de los datos

2.1. Variables

Las variables desempeñan un papel esencial en la estadística, ya que constituyen las características que pueden variar de un individuo a otro en una población. Como señala Venuti et al. (2015), una variable se define como cualquier característica que exhibe variabilidad entre los miembros individuales de una población. En el contexto de los seres humanos, existen numerosos ejemplos de variables, entre los que se incluyen la altura, el peso, el número de hermanos, el sexo, el estado civil y el color de ojos.

Es importante destacar que estas variables se pueden clasificar en dos categorías principales: variables cuantitativas (o numéricas) y variables cualitativas (o categóricas), según el tipo de información que proporcionen. Las variables cuantitativas, como la altura, el peso y el número de hermanos, ofrecen mediciones numéricas y permiten realizar cálculos matemáticos. Por otro lado, las variables cualitativas, como el sexo, el estado civil y el color de ojos, representan información no numérica y se utilizan para describir categorías o grupos.

Dentro de las **variables cuantitativas,** se pueden distinguir dos subtipos: variables discretas y variables continuas. Las variables discretas, como el número de hijos en una familia o la cantidad de accidentes de tráfico en una carretera en diferentes días, se caracterizan por su naturaleza discreta y resultan de un conteo. Estas variables solo pueden tomar un conjunto finito o infinito contable de valores, generalmente números enteros, como 0, 1, 2, 3, y así sucesivamente.

En contraste, las **variables continuas**, como la longitud, el peso o la temperatura, pueden medirse con un alto grado de precisión y no están limitadas a valores discretos. Como señala Barton & Pantelides (1994), estas variables pueden

expresarse con una precisión prácticamente ilimitada, lo que las diferencia fundamentalmente de las variables discretas. No existe una unidad indivisible en las variables continuas, y su rango de valores posibles es infinitamente grande.

2.1.1. Escalas

En el ámbito de la estadística, es esencial no solo considerar si una variable es cualitativa o cuantitativa, sino también entender la escala en la que se define, ya que esto aporta estructura y significado a la variable. La escala de una variable determina cómo se organizan y comprenden sus categorías.

Cuando nos referimos a variables cualitativas, es importante reconocer que las categorías en las que se dividen pueden o no tener un orden natural. Por ejemplo, las categorías profesionales no siguen un orden inherente. Cuando las categorías no tienen un orden intrínseco, se dice que la variable cualitativa está definida en una escala nominal, donde "nominal" denota que las categorías son simplemente nombres sin un orden subyacente. En cambio, cuando las categorías tienen un orden significativo, estamos tratando con una escala ordinal. Ejemplos de variables ordinales incluyen niveles de educación (como bajo, alto) o la "fuerza de opinión" sobre una propuesta (favorable, indiferente, en contra), así como las posiciones al final de una carrera (primero, segundo, etc.).

En el caso de variables cuantitativas, ya sean discretas o continuas, se utilizan escalas específicas que pueden ser de intervalo o de razón. Una variable cuantitativa se clasifica como una escala de intervalo cuando las diferencias entre las mediciones son significativamente comparables, aunque no lo sean las proporciones entre ellas. Por otro lado, cuando las diferencias y las proporciones entre las mediciones son comparables, estamos en presencia de una escala de razón. La diferencia fundamental radica en la existencia de puntos de cero absolutos significativos en una escala de razón, lo que permite que las proporciones sean significativas.

Un ejemplo que ilustra esto es la medición de la temperatura en el sistema Celsius, que se considera una variable de intervalo. Por otro lado, la altura de una persona se clasifica como una variable de razón debido a la presencia de un punto de cero absoluto significativo, como lo es la ausencia total de altura.

2.2. Organización de los datos

Cuando nos enfrentamos al análisis de datos, resulta esencial comprender cómo se organizan y representan estos datos. Los datos son el resultado de observar los valores de una o varias variables en individuos, objetos o unidades de estudio, y cada valor individual se denomina observación. La recopilación de todas estas observaciones para determinadas variables constituye lo que se conoce como conjunto de datos o matriz de datos (Van Buuren et al., 2006).

En el contexto de las variables cualitativas, es común utilizar la codificación mediante la asignación de números a las diferentes categorías. Esta práctica permite transformar los datos categóricos en datos numéricos de manera superficial, lo que facilita su manipulación y análisis. Por ejemplo, en el caso del estado civil, se podría asignar los números 1, 2, 3 y 4 a las categorías soltero, casado, viudo y divorciado, respectivamente.

Sin embargo, es fundamental comprender que, a pesar de la codificación numérica, la naturaleza de estos datos sigue siendo nominal. La escala nominal implica que las categorías no tienen un orden intrínseco ni propiedades matemáticas significativas más allá de la etiqueta numérica asociada a ellas. En este sentido, la asignación de números a las categorías es meramente una conveniencia para su manejo, pero no implica una relación de orden o jerarquía entre las categorías (Anderberg, 1973).

Por lo tanto, al analizar datos codificados de esta manera, es fundamental recordar que la naturaleza nominal de las variables cualitativas permanece intacta, y cualquier interpretación debe realizarse con consideración a esta característica fundamental.

Los datos numéricos codificados no presentan las propiedades de los números utilizados en la aritmética convencional. Por ejemplo, en el contexto de los códigos de estado civil, no podemos establecer relaciones de orden como 3 > 1 o 2 < 4, ni realizar operaciones aritméticas como 2 - 1 = 4 - 3 o 1 + 3 = 4. Esto subraya la importancia de siempre verificar la validez del tratamiento matemático de los datos estadísticos.

Los datos se presentan en forma de matriz, específicamente, una matriz de datos. Cada variable en particular se organiza en una columna, de manera que los valores correspondientes a dicha variable constituyen dicha columna en la matriz de datos. Por otro lado, una observación, que representa las mediciones obtenidas de la unidad de muestreo, se dispone en una fila dentro de la matriz de datos.

Para ilustrar esto con mayor precisión, consideremos un escenario en el que se tienen k variables y n observaciones, donde n representa el tamaño de la muestra. En este contexto, la estructura del conjunto de datos deberá seguir el siguiente formato:

$$\begin{array}{cc} & \text{Variables} \\ \text{Unidades Muestrales} & \begin{pmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1k} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2k} \\ x_{31} & x_{32} & x_{33} & \cdots & x_{3k} \\ & \vdots & & \ddots & \\ x_{n1} & x_{n2} & x_{n3} & \cdots & x_{nk} \end{pmatrix} \end{array}$$

donde x_{ij} es un valor de la variable recogido en la observación, i = 1, 2, ... n y j = 1, 2, . . . , k.

3. Descripción de datos mediante tablas y gráficos

3.1. Variable cualitativa

Una variable cualitativa se define como una característica que puede ser categorizada en diferentes clases o categorías discretas. La cuantificación de la cantidad de observaciones que se asigna a una categoría específica de esta variable se conoce en términos precisos como la frecuencia o el recuento de dicha categoría. Esta información se presenta de manera sistemática en una tabla denominada distribución de frecuencias, la cual enumera de forma exhaustiva todas las categorías junto con sus correspondientes frecuencias.

Adicionalmente a la descripción de las frecuencias, en ciertas circunstancias, surge la necesidad de comprender la proporción relativa de una categoría específica en relación con el conjunto total de observaciones. Para calcular esta medida, se realiza una operación aritmética sencilla: se divide la frecuencia correspondiente a la categoría en cuestión entre el número total de observaciones, y luego se multiplica el resultado por 100. Esta medida, expresada en forma decimal, se conoce comúnmente como la frecuencia relativa de la categoría.

En apoyo a esta definición, el eminente investigador Korsunsky et al. (2019) ha subrayado que la frecuencia relativa es una métrica esencial en el análisis de variables cualitativas, ya que permite una comparación precisa de la relevancia de cada categoría dentro del conjunto de datos. Asimismo, el destacado académico Hyland (1996) ha enfatizado que esta medida desempeña un papel fundamental en la interpretación y la comunicación de resultados en investigaciones científicas, ya que representa la proporción de observaciones asociadas con una categoría específica.

$$\text{Frecuencia relativa de la clase} = \frac{\text{Frecuencia en la clase}}{\text{Número de observaciones}}$$

Ejercicio:

Supongamos que tenemos una muestra de 150 estudiantes de una escuela y deseamos calcular la frecuencia relativa de la categoría "Estudiantes que aprobaron el examen de matemáticas". De los 150 estudiantes, 112 aprobaron el examen.

Para calcular la frecuencia relativa de esta categoría, utilizamos la fórmula:

Frecuencia relativa de la clase = Frecuencia en la base / Número de observaciones

En este caso:

- Frecuencia relativa de "Estudiantes que aprobaron el examen de matemáticas" = 112 (estudiantes que aprobaron) / 150 (número total de estudiantes en la muestra)
- Frecuencia relativa de "Estudiantes que aprobaron el examen de matemáticas" = 0.7467 (aproximadamente)

Por lo tanto, la frecuencia relativa de la categoría "Estudiantes que aprobaron el examen de matemáticas" es aproximadamente 0.7467 o 74.67%. Esto significa que alrededor del 74.67% de los estudiantes en la muestra aprobaron el examen de matemáticas.

Una tabla que incluye todas las categorías junto con sus frecuencias relativas se denomina "distribución de frecuencias relativas". Las frecuencias relativas proporcionan información altamente relevante sobre la estructura de los datos. Además, es importante señalar el tamaño de la muestra, ya que esto sirve como indicador de la confiabilidad de las frecuencias relativas. Cabe destacar que la suma de las frecuencias relativas siempre es igual a 1, lo que equivale al 100%.

La frecuencia acumulada, también conocida como frecuencia relativa acumulada, se obtiene sumando las frecuencias (o frecuencias relativas) de todas las categorías hasta llegar a una categoría específica. En el caso de las variables cualitativas, es importante destacar que las frecuencias acumuladas solo tienen sentido en el contexto de variables ordinales y no para las nominales.

La representación gráfica de datos cualitativos se lleva a cabo mediante gráficos circulares, gráficos de barras horizontales o gráficos de barras verticales.

Un gráfico circular consiste en un círculo dividido en segmentos de tarta, donde cada segmento es proporcional a las frecuencias relativas de las categorías. Para determinar el ángulo de cualquier categoría, simplemente multiplicamos la frecuencia relativa correspondiente por 360 grados, que representa el círculo completo.

Por otro lado, un gráfico de barras horizontales presenta las categorías en el eje horizontal y las frecuencias (o frecuencias relativas) en el eje vertical. Cada categoría se representa mediante una barra vertical cuya altura corresponde a la frecuencia (o frecuencia relativa) de la categoría. En un gráfico de barras, las barras no se superponen entre sí. En el caso de un gráfico de barras verticales, las categorías se disponen en el eje vertical y las frecuencias de las categorías en el eje horizontal.

Los datos nominales se representan mejor mediante un gráfico circular y los datos ordinales mediante un gráfico de barras horizontales o verticales.

Ejemplo. Supongamos que los grupos sanguíneos de 40 personas son los siguientes

O O A B A O A A A O B O B O A O O A A A A AB A B A A O O A O A A A O O O AB

Resumiendo, los datos en una tabla de frecuencias utilizando SPSS:

Analizar -> Estadística descriptiva -> Frecuencias,

Analizar -> Tablas personalizadas -> Tablas de frecuencias

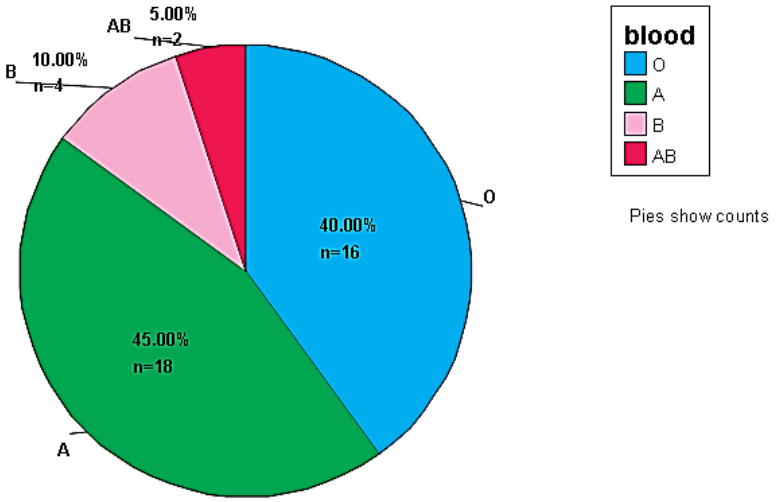
Tabla 1: Distribución de frecuencias de los tipos de sangre

SANGRE	Estadísticas	
	Frecuencia	**Porcentaje**
O	16	40.0
A	18	45.0
B	4	10.0
AB	2	5.0
Total	40	100.0

Presentación gráfica de datos en SPSS:

Gráficos -> Interactivos -> Tarta -> Simples,

Gráficos -> Interactivos -> Barra



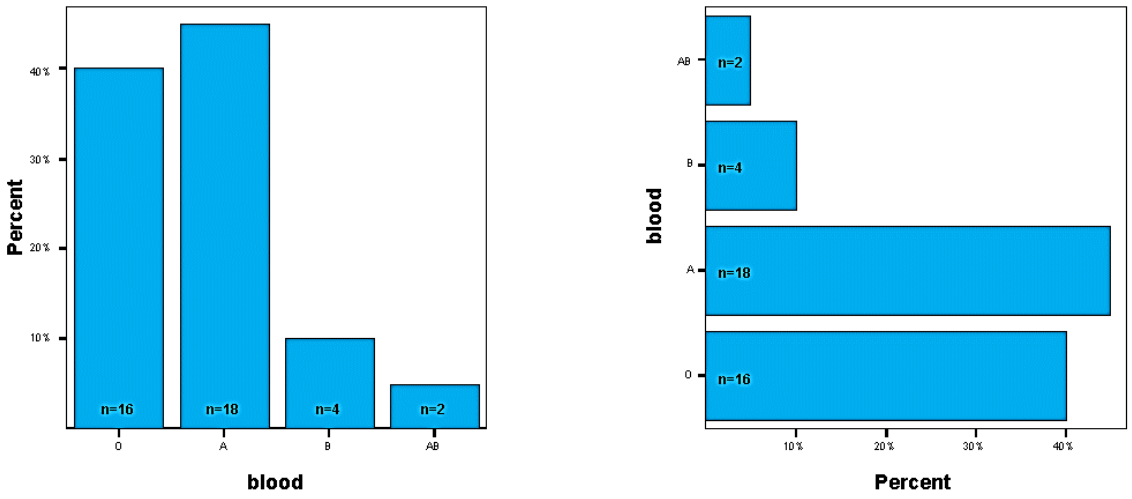


Figura 2: Tablas de grupos sanguíneos

3.2. Variable cuantitativa

La presentación de datos relacionados con una variable cuantitativa puede llevarse a cabo mediante la construcción de una distribución de frecuencias. En el caso de una variable discreta que posee un número limitado de valores distintos, la información puede ser resumida de manera análoga a las variables cualitativas, mediante la creación de una tabla de frecuencias. Sin embargo, en lugar de categorías cualitativas, esta tabla de frecuencias contendrá las diversas medidas numéricas que se encuentran en el conjunto de datos discreto, junto con sus respectivas frecuencias.

En situaciones en las que la variable discreta puede tomar una gran cantidad de valores diferentes o cuando se trata de una variable cuantitativa continua, se hace necesario agrupar los datos en clases o categorías antes de construir la tabla de frecuencias. Los pasos fundamentales en el proceso de agrupación de una variable cuantitativa en clases son los siguientes:

a) Identificar los valores mínimo y máximo presentes en el conjunto de datos de la variable.

b) Seleccionar intervalos de igual longitud que abarquen todo el rango entre el valor mínimo y máximo, sin superponerse entre sí. Estos intervalos se conocen como "intervalos de clase", y sus puntos finales se denominan "límites de clase".
c) Contar el número de observaciones del conjunto de datos que se encuentran dentro de cada intervalo de clase. Esta cantidad se conoce como "frecuencia de clase".
d) Calcular las frecuencias relativas de cada clase dividiendo la frecuencia de clase entre el número total de observaciones en el conjunto de datos.

En adición, es importante destacar que el número ubicado en el centro de cada clase se denomina "marca de clase", mientras que el número que se halla en medio entre el límite superior de una clase y el límite inferior de la siguiente se conoce como "límite real de la clase". La elección del número de intervalos en la agrupación de datos depende de la cantidad de observaciones, y suele variar entre 5 y 15 clases, aunque un número menor de intervalos es preferible cuando la cantidad de datos es limitada, y se pueden utilizar más de 15 intervalos cuando el conjunto de datos es extenso.

La representación gráfica de datos cuantitativos se logra comúnmente mediante histogramas o gráficos de barras, tanto en su versión horizontal como vertical. El histograma, en particular, se diferencia de un gráfico de barras en que las barras se encuentran adyacentes entre sí. Este tipo de gráfico se utiliza cuando los datos han sido previamente agrupados, y muestra las frecuencias o frecuencias relativas (expresadas en porcentajes) de cada intervalo de clase.

Si los datos cuantitativos son discretos y solo presentan un número reducido de valores posibles, se recomienda utilizar un gráfico de barras para su representación. Además, en situaciones donde sea más apropiado utilizar intervalos de clase desiguales para la construcción de la tabla de frecuencias, también es aconsejable representar visualmente los datos mediante un gráfico de barras.

Ejemplo. Edad (en años) de 102 personas:

34,67,40,72,37,33,42,62,49,32,52,40,31,19,68,55,57,54,37,32,

54,38,20,50,56,48,35,52,29,56,68,65,45,44,54,39,29,56,43,42,

22,30,26,20,48,29,34,27,40,28,45,21,42,38,29,26,62,35,28,24,

44,46,39,29,27,40,22,38,42,39,26,48,39,25,34,56,31,60,32,24,

51,69,28,27,38,56,36,25,46,50,36,58,39,57,55,42,49,38,49,36,

48,44

Resumir los datos en una tabla de frecuencias utilizando SPSS:

Analizar -> Estadísticas Descriptivas -> Frecuencias,

Analizar -> Tablas Personalizadas -> Tablas de Frecuencias

Tabla 2: Distribución de frecuencias de la edad de las personas

Distribución de frecuencias de la edad de las personas

Intervalo de Valores	Frecuencia	Porcentaje	Porcentaje Acumulativo
18 – 22	6	5.9%	5.9%
23 – 27	10	9.8%	15.7%
28 – 32	14	13.7%	29.4%
33 – 37	11	10.8%	40.2%
38 – 42	19	18.6%	58.8%
43 – 47	8	7.8%	66.7%
48 – 52	12	11.8%	78.4%
53 – 57	12	11.8%	90.2%
58 – 62	4	3.9%	94.1%
63 – 67	2	2.0%	96.1%
68 – 72	4	3.9%	100.0%
Total	**102**	**100.0%**	

Presentación gráfica de datos en SPSS:

Gráficos -> Interactivo -> Histograma,

Gráficos -> Histograma

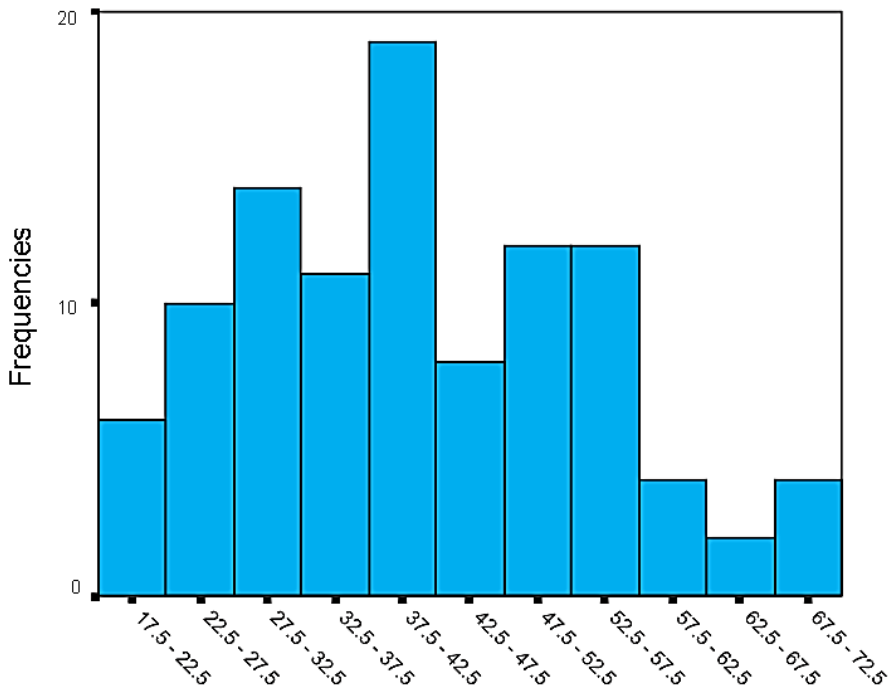


Figure 3: Histograma para la edad de las personas

Ejemplo. Precios de los perros calientes ($/oz.):

0,11,0,17,0,11,0,15,0,10,0,11,0,21,0,20,0,14,0,14,0,23,0,25,0,07,

0.09,0.10,0.10,0.19,0.11,0.19,0.17,0.12,0.12,0.12,0.10,0.11,0.13,

0.10,0.09,0.11,0.15,0.13,0.10,0.18,0.09,0.07,0.08,0.06,0.08,0.05,

0.07,0.08,0.08,0.07,0.09,0.06,0.07,0.08,0.07,0.07,0.07,0.08,0.06,

0.07,0.06

Tabla de frecuencias:

Cuadro 3: Distribución de frecuencias de los precios de los perros calientes

Frecuencias de los precios de los perritos calientes ($/oz.)

Intervalo de Valores	Frecuencia	Porcentaje	Porcentaje Acumulativo
0.031 - 0.06	5	9.3%	9.3%
0.061 - 0.09	19	35.2%	44.4%
0.091 - 0.12	15	27.8%	72.2%
0.121 - 0.15	6	11.1%	83.3%
0.151 - 0.18	3	5.6%	88.9%
0.181 - 0.21	4	7.4%	96.3%
0.211 - 0.24	1	1.9%	98.1%
0.241 - 0.27	1	1.9%	100.0%
Total	**54**	**100.0%**	

Cuadro 4: Distribución de frecuencias de los precios de los perros calientes (excluidos los extremos izquierdos e incluidos los derechos)

Frecuencias de los precios de los perritos calientes ($/oz.)

Intervalo de Valores	Frequencia	Porcentaje	Porcentaje Acumulativo
0.03-0.06	5	9.3	9.3
0.06-0.09	19	35.2	44.4
0.09-0.12	15	27.8	72.2
0.12-0.15	6	11.1	83.3
0.15-0.18	3	5.6	88.9
0.18-0.21	4	7.4	96.3
0.21-0.24	1	1.9	98.1
0.24-0.27	1	1.9	100.0
Total	**54**	**100.0**	

Presentación gráfica de los datos:

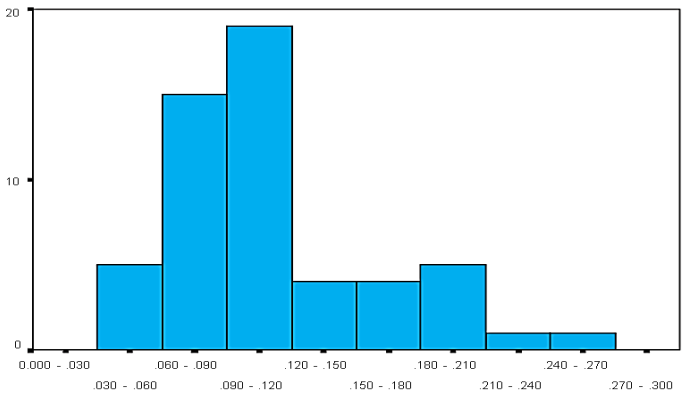


Figura 4: Histograma de precios.

Una alternativa para resumir los precios de los perritos calientes mediante una tabla de frecuencias. Inicialmente, se advierte que el precio mínimo de los perritos calientes es de 0,05. En consecuencia, se decide agrupar los valores observados de 0.05 y 0.06 en un mismo intervalo de clase, así como los valores observados de 0.07 y 0.08, y así sucesivamente. A continuación, se eligen los límites de clase de manera que representen valores medios entre 0,06 y 0,07, y así sucesivamente. Esto conduce a la siguiente tabla de frecuencias:

Table 5: Distribución de frecuencias de los precios de los perros calientes

Frecuencias de los precios de los perros calientes ($/oz.)

Intervalo de Valores	Frequencia	Porcentaje	Porcentaje Acumulativo
0.045-0.065	5	9.3	9.3
0.065-0.085	15	27.8	37.0
0.085-0.105	10	18.5	55.6
0.105-0.125	9	16.7	72.2
0.125-0.145	4	7.4	79.6
0.145-0.165	2	3.7	83.3
0.165-0.185	3	5.6	88.9
0.185-0.205	3	5.6	94.4
0.205-0.225	1	1.9	96.3
0.225-0.245	1	1.9	98.1
0.245-0.265	1	1.9	100.0
Total	**54**	**100.0**	

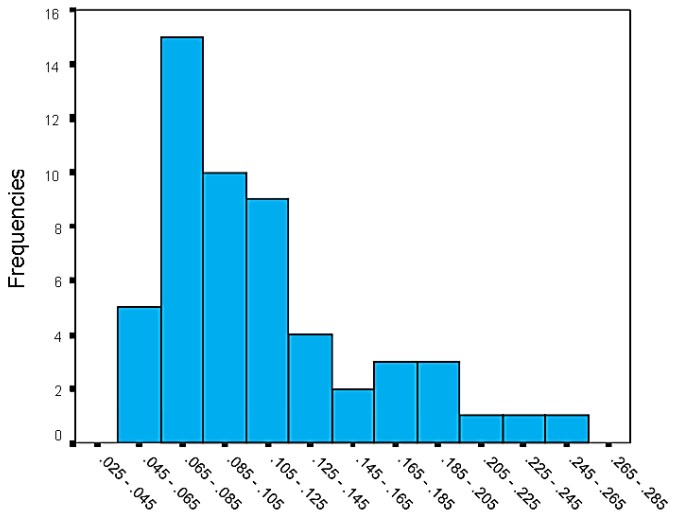


Figura 5: Histograma de precios

Otros tipos de representación gráfica de datos cuantitativos son

a) Dotplot

Gráficos -> Interactivos -> Dot

b) Diagrama de tallo y hoja de sólo hoja

Analizar -> Estadística descriptiva -> Explorar

c) Polígono de frecuencias y de frecuencias relativas (Gráficos -> Interactivo -> Línea)
d) Ojivas para frecuencias acumuladas y para frecuencias relativas acumuladas (Gráficos -> Interactivo -> Línea)

3.3. Distribuciones muestrales y poblacionales

El concepto de distribuciones de frecuencias aplicado tanto a una población como a muestras tomadas de dicha población. El primer tipo se conoce como la distribución poblacional de una variable, mientras que el segundo tipo se denomina distribución muestral. De alguna manera, la distribución muestral se asemeja a una imagen borrosa de la distribución poblacional. A medida que el tamaño de la muestra aumenta, la frecuencia relativa de la muestra en cualquier intervalo de clase tiende a acercarse más a la frecuencia relativa real de la población. De esta manera, la imagen se vuelve más nítida, y la distribución de la muestra se aproxima más a la distribución de la población.

Este concepto encuentra respaldo en la obra de Kahneman & Tversky (1972) quien subrayó la importancia de entender que la distribución muestral se convierte en un reflejo más fiel de la distribución poblacional a medida que se incrementa el tamaño de la muestra. En otras palabras, a través de un aumento en el tamaño muestral, se logra una mayor precisión en la representación de la población de interés.

Cuando se trabaja con variables continuas, existe la flexibilidad de seleccionar intervalos de clase en la construcción de la distribución de frecuencias y en la creación de un histograma, con la posibilidad de hacer estos intervalos tan estrechos como se necesite. Sin embargo, es crucial tener en cuenta que a medida que el tamaño de la muestra crece de manera indefinida y, simultáneamente, se incrementa el número de intervalos de clase, reduciendo su amplitud, el aspecto del histograma de la muestra evoluciona gradualmente hacia una curva suave. Estas curvas suaves, conocidas como distribuciones de población, se emplean para representar la distribución de una población.

Este concepto de transición de histograma a curva suave se encuentra respaldado por las investigaciones de Bishara & Hittner (2012), quien exploró las propiedades de las distribuciones de población en el contexto de variables continuas y destacó cómo el aumento en el tamaño de la muestra puede transformar la representación visual de los datos.

En la Figura 6, se presentan dos histogramas basados en muestras, uno generado a partir de una muestra con un tamaño de 100 y otro construido a partir de una muestra con un tamaño de 2000. También se incluye una curva suave que representa la distribución de la población. Esta representación gráfica ilustra cómo, a medida que aumenta el tamaño de la muestra, la forma del histograma se asemeja cada vez más a la distribución de la población subyacente, lo que resalta la utilidad de las curvas suaves en la representación de distribuciones poblacionales.

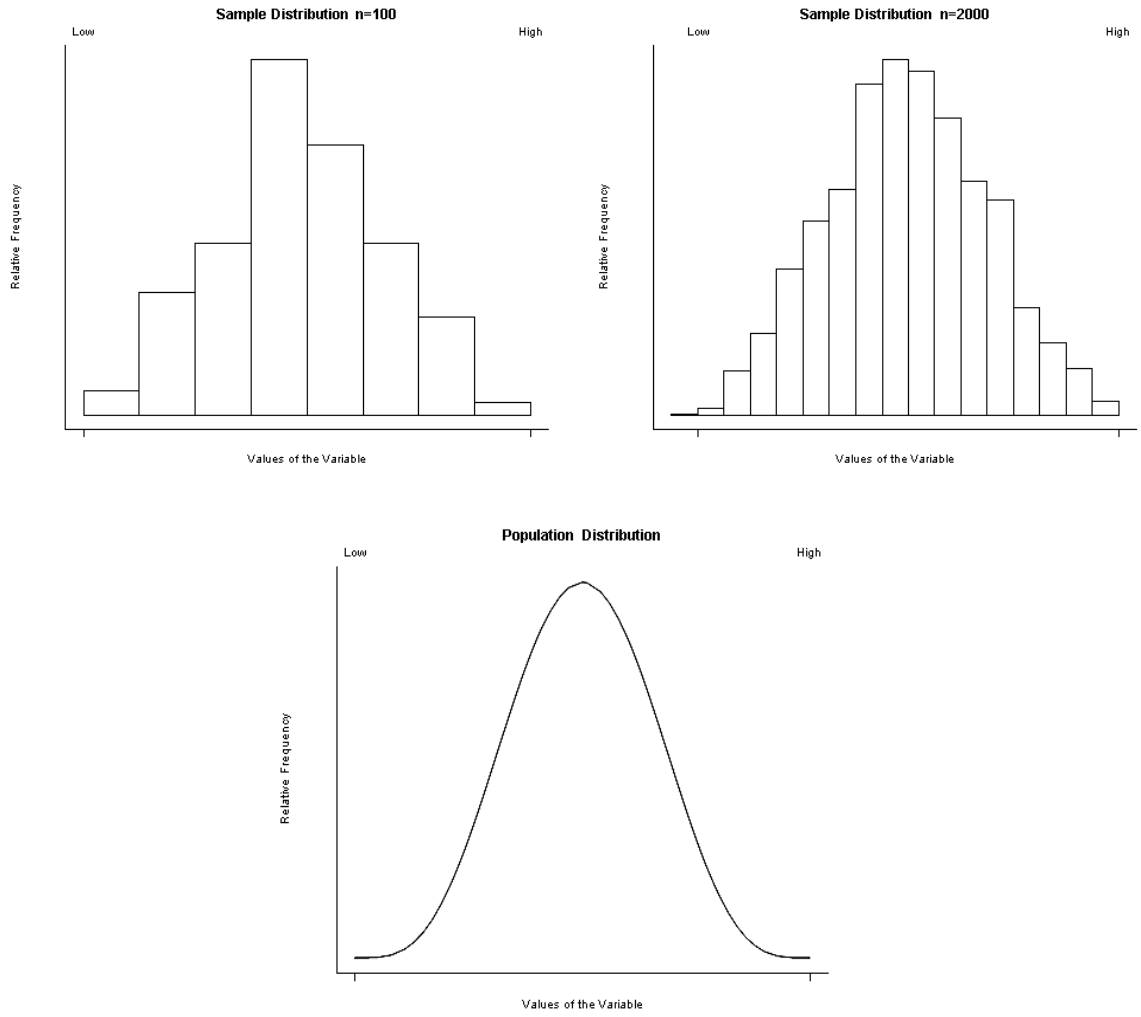


Figura 6: Distribuciones de la muestra y la población

Una de las maneras de resumir la característica de una muestra tomada de una distribución de población es describir su forma. La forma de una distribución es un indicador fundamental que diferencia significativamente un conjunto de datos cuya distribución presenta una forma en campana de otro conjunto que exhibe una forma en U, por mencionar un ejemplo.

Es importante destacar que las distribuciones en forma de campana y en forma de U, como se ilustran en la Figura 7, se caracterizan por ser simétricas. Esto significa que sus valores se distribuyen de manera equitativa a ambos lados de un punto central, lo que resulta en una simetría perfecta. Por otro lado, cuando una

distribución no presenta esta simetría, se dice que está sesgada, y la dirección del sesgo se describe como "a la derecha" o "a la izquierda" en función de cuál de las colas de la distribución sea más larga.

La importancia de comprender la forma de una distribución y su simetría o sesgo se refleja en la capacidad para interpretar adecuadamente los datos y tomar decisiones informadas. En este contexto, MacKinnon et al. (2004) subrayan la relevancia de la forma de la distribución en el análisis estadístico y resalta cómo las distribuciones simétricas y sesgadas pueden requerir enfoques analíticos distintos para una comprensión más precisa de los datos.

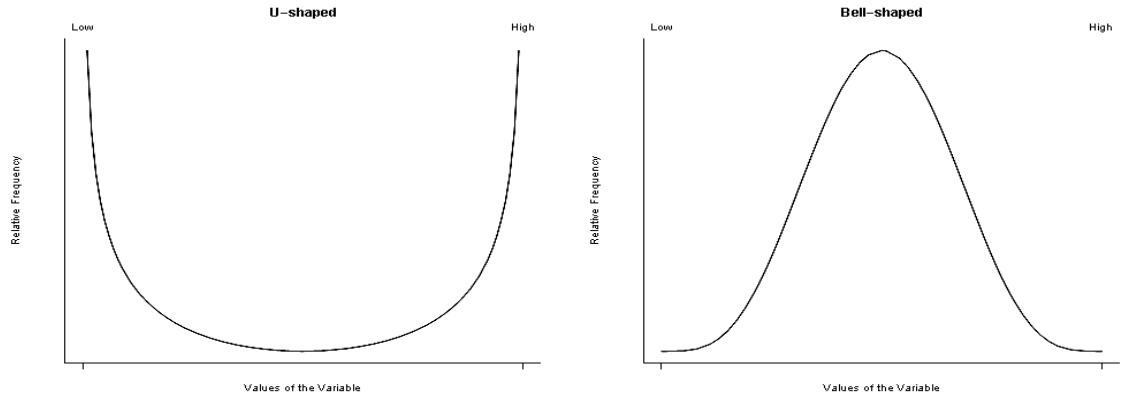


Figura 7: Distribuciones de frecuencias en forma de U y de campana

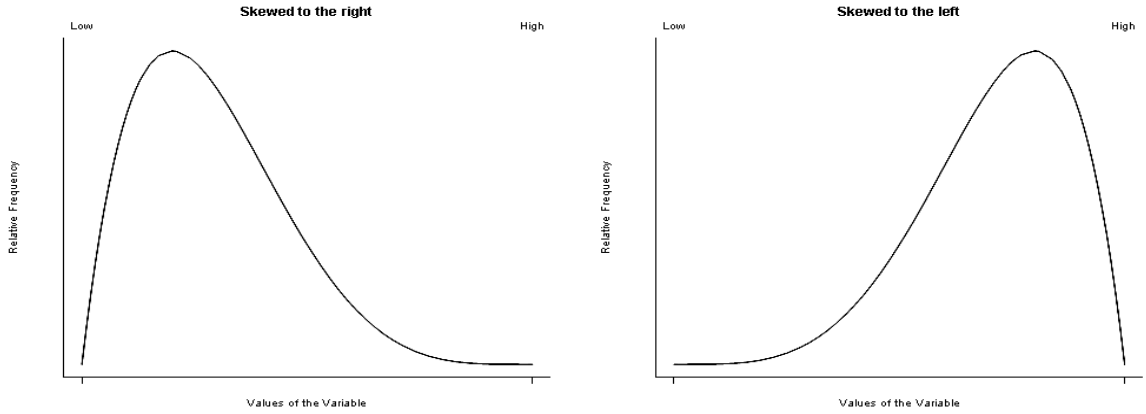


Figura 8: Distribuciones de frecuencias asimétricas

4. Medidas de tendencia central

Las medidas de tendencia central desempeñan un papel fundamental al proporcionar información esencial sobre la ubicación o el valor típico de una variable dentro de un conjunto de datos. Estas medidas, a menudo referidas como "promedios," son un punto de referencia esencial para comprender la estructura y distribución de los datos (Rayat, 2018).

4.1. Media Aritmética (Promedio):

La media aritmética es una de las medidas de tendencia central más utilizadas y se denota como $\bar{x}$ se calcula sumando todos los valores en el conjunto de datos y luego dividiendo esta suma por el número total de observaciones. La fórmula general es:

$$\bar{x} = \frac{\sum x}{N}$$

Donde $\sum \boldsymbol{x}$ representa la suma de todos los valores y *N* es el número total de observaciones. La media es aplicable únicamente a datos cuantitativos, ya que involucra operaciones matemáticas como la adición y la división.

4.2. La moda

La moda muestral de una variable cualitativa o cuantitativa discreta es el valor de la variable que aparece con mayor frecuencia en un conjunto de datos. A continuación, se ofrece una definición más exacta de la moda.

La **moda** obtiene la frecuencia de cada valor observado de la variable en un dato y anota la mayor frecuencia.

1. Si la mayor frecuencia es 1 (es decir, ningún valor aparece más de una vez), la variable no tiene moda.
2. Si la mayor frecuencia es 2 o superior, cualquier valor que aparezca con esa mayor frecuencia se denomina moda muestral de la variable.

Para obtener la(s) moda(s) de una variable, primero construimos una distribución de frecuencias para los datos utilizando clases basadas en un único valor. A continuación, la(s) moda(s) puede(n) determinarse fácilmente a partir de la distribución de frecuencias.

Ejemplo. La tabla de frecuencias de los grupos sanguíneos de 40 personas.

Ver en la tabla de frecuencias que la moda de los tipos de sangre es A. La moda en SPSS:

Analizar -> Estadística descriptiva -> Frecuencias

Tabla 6: Distribución de frecuencias de los tipos de sangre

SANGRE	Estadísticas	
	Frecuencia	Porcentaje
O	16	40.0
A	18	45.0
B	4	10.0
AB	2	5.0
Total	**40**	**100.0**

Cuando se lleva a cabo la medición de una variable continua (o incluso una variable discreta con una amplia gama de valores), como la altura o el peso de una persona, se encuentra una peculiaridad interesante: cada medición puede ser diferente de las demás. En tal situación, no existe un valor que se repita con mayor frecuencia, y, por lo tanto, no se puede identificar una moda, ya que cada valor observado tiene una frecuencia de 1.

Sin embargo, en muchas ocasiones, para facilitar el análisis de datos y obtener una comprensión más significativa de la distribución, los datos pueden agruparse en intervalos de clase. Aquí es donde entra en juego la noción de moda basada en frecuencias de clase. En el contexto de variables cuantitativas agrupadas, la moda se define como el intervalo de clase que contiene la frecuencia más alta.

Tomemos un ejemplo concreto para ilustrar este concepto. Supongamos que tenemos una tabla de frecuencias que registra los precios de los perros calientes en dólares por onza ($/oz). Dentro de esta tabla, observamos diferentes intervalos de clase junto con sus respectivas frecuencias. En este escenario, la clase de moda sería el intervalo de clase que tiene la frecuencia más alta. Por ejemplo, si en la tabla se observa que el intervalo de clase "0,065-0,085" tiene la frecuencia más alta en comparación con otros intervalos, entonces esta sería la clase modal.

La identificación de la clase modal en datos agrupados proporciona una representación resumida de la ubicación central de los valores en la distribución y puede ser valiosa para comprender dónde se concentran los datos en un conjunto de datos continuo o discreto agrupado en intervalos de clase.

Tabla 7: Distribución de frecuencia de los precios de los perros calientes

Frecuencias de los precios de perros calientes ($/oz.)

Intervalo de Valores	Frequencia	Porcentaje	Porcentaje Acumulativo
0.045-0.065	5	9.3	9.3
0.065-0.085	15	27.8	37.0
0.085-0.105	10	18.5	55.6
0.105-0.125	9	16.7	72.2
0.125-0.145	4	7.4	79.6
0.145-0.165	2	3.7	83.3
0.165-0.185	3	5.6	88.9
0.185-0.205	3	5.6	94.4
0.205-0.225	1	1.9	96.3
0.225-0.245	1	1.9	98.1
0.245-0.265	1	1.9	100.0
Total	**54**	**100.0**	

4.3. La mediana

La mediana muestral de una variable cuantitativa es el valor de la variable en un conjunto de datos que divide el conjunto de valores observados por la mitad, de forma que los valores observados en una mitad son menores o iguales que el valor

de la mediana y los valores observados en la otra mitad son mayores o iguales que el valor de la mediana. Para obtener la mediana de la variable, ordenamos los valores observados en un conjunto de datos en orden creciente y, a continuación, determinamos el valor medio de la lista ordenada.

Ordene los valores observados de la variable en un conjunto de datos en orden creciente.

1. Si el número de observación es impar, entonces la mediana muestral es el valor observado exactamente en el centro de la lista ordenada.

2. Si el número de observaciones es par, la mediana de la muestra es el número situado a medio camino entre los dos valores observados centrales de la lista ordenada.

En ambos casos, si n denota el número de observaciones de un conjunto de datos, entonces la mediana muestral está en la posición $\frac{n+1}{2}$ en la lista ordenada.

Ejemplo. 7 participantes en una carrera ciclista tuvieron los siguientes tiempos finales en minutos: 28,22,26,29,21,23,24.

¿Cuál es la mediana?

Ejemplo. 8 participantes en la carrera de bicicletas tuvieron los siguientes tiempos finales en minutos: 28,22,26,29,21,23,24,50.

¿Cuál es la mediana? La mediana en SPSS:

Analizar -> Estadística Descriptiva -> Frecuencias

La mediana es un valor "central" - hay tantos valores mayores que él como menores que él.

4.4. La media

La medida del centro más utilizada para las variables cuantitativas es la media (aritmética) de la muestra. Cuando la gente habla de sacar una media, es a la media a la que se refiere con más frecuencia.

La media muestral de la variable es la suma de los valores observados en un dato dividida por el número de observaciones.

Ejemplo. 7 participantes en una carrera ciclista tuvieron los siguientes tiempos finales en minutos: 28,22,26,29,21,23,24.

¿Cuál es la media?

Ejemplo. 8 participantes en la carrera ciclista tuvieron los siguientes tiempos finales en minutos: 28,22,26,29,21,23,24,50.

¿Cuál es la media? La media en SPSS:

Analizar -> Estadística Descriptiva -> Frecuencias,

Analizar -> Estadística Descriptiva -> Descriptivos

Para presentar eficazmente las ideas y los cálculos asociados, es conveniente representar las variables y los valores observados de las variables mediante símbolos para evitar que la discusión quede anclada a un conjunto específico de números. Así pues, se utiliza x para denotar la variable en cuestión y, a continuación, el símbolo xi denota la i-ésima observación de esa variable en el conjunto de datos.

Si el tamaño de la muestra es n, la media de la variable x es

$$\frac{X_1 + X_2 + X_3 \dots X_n}{n}$$

Para simplificar aún más la escritura de una suma, se utiliza la letra griega $\sum$ (sigma) como abreviatura. La suma $x_1 + x_2 + x_3 + - - - + x_n$ se expresa como

$$\sum_{i=1}^{n} x_i$$

y se lee como "la suma de todos los x_i con i comprendido entre 1 y n". Por lo tanto, se define formalmente la media de la siguiente manera.

La media muestral de la variable es la suma de los valores observados x_1, x_2, x_3, ..., x_n en un dato dividida por el número de observaciones n. La media muestral se denota por $\bar{X}$, y se expresa de forma operativa,

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} \text{ ó } \frac{\sum x_i}{n}$$

4.5. ¿Qué medida elegir?

La moda debe utilizarse para calcular la medida central de la variable cualitativa. Cuando la variable es cuantitativa con distribución simétrica, la media es la medida adecuada del centro. En el caso de una variable cuantitativa con distribución sesgada, la mediana es una buena elección para la medida del centro. Esto está relacionado con el hecho de que la media puede estar muy influida por una observación que se aleje del resto de los datos, denominada **valor atípico**.

Hay que tener en cuenta que la moda, la mediana y la media muestrales de la variable en cuestión tienen las correspondientes medidas poblacionales de centro, es decir, que la variable en cuestión tiene también la moda, la mediana y la media poblacionales, que son todas desconocidas. Entonces, la moda, la mediana y la media muestrales pueden utilizarse para **estimar** los valores de estos valores poblacionales desconocidos correspondientes.

5. Medidas de variación

Además de localizar el centro de los valores observados de la variable en los datos, otro aspecto importante de un estudio descriptivo de la variable es medir numéricamente el grado de variación en torno al centro. Dos conjuntos de datos de la misma variable pueden mostrar posiciones similares del centro, pero pueden ser notablemente diferentes con respecto a la variabilidad.

Al igual que existen distintas medidas del centro, también existen distintas medidas de la variación. En este apartado examinaremos tres de las medidas de variación más utilizadas: **el rango muestral**, el **rango intercuartílico muestral** y la **desviación típica muestral**. Las medidas de variación se utilizan sobre todo para variables cuantitativas.

5.1. Rango

El rango de la muestra se obtiene calculando la diferencia entre el mayor valor observado de la variable en un conjunto de datos y el menor.

El **rango muestral** de la variable es la diferencia entre sus valores máximo y mínimo en un conjunto de datos:

Rango = Máx - Mín.

El rango muestral de la variable es bastante fácil de calcular. Sin embargo, al utilizar el rango, se ignora una gran cantidad de información, es decir, sólo se consideran los valores mayor y menor de la variable; los demás valores observados no se tienen en cuenta. También hay que señalar que el rango nunca puede disminuir, sino aumentar, cuando se incluyen observaciones adicionales en el conjunto de datos y que, en cierto sentido, el rango es demasiado sensible al tamaño de la muestra.

Ejemplo. 7 participantes en una carrera ciclista tuvieron los siguientes tiempos finales en minutos: 28,22,26,29,21,23,24.

¿Cuál es el rango?

Ejemplo. 8 participantes en la carrera ciclista tuvieron los siguientes tiempos finales en minutos: 28,22,26,29,21,23,24,50.

¿Cuál es el rango?

Ejemplo. Precios de los perros calientes ($/oz.):

0.11,0.17,0.11,0.15,0.10,0.11,0.21,0.20,0.14,0.14,0.23,0.25,0.07,

0.09,0.10,0.10,0.19,0.11,0.19,0.17,0.12,0.12,0.12,0.10,0.11,0.13,

0.10,0.09,0.11,0.15,0.13,0.10,0.18,0.09,0.07,0.08,0.06,0.08,0.05,

0.07,0.08,0.08,0.07,0.09,0.06,0.07,0.08,0.07,0.07,0.07,0.08,0.06,

0.07,0.06

El rango en SPSS:

Analizar -> Estadística Descriptiva -> Frecuencias,

Analizar -> Estadística Descriptiva -> Descriptivos

Tabla 8: Rango de precios de los perros calientes

Rango de precios de los perros calientes

	N	Rango	Mínimo	Máximo
Precio ($/oz)	54	.20	.05	.25
Valid N (lista)	54			

5.2. Rango intercuartílico

Antes de poder definir el rango intercuartílico de la muestra, hay que definir primero los **percentiles**, los **deciles** y los **cuartiles** de la variable en un conjunto de datos. Como se ha visto, la mediana de la variable divide los valores observados en dos partes iguales: el 50% inferior y el 50% superior. Los percentiles de la variable dividen los valores observados en centésimas, es decir, 100 partes iguales. A grandes rasgos, el primer percentil, P_1, es el número que divide el 1% inferior de los valores observados del 99% superior; el segundo percentil, P_2, es el número que divide el 2% inferior de los valores observados del 98% superior; y así sucesivamente. La mediana es el percentil 50.

Los deciles de la variable dividen los valores observados en décimas, o 10 partes iguales. La variable tiene nueve deciles, denotados por D_1, D_2, . . . , D_9. El primer decil D_1 es el percentil 10, el segundo decil D_2 es el percentil 20, y así sucesivamente.

Los percentiles más utilizados son los cuartiles. Los cuartiles de la variable dividen los valores observados en cuartos, o 4 partes iguales. La variable tiene tres cuartiles, denominados Q_1, Q_2 y Q_3. A grandes rasgos, el primer cuartil, Q_1, es el número que divide el 25% inferior de los valores observados del 75% superior; el segundo cuartil, Q_2, es la mediana, que es el número que divide el 50% inferior de los valores observados del 50% superior; y el tercer cuartil, Q_3, es el número que divide el 75% inferior de los valores observados del 25% superior.

En este punto, nuestras definiciones intuitivas de percentiles y deciles serán suficientes. Sin embargo, es necesario definir los cuartiles con mayor precisión, lo que se hace a continuación.

Sea n el número de observaciones de un conjunto de datos. Ordene los valores observados de una variable en un conjunto de datos en orden creciente.

1. El primer cuartil Q_1 está en la posición $\frac{n+1}{4}$,
2. El segundo cuartil Q_2 (la mediana) está en la posición $\frac{n+1}{2}$,
3. El tercer cuartil Q_3 está en la posición $\frac{3(n+1)}{4}$,

en la lista ordenada.

Si una posición no es un número entero, se utiliza la interpolación lineal.

A continuación, definimos el rango intercuartílico de la muestra. Dado que el rango intercuartílico se define utilizando cuartiles, es la medida de variación preferida cuando se utiliza la mediana como medida del centro (es decir, en caso de distribución sesgada).

El **rango intercuartílico** muestral de la variable, denotado IQR, es la diferencia entre el primer y el tercer cuartil de la variable, es decir

$$IQR = Q_3 - Q_1.$$

A grandes rasgos, el IQR da el rango del 50% medio de los valores observados.

El rango intercuartílico de la muestra representa la longitud del intervalo cubierto por la mitad central de los valores observados de la variable. Esta medida de variación no se ve alterada si una pequeña fracción de los valores observados son muy grandes o muy pequeños.

Ejemplo. 7 participantes en una carrera ciclista tuvieron los siguientes tiempos finales en minutos: 28,22,26,29,21,23,24.

¿Cuál es el rango intercuartílico?

Ejemplo. 8 participantes en la carrera ciclista tuvieron los siguientes tiempos finales en minutos: 28,22,26,29,21,23,24,50.

¿Cuál es el rango intercuartílico?

Ejemplo. El rango intercuartílico para los precios de los perros calientes ($/oz.) en SPSS:

Analizar -> Estadística Descriptiva -> Explorar

Tabla 9: Rango intercuartílico de los precios de los perros calientes

Rango intercuartílico de los precios de los perros calientes

Estadístico	Valor
Rango Intercuartílico	0.0625 $/oz

5.2.1. Resumen de cinco números y boxplots

Mínimo, máximo y cuartiles proporcionan información sobre el centro y la variación de la variable de una forma compacta y agradable. Escritos en orden creciente, constituyen lo que se denomina el **resumen de cinco números** de la variable.

El resumen de cinco números de la variable consiste en el mínimo, el máximo y los cuartiles escritos en orden creciente:

Mín, Q1, Q2, Q3, Máx.

Un diagrama de caja, también conocido como boxplot, es una herramienta gráfica ampliamente utilizada en estadísticas para resumir y visualizar la distribución, la variabilidad y los valores atípicos en un conjunto de datos. Este gráfico se basa en un conjunto de cinco números resumen, que proporcionan información esencial sobre la variable en cuestión. Hay dos tipos principales de boxplots en uso común: el boxplot estándar y el boxplot modificado, y la elección entre ellos depende de la necesidad de detectar valores atípicos.

Hubert & Vandervieren (2008) introdujo el concepto del boxplot como una forma eficaz de representar gráficamente la distribución de un conjunto de datos. El boxplot es especialmente útil para identificar de manera visual la mediana (el valor en el centro de los datos) y los cuartiles, que dividen el conjunto de datos en cuatro partes iguales.

Ahora, en lo que respecta al boxplot modificado, la diferencia clave radica en cómo se manejan los valores atípicos potenciales. En un boxplot estándar, los valores atípicos se tratan como observaciones individuales y se representan de manera discreta si están por encima o por debajo de ciertos umbrales, conocidos como límites de valores atípicos (Kürzl, 1988). En cambio, en un boxplot modificado, estos valores atípicos potenciales se representan de manera individual, lo que facilita su identificación en el gráfico.

Para construir un diagrama de caja estándar o modificado, se siguen los siguientes pasos:

1. Calcule el primer cuartil (Q1) y el tercer cuartil (Q3) del conjunto de datos.
2. Calcule el rango intercuartílico (RIC) como la diferencia entre Q3 y Q1.

3. Determine los límites inferior y superior de los bigotes. En un boxplot estándar, los límites se calculan como 1.5 veces el RIC por debajo de Q1 y por encima de Q3.
4. Identifique los valores que están por debajo del límite inferior o por encima del límite superior de los bigotes en un boxplot estándar. En un boxplot modificado, represente individualmente todos los valores atípicos potenciales.
5. Dibuje la caja que se extiende desde Q1 hasta Q3, con una línea que representa la mediana en su interior. Luego, dibuje los bigotes desde los límites inferior y superior, y represente los valores atípicos según el tipo de boxplot que esté creando.

Para construir un boxplot

1. Determine el resumen de cinco números

2. Dibuje un eje horizontal (o vertical) en el que puedan situarse los números obtenidos en el paso 1. Sobre este eje, marque los cuartiles y el mínimo y el máximo con líneas verticales (horizontales).

3. Conecte los cuartiles entre sí para formar una caja y, a continuación, conecte la caja al mínimo y al máximo con líneas.

El boxplot modificado puede construirse de forma similar; excepto que los valores atípicos potenciales se identifican primero y se trazan individualmente y los valores mínimo y máximo en el boxplot se sustituyen por los valores adyacentes, que son las observaciones más extremas que no son valores atípicos potenciales.

Ejemplo. 7 participantes en una carrera ciclista tuvieron los siguientes tiempos finales en minutos: 28,22,26,29,21,23,24.

Construya el diagrama de caja.

Ejemplo. El resumen de cinco números y boxplot para precios de perros calientes ($/oz.) en SPSS:

Analizar -> Estadística Descriptiva -> Descriptivos

Tabla 10: Resumen de cinco números de los precios de los perritos calientes

Resumen de cinco números

N	Valid	54
	Missing	0
Mediana		.1000
Mínimo		.05
Máximo		.25
Percentiles	25	.0700
	50	.1000
	75	.1325

Gráficos -> Interactivos -> Boxplot,

Gráficos -> Boxplot

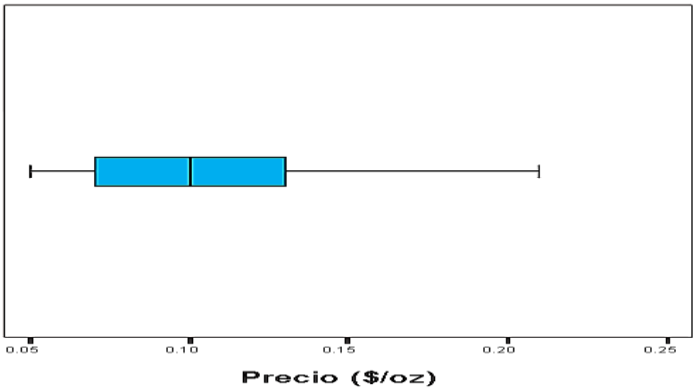


Figura 9: Boxplot de los precios de los perros calientes

5.3. Desviación típica o desviación estándar

La desviación típica de la muestra es la medida de variabilidad más utilizada, aunque no es tan fácil de entender como los rangos. Puede considerarse como una especie de media de las desviaciones absolutas de los valores observados con respecto a la media de la variable en cuestión.

Para una variable x, la desviación típica muestral, denotada por s_x (o cuando no hay confusión, simplemente por s), es

$$S_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Dado que la desviación típica se define utilizando la media muestral $\bar{x}$ de la variable x, es la medida de variación preferida cuando se utiliza la media como medida del centro (es decir, en caso de distribución simétrica). Obsérvese que la desviación típica es siempre un número positivo, es decir, $s_x \geq 0$.

En una fórmula de la desviación típica, la suma de las desviaciones al cuadrado de la media,

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2,$$

Se denomina suma de desviaciones al cuadrado y proporciona una medida de la desviación total de la media para todos los valores observados de la variable. Una vez que la suma de desviaciones al cuadrado se divide por n - 1, obtenemos

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1},$$

Que se denomina varianza muestral. La desviación típica muestral tiene las siguientes fórmulas alternativas:

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \qquad (1)$$

$$s_x = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}} \qquad (2)$$

$$s_x = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 / n}{n-1}} \qquad (3)$$

Las fórmulas (2) y (3) son útiles desde el punto de vista computacional. En el cálculo manual, el uso de estas fórmulas alternativas suele reducir el trabajo aritmético, sobre todo cuando $\bar{x}$ resulta ser un número con muchos decimales.

Cuanto mayor sea la variación de los valores observados, mayor será la desviación típica de la variable en cuestión. Así pues, la desviación típica satisface

el criterio básico de una medida de variación y, como se ha dicho, es la medida de variación más utilizada. Sin embargo, la desviación típica tiene sus inconvenientes. Por ejemplo, sus valores pueden verse muy afectados por unas pocas observaciones extremas.

Ejemplo. 7 participantes en una carrera ciclista tuvieron los siguientes tiempos finales en minutos: 28,22,26,29,21,23,24.

¿Cuál es la desviación típica de la muestra?

Ejemplo. La desviación típica de los precios de los perros calientes ($/oz.) en SPSS:

Analizar -> Estadística Descriptiva -> Frecuencias,

Analizar -> Estadística Descriptiva -> Descriptivos

Cuadro 11: Desviación típica de los precios de los perros calientes

Desviación típica de los precios de los perros calientes

	N	Media	Desviación St	Varianza
Precio ($/oz)	54	.1113	.04731	.002
Valid N (lista)	54			

5.3.1. Regla empírica para distribuciones simétricas

Para las distribuciones simétricas en forma de campana (como la distribución normal), la regla empírica relaciona la desviación típica con la proporción de los valores observados de la variable en un conjunto de datos que se encuentran en un intervalo alrededor de la media $\bar{x}$.

Regla empírica para una distribución simétrica en forma de campana, aproximadamente

68% de los valores se encuentran dentro de $\bar{x} \pm s_x$,

95% de los valores se encuentran dentro de $\bar{x} \pm 2s_x$,

99,7% de los valores se encuentran dentro de $\bar{x} \pm 3s_x$.

En el contexto de datos agrupados, donde los valores originales se han clasificado en intervalos de clase, calcular la desviación media, la varianza y la desviación típica requiere un enfoque ligeramente diferente en comparación con datos no agrupados. Aquí, presentaré cómo calcular estas medidas para datos agrupados:

5.4. Desviación Media en Datos Agrupados:

La desviación media en datos agrupados mide la dispersión promedio de los datos en relación con la media de la variable cuantitativa agrupada. Para calcularla, sigue estos pasos:

- Identifica el punto medio (marca de clase) de cada intervalo de clase.
- Calcula la desviación entre la marca de clase y la media de la variable (diferencia absoluta).
- Multiplica cada desviación por la frecuencia de su intervalo de clase.
- Suma todas estas desviaciones ponderadas y divide por el número total de observaciones.

La fórmula para la desviación media en datos agrupados es:

Desviación Media$-\frac{1}{N}\sum_{i-1}^{k} fi\ |Xi - \mu|$

Donde:

- N es el número total de observaciones.
- k es el número de intervalos de clase.
- f_i es la frecuencia de la clase
- X_i es la marca de clase del intervalo
- μ es la media de la variable.

5.5. Varianza en Datos Agrupados:

La varianza en datos agrupados mide la dispersión de los datos cuantitativos en relación con la media cuantificando la media de los cuadrados de las desviaciones entre los valores y la media. Para calcularla, sigue estos pasos:

- Calcula las desviaciones entre la marca de clase y la media de la variable.
- Cuadra cada desviación.
- Multiplica cada cuadrado de desviación por la frecuencia de su intervalo de clase.
- Suma todos estos valores y divide por el número total de observaciones.

La fórmula para la varianza en datos agrupados es:

Varianza $-\frac{1}{N}\sum_{i-1}^{k} fi\,(Xi-\mu)^2$

5.6. Desviación Típica en Datos Agrupados:

La desviación típica en datos agrupados es simplemente la raíz cuadrada de la varianza. Esta medida proporciona una estimación de la dispersión promedio de los datos en relación con la media, en las mismas unidades que la variable original. La fórmula es:

Desviación Típica $-\sqrt{\text{Varianza}}$

Calcular estas medidas en datos agrupados es fundamental para comprender la variabilidad de una variable cuantitativa cuando se trabaja con intervalos de clase en lugar de valores individuales. Estas fórmulas te permitirán cuantificar la dispersión de los datos de manera más precisa y tomar decisiones informadas en tu análisis estadístico.

5.7. Estadísticas de la muestra y parámetros de la población

De las medidas de centro y variación, la media muestral $\bar{x}$ y la desviación típica muestral s son las que se comunican con más frecuencia. Dado que sus valores dependen de la muestra seleccionada, varían de una muestra a otra. En este

sentido, se denominan variables aleatorias para destacar que sus valores varían en función de la muestra seleccionada. Sus valores son desconocidos antes de elegir la muestra. Una vez seleccionada la muestra y calculados, se convierten en estadísticos muestrales conocidos.

Distinguir regularmente entre los estadísticos muestrales y las medidas correspondientes para la población. Antes, se introdujo el parámetro de una medida resumen de la población. Un estadístico describe una muestra, mientras que un parámetro describe la población de la que se ha extraído la muestra.

Notación de los parámetros. Sean μ y σ la media y la desviación típica de una variable para la población.

La media de la población es la media de las medidas de la población. La desviación típica de la población describe la variación de las medidas de la población con respecto a la media de la población.

Mientras que los estadísticos $\bar{x}$ son variables s, cuyos valores dependen de la muestra elegida, los parámetros μ y σ son constantes. Esto se debe a que μ y σ se refieren sólo a un grupo concreto de mediciones, a saber, las mediciones de toda la población. Por supuesto, los valores de los parámetros suelen ser desconocidos, razón por la que se toman muestras y se calculan los estadísticos muestrales como estimaciones de sus valores. Es decir, hacer inferencias sobre parámetros desconocidos (como μ y σ) utilizando estadísticos muestrales (como $\bar{x}$ y s).

5.8. Varianza

La variabilidad en un conjunto de datos se analiza mediante la utilización de dos medidas estadísticas clave: la varianza y la desviación estándar o desviación típica. Estas métricas son esenciales para comprender la dispersión de los datos y evaluar su consistencia.

La varianza cuantifica la magnitud de las diferencias entre los valores de un conjunto de datos y su valor promedio. En el caso de datos agrupados, la varianza

se calcula teniendo en cuenta la agrupación de datos en intervalos o clases. Este enfoque permite estimar la variabilidad dentro de cada intervalo y luego combinar estas variabilidades para obtener la varianza total.

5.9. Desviación estándar

La desviación estándar, también conocida como desviación típica, es la raíz cuadrada de la varianza. Proporciona una medida de la dispersión que se encuentra en la misma escala que los datos originales. En el contexto de datos agrupados, la desviación estándar se calcula de manera similar a la varianza, pero se toma la raíz cuadrada de la varianza calculada en lugar de utilizar directamente la varianza.

6. Distribuciones de probabilidad

Los métodos estadísticos inferenciales se basan en el uso de datos de muestra con el fin de realizar predicciones acerca de los valores de parámetros descriptivos cruciales que caracterizan la población de interés. Así mismo, se aborda la concepción de los parámetros como magnitudes conocidas, lo cual constituye un enfoque artificial. En realidad, los valores de estos parámetros suelen ser desconocidos, de lo contrario, no se requeriría la aplicación de métodos inferenciales.

No obstante, gran parte de los métodos inferenciales se fundamentan en la comparación entre las estadísticas observadas en la muestra y los valores esperados si los parámetros fueran establecidos en cifras específicas. En caso de que los datos no concuerden de manera coherente con estos valores particulares de los parámetros, se infiere que los valores reales de dichos parámetros difieren en cierta medida. Este proceso permite obtener estimaciones más precisas y adecuadas de los parámetros poblacionales, a pesar de la incertidumbre inherente a su naturaleza desconocida.

6.1. Distribuciones de probabilidad

Para abordar el concepto de probabilidad, se adoptará un enfoque basado en frecuencias relativas. Se puede considerar un experimento hipotético que consiste

en una sucesión extensa de observaciones repetidas de un fenómeno aleatorio. Cada observación puede conducir o no a la ocurrencia de un resultado específico. La probabilidad asociada a dicho resultado se establece como la frecuencia relativa de su aparición a lo largo de un período prolongado.

La probabilidad de un resultado concreto es la proporción de veces que ese resultado se produciría en una serie larga de observaciones repetidas.

Una representación simplificada de un experimento de este tipo es una secuencia muy larga de lanzamientos de una moneda, cuyo resultado de interés es que salga cara. Cualquier lanzamiento puede o no dar como resultado una cara. Si la moneda está equilibrada, un resultado básico de la probabilidad, llamado **ley de los grandes números**, implica que la proporción de lanzamientos que resultan en una cara tiende a 1/2 a medida que aumenta el número de lanzamientos. Así, la probabilidad de que salga cara en cualquier lanzamiento de la moneda es igual a 1/2.

En la mayoría de las ocasiones, nos enfrentamos a variables que producen resultados numéricos. Una variable que puede adquirir al menos dos valores numéricos distintos en una serie prolongada de observaciones repetidas recibe el nombre de variable aleatoria.

Una variable aleatoria es una variable cuyo valor es el resultado numérico de un fenómeno aleatorio.

Usualmente, se representan las variables aleatorias mediante letras mayúsculas ubicadas hacia el final del alfabeto, como X o Y. Dentro de la variable aleatoria X, ciertos valores pueden presentar una probabilidad mayor que otros. La distribución de probabilidad de X registra los resultados potenciales, acompañados de sus respectivas probabilidades, que dicha variable X podría asumir.

La distribución de probabilidad de una variable aleatoria discreta X asigna una probabilidad a cada uno de los posibles valores de la variable. Cada probabilidad es un número entre 0 y 1, y la suma de las probabilidades de todos los valores

posibles es igual a 1. Sea x_i, i = 1, 2, . . . , k, denota un posible resultado de la variable aleatoria X, y P (X = x_i) = P (x_i) = p_i denota la probabilidad de ese resultado. Entonces

$$0 \leq P(x_i) \leq 1 \quad \text{y} \quad \sum_{i=1}^{k} P(x_i) = 1$$

Cada probabilidad está comprendida entre 0 y 1, y puesto que la probabilidad total es igual a 1.

Distribución de probabilidad de una variable aleatoria discreta. Una variable aleatoria discreta X tiene un número contable de valores posibles. La distribución de probabilidad de X enumera los valores y sus probabilidades:

Valor de X	X_1	X_2	X_3	…	X_k
Probabilidad	$P(X_1)$	$P(X_2)$	$P(X_3)$	…	$P(X_k)$

Las probabilidades P(x_i) deben cumplir dos requisitos:

1. Cada probabilidad P(x_i) es un número comprendido entre 0 y 1.

2. P(x_1) + P(x_2) + - - - + P(x_k) = 1.

Para representar la distribución de probabilidad de una variable aleatoria discreta, podemos emplear un histograma de probabilidad. Además, podemos calcular la probabilidad de cualquier evento, como P (X ≤ xi) o P (xi ≤ X ≤ xj), donde i ≤ j, sumando las probabilidades P(xi) de los valores específicos xi que conforman dicho evento.

Para ilustrar esto, consideremos el siguiente ejemplo. Un instructor en una clase numerosa asigna calificaciones de la siguiente manera: el 15% de los estudiantes obtiene una calificación de 5 = excelente, el 20% recibe una calificación de 4 = muy bueno, el 30% obtiene una calificación de 3 = bueno, el 20% recibe una calificación de 2 = satisfactorio, el 10% recibe una calificación de 1 = suficiente y el 5% recibe

una calificación de 0 = fracaso. Si seleccionamos al azar a un estudiante de esta clase, la calificación del estudiante es una variable aleatoria X. Aunque el valor de X cambia en cada selección, siempre será uno de los siguientes: 0, 1, 2, 3, 4 o 5.

¿Cuál es la distribución de probabilidad de X? Dibuja un histograma de probabilidad para X.

Cuál es la probabilidad de que el alumno haya obtenido 4=muy bueno o mejor, es decir $P(X \geq 4)$?

La variable aleatoria *continua* X, por otro lado, toma todos los valores en algún intervalo de números entre a y b. Es decir, la variable aleatoria continua tiene un continuo de posibles valores que puede tener. Sean x_1 y x_2, $x_1 \leq x_2$, los posibles resultados de la variable aleatoria X que pueden tener valores en el intervalo de números comprendido entre a y b. Entonces, claramente, tanto x_1 como x_2 pertenecen al intervalo de a y b, es decir,

$$x_1 \in [a, b] \text{ y } x_2 \in [a, b],$$

Los valores individuales x1 y x2 definen un intervalo numérico [x1, x2]. En el contexto de una variable aleatoria continua X, su distribución de probabilidad asigna una probabilidad a cada uno de estos intervalos potenciales [x1, x2]. La probabilidad de que X se encuentre dentro de un intervalo específico [x1, x2] es un valor comprendido entre 0 y 1. Además, la probabilidad asociada al intervalo completo [a, b], que incluye todos los valores posibles, equivale a 1. En otras palabras, se establece como requisito que

$$0 \leq P(x_1 \leq X \leq x_2) \leq 1 \text{ y } P(a \leq X \leq b) = 1.$$

Distribución de probabilidad de una variable aleatoria continua. Una variable aleatoria continua X toma todos los valores en un intervalo de números [a, b]. La distribución de probabilidad de X describe las probabilidades $P(x_1 \leq X \leq x_2)$ de todos los intervalos posibles de números $[x_1, x_2]$.

Las probabilidades $P(x_1 \leq X \leq x_2)$ deben satisfacer dos requisitos:

1. Para todo intervalo $[x_1, x_2]$, la probabilidad $P(x_1 \leq X \leq x_2)$ es un número comprendido entre 0 y 1.

2. $P(a \leq X \leq b) = 1$.

El modelo de probabilidad de una variable aleatoria continua asigna probabilidades a intervalos de resultados en lugar de a resultados individuales. De hecho, todas las distribuciones de probabilidad continuas asignan probabilidad 0 a cada resultado individual.

La distribución de probabilidad de una variable aleatoria continua se representa mediante una **curva de densidad**. Una curva de densidad es una curva continua suave que tiene un área exactamente 1 debajo de ella, como las curvas que representan la distribución de la población en la sección 3.3. De hecho, la distribución de la población de una variable aleatoria continua se representa mediante una curva de densidad. De hecho, la distribución poblacional de una variable es, equivalentemente, la distribución de probabilidad del valor de esa variable para un sujeto seleccionado aleatoriamente de la población.

Ejemplo.

Probabilidades de una variable aleatoria continua

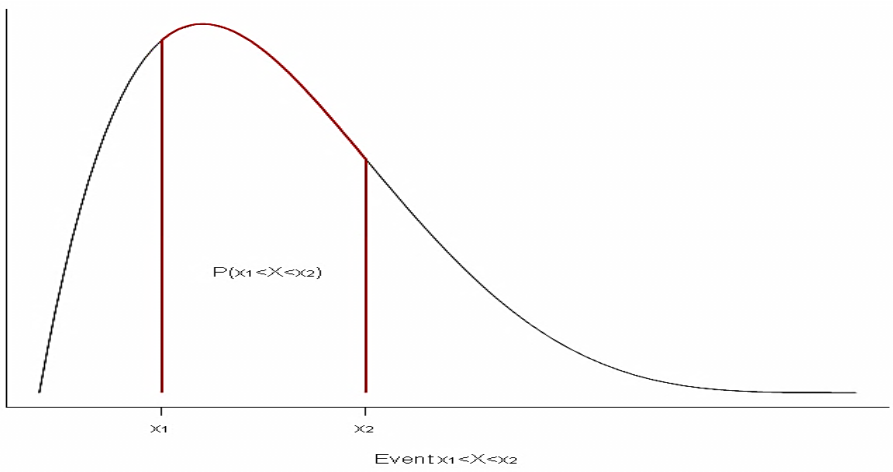


Figura 10: La distribución de probabilidad de una variable aleatoria continua asigna probabilidades como áreas bajo una curva de densidad.

6.2. Media y desviación típica de una variable aleatoria

Como la distribución de una población, la distribución de probabilidad de una variable aleatoria posee parámetros que caracterizan tanto su tendencia central como su variabilidad. La media se encarga de representar la tendencia central, mientras que la desviación típica describe la variabilidad de la variable aleatoria X. Estos parámetros son los valores que estas medidas asumirían a lo largo del tiempo si lleváramos a cabo observaciones repetidas de los valores generados por la variable aleatoria X.

La media y la desviación típica de la variable aleatoria discreta se definen de la siguiente manera.

Supongamos que X es una variable aleatoria discreta cuya distribución de probabilidad es

Valor de X	X_1	X_2	X_3	…	X_k
Probabilidad	$P(X_1)$	$P(X_2)$	$P(X_3)$	…	$P(X_k)$

La media de la variable aleatoria discreta X es

$$\mu = x_1P(x_1) + x_2P(x_2) + x_3P(x_3) + \cdot\cdot\cdot + x_kP(x_k)$$

$$= \sum_{i=1}^{k} x_1 P(x_i)$$

La media μ también se denomina valor esperado de X y se denota por E(X).

Desviación típica de una variable aleatoria discreta. Supongamos que X es una variable aleatoria discreta cuya distribución de probabilidad es

Value of X	x_1	x_2	x_3	. . .	x_k
Probability	$P(x_1)$	$P(x_2)$	$P(x_3)$	. . .	$P(x_k)$

y que μ es la media de X. La varianza de la variable aleatoria discreta X es

$$\sigma^2 = (x_1 - \mu)^2 P(x_1) + (x_2 - \mu)^2 P(x_2) + (x_3 - \mu)^2 P(x_3) + \cdots + (x_k - \mu)^2 P(x_k)$$

$$= \sum_{i=1}^{k} (x_i - \mu) P(x_i)$$

La desviación típica σ de X es la raíz cuadrada de la varianza.

Ejemplo. En un experimento sobre el comportamiento de niños pequeños, se coloca a cada sujeto en un área con cinco juguetes. La respuesta de interés es el número de juguetes con los que juega el niño. Experimentos anteriores con muchos sujetos han demostrado que la distribución de probabilidad del número X de juguetes con los que juega es la siguiente:

Núm de juguetes x_i	0	1	2	3	4	5
Probabilidad $P(x_i)$	0.03	0.16	0.30	0.23	0.17	0.11

Calcular la media μ y la desviación típica σ.

La media y la desviación típica de una variable aleatoria continua pueden calcularse, pero para ello se requieren matemáticas más avanzadas, por lo que no las consideramos en este libro.

6.3. Distribución normal

Cuando una variable aleatoria continua se representa gráficamente mediante una curva de densidad en forma de campana, se clasifica como una distribución normal. Esta distribución ostenta un papel fundamental en estadística debido a su capacidad para aproximarse eficazmente a las distribuciones de numerosas variables. Con frecuencia, los histogramas de conjuntos de datos muestrales exhiben una forma de campana, lo que nos lleva a inferir que la variable posee una distribución aproximadamente normal.

Sin embargo, la razón primordial de su relevancia radica en que la mayoría de los métodos estadísticos inferenciales emplean las propiedades inherentes a la

distribución normal, incluso en situaciones en las que los datos de la muestra no siguen una distribución de forma campaniforme.

Una variable aleatoria continua X que sigue una distribución normal tiene dos parámetros: la media μ y la desviación típica σ.

Distribución normal. Se dice que una variable aleatoria continua X está distribuida normalmente o tiene una distribución normal si su curva de densidad es una curva simétrica en forma de campana, caracterizada por su media μ y su desviación típica σ. Para cada número fijo z, la probabilidad concentrada en el intervalo [μ - zσ, μ + zσ] es la misma para todas las distribuciones normales. En particular, las probabilidades se mantienen

$$P(\mu - \sigma < X < \mu + \sigma) = 0.683 \qquad (4)$$

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.954 \qquad (5)$$

$$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.997 \qquad (6)$$

Una variable aleatoria X que sigue una distribución normal con una media de μ y una desviación típica de σ se denota por $X \sim N(\mu, \sigma)$.

Existen otras curvas de densidad simétricas en forma de campana que no son normales. Las curvas de densidad normales se especifican mediante una ecuación particular. La altura de la curva de densidad en cualquier punto x viene dada por la función de densidad

$$f(x) = \frac{1}{\sigma\sqrt{2\Pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \qquad (7)$$

Aunque este hecho constituye la base del trabajo matemático relacionado con la distribución normal, no se recurre a su utilización directa. Es importante señalar que la función de densidad queda plenamente determinada por los parámetros μ y σ.

Ejemplo.

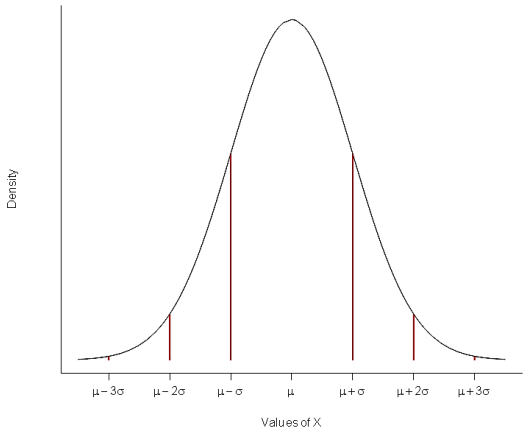


Figura 11. Distribución normal.

Distribución normal estándar. Se dice que una variable aleatoria continua Z tiene una distribución normal estándar si Z tiene una distribución normal con media $\mu = 0$ y desviación estándar $\sigma = 1$, es decir, $Z \sim N(0, 1)$.

La tabla normal estándar puede utilizarse para calcular probabilidades relativas a la variable aleatoria Z. La tabla normal estándar da el área a la izquierda de un valor especificado de z bajo la curva de densidad:

$$P(Z \leq z) = \text{Área bajo la curva a la izquierda de } z.$$

Para la probabilidad de un intervalo [a, b]:

$$P(a \leq Z \leq b) = [\text{Área a la izquierda de } b] - [\text{Área a la izquierda de } a].$$

A partir de la simetría de la distribución normal estándar en torno a 0 se pueden observar las siguientes propiedades:

(a) $P(Z \leq 0) = 0{,}5$,

(b) $P(Z \leq -z) = 1 - P(Z \leq z) = P(Z \geq z)$.

Ejemplo.

(a) Calcule $P(-0{,}155 < Z < 1{,}60)$.

(b) Localice el valor z que satisface P (Z > z) = 0,25.

Si la variable aleatoria X se distribuye como X N(μ, σ), entonces la variable normalizada

$$Z = \frac{X - \mu}{\sigma} \qquad (8)$$

tiene la distribución normal estándar. Es decir, si X se distribuye como X ~

N (μ, σ), entonces

$$P\ (a \leq X \leq b) = P\left(\frac{a - \mu}{\sigma} \leq Z \leq \frac{b - \mu}{\sigma}\right) \qquad (9)$$

En esta expresión, Z sigue una distribución normal estándar. Esta característica inherente a la distribución normal nos concede la capacidad de trasladar el problema de probabilidad relacionado con X a uno que concierne a Z.

Ejemplo. El número de calorías de una ensalada del menú del almuerzo se distribuye normalmente con una media μ = 200 y una desviación típica σ = 5. Encuentre la probabilidad de que la ensalada que usted elija tenga una distribución normal estándar. Encuentre la probabilidad de que la ensalada seleccionada contenga

(a) Más de 208 calorías.

(b) Entre 190 y 200 calorías.

7. Distribuciones muestrales

7.1. Distribuciones muestrales

La inferencia estadística extrae conclusiones sobre la población a partir de los datos. Los datos se resumen mediante estadísticos como la media muestral y la desviación típica muestral. Cuando los datos se obtienen mediante muestreo aleatorio o experimentación aleatoria, un estadístico es una variable aleatoria que obedece a las leyes de la teoría de la probabilidad. El vínculo entre la probabilidad y los datos está formado por las distribuciones de muestreo de los estadísticos. Una

distribución muestral muestra cómo variaría una estadística en la producción repetida de datos.

Una distribución de muestral es una distribución de probabilidades que determina las probabilidades de los posibles valores de una estadística muestral (Edgington, 1969).

Cada estadística tiene una distribución muestral. Una distribución muestral es simplemente un tipo de distribución de probabilidad. A diferencia de las distribuciones estudiadas hasta ahora, una distribución muestral no se refiere a observaciones individuales, sino a los valores de la estadística calculados a partir de esas observaciones, muestra tras muestra.

La distribución muestral captura la variabilidad inherente al proceso de muestreo, que se origina al recolectar datos y emplear estadísticas muestrales para estimar los parámetros. En esencia, representa la distribución de probabilidad de una estadística particular, la cual se deriva a partir de tomar múltiples muestras de tamaño n y calcular el valor de la estadística en cada instancia. Usualmente, la forma de la distribución muestral se conoce de manera teórica. De este modo, adquirimos la capacidad de formular afirmaciones probabilísticas sobre el valor de dicha estadística en una muestra de tamaño fijo n.

7.2. Distribuciones muestrales de las medias muestrales

Dado que la media muestral se utiliza mucho, su distribución muestral merece una atención especial. En primer lugar, consideraremos la media y la desviación típica de la media muestral.

Al efectuar una selección de una muestra aleatoria simple de tamaño n a partir de la población, se procede a medir la variable X en cada sujeto incluido en la muestra. Los datos resultantes se componen de observaciones relacionadas a n variables aleatorias, denotadas como X1, X2, ... Xn. Cada Xi constituye una medición correspondiente a un individuo seleccionado de forma aleatoria de la población, lo que implica que Xi es en sí misma una variable aleatoria con una

distribución de probabilidad equivalente a la distribución poblacional de la variable X.

Cuando la población es sustancial en comparación con la muestra, se vuelve plausible considerar que X1, X2, ... Xn son variables aleatorias independientes que siguen la misma distribución de probabilidad. Este se convierte en nuestro modelo de probabilidad para las mediciones tomadas en cada individuo dentro de una muestra aleatoria simple.

La media muestral de una muestra aleatoria simple de tamaño n es

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Se nota que en este momento estamos empleando la notación $\bar{X}$ para representar la media muestral, con la finalidad de resaltar que $\bar{X}$...

Es una variable aleatoria. Una vez observados los valores de las variables aleatorias $X_1, X_2, \ldots X_n$, es decir, tenemos los valores $x_1, x_2, \ldots x_n$, podemos calcular la media muestral $\bar{X}$ de la forma habitual.

Si la variable poblacional X tiene una media poblacional μ, μ también es la media de cada observación X_i. Por lo tanto, por la regla de adición para medias de variables aleatorias,

$$\mu\bar{X} = E(\bar{X}) = E = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

$$= \frac{E(X_1 + X_2 + \cdots + X_n)}{n}$$

$$= \frac{E(X_1) + (EX_2) + \cdots + (EX_n)}{n}$$

$$= \frac{\mu X_1 + \mu X_2 + \cdots + \mu X_n}{n}$$

$$= \frac{\mu + \mu + \cdots + \mu}{n}$$

$$= \mu$$

Es decir, la media de $\bar{X}$ es la misma que la media poblacional μ de la variable X. Además, según la regla de adición de varianzas de variables aleatorias independientes, $\bar{X}$ tiene la varianza

$$\sigma_{\bar{x}}^2 = \frac{\sigma_{x_1}^2 + \sigma_{x_2}^2 + \cdots + \sigma_{x_n}^2}{n^2}$$

$$= \frac{\sigma^2 + \sigma^2 + \cdots + \sigma^2}{n^2}$$

$$= \frac{\sigma^2}{n}$$

y por tanto la desviación típica de $\bar{X}$ es

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

La desviación típica de $\bar{X}$ también se denomina error típico de $\bar{X}$.

Para una muestra aleatoria de tamaño n de una población que tiene media μ y desviación típica σ, la distribución de la media muestral $\bar{X}$ tiene media $\mu\bar{X}$ = μ y desviación típica es decir, error típico $\sigma\bar{X} = \frac{\sigma}{\sqrt{n}}$. (Akers, 2018).

La media y el error típico de $\bar{X}$ muestran que la media muestral $\bar{X}$ tiende a estar más cerca de la media poblacional μ para valores mayores de n, ya que la distribución de muestreo se vuelve menos uniforme en torno a μ. Esto concuerda con nuestra intuición de que las muestras más grandes proporcionan estimaciones más precisas de las características de la población.

Ejemplo. Considerar la siguiente distribución poblacional de la variable

X:

Valores de X	**2**	**3**	**4**

Frecuencias relativas de X	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

y que X_1 y X_2 sean variables aleatorias que siguen la distribución de probabilidad de la distribución poblacional de X.

(a) Verifica que la media poblacional y la varianza poblacional son

$$\mu=3, \qquad \sigma^2=\frac{2}{3}$$

(b) Construye la distribución de probabilidad de la media muestral $\bar{X}$.

(c) Calcule la media y la desviación típica de la media muestral $\bar{X}$.

Previo a esto, se han examinado las medidas centrales y la dispersión de la distribución de probabilidad correspondiente a la media muestral $\bar{X}$, sin embargo, aún no se ha abordado su configuración específica. La configuración de la distribución de $\bar{X}$ está intrínsecamente ligada a la estructura de la distribución de la población subyacente. Un caso particular de interés se da cuando la distribución de la población sigue una distribución normal.

Distribución de la media muestral. Supongamos que una variable X de una población se distribuye normalmente con media μ y desviación típica σ (Rayat, 2018)

Entonces, para muestras de tamaño n, la media muestral $\bar{X}$ también se distribuye normalmente y tiene media μ y desviación típica $\frac{\sigma}{\sqrt{n}}$. Es decir, si $X \sim N(\mu, \sigma)$, entonces $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$.

Ejemplo. Considere una población normal con media μ = 82 y desviación típica σ = 12.

(a) Si se selecciona una muestra aleatoria de tamaño 64, ¿cuál es la probabilidad de que la media muestral $\bar{X}$ esté comprendida entre 80,8 y 83,2?

(b) Con una muestra aleatoria de tamaño 100, ¿cuál es la probabilidad de que la media muestral $\bar{X}$ esté comprendida entre 80,8 y 83,2?

En el muestreo de una población no normal, la distribución de $\bar{X}$ depende de cuál sea la distribución poblacional de la variable X. Un resultado sorprendente, conocido como teorema del límite central, afirma que cuando el tamaño de la muestra n es grande, la distribución de probabilidad de la media muestral $\bar{X}$ es aproximadamente normal, independientemente de la forma de la distribución poblacional.

Teorema del límite central. Cualquiera que sea la distribución poblacional de la variable X, la distribución de probabilidad de la media muestral $\bar{X}$ es aproximadamente normal cuando n es grande. Es decir, cuando n es grande, entonces

$$\bar{X} = aproximadamente\ N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

En la práctica, la aproximación normal para $\bar{X}$ suele ser adecuada cuando n es superior a 30. El teorema del límite central nos permite utilizar cálculos de probabilidad normal para responder a preguntas sobre medias muestrales de muchas observaciones, incluso cuando la distribución de la población no es normal.

Ejemplo.

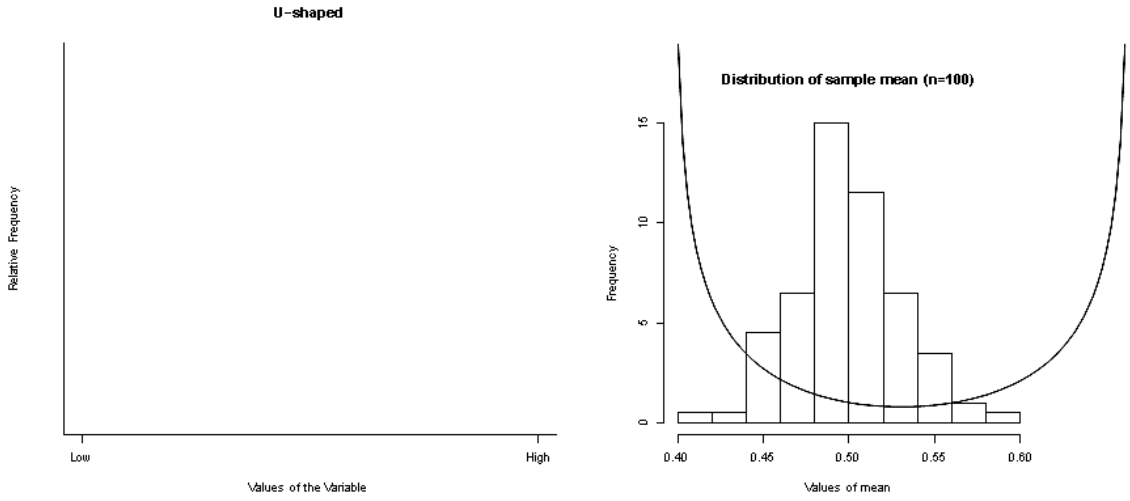


Figura 12: Distribuciones de frecuencias en forma de U y media muestral con n = 100

8. Estimación

En esta sección, se explorará el proceso de estimación, que implica el empleo de datos de muestra con el propósito de calcular parámetros poblacionales de los cuales carecemos de información directa. En el ámbito de la inferencia estadística, se utilizan datos de muestra para desarrollar dos categorías de estimadores de parámetros.

Una estimación puntual se caracteriza por ser un valor único, derivado de los datos, que representa la mejor conjetura respecto al parámetro poblacional desconocido. En contraste, una estimación de intervalo se compone de un rango de valores alrededor de la estimación puntual, el cual se considera como un intervalo de confianza que contiene el parámetro con un nivel de confianza específico.

8.1. Estimación puntual

El objetivo de la estimación puntual es calcular, a partir de los datos de la muestra, un único número que probablemente se aproxime al valor desconocido del parámetro poblacional. Se supone que la información disponible es una muestra aleatoria X_1, X_2, . . . X_n de tamaño n tomada de la población. El objetivo es formular un estadístico tal que su valor calculado a partir de los datos de la muestra refleje lo más fielmente posible el valor del parámetro poblacional.

Un estimador puntual de un parámetro poblacional desconocido es un estadístico que estima el valor de dicho parámetro. Una estimación puntual de un parámetro es el valor de un estadístico que se utiliza para estimar el parámetro (Steiger, 1990).

Por ejemplo, para estimar una media poblacional μ, quizá el estimador puntual más intuitivo sea la media muestral:

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Una vez que los valores observados x1, x2, . . . , xn de las variables aleatorias Xi, podemos calcular realmente el valor observado de la media muestral $\bar{X}$, que se denomina estimación puntual de μ.

Un buen estimador puntual de un parámetro es aquel cuya distribución muestral está centrada en el parámetro y tiene un error estándar lo más pequeño posible. Un estimador puntual se denomina insesgado si su distribución muestral se centra alrededor del parámetro en el sentido de que el parámetro es la media de la distribución.

Por ejemplo, la media de la distribución muestral de la media muestral $\bar{X}$ es igual a μ. Por lo tanto, $\bar{X}$ es un estimador insesgado de la media poblacional μ.

Una segunda propiedad preferible para un estimador es un error estándar pequeño. Se dice que un estimador cuyo error estándar es menor que el de otros posibles estimadores es eficiente. Un estimador eficiente es deseable porque, en promedio, se acerca más que otros estimadores al parámetro. Por ejemplo, puede demostrarse que, en una distribución normal, la media muestral es un estimador eficiente y, por tanto, tiene un error estándar menor que, por ejemplo, la mediana muestral.

8.1.1. Estimadores puntuales de la media y la desviación típica de la población

La media muestral $\bar{X}$ es el estimador puntual obvio de una media poblacional μ. De hecho, $\bar{X}$ es insesgado y relativamente eficiente para la mayoría de las distribuciones poblacionales. de la población. Es el estimador puntual, denotado por $\hat{\mu}$, utilizado en este texto:

$$\hat{\mu} = \bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Además, la desviación típica muestral s es la estimación puntual más popular de la desviación típica poblacional σ. Es decir

$$\hat{\mu} = s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

8.2. Intervalo de confianza

Para la estimación puntual, un único número se sitúa en primer plano aunque se adjunte un error estándar. En cambio, a menudo es más deseable producir un intervalo de valores que probablemente contenga el verdadero valor del parámetro desconocido.

Una **estimación de intervalo de confianza** de un parámetro consiste en un intervalo de números obtenidos a partir de una estimación puntual del parámetro junto con un porcentaje que especifica el grado de confianza que tenemos en que el parámetro se encuentra en el intervalo. El porcentaje de confianza se denomina **nivel de confianza**.

Un intervalo de confianza para un parámetro es un rango de números dentro del cual se cree que se encuentra el parámetro. La probabilidad de que el intervalo de confianza contenga el parámetro se denomina coeficiente de confianza. Se trata de un número elegido cercano a 1, como 0,95 o 0.99 (Tamhane, 1977).

8.2.1. Intervalo de confianza para μ cuando se conoce σ

En primer lugar, nos centraremos en la construcción de un intervalo de confianza para la media poblacional μ suponiendo que la variable poblacional X se distribuye normalmente y se conoce su desviación típica σ.

Recordemos que cuando la población se distribuye normalmente, la distribución de $\bar{X}$ también es normal, es decir, $\bar{X} \sim$ N(μ, $\frac{\sigma}{\sqrt{n}}$). La tabla normal muestra que la probabilidad es 0,95 de que una variable aleatoria normal se encuentre dentro de 1,96 desviaciones típicas de su media. Para $\bar{X}$, tenemos entonces

$$P\left(\mu - 1.96\frac{\sigma}{\sqrt{n}} < \bar{x} < \mu + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95.$$

Ahora la relación

$$\mu - 1.96\frac{\sigma}{\sqrt{n}} < \bar{x}\ es\ igual\ a\ \mu < \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$$

Y

$$\bar{x} < \mu + 1.96\frac{\sigma}{\sqrt{n}} \quad es\ igual\ a \quad \bar{x} - 1.96\frac{\sigma}{\sqrt{n}} < \mu$$

Por lo tanto, la declaración de probabilidad

$$P\left(\mu - 1.96\frac{\sigma}{\sqrt{n}} < \bar{x} < \mu + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

También puede ser expresada como

$$P\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

Esta segunda forma nos dice que el intervalo aleatorio

$$\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

Incluirá el parámetro desconocido con una probabilidad de 0,95. Dado que se supone que σ es conocido, tanto el punto final superior como el inferior pueden calcularse en cuanto se disponga de los datos de la muestra. Por lo tanto, decimos que el intervalo.

$$\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

Un **intervalo de confianza del 95%** para μ cuando la variable poblacional X se distribuye normalmente y se conoce σ.

No necesitamos considerar siempre intervalos de confianza a la elección de un nivel de confianza del 95%. Es posible que deseemos especificar un nivel de probabilidad diferente. Denotamos esta probabilidad por 1 - α y hablamos de un nivel de confianza del 100(1 - α) %. El único cambio consiste en sustituir 1,96 por zα/2, donde $z_{\alpha/2}$ es un número tal que P ($-z_{\alpha/2} < Z < z_{\alpha/2}$) = 1 - α cuando Z ~ N(0, 1).

Cuando la variable poblacional X se distribuye normalmente y σ es conocida, un intervalo de confianza del 100(1 - α) % para μ viene dado por

$$\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Ejemplo. Dada una muestra aleatoria de 25 observaciones de una población normal para la que μ es desconocida y σ = 8, se calcula que la media muestral es $\bar{X}$ = 42,7. Construya intervalos de confianza del 95% y del 99% para μ. Construya intervalos de confianza del 95% y del 99% para μ (Foster, 2014).

8.2.2. Intervalo de confianza de muestra grande para μ

Consideremos ahora una situación más realista en la que se desconoce la desviación típica de la población σ. Requerimos que el tamaño de la muestra n sea grande y, por tanto, el teorema del límite central nos dice que la declaración de probabilidad.

$$\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = 1 - \alpha$$

se cumple aproximadamente, sea cual sea la distribución de población subyacente. Además, como n es grande, sustituir $\frac{\sigma}{\sqrt{n}}$ por es estimador $\frac{s}{\sqrt{n}}$ no afecta apreciablemente el enunciado de probabilidad anterior. Por lo tanto, tenemos lo siguiente hecho clave.

Cuando n es grande y σ se desconoce, un intervalo de confianza del 100(1 _ α)% para μ viene dado por

$$\left(\bar{X} - z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma s}{\sqrt{n}}\right)$$

donde s es la desviación típica de la muestra.

8.2.3. Intervalo de confianza de muestra pequeña para μ

Cuando la variable poblacional X se distribuye normalmente con una media μ y una desviación estándar σ, entonces la variable estandarizada.

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Tiene la distribución normal estándar Z ~ N(0, 1). Sin embargo, si consideramos la relación

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

La variable aleatoria t tiene la **distribución t de Student con n - 1 grados de libertad**.

Sea $t_{\alpha/2}$ un número tal que P ($-t_{\alpha/2} < t < t_{\alpha/2}$) = 1 - α cuando t tiene la distribución t de Student con n - 1 grados de libertad (véase la tabla t). Por lo tanto, tenemos las siguientes declaraciones de probabilidad equivalentes:

$$P\left(-t_{\frac{\alpha}{2}} < t < t_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

$$P\left(-t_{\frac{\alpha}{2}} < \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} < t_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

$$P\left(\bar{X} - t_{\frac{\alpha}{2}} < \frac{s}{\sqrt{n}} < \mu < \bar{X} + t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

La última expresión nos da el siguiente intervalo de confianza de muestra pequeña para μ.

Cuando la variable poblacional X se distribuye normalmente y σ es desconocida, un intervalo de confianza del 100(1 - α)% para μ viene dado por

$$\bar{X} - t_{\frac{\alpha}{2}} < \frac{s}{\sqrt{n}}, \bar{X} + t_{\frac{\alpha}{2}} < \frac{s}{\sqrt{n}}$$

donde $t_{\alpha/2}$ es el punto superior α/2 de la distribución t de Student con n - 1 grados de libertad

Ejemplo. Considere una muestra aleatoria de una población normal para la que μ y σ son desconocidas:

10, 7, 15, 9, 10, 14, 9, 9, 12, 7.

Construya intervalos de confianza del 95% y del 99% para μ.

Ejemplo. Suponga que los tiempos de llegada en la carrera de bicicletas siguen la distribución normal con μ y σ desconocidas. Considere que 7 participantes en la carrera de bicicletas tuvieron los siguientes tiempos finales en minutos:

28, 22, 26, 29, 21, 23, 24.

Construye un intervalo de confianza del 90% para μ.

Analizar -> Estadística descriptiva -> Explorar

Tabla 12: Intervalo de confianza del 90% para μ de los tiempos de llegada en la carrera ciclista

Descriptivos

Estadístico			Estadístico	Error St
bike7	Media		24.7143	1.14879
	90% Intervalo de Confianza para-Media	Límite inferior	22.4820	
		Límite inferior	26.9466	

9. Comprobación de hipótesis

9.1. Hipótesis

Un objetivo habitual en muchos estudios es comprobar si los datos concuerdan con determinadas predicciones. Estas predicciones son hipótesis sobre variables medidas en el estudio.

Una **hipótesis** es una afirmación sobre alguna característica de una variable o un conjunto de variables (MacCorquodale & Meehl, 1948).

Las hipótesis surgen de la teoría que impulsa la investigación. Cuando una hipótesis se refiere a características de una población, como los parámetros poblacionales, se pueden utilizar métodos estadísticos con datos de muestra para comprobar su validez.

Una prueba de significación es una forma de probar estadísticamente una hipótesis comparando los datos con los valores predichos por la hipótesis. Los datos que se alejan de los valores predichos proporcionan pruebas en contra de la hipótesis. Todas las pruebas de significación constan de cinco elementos: supuestos, hipótesis, estadístico de prueba, valor p y conclusión.

Todas las pruebas de significación requieren ciertos supuestos para que sean válidas. Estos supuestos se refieren, por ejemplo, al tipo de datos, la forma de distribución de la población, el método de muestreo y el tamaño de la muestra.

Una prueba de significación considera dos hipótesis sobre el valor de un parámetro de la población: la hipótesis nula y la hipótesis alternativa.

La **hipótesis nula** H_0 es la hipótesis que se comprueba directamente. Suele ser una afirmación de que el parámetro tiene un valor que corresponde, en cierto sentido, a ningún efecto. La **hipótesis alternativa** H_a es una hipótesis que contradice la hipótesis nula. Esta hipótesis afirma que el parámetro cae en algún conjunto de valores alternativo al que especifica la hipótesis nula (Masson, 2011).

Una prueba de significación analiza la fuerza de las pruebas de la muestra contra la hipótesis nula. La prueba se realiza para investigar si los datos contradicen la hipótesis nula, lo que sugiere que la hipótesis alternativa es cierta. La hipótesis alternativa se considera aceptable si los datos de la muestra no concuerdan con la hipótesis nula. Es decir, se apoya la hipótesis alternativa si la hipótesis nula parece ser incorrecta. Las hipótesis se formulan antes de recoger o analizar los datos.

El estadístico de prueba es un estadístico calculado a partir de los datos de la muestra para probar la hipótesis nula. Este estadístico suele consistir en una estimación puntual del parámetro al que se refieren las hipótesis.

Utilizando la distribución muestral del estadístico de prueba, calculamos la probabilidad de que se produzcan valores del estadístico como el observado si la hipótesis nula fuera cierta. Esto proporciona una medida de lo inusual que es el valor observado de la estadística de prueba en comparación con lo que predice H_0.

Es decir, consideramos el conjunto de posibles valores del estadístico de prueba que proporcionan al menos tanta evidencia contra la hipótesis nula como el estadístico de prueba observado. Este conjunto se forma con referencia a la hipótesis alternativa: los valores que proporcionan pruebas más sólidas contra la hipótesis nula son los que proporcionan pruebas más sólidas a favor de la hipótesis alternativa. El valor p es la probabilidad, si H_0 fuera cierta, de que el estadístico de prueba se encontrara en este conjunto de valores.

El **valor p** es la probabilidad, cuando H_0 es verdadera, de un valor de la estadística de prueba al menos tan contradictorio con H_0 como el valor realmente observado. Cuanto menor sea el valor p, más fuertemente contradicen los datos a H_0 (Gibbons & Pratt, 1975).

El valor p resume la evidencia de los datos sobre la hipótesis nula. Un valor p de moderado a grande significa que los datos son coherentes con H_0. Por ejemplo, un valor p como 0,3 o 0,8 indica que los datos observados no serían inusuales si H0 fuera cierta. Pero un valor p como 0,001 significa que esos datos serían muy poco probables, si H_0 fuera cierta. Esto proporciona una prueba contundente contra H_0.

El valor p es el resultado principal de una prueba de significación. Un observador de los resultados de la prueba puede juzgar el alcance de las pruebas contra H_0. A veces es necesario tomar una decisión formal sobre la validez de H_0. Si el valor p es suficientemente pequeño, se rechaza H_0 y se acepta H_a. Sin embargo, la conclusión siempre debe incluir una interpretación de lo que el valor p o la decisión sobre H_0 nos dice sobre la pregunta original que motivó la prueba. La mayoría de los estudios requieren un valor p muy pequeño, como $p \leq 0,05$, antes de concluir que los datos contradicen suficientemente H0 como para rechazarla. En tales casos, se dice que los resultados son significativos al nivel 0,05. Esto significa que si la hipótesis nula fuera cierta, la probabilidad de obtener resultados tan extremos como en los datos de la muestra no sería superior al 5%.

9.2. Prueba de significación para una media poblacional μ

En correspondencia con los intervalos de confianza para μ, presentamos ahora tres pruebas de significación diferentes sobre la media poblacional μ. Las hipótesis son todas iguales en estas pruebas, pero el estadístico de prueba utilizado varía en función de las suposiciones que hayamos hecho.

9.2.1. Prueba de significación para μ cuando se conoce σ

9.2.1.1. Supuestos

Sea una variable poblacional X normalmente distribuida con la media μ desconocida y la desviación típica σ conocida.

9.2.1.2. Hipótesis

Se considera que la hipótesis nula tiene la forma

$$H_0 : \mu = \mu_0$$

donde μ_0 es un número determinado. En otras palabras, el valor hipotético de μ en H_0 es un valor único.

La hipótesis alternativa se refiere a valores de los parámetros alternativos al de la hipótesis nula. La forma más común de hipótesis alternativa es

$$H_a : \mu \neq \mu_0$$

Esta hipótesis alternativa se denomina bifronte, ya que incluye valores que caen tanto por debajo como por encima del valor μ_0 que figura en H_0

9.2.1.3. Estadística de prueba

La media muestral $\bar{X}$ estima la media poblacional μ. Si $H_0 : \mu = \mu_0$ es cierta, entonces el centro de la distribución muestral de $\bar{X}$ debe ser el número μ_0. La prueba sobre H_0 es la distancia del valor muestral $\bar{X}$ del valor de la hipótesis nula μ_0, relativa al error estándar. Un valor observado $\bar{X}$ que cae muy lejos en la cola de esta

distribución muestral de $\bar{X}$ pone en duda sobre la validez de H_0, porque sería improbable observar un valor $\bar{X}$ muy alejado de μ_0 si realmente $\mu = \mu_0$.

El estadístico de prueba es el estadístico Z

$$Z = \frac{\bar{X} - \mu_o}{\sigma/\sqrt{n}}$$

Cuando H_0 es verdadera, la distribución muestral del estadístico Z es normal estándar normal estándar, $Z \sim N(0, 1)$. Cuanto más se aleje el valor observado $\bar{X}$ de μ_0, mayor es el valor absoluto del valor observado z del estadístico Z.

Por lo tanto, cuanto mayor sea el valor de |z|, más fuerte será la evidencia contra H_0.

9.2.1.4. Valor p

Calcular el p-valor suponiendo que H_0 es cierta. Es decir, damos el beneficio de la duda a la hipótesis nula, analizando la probabilidad que tendrían los datos observados si esa hipótesis fuera cierta. El valor p es la probabilidad de que el estadístico Z sea al menos tan grande en valor absoluto como el valor observado z del estadístico Z. Esto significa que p es la probabilidad de que $\bar{X}$ tenga un valor al menos tan alejado de μ_0 en cualquier dirección como el valor observado $\bar{X}$. Es decir, sea z el valor observado del estadístico Z:

$$z = \frac{\bar{x} - \mu_o}{\sigma/\sqrt{n}}$$

Entonces, el valor p es la probabilidad

$$2 \cdot P\,(Z \geq |z|) = p,$$

donde $Z \sim N(0, 1)$.

9.2.1.5. Conclusión

En el cierre de este estudio, resulta fundamental proporcionar el valor p, ya que este dato permite a otros evaluadores ponderar la robustez de la evidencia

presentada. Cuanto menor sea el valor p, más sólida será la evidencia en contra de la hipótesis nula (H0) y a favor de la hipótesis alternativa (Ha). En el caso de un valor p extremadamente bajo, como 0,01 o menos, se justifica un rechazo enfático de la hipótesis nula H0 en favor de Ha. Si el valor p cae en el rango de 0,01 a 0,05, la evidencia respalda igualmente el rechazo de la hipótesis nula H0 en favor de Ha. Por último, cuando el valor p supera el umbral de 0,05, concluimos que existe suficiente evidencia para aceptar la hipótesis nula H0.

Ejemplo. Dada una muestra aleatoria de 25 observaciones de una población normal para la que μ es desconocida y σ = 8, la media muestral se calcula $\bar{x}$ = 42,7. Pruebe la hipótesis $H_0 : \mu = \mu_0 = 35$ para μ contra hipótesis alternativa $Ha : \mu \neq \mu_0$.

9.2.2. Prueba de significación para μ en muestras grandes

Ahora se supone que el tamaño de la muestra n es grande (n 50) y que σ es desconocida. Las hipótesis son similares a las anteriores:

$$H_0 : \mu = \mu_0 \quad \text{y} \quad Ha : \mu \neq \mu_0.$$

La estadística de la prueba en el caso de la muestra grande es la siguiente estadística Z

$$Z = \frac{\bar{X} - \mu_o}{s/\sqrt{n}}$$

Donde s es la desviación típica de la muestra. Debido al teorema del límite central, el estadístico Z anterior sigue ahora aproximadamente la distribución normal estándar si H_0 es cierta, véase la correspondencia con el intervalo de confianza de la muestra grande para μ. Por lo tanto, el valor p es de nuevo la probabilidad

$$2 \cdot P\,(Z \geq |z|) = p,$$

donde Z aproximadamente N(0, 1), y las conclusiones pueden hacerse de forma similar a la anterior.

9.2.3. Prueba de significación para μ en muestras pequeñas

En una situación de muestra pequeña, suponemos que la población se distribuye normalmente con una media μ y una desviación típica σ desconocidas. De nuevo, las hipótesis se formulan como:

$$H_0 : \mu = \mu_0 \qquad \text{y} \qquad H_a : \mu \neq \mu_0.$$

La estadística de la prueba se basa ahora en la distribución t de Student. El estadístico t

$$t = \frac{\bar{X} - \mu_o}{s/\sqrt{n}}$$

tiene la distribución t de Student con n - 1 grados de libertad si H_o es verdadera. Sea t el valor observado del estadístico t. Entonces el valor p es la probabilidad

$$2 \cdot P\,(t \geq |t\ \ |) = p,$$

Las conclusiones vuelven a ser similares a las de los casos anteriores.

Ejemplo. Considere una muestra aleatoria de una población normal para la que μ y σ son desconocidas:

10, 7, 15, 9, 10, 14, 9, 9, 12, 7.

Pruebe las hipótesis $H_0 : \mu = \mu_0 = 7$ y $H_0 : \mu = \mu_0 = 10$ para μ contra la hipótesis alternativa de dos caras $Ha : \mu \neq \mu_0$.

Ejemplo. Supongamos que los tiempos de llegada en la carrera de bicicletas siguen la distribución normal con μ y σ desconocidos. Considere que 7 participantes en la carrera de bicicletas tuvieron los siguientes tiempos finales en minutos:

28, 22, 26, 29, 21, 23, 24.

Pruebe la hipótesis $H_0 : \mu = \mu_0 = 28$ para μ contra la hipótesis alternativa de dos caras $Ha : \mu \neq \mu_0$.

Analizar -> Comparar medias -> Prueba T de una muestra

Cuadro 13: Prueba t para $H_0 : \mu = \mu_0 = 28$ frente a $Ha : \mu \neq \mu 0$.

Prueba de una muestra.

	Valor de Prueba = 28					
	t	df	Sig. (2-tailed)	Diferencia Media	95% Intervalo de Confianza de la Diferencia	
					Baja	Alta
bike7	-2.860	6	.029	-3.28571	-6.0967	-.4747

10. Resumen de datos bivariantes

Hasta este punto, se ha abordado de manera exhaustiva la descripción y la inferencia estadística concernientes a una sola variable. Sin embargo, en la mayoría de los estudios estadísticos, entra en juego más de una variable. En este segmento del análisis, se explora con detenimiento la relación entre dos variables específicas. Estos datos bivariantes consisten en los valores observados de dichas variables y pueden manifestar tanto una naturaleza cualitativa como cuantitativa. Es decir, es factible que ambas variables sean cualitativas o cuantitativas, y también puede darse el caso en que una sea cualitativa mientras que la otra sea cuantitativa. A lo largo de esta sección, examinaremos exhaustivamente todas estas posibilidades.

10.1. Variables cualitativas

Los datos cualitativos bivariantes resultan de los valores observados de las dos variables cualitativas. En la sección 3.1, en un caso de una sola variable cualitativa, la distribución de frecuencias de la variable se presentó mediante una tabla de frecuencias. En el caso de dos variables cualitativas, la distribución conjunta de las variables puede resumirse en forma de tabla de frecuencias bidireccional.

En una tabla de frecuencias bidireccional, las clases (o categorías) de una variable (denominada variable de fila) se marcan a lo largo del margen izquierdo, las de la otra (denominada variable de columna) a lo largo del margen superior, y los recuentos de frecuencia se registran en las celdas. El resumen de datos bivariantes mediante una tabla de frecuencias bidireccional se denomina tabulación

cruzada o clasificación cruzada de los valores observados. En terminología estadística, las tablas de frecuencias bidireccionales también se denominan tablas de continencia.

La tabla de frecuencias más sencilla es la tabla de frecuencias 2 X 2, en la que cada variable tiene sólo dos clases. De forma similar, puede haber tablas 2 X 3, 3 X 3, etc., en las que el primer número indica el número de filas que tiene la tabla y el segundo el número de columnas.

Ejemplo. Los grupos sanguíneos y el sexo de 40 personas son los siguientes:

(O,Masculino),(O,Femenino),(A,Femenino),(B,Masculino),(A,Femenino),(O,Femenino),(A,Masculino),(A,Masculino),(A,Femenino),(O,Masculino),(B,Masculino),(O,Masculino),(A,Femenino),(O,Masculino),(A,Femenino),(A,Femenino),(A,Femenino),(A,Masculino),(A,Masculino),(AB,Femenino),(A,Femenino),(B,Femenino),(A,Masculino),(A,Femenino),(O,Masculino),(A,Femenino),(O,Masculino),(O,Femenino),(A,Femenino),(A,Masculino),(A,Masculino),(O,Masculino),(A,Masculino),(O,Femenino),(O,Femenino),(AB,Masculino).

Resumir los datos en una tabla de frecuencias bidireccional utilizando SPSS:

Analizar -> Estadísticas Descriptivas -> Tablas Cruzadas,

Analizar -> Tablas Personalizadas -> Tablas de Frecuencias

Tabla 14: Distribución de frecuencias de los grupos sanguíneos y el sexo

Tabulación cruzada de sangre y género.

		GENERO	
		Masculino	**Femenino**
SANGRE	O	11	5
	A	8	10
	B	2	2
	AB	1	1

Suponer que una variable cualitativa tiene i clases y la otra j clases. Entonces, la distribución conjunta de las dos variables puede resumirse mediante una tabla de frecuencias i × j. Si el tamaño de la muestra es n y la celda ij tiene una frecuencia f_{ij}, la frecuencia relativa de la celda ij es

$$Frecuencia\ relativa\ de\ una\ célula\ ij = \frac{Frecuencia\ en\ la\ celda\ ij}{Número\ total\ de\ observaciones} = \frac{f_{ij}}{n}$$

Los porcentajes son también frecuencias relativas multiplicadas por 100.

A partir de la tabla de frecuencias bidireccional, calcular los totales (marginales) de fila y columna. Para la fila i, el total de fila f_i es

fi- = fi1 + fi2 + fi3 + - - - + fij,

y de forma similar para la j-ésima columna, el total de la columna f_j es

$$f_{.j} = f_{1j} + f_{2j} + f_{3j} + \cdot\cdot\cdot + f_{ij}.$$

Tanto los totales de fila como los de columna tienen una propiedad obvia; $n = \sum_{k=1}^{n} f_k = \sum_{k=1}^{n} f_{\cdot k.}$

A partir de los totales de filas y columnas, calcular las frecuencias relativas por filas y las frecuencias relativas por columnas. Para la celda ij, la frecuencia relativa por fila i es

$$\text{frecuencia relativa por fila de una celda ij} = \frac{f_{ij}}{f_{i.}}$$

y la frecuencia relativa por columna j es

$$\text{frecuencia relativa por columna de una celda ij} = \frac{f_{ij}}{f_{i.}}$$

Las frecuencias relativas por fila, representadas por i, proporcionan información acerca de la distribución condicional de la variable en la columna correspondiente al valor i de la variable en la fila. En otras palabras, estas frecuencias relativas por fila ofrecen una respuesta a la interrogante sobre cuál es la distribución de la variable en la columna cuando el valor observado de la variable en la fila es i. De

manera análoga, las frecuencias relativas por columna, representadas por j, ofrecen una visión de la distribución condicional de la variable en la fila para el valor j de la variable en la columna.

Adicionalmente, podemos definir los totales relativos de fila con respecto al total general y los totales relativos de columna con respecto al total general, los cuales corresponden al i-ésimo total de fila y al j-ésimo total de columna, respectivamente.

$$\frac{f_{i.}}{n}, \qquad \frac{f_{.j}}{n}$$

Respectivamente

Ejemplo. Continuemos con el ejemplo del grupo sanguíneo y el sexo:

Cuadro 15: Porcentajes de filas de los grupos sanguíneos y el sexo

Tabulación cruzada de sangre y género.

			GENERO		Total
			Masculino	**Femenino**	
SANGRE	O	Count	11	5	**16**
		% within SANGRE	68.8%	31.3%	**100.0%**
	A	Count	8	10	**18**
		% within SANGRE	44.4%	55.6%	**100.0%**
	B	Count	2	2	**4**
		% within SANGRE	50.0%	50.0%	**100.0%**
	AB	Count	1	1	**2**
		% within SANGRE	50.0%	50.0%	**100.0%**
Total		Count	22	18	**40**
		% within BLOOD	**55.0%**	**45.0%**	**100.0%**

Cuadro 16: Porcentajes en columna de los grupos sanguíneos y el sexo

Tabulación cruzada de sangre y género.

			GENERO		**Total**
			Masculino	**Femenino**	
SANGRE	O	Count	11	5	**16**
		% within GENERO	50.0%	27.8%	**40.0%**
	A	Count	8	10	**18**
		% within GENERO	36.4%	55.6%	**45.0%**
	B	Count	2	2	**4**

		% within GENERO	9.1%	11.1%	**10.0%**
	AB	Count	1	1	**2**
		% within GENERO	4.5%	5.6%	**5.0%**
Total		Count	22	18	**40**
		% within GENERO	**100.0%**	**100.0%**	**100.0%**

En los ejemplos previos, hemos computado los porcentajes tanto por fila como por columna, es decir, las distribuciones condicionales de la variable en la columna para un valor específico de la variable en la fila, y viceversa. La cuestión que se plantea ahora es: ¿cuál es la razón detrás de este cálculo de las diversas distribuciones condicionales, y cuáles de estas distribuciones condicionales debemos utilizar en el análisis?

El propósito fundamental de calcular estas distribuciones condicionales radica en la identificación de una posible asociación entre las variables en las filas y columnas del conjunto de datos. Si los porcentajes por fila difieren significativamente entre sí, entonces se refleja una variabilidad en las distribuciones condicionales de la variable en la columna en cada fila. Este escenario nos sugiere la existencia de una asociación entre las variables, indicando que el valor de la variable en la fila incide en el valor de la variable en la columna. De manera análoga, si los porcentajes por columna presentan claras disparidades entre las distintas columnas, las distribuciones condicionales de la variable en la fila varían en cada columna. Esto también sugiere la presencia de una asociación entre las variables, señalando que el valor de la variable en la columna afecta al valor de la variable en la fila.

La dirección de la asociación depende de las formas de las distribuciones condicionales. Si los porcentajes de fila (o los porcentajes de columna) son bastante similares de fila a fila (o de columna a columna), entonces no hay asociación entre las variables y decimos que las variables son independientes.

La utilización de los porcentajes de filas y columnas para la inferencia de una posible asociación depende de cuál sea la variable de respuesta y cuál la variable explicativa. Demos primero una definición más general para la variable de respuesta y la variable explicativa.

Una **variable de respuesta** mide el resultado de un estudio. Una **variable explicativa** intenta explicar los resultados observados.

En muchos casos ni siquiera es posible identificar cuál es la variable de respuesta y cuál la variable explicativa. En ese caso, utilizar porcentajes de fila o de columna para averiguar si existe asociación entre las variables o no. Si ahora descubrimos que hay asociación entre las variables, no decir que una variable esté causando cambios en otra variable, es decir, la asociación no implica causalidad.

En otra perspectiva, cuando identificamos que la variable en la fila actúa como la variable de respuesta y la variable en la columna desempeña el papel de variable explicativa, nos enfrentamos al escenario en el cual debemos comparar las distribuciones condicionales de la variable en la fila para las distintas categorías de la variable en la columna. Este proceso de comparación se lleva a cabo con el objetivo de discernir la existencia de asociación y, posiblemente, de causalidad entre las variables en consideración.

De manera correspondiente, cuando determinamos que la variable en la columna asume la función de variable de respuesta y la variable en la fila actúa como variable explicativa, se hace necesario comparar las distribuciones condicionales de la variable en la columna. Sin embargo, es vital destacar que, especialmente en el contexto de dos variables cualitativas, debemos ejercer un riguroso discernimiento en cuanto a si la asociación entre las variables conlleva necesariamente una relación causal.

Los datos cualitativos bivariantes se presentan mejor gráficamente mediante gráficos de barras agrupadas o apiladas. También pueden ser informativos los gráficos circulares divididos por diferentes categorías de una variable (denominados gráficos circulares trazados).

Ejemplo. ... continúe con el ejemplo del grupo sanguíneo y el sexo:

Gráficos -> Interactivos -> Barra,

Gráficos -> Interactivos -> Tarta -> Gráfico

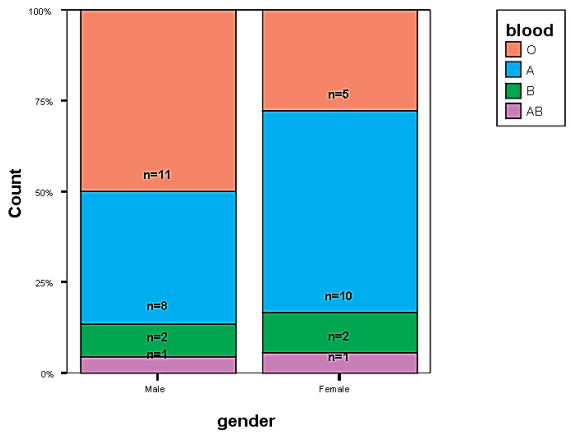


Figura 13: Gráfico de barras apiladas para el grupo sanguíneo y el sexo

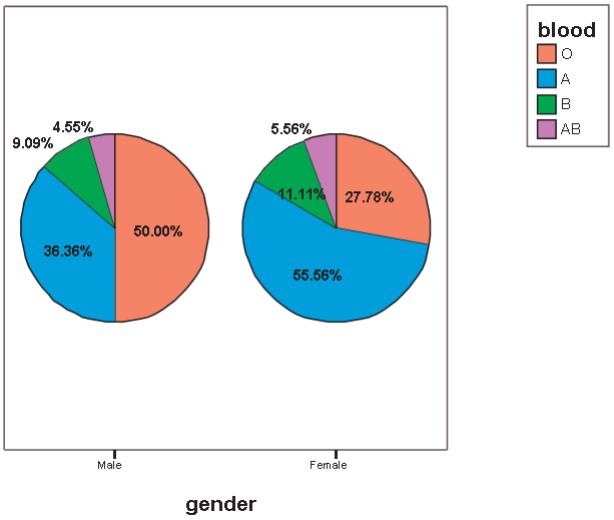


Figura 14: Gráfico circular del grupo sanguíneo y el sexo

10.2. Variable cualitativa y variable cuantitativa

Cuando se está en la situación en la que una de las variables es de naturaleza cualitativa y la otra es cuantitativa, es recomendable emplear una tabla de frecuencias bidireccional como medio para investigar la posible existencia de asociación entre ambas variables. No obstante, en este caso, es esencial realizar

previamente una agrupación de la variable cuantitativa en clases, tal como se detalló en el apartado 3.2. Posteriormente, se procede a presentar la distribución conjunta de las variables en la mencionada tabla de frecuencias bidireccional.

La inferencia subsecuente se sustenta en las distribuciones condicionales derivadas de la tabla de frecuencias bidireccional. Es importante resaltar que, especialmente cuando resulta evidente que una de las variables actúa como variable de respuesta y la otra como variable explicativa, la tabla de frecuencias bidireccional se convierte en una herramienta crucial para evaluar si existe una asociación significativa entre las variables en estudio.

EJEMPLO 10.4. Precios y tipos de perros calientes:

Tabla 17: Porcentajes en columnas de precios y tipos de perros calientes

Precios y tipos de perros calientes.

			Tipo			Total
			Ternera	Carne	Aves	
Precios	-.08	Count	1	3	16	**20**
		% within Tipo	5.0%	17.6%	94.1%	**37.0%**
	0.081 - 0.14	Count	10	12	1	**23**
		% within Tipo	50.0%	70.6%	5.9%	**42.6%**
	0.141 -	Count	9	2		**11**
		% within Tipo	45.0%	11.8%		**20.4%**
Total		**Count**	**20**	**17**	**17**	**54**
		% within Tipo	**100.0%**	**100.0%**	**100.0%**	**100.0%**

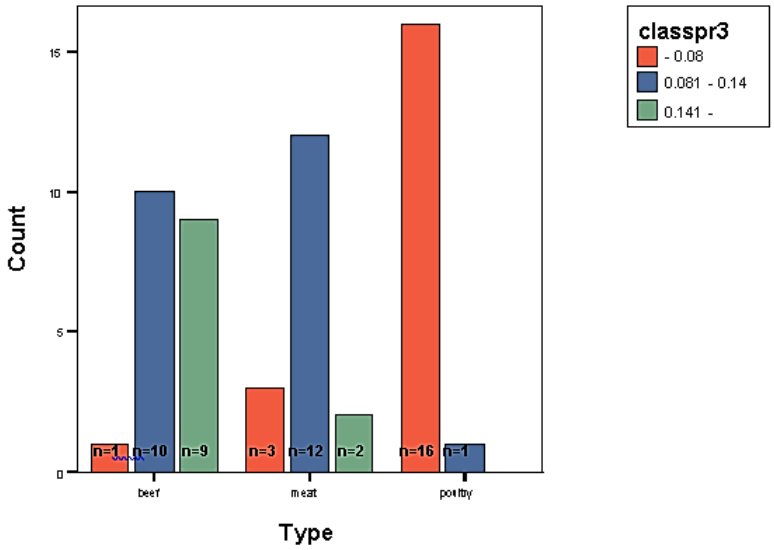


Figura 15: Gráfico de barras agrupadas para precios y tipos de perros calientes

Normalmente, en caso de que una variable sea cualitativa y la otra cuantitativa, nos interesa saber cómo se distribuye la variable cuantitativa en las distintas clases de la variable cualitativa, es decir, cuál es la distribución condicional de la variable cuantitativa para un valor específico de la variable cualitativa y si estas distribuciones condicionales varían en cada clase de la variable cualitativa. Al analizar las distribuciones condicionales de este modo, suponemos que la variable cuantitativa es la variable de respuesta y la cualitativa la variable explicativa.

Ejemplo. Se pesaron 198 recién nacidos y se recogió información sobre el gender y el peso:

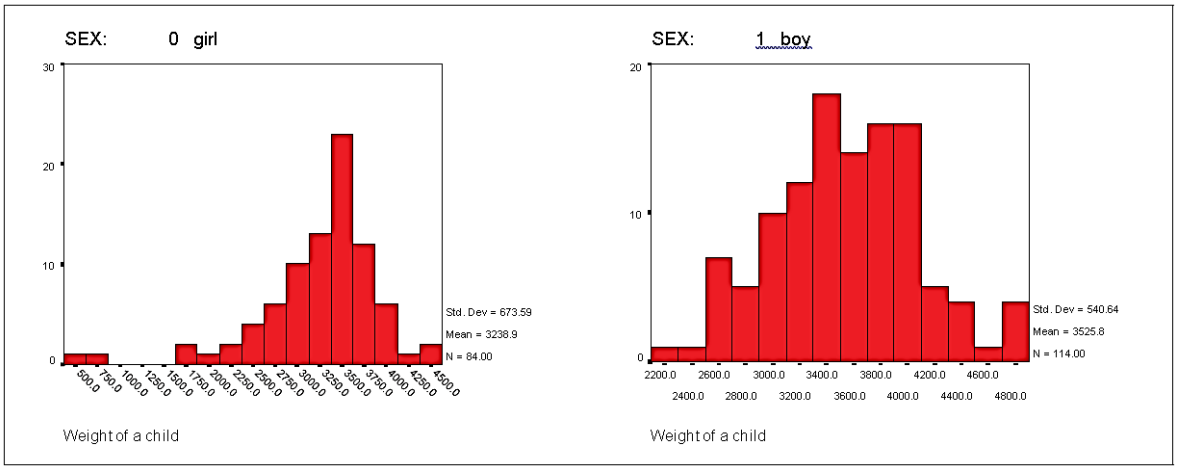
Gender	Weight
boy	4870
girl	3650
girl	3650
girl	3650
girl	2650
girl	3100
boy	3480
girl	3600
boy	4870

. | .

Los histogramas muestran las distribuciones condicionales del peso:

Datos -> Dividir archivo -> (Comparar grupos) y luego

Gráficos -> Histograma



Cuando la variable dependiente es cuantitativa y la variable independiente es cualitativa, se requiere la comparación de las distribuciones condicionales de la variable cuantitativa mediante medidas específicas que caractericen dichas distribuciones. Como se ha abordado previamente en secciones anteriores, para caracterizar la distribución de la variable en cuestión, es posible utilizar medidas de tendencia central y medidas de dispersión.

En el mismo sentido, las distribuciones condicionales pueden caracterizarse calculando medidas condicionales de tendencia central y medidas condicionales de dispersión a partir de los valores observados de la variable dependiente, considerando un valor específico de la variable independiente. Concretamente, estas medidas condicionales de tendencia central son conocidas como medias muestrales condicionales y medianas muestrales condicionales. Del mismo modo, las medidas condicionales de dispersión pueden denominarse rango muestral

condicional, rango intercuartílico muestral condicional y desviación muestral condicional.

Estas medidas condicionales de tendencia central y dispersión adquieren un papel fundamental en la evaluación de la existencia de asociación, e incluso causalidad, entre las variables en estudio. Por ejemplo, si se observa una marcada diferencia en los valores de las medias condicionales de la variable cuantitativa en cada categoría de la variable cualitativa, se puede inferir la presencia de una asociación entre estas variables. En el caso de distribuciones condicionales simétricas, es esencial calcular y comparar las medias condicionales y las desviaciones condicionales. En contraste, cuando las distribuciones condicionales exhiben asimetría, se requiere el empleo de las medianas condicionales y los intercuartiles condicionales para el análisis pertinente.

Ejemplo. Cálculo de medias condicionales y desviaciones estándar condicionales para el peso de 198 recién nacidos en función del sexo en SPSS:

Analizar -> Comparar medias -> Medias

Tabla 18: Medias y desviaciones estándar condicionales para el peso de los recién nacidos

Grupo Medias y desviaciones estándar.

Género del niño	Media	N	Desviación St
Chica	3238.93	84	**673.591**
Chico	3525.78	114	**540.638**
Total	3404.09	198	615.648

Cálculo de otras medidas de centro y variación para el peso de 198 recién nacidos en función del sexo en SPSS:

Analizar -> Estadística descriptiva -> Explorar

Cuadro 19: Otras medidas de centro y variación para el peso de los recién nacidos

Género	Estadística	Valor	Error Estándar

Chica	Media	3238.93	73.495
	Intervalo de Confianza al 95%	3092.75	-
	Media (Trimmed al 5%)	3385.11	-
	Mediana	3289.74	-
	Varianza	3400.00	-
	Desviación Estándar	453725.3	-
	Mínimo	673.591	-
	Máximo	510	-
	Rango	4550	-
	Rango Intercuartil	4040	-
	Asimetría (Skewness)	572.50	-1.565
	Kurtosis	4.155	0.263
Chico	Media	3525.78	50.635
	Intervalo de Confianza al 95%	3425.46	-
	Media (Trimmed al 5%)	3626.10	-
	Mediana	3517.86	-
	Varianza	3500.00	-
	Desviación Estándar	292289.1	-
	Mínimo	540.638	-
	Máximo	2270	-
	Rango	4870	-
	Rango Intercuartil	2600	-
	Asimetría (Skewness)	735.00	0.134
	Kurtosis	-	0.226

Gráficamente, la mejor forma de ilustrar las distribuciones condicionales de la variable cuantitativa es dibujar gráficos de caja de cada distribución condicional. También las barras de error son la mejor forma de describir gráficamente si las medias condicionales difieren realmente entre sí.

Ejemplo. Construir gráficos de caja para el peso de 198 recién nacidos en función del sexo en SPSS:

Gráficos -> Interactivos -> Boxplot

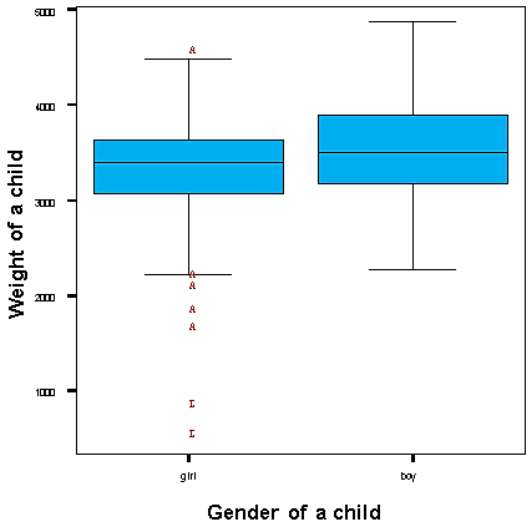


Figura 17: Gráficos de caja para el peso de los recién nacidos

Construcción de barras de error para el peso de 198 recién nacidos en función del sexo en SPSS:

Gráficos -> Interactivos -> Barra de error

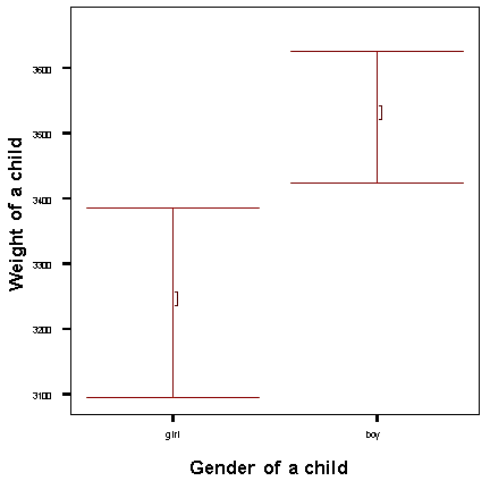


Figura 18: Barras de error para el peso de los recién nacidos

10.3. Variables cuantitativas

Cuando ambas variables son cuantitativas, pueden aplicarse los métodos presentados anteriormente para detectar una posible asociación de las variables.

En primer lugar, se pueden agrupar ambas variables y, a continuación, presentar la distribución conjunta mediante una tabla de frecuencias bidireccional. También es posible agrupar sólo una de las variables y luego comparar las medidas condicionales de centro y variación de la otra variable para averiguar la posible asociación.

Pero cuando ambas variables son cuantitativas, la mejor forma de ver gráficamente la relación entre ellas es construir un diagrama de dispersión. El diagrama de dispersión ofrece información visual sobre la cantidad y la dirección de la asociación, o correlación, como se denomina en el caso de las variables cuantitativas. La construcción de gráficos de dispersión y el cálculo de los coeficientes de correlación se estudian con más detenimiento en la siguiente sección.

11. Diagrama de dispersión y coeficiente de correlación

11.1. Gráfico de dispersión

La forma más eficaz de mostrar la relación entre dos variables cuantitativas es un gráfico de dispersión. Un gráfico de dispersión muestra la relación entre dos variables cuantitativas medidas en los mismos individuos. Los valores de una variable aparecen en el eje horizontal y los valores de la otra variable aparecen en el eje vertical. Cada individuo de los datos aparece como un punto en el gráfico fijado por los valores de ambas variables para ese individuo. La variable explicativa, si la hay, se representa siempre en el eje horizontal (eje x) de un diagrama de dispersión. Como recordatorio, normalmente llamar a la variable explicativa x y a la variable de respuesta y. Si no hay distinción entre explicativa y de respuesta, cualquiera de las variables puede ir en el eje horizontal.

Ejemplo. La altura y el peso de 10 personas son los siguientes:

Altura	Peso
158	48
162	57
163	57

170	60
154	45
167	55
177	62
170	65
179	70
179	68

Gráfico de dispersión en SPSS:

Gráficos -> Interactivos -> Diagrama de dispersión

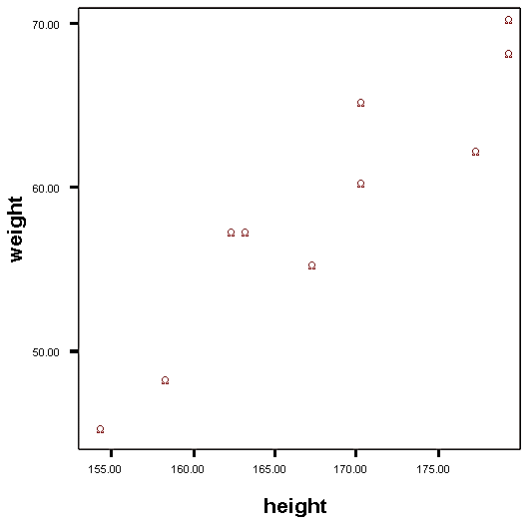


Figura 19: Diagrama de dispersión de la altura y el peso

Para abordar la interpretación de un diagrama de dispersión de manera rigurosa, es crucial en primer lugar buscar un patrón general que permita desvelar la dirección, forma y magnitud de la relación entre las dos variables en estudio.

En el contexto de esta interpretación, dos variables se consideran asociadas de manera positiva cuando los valores que superan la media de una de ellas tienden a coexistir con los valores que también superan la media de la otra variable, mientras que los valores inferiores a la media suelen encontrarse en conjunto. Por el contrario, una asociación negativa se presenta cuando los valores que exceden la

media en una variable se asocian con valores por debajo de la media en la otra, y viceversa.

Es relevante destacar que uno de los patrones más importantes que podemos encontrar en la representación gráfica de las relaciones entre variables es la linealidad. En las relaciones lineales, los puntos del diagrama de dispersión tienden a conformar un patrón rectilíneo. Sin embargo, no debemos pasar por alto otras formas de relación, como las curvas o los conglomerados, que también pueden ofrecer valiosa información.

La fuerza de la relación se encuentra intrínsecamente ligada a la proximidad de los puntos en el diagrama de dispersión respecto a una forma simple, como una línea recta. Cuanto más cercanos estén los puntos a esta forma, mayor será la fuerza de la relación entre las variables estudiadas.

11.2. Coeficiente de correlación

El diagrama de dispersión se erige como una herramienta visual de inestimable valor para analizar la naturaleza de la relación que existe entre los valores de las variables x e y en un conjunto de datos bivariados. En numerosas ocasiones, al observar el diagrama, se percibe una tendencia donde los puntos parecen distribuirse en torno a una línea recta, generando una impresión visual de esta relación.

No obstante, para trascender la impresión visual y cuantificar con rigor la proximidad de la dispersión de puntos respecto a una relación lineal, se recurre a una medida numérica fundamental conocida como el coeficiente de correlación de la muestra. Este coeficiente, de naturaleza cuantitativa, proporciona una evaluación precisa del grado de correlación entre las variables x e y, permitiendo así una comprensión más profunda de la relación presente en los datos.

El coeficiente de correlación de la muestra, al calcularse, ofrece una medida objetiva que oscila entre -1 y 1, indicando la dirección y fuerza de la relación. Un valor cercano a 1 denota una correlación positiva fuerte, mientras que un valor

cercano a -1 señala una correlación negativa fuerte. Por otro lado, un valor cercano a 0 sugiere una ausencia de correlación lineal.

El **coeficiente de correlación** de la muestra, denotado por r (o en algunos casos r_{xy}), es una medida de la fuerza de la relación lineal entre las variables x e y.

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \qquad (10)$$

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y})}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\sqrt{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}} \qquad (11)$$

$$r = \frac{\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{s_x s_y} \qquad (12)$$

$$r = \frac{s_{xy}}{\sqrt{s_{xx}}\sqrt{s_{yy}}}, \qquad (13)$$

Donde

$$s_{xx} = \sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = (n-1)s_x^2,$$

$$s_{yy} = \sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = (n-1)s_y^2,$$

$$s_{xy} = \sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

Las cantidades S_{xx} y S_{yy} son las sumas de las desviaciones al cuadrado de los valores observados x y los valores observados y, respectivamente. S_{xy} es la suma de los productos cruzados de las desviaciones x con las desviaciones y.

Ejemplo. ... continuación.

Altura	Peso	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
158	48	-9.9	98.01	-10.7	114.49	**105.93**
162	57	-5.9	34.81	-1.7	2.89	**10.03**
163	57	-4.9	24.01	-1.7	2.89	**8.33**
170	60	2.1	4.41	1.3	1.69	**2.73**
154	45	-13.9	193.21	-13.7	187.69	**190.43**

167	55	-0.9	0.81	-3.7	13.69	**3.33**
177	62	9.1	82.81	3.3	10.89	**30.03**
170	65	2.1	4.41	6.3	39.69	**13.23**
179	70	11.1	123.21	11.3	127.69	**125.43**
179	68	11.1	123.21	9.3	86.49	**103.23**
			688.9		**588.1**	**592.7**

Esto da el coeficiente de correlación como

$$r = \frac{592.7}{\sqrt{688.9}\sqrt{588.1}} = 0.9311749$$

Coeficiente de correlación en SPSS:

Analizar -> Correlacionar -> Bivariante

Tabla 20: Coeficiente de correlación entre la altura y el peso

Correlaciones.

Aquí tienes una tabla mejorada:

	ALTURA	**PESO**
	Pearson Correlation	Pearson Correlation
	1	0.931
N	10	10

En esta tabla, se presenta la correlación de Pearson entre las variables ALTURA y PESO, con sus respectivos valores. La inclusión de las etiquetas "ALTURA" y "PESO" en la parte superior de las columnas facilita la identificación de las variables en cuestión.

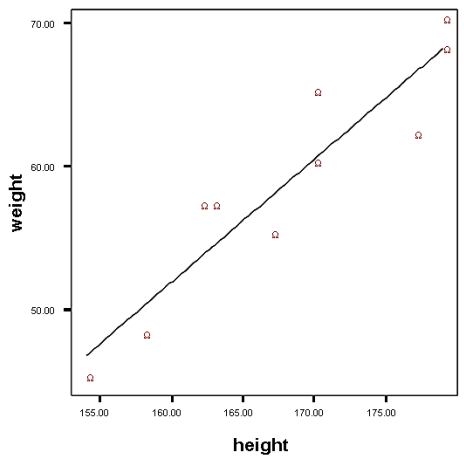


Figura 20: Diagrama de dispersión con línea lineal

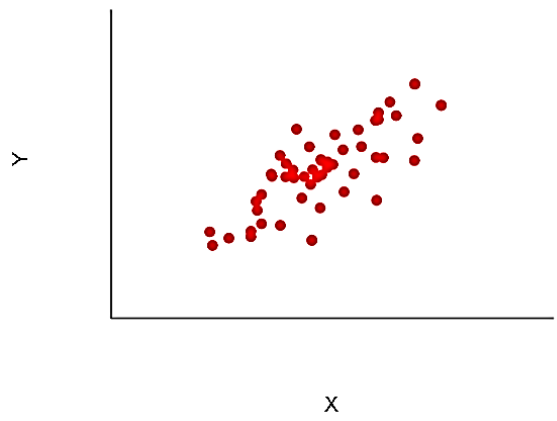
A continuación, se presentan algunas características cruciales del coeficiente de correlación, detalladas en un contexto científico:

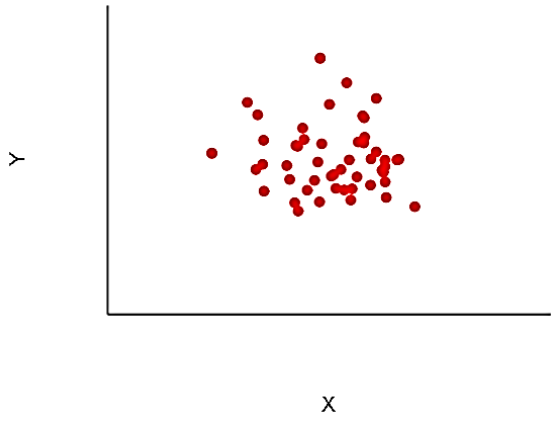
1. **Indicador de Asociación:** El coeficiente de correlación (r) desempeña un papel esencial al revelar la naturaleza de la relación entre dos variables. Un valor de r positivo indica una asociación positiva entre las variables, lo que significa que cuando una variable aumenta, la otra tiende a aumentar también. Por otro lado, un valor de r negativo indica una asociación negativa, lo que sugiere que cuando una variable aumenta, la otra tiende a disminuir.
2. **Rango de Valores:** La correlación r siempre se encuentra dentro del intervalo de -1 a 1. Los valores cercanos a 0 indican una relación lineal extremadamente débil, mientras que aquellos que se alejan de 0 hacia -1 o 1 denotan una relación lineal más fuerte. Los valores de r próximos a -1 o 1 indican que los puntos en un diagrama de dispersión se acercan notablemente a una línea recta. Solo en casos de una relación lineal perfecta, donde los puntos se alinean perfectamente en una línea recta, se alcanzan los valores extremos de r = -1 o r = 1.
3. **Invariabilidad ante Cambios de Unidades:** Una característica relevante del coeficiente de correlación es su insensibilidad a los cambios en las unidades de medida de las variables. Esto se debe a que r se calcula utilizando valores

normalizados de las observaciones (es decir, los valores xi - x̄ e yi - ȳ). Por lo tanto, cambiar de centímetros a pulgadas o de kilogramos a libras no altera la correlación entre las variables, ya que r carece de unidades de medida. Es simplemente un número que oscila entre -1 y 1.

4. **Limitación a Relaciones Lineales:** Importante destacar que el coeficiente de correlación mide exclusivamente la fuerza de una relación lineal entre dos variables. No tiene la capacidad de describir relaciones curvas, por intensas que sean. Por lo tanto, cuando se trata de relaciones no lineales, el coeficiente de correlación puede no ser indicativo de la relación subyacente.
5. **Sensibilidad a Valores Atípicos:** Al igual que la media y la desviación estándar, la correlación r puede verse significativamente influenciada por valores atípicos o extremos en el diagrama de dispersión. En presencia de observaciones inusuales, se debe utilizar r con precaución, ya que estos valores atípicos pueden distorsionar la magnitud de la correlación.

Examen. ¿Cuáles son los coeficientes de correlación en los casos siguientes?





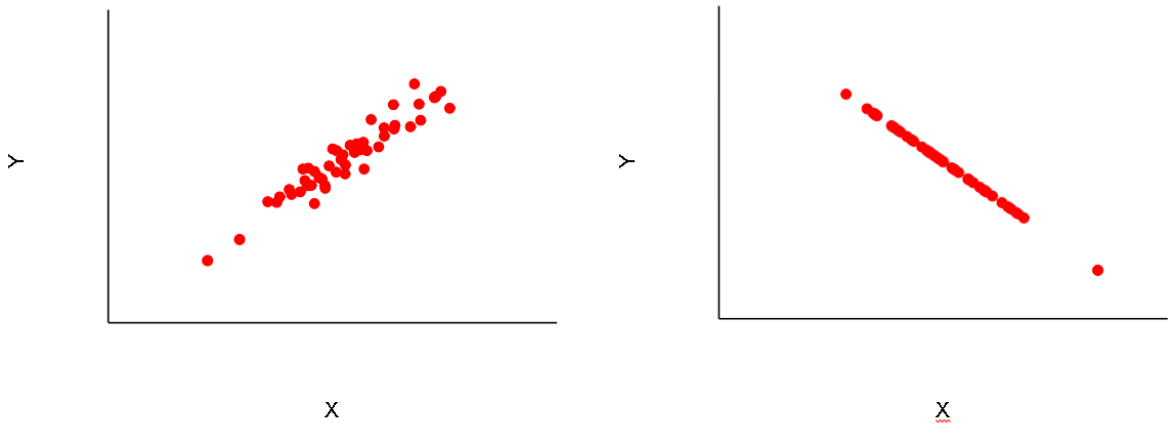


Figura 21: Ejemplo de gráficos de dispersión

Ejemplo. ¿Cómo interpretar estos diagramas de dispersión?

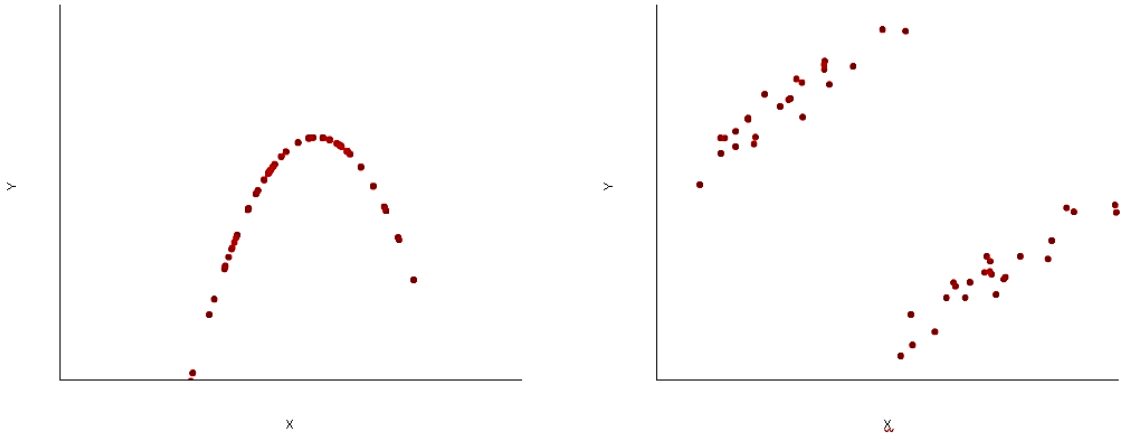


Figura 22: Ejemplo de gráficos de dispersión

En el ámbito científico, es esencial comprender que una correlación fuerte entre dos variables no necesariamente implica una relación causal entre ellas. El coeficiente de correlación, en particular, se limita a evaluar la magnitud de la asociación lineal entre estas variables, sin considerar la distinción entre variables explicativas y variables de respuesta.

Además, es importante reconocer que dos variables pueden exhibir una correlación significativa debido a la influencia de otras variables, denominadas "variables ocultas", que afectan a ambas variables en cuestión. Estas variables ocultas pueden actuar como factores confusos, creando una aparente correlación entre las dos variables originales, incluso cuando no existe una relación causal directa entre ellas.

El coeficiente de correlación muestral, también conocido como coeficiente de correlación de Pearson, se aplica en casos en los que ambas variables son cuantitativas y, por lo menos, se definen en una escala de intervalo. Es importante tener en cuenta que cuando se trabaja con variables cualitativas en una escala ordinal, el coeficiente de correlación de Spearman se convierte en una herramienta útil para medir la asociación entre estas variables. El coeficiente de Spearman se basa en la clasificación de los sujetos, lo que lo hace adecuado para datos ordinales.

Sin embargo, cabe destacar que una descripción exhaustiva de las propiedades del coeficiente de correlación de Spearman escapa al alcance de este libro. En consecuencia, en esta obra, nos centramos en proporcionar una comprensión sólida del coeficiente de correlación de Pearson y su aplicación en el análisis de datos cuantitativos, dejando la discusión detallada del coeficiente de Spearman para investigaciones más especializadas en el tema.

12. Covarianza:

La covarianza es una medida que evalúa la relación lineal entre dos variables cuantitativas. Indica si las variables tienden a aumentar o disminuir juntas. Una covarianza positiva sugiere una relación directa, mientras que una covarianza negativa indica una relación inversa. Sin embargo, la covarianza no es fácil de interpretar en términos de fuerza o dirección de la relación, ya que su magnitud depende de las escalas de las variables. Es una medida útil para explorar la relación inicial entre dos variables antes de aplicar técnicas de regresión.

13. Regresión Lineal:

La regresión lineal es una técnica estadística que se utiliza para modelar la relación entre una variable dependiente y una o más variables independientes. Esta técnica asume una relación lineal entre las variables, lo que permite predecir el valor de la variable dependiente en función de los valores de las variables independientes. La regresión lineal es ampliamente utilizada en la investigación y el análisis de datos para entender y predecir fenómenos en diversos campos, como la economía, la ciencia social y la biología.

14. Regresión Múltiple:

La regresión múltiple es una extensión de la regresión lineal simple y permite modelar la relación entre una variable dependiente y dos o más variables independientes. Esto es útil cuando se quiere evaluar cómo múltiples variables independientes afectan a una variable dependiente. La regresión múltiple es especialmente valiosa en investigaciones donde varios factores pueden influir en un resultado.

15. Prueba de Chi-Cuadrada:

La prueba de chi-cuadrada es una herramienta estadística utilizada para determinar si existe una asociación significativa entre dos variables categóricas. Se utiliza para evaluar si las diferencias observadas entre las frecuencias observadas y las frecuencias esperadas son estadísticamente significativas. Esta prueba es comúnmente utilizada en investigaciones de salud pública, encuestas y estudios de comportamiento.

16. Referencias

1. Akers, D. M. D. (2018). *Exploring, Analysing and Interpeting Data with Minitab 18: First Edition* (First Edition). Minitab 18.
2. Anderberg, D. M. R. (1973). *Cluster Analysis for Applications: Probability and Mathematical Statistics*. Academic Press, INC.
3. Barton, P. I., & Pantelides, C. C. (1994). Modeling of combined discrete/continuous processes. *AIChE Journal*, *40*(6), 966–979. https://doi.org/10.1002/aic.690400608
4. Bechtold, W. A., & Patterson, P. L. (2005). *The Enhanced Forest Inventory and Analysis Program--national Sampling Design and Estimation Procedures*. Southern Research station.
5. Bishara, A. J., & Hittner, J. B. (2012). Testing the significance of a correlation with nonnormal data: Comparison of Pearson, Spearman, transformation, and resampling approaches. *Psychological Methods*, *17*(3), 399–417. https://doi.org/10.1037/a0028087
6. Byrne, G. (2007). A Statistical Primer: Understanding Descriptive and Inferential Statistics. *Evidence Based Library and Information Practice*, *2*(1), 32–47. https://doi.org/10.18438/B8FW2H
7. Cobb, G. W., & Moore, D. S. (1997). Mathematics, Statistics, and Teaching. *The American Mathematical Monthly*, *104*(9), 801–823. https://doi.org/10.1080/00029890.1997.11990723
8. Crawford, J. R., Garthwaite, P. H., & Porter, S. (2010). Point and interval estimates of effect sizes for the case-controls design in neuropsychology: Rationale, methods, implementations, and proposed reporting standards. *Cognitive Neuropsychology*, *27*(3), 245–260. https://doi.org/10.1080/02643294.2010.513967
9. Edgington, E. S. (1969). Approximate Randomization Tests. *The Journal of Psychology*, *72*(2), 143–149. https://doi.org/10.1080/00223980.1969.10543491
10. Etikan, I. (2016). Comparison of Convenience Sampling and Purposive Sampling. *American Journal of Theoretical and Applied Statistics*, *5*(1), 1. https://doi.org/10.11648/j.ajtas.20160501.11

11. Foster, C. (2014). Confidence Trick: The Interpretation of Confidence Intervals. *Canadian Journal of Science, Mathematics and Technology Education*, *14*(1), 23–34. https://doi.org/10.1080/14926156.2014.874615
12. Gibbons, J. D., & Pratt, J. W. (1975). P-values: Interpretation and Methodology. *The American Statistician*, *29*(1), 20–25. https://doi.org/10.1080/00031305.1975.10479106
13. Hernán, M. A., Hsu, J., & Healy, B. (2019). A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks. *CHANCE*, *32*(1), 42–49. https://doi.org/10.1080/09332480.2019.1579578
14. Hubert, M., & Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Computational Statistics & Data Analysis*, *52*(12), 5186–5201. https://doi.org/10.1016/j.csda.2007.11.008
15. Hyland, K. (1996). Writing Without Conviction? Hedging in Science Research Articles. *Applied Linguistics*, *17*(4), 433–454. https://doi.org/10.1093/applin/17.4.433
16. Johnson, R. A., & Bhattacharyya, G. K. (1992). *Estadística: Principios y métodos* (2nd Edition). Wiley.
17. Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430–454. https://doi.org/10.1016/0010-0285(72)90016-3
18. Keoleian, G. A., & Menerey, D. (1994). Sustainable Development by Design: Review of Life Cycle Design and Related Approaches. *Air & Waste*, *44*(5), 645–668. https://doi.org/10.1080/1073161X.1994.10467269
19. Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P., & Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, *16*(12), 1289–1296. https://doi.org/10.1038/s41592-019-0619-0
20. Kürzl, H. (1988). Exploratory data analysis: recent advances for the interpretation of geochemical data. *Journal of Geochemical Exploration*, *30*(1–3), 309–322. https://doi.org/10.1016/0375-6742(88)90066-0

21. Levy, D. P. S., & Lemeshow, S. (2008). *Sampling of Populations: Methods and Applications* (Fourth edition). Published by John Wilwy & Sons.
22. Livi-Bacci, M. (2007). *Introduzzione alla demografia* (3era edición impresa). Ariel.
23. MacCorquodale, K., & Meehl, P. E. (1948). On a distinction between hypothetical constructs and intervening variables. *Psychological Review*, *55*(2), 95–107. https://doi.org/10.1037/h0056029
24. MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence Limits for the Indirect Effect: Distribution of the Product and Resampling Methods. *Multivariate Behavioral Research*, *39*(1), 99–128. https://doi.org/10.1207/s15327906mbr3901_4
25. Marshall, G., & Jonker, L. (2010). An introduction to descriptive statistics: A review and practical guide. *Radiography*, *16*(4), e1–e7. https://doi.org/10.1016/j.radi.2010.01.001
26. Martínez Bencardino, C. (2019). *Estadística Básica Aplicada* (Quinta edición). Ecoe Ediciones Limitada.
27. Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods*, *43*(3), 679–690. https://doi.org/10.3758/s13428-010-0049-5
28. O'Fallon, B. D. (2013). ACG: rapid inference of population history from recombining nucleotide sequences. *BMC Bioinformatics*, *14*(1), 40. https://doi.org/10.1186/1471-2105-14-40
29. Pimienta Lastra, R. (2000). Encuestas probabilísticas vs. no probabilísticas. *Política y Cultura*, *13*, 263–276. http://www.redalyc.org/articulo.oa?id=26701313
30. Preacher, K. J., & Kelley, K. (2011). Effect size measures for mediation models: Quantitative strategies for communicating indirect effects. *Psychological Methods*, *16*(2), 93–115. https://doi.org/10.1037/a0022658

31. Rayat, C. S. (2018). Measures of Central Tendency. In *Statistical Methods in Medical Research* (pp. 33–46). Springer Singapore. https://doi.org/10.1007/978-981-13-0827-7_6
32. Ross, S. M. (2018). *Introducción a la Estadística*. Editorial Reverté, S.A.
33. Silva, A. C., Viseu, F., Aires, A. P., & Neto, T. B. (2022). Investigative Activities for Statistics Learning with 1st Grade Portuguese Students. *Education Sciences*, *12*(10), 665. https://doi.org/10.3390/educsci12100665
34. Steiger, J. H. (1990). Structural Model Evaluation and Modification: An Interval Estimation Approach. *Multivariate Behavioral Research*, *25*(2), 173–180. https://doi.org/10.1207/s15327906mbr2502_4
35. Tamhane, A. C. (1977). Multiple comparisons in model i one-way anova with unequal variances. *Communications in Statistics - Theory and Methods*, *6*(1), 15–32. https://doi.org/10.1080/03610927708827466
36. Van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, *76*(12), 1049–1064. https://doi.org/10.1080/10629360600810434
37. Venuti, L., Bouvier, J., Irwin, J., Stauffer, J. R., Hillenbrand, L. A., Rebull, L. M., Cody, A. M., Alencar, S. H. P., Micela, G., Flaccomio, E., & Peres, G. (2015). UV variability and accretion dynamics in the young open cluster NGC 2264. *Astronomy & Astrophysics*, *581*, A66. https://doi.org/10.1051/0004-6361/201526164
38. Weiss, N. A. (2016). *Instructory Statistics* (Tenth edition). Publishing as Pearson.
39. Wild, C. J., & Pfannkuch, M. (1999). Statistical Thinking in Empirical Enquiry. *International Statistical Review*, *67*(3), 223–248. https://doi.org/10.1111/j.1751-5823.1999.tb00442.x
40. Willemse, I. (2009). *Statistical Methods and Calculation Skills* (Third edition). Juta & Co Ltd.

41. Zieffler, A., Garfield, J., Delmas, R., & Reading, C. (2008). A framesork to support research on informal inferential reasoning. *Statistics Education Research Journal*, *7*(2), 40–58. https://doi.org/10.52041/serj.v7i2.469
42. Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, *1*(1), 3–14. https://doi.org/10.1111/j.2041-210X.2009.00001.x