

Estadística para Ciencias Sociales

Julián Cruz

Table of contents

Comencemos

```
# justificar los textos vía css
# Pruebas de hipótesis: analíticas paramétricas, analíticas no paramétricas, computacionales
# Poner todas las fórmulas en cajas
# Variables cualitativas, cómo se ven en un gráfico, cuáles son sus elementos visuales, cómo
# Hacer capítulo de preparación de los datos.
#### Completar la fórmula de las estadísticas descriptivas
# Actualizar los covers para que coincidan con los capítulos
```

El Laboratorio Interdisciplinar de Medición Social (LABMEDISOCIAL) se complace en presentar estas notas de clase sobre estadística en ciencias sociales.

La estadística, en su esencia, constituye un puente entre la observación empírica y la construcción de conocimiento científico. En el ámbito de las ciencias sociales, su aplicación trasciende el análisis de datos para convertirse en una herramienta indispensable en la formulación de teorías, la validación de hipótesis y la toma de decisiones fundamentadas. Este libro se concibe como una contribución a ese esfuerzo, ofreciendo una integración rigurosa de fundamentos teóricos y aplicaciones prácticas, sustentada en la investigación social.

En investigación social, el uso adecuado de métodos estadísticos para comprender la estructura y operatividad de un fenómeno determinado, constituye una ventaja del investigador en un entorno competitivo de alto desempeño. Una gran variedad de procesos de planeación y evaluación de actividades gubernamentales, administrativas, económicas y financieras, se basan en resultados obtenidos mediante el análisis estadístico de los fenómenos en ellos involucrados.

Además, dado el crecimiento exponencial de las fuentes de información y el desarrollo acelerado de las herramientas tecnológicas, es apropiado disponer de una sólida fundamentación conceptual y práctica que le permita transformar y comprender grandes cantidades de información.

Material

Los contenidos han sido generados utilizando modelos de generación de texto y consultando algunos materiales que se comparten a continuación. Es importante prever que **algunos de**

estos **vínculos pueden caducar con el tiempo**, por lo cual no es raro que se pierda el acceso a los materiales.

[OpenIntro Statistics](#)

[Estadística para todos - Diana Kelmansky](#)

[Discourse of Objectivity](#)

[To explain or to predict](#)

[Escalas de medición - Jorge Coronado Padilla](#)

[Técnicas de Muestreo sobre una Población a Estudio - Tamara Otzen, Carlos Manterola](#)

[OpenIntro Statistics - David Diez](#)

[Análisis bivariante - Jordi Mas Elias](#)

[Handbook on Constructing Composite Indicators: Methodology and User Guide](#)

[Diseño y validación de instrumentos de medición](#)

[Principal Component Analysis](#)

Recopilación

La tarea de recopilar y organizar la información ha sido llevada a cabo por Julián Cruz.

Científico de datos, profesional en estadística y magíster en ciencias, con más de 15 años de experiencia demostrada en analítica y ciencia de datos. Su perfil contempla desde liderazgo de programas de capacitación y gestión del cambio, hasta ejecución proyectos de base tecnológica. A través de esta experiencia ha desarrollado diferentes competencias, como la orientación al valor en toma de decisiones, la conformación y desarrollo de equipos de alto desempeño y la negociación integradora.

Part I

Estadística descriptiva

En esta primera sección se examina la estadística descriptiva. El contenido ha sido seleccionado cuidadosamente con el fin de agregar valor a los análisis cuantitativos que se proponen al interior de las investigaciones en ciencias sociales.

Introducción

Este primer tema introduce a los estudiantes en los conceptos básicos de la estadística, comenzando por una definición de estadística y su relevancia en el análisis de datos. Se exploran los **tipos de variables**, que incluyen **variables cualitativas** (nominales y ordinales) y **cuantitativas** (discretas y continuas). A continuación, se aborda la **tipología de preguntas** en encuestas, diferenciando entre preguntas abiertas, cerradas y escalas, y cómo estas se relacionan con la medición de las variables. También se presentan los **tipos de encuestas**, clasificadas según el método de recolección (presencial, telefónica, online, etc.), su estructura (estructuradas, semi-estructuradas, no estructuradas) y los contextos en los que se aplican, lo cual es fundamental para la adecuada recolección de datos en cualquier estudio estadístico.

Muestreo

En este tema se abordan los fundamentos del **muestreo**, que es el proceso de seleccionar una parte representativa de una población para inferir conclusiones sobre el total. Se explican dos grandes tipos de muestreo: **probabilístico** y **no probabilístico**. En el muestreo probabilístico, cada elemento de la población tiene una probabilidad conocida y no nula de ser seleccionado, y se estudian técnicas como el **muestreo aleatorio simple**, el **muestreo multietápico**, el **muestreo estratificado** y el **muestreo por conglomerados**. En el muestreo no probabilístico, la selección no se basa en probabilidades, sino en criterios subjetivos o conveniencia, como el **muestreo por conveniencia**, **muestreo por cuotas** o **bola de nieve**. Este tema profundiza en las ventajas y limitaciones de cada enfoque, subrayando la importancia de la representatividad de la muestra y cómo esto afecta los resultados y su generalización.

Estadística descriptiva para una variable

La **estadística descriptiva para una variable** se centra en la síntesis y descripción de los datos mediante medidas numéricas y representaciones gráficas. Se examinan las **medidas de tendencia central** como la media, la mediana y la moda, que resumen el comportamiento típico de la variable. También se analizan las **medidas de dispersión** como el rango, la varianza y la desviación estándar, que describen la variabilidad en los datos. El tema incluye el uso de **gráficos** como histogramas, gráficos de barras y diagramas de caja para visualizar la distribución de los datos. Estas herramientas permiten una comprensión clara de cómo los datos están distribuidos y proporcionan una base sólida para la inferencia estadística.

Estadística descriptiva para dos o más variables

Este tema amplía el análisis de la estadística descriptiva a dos o más variables, abordando técnicas para analizar relaciones y asociaciones entre ellas. Se introducen las **tablas de contingencia** para estudiar la relación entre dos variables cualitativas, así como las **medidas de asociación** como el **coeficiente de correlación de Pearson** y **Spearman** para variables cuantitativas. También se discuten las **gráficas de dispersión** para visualizar relaciones entre dos variables cuantitativas y cómo se puede ajustar una **recta de regresión** para predecir una variable en función de otra. Además, se examinan los **coeficientes de correlación parcial** y **análisis multivariado** para explorar cómo varias variables se relacionan simultáneamente, lo que permite una comprensión más profunda de las interacciones dentro de los datos.

1 Preliminares

La estadística, en su concepción moderna, es el resultado de un extenso desarrollo histórico y epistemológico que refleja la evolución del pensamiento humano en su búsqueda de patrones y significados en la incertidumbre. Sus raíces se remontan a la antigüedad, cuando civilizaciones como las babilónicas y egipcias utilizaban conteos y registros para la administración y predicción. Sin embargo, fue en el Renacimiento, con la formalización del pensamiento probabilístico por figuras como Gerolamo Cardano y Pierre de Fermat, cuando emergieron las bases matemáticas que darían forma a su carácter científico.

La epistemología de la estadística comenzó a consolidarse en el siglo XVIII, con las contribuciones de pensadores como Thomas Bayes, quien introdujo un enfoque probabilístico basado en la actualización del conocimiento, y Carl Friedrich Gauss, cuya teoría de errores cimentó el papel de la estadística en la inferencia científica. A medida que el método científico se institucionalizaba, la estadística se convirtió en un vehículo para traducir fenómenos observables en explicaciones cuantitativas, adoptando principios de objetividad, reproducibilidad y rigor.

El siglo XX marcó una transformación paradigmática, impulsada por avances en la teoría de conjuntos, el cálculo y la computación. La estadística inferencial y la probabilidad frecuentista, promovidas por Ronald A. Fisher, Jerzy Neyman y Egon Pearson, coexistieron con enfoques bayesianos, destacando debates epistemológicos sobre la naturaleza del azar y la inferencia. Esta diversidad metodológica enriqueció su aplicabilidad, desde el diseño de experimentos hasta el análisis de sistemas complejos en las ciencias sociales.

Hoy, la estadística es no solo un instrumento técnico, sino también un lenguaje epistemológico que da forma a cómo entendemos la realidad. Su desarrollo histórico subraya la tensión entre el dato y la interpretación, entre el caos del mundo empírico y la búsqueda de orden intelectual.

1.1 La verdad y la objetividad

La necesidad de procesar información es inherente al ser humano, un reflejo de su insaciable curiosidad por comprender el mundo que lo rodea. En un universo complejo e incierto, la información se presenta como un recurso esencial para la supervivencia, la toma de decisiones y la construcción de conocimiento. Sin procesarla, los datos brutos carecen de significado; es mediante el análisis y la síntesis que transformamos fragmentos dispersos en narrativas coherentes que guían nuestra acción. Esta capacidad no solo nos define como especie, sino que también alimenta nuestra aspiración más profunda: la búsqueda de la verdad.

1.1.1 ¿Existe la verdad?

La existencia de la verdad es una cuestión filosófica que ha desafiado a pensadores desde tiempos antiguos. Para Platón, la verdad era una forma ideal, un conocimiento absoluto al que solo se accedía mediante la razón. En contraste, las tradiciones empiristas, como las de Aristóteles o Locke, vincularon la verdad con la experiencia y la observación. Sin embargo, estas concepciones comparten una premisa: que la verdad es algo objetivo, independiente del observador.

En un sentido práctico, la verdad parece manifestarse en dos planos: como correspondencia con la realidad y como coherencia lógica dentro de un sistema de pensamiento. La primera implica que una afirmación es verdadera si refleja fielmente un hecho, mientras que la segunda sugiere que la verdad puede ser relativa a los marcos epistemológicos que la sustentan. Esto lleva a una paradoja: aunque aspiramos a la objetividad, nuestra percepción de la verdad está inevitablemente condicionada por nuestra cultura, lenguaje y experiencia.

1.2 Verdad y objetividad en la modernidad

La modernidad abrazó un enfoque racionalista y científico de la verdad, fundado en la fe en el progreso y en la capacidad humana de descubrir leyes universales que rigen la naturaleza y la sociedad. Este paradigma, influido por el empirismo y el método científico, promovió una visión de la verdad como algo accesible a través del rigor metodológico y la acumulación de evidencias. La Revolución Científica y la Ilustración, con pensadores como Descartes, Kant y Newton, consolidaron esta noción, priorizando la objetividad y la replicabilidad como criterios de validez.

Sin embargo, esta confianza en la verdad objetiva también trajo tensiones. Por un lado, el poder del conocimiento científico revolucionó la tecnología, la medicina y la economía. Por otro, redujo la complejidad humana y social a ecuaciones y categorías simplificadas, marginando otras formas de saber, como las éticas, estéticas y espirituales. La modernidad trató la verdad como una meta absoluta, pero al hacerlo, a menudo ignoró las dimensiones subjetivas y plurales de la experiencia humana.

La objetividad en la modernidad se concibe como un ideal epistemológico central, que busca eliminar la influencia de valores, emociones o prejuicios individuales en la producción de conocimiento. Este concepto está profundamente enraizado en el contexto histórico, filosófico y científico de la modernidad, un período caracterizado por la confianza en la razón, el método científico y el progreso.

1.2.1 El surgimiento de la objetividad moderna

La modernidad, influenciada por la Ilustración y la Revolución Científica, estableció la objetividad como el estándar para alcanzar verdades universales. Filósofos como Descartes promovieron una visión racionalista en la que el pensamiento lógico y metódico podía superar las limitaciones de la percepción subjetiva. La objetividad se convirtió en un mecanismo para garantizar que el conocimiento no estuviera contaminado por creencias personales, ideologías o tradiciones.

En este contexto, la ciencia se posicionó como el paradigma del conocimiento objetivo. Galileo, Newton y otros pioneros de la ciencia moderna formularon leyes universales que prometían describir la naturaleza de manera imparcial. El método científico, con su énfasis en la observación sistemática, la experimentación controlada y la replicabilidad, se estableció como el camino hacia un conocimiento fiable y verificable.

1.2.2 Características de la objetividad en la modernidad

1. **Universalidad:** La objetividad buscaba verdades aplicables a todos los contextos, independientemente de las diferencias culturales, históricas o individuales. Por ejemplo, las leyes de la física se consideraron válidas tanto en Europa como en otras partes del mundo.
2. **Neutralidad:** La modernidad aspiraba a un conocimiento libre de intereses particulares, de manera que la investigación científica pudiera ser una empresa independiente de valores políticos, éticos o sociales.
3. **Rigor metódico:** La objetividad requería procedimientos estandarizados que redujeran al mínimo la influencia de sesgos personales. Esto incluía el uso de instrumentos de medición, controles experimentales y protocolos reproducibles.
4. **Despersonalización:** En aras de la objetividad, se promovió una separación entre el investigador y el objeto de estudio. El científico debía actuar como un observador distante, dejando a un lado emociones, creencias y deseos personales.

1.2.3 Críticas y tensiones

Aunque la objetividad en la modernidad permitió avances extraordinarios en ciencia y tecnología, también generó críticas y tensiones. Una de las principales críticas es que la objetividad moderna a menudo enmascara los contextos sociales, políticos y culturales que inevitablemente influyen en la producción de conocimiento. Por ejemplo, Donna Haraway argumenta que todo conocimiento es situado, es decir, que está condicionado por la perspectiva y el contexto del observador, incluso en disciplinas científicas.

Asimismo, la idea de neutralidad ha sido cuestionada. La investigación científica no está aislada de los intereses económicos y políticos: qué se investiga, cómo se financia y cómo se aplica el conocimiento son decisiones influenciadas por valores y estructuras de poder. En este sentido, la objetividad moderna ha sido vista como una aspiración más que como una realidad alcanzable.

La objetividad en la modernidad representó un ideal revolucionario que buscaba superar la subjetividad individual y alcanzar verdades universales mediante el rigor metodológico y la neutralidad. Aunque permitió el desarrollo de un conocimiento confiable y replicable, sus limitaciones y tensiones muestran que incluso los estándares más altos de objetividad no están exentos de condicionamientos humanos. Este reconocimiento no invalida su valor, sino que nos invita a reflexionar críticamente sobre cómo construimos conocimiento en contextos históricos y sociales específicos.

1.3 Verdad y objetividad en la postmodernidad

La postmodernidad desafió la narrativa hegemónica de verdad universal, cuestionando la existencia misma de una verdad universal. En este contexto, pensadores como Michel Foucault, Jean-François Lyotard y Jacques Derrida argumentaron que lo que llamamos “verdad” es una construcción social, moldeada por relaciones de poder, discursos dominantes y contextos históricos. Según este enfoque, no hay una verdad única, sino una multiplicidad de verdades, cada una dependiente de los valores, intereses y perspectivas que la producen.

Este giro epistemológico desestabilizó las certezas modernas, abriendo espacio para voces antes excluidas y reconociendo la importancia de la narrativa, la identidad y la subjetividad. Sin embargo, también planteó un desafío: si toda verdad es relativa, ¿cómo distinguimos entre lo válido y lo falso? En la postmodernidad, la verdad se trata no como un punto fijo, sino como un proceso dinámico de negociación entre perspectivas, un diálogo interminable que refleja la diversidad y la complejidad de la experiencia humana.

La búsqueda de la verdad es tanto una necesidad como un desafío. Procesamos información para darle sentido al caos, pero nuestras herramientas y paradigmas siempre están limitados por el contexto histórico y cultural. La modernidad trató la verdad como algo que podía descubrirse y validarse de manera objetiva, mientras que la postmodernidad nos recuerda que nuestras nociones de verdad están intrínsecamente ligadas a nuestras perspectivas. En última instancia, procesar información y buscar la verdad son actos profundamente humanos, que reflejan no solo nuestra capacidad intelectual, sino también nuestra vulnerabilidad, creatividad y deseo de trascender.

La ciencia, en su esencia, se erige como un esfuerzo humano para interpretar el mundo, generando conocimiento estructurado que busca explicar fenómenos y resolver problemas. Para abordar cuestiones como el significado de la información, el carácter político de las ciencias,

su función instrumental y la naturaleza de los modelos científicos, es necesario explorar sus fundamentos epistemológicos y sociales.

1.3.1 La Objetividad en la Postmodernidad

La postmodernidad cuestiona profundamente la noción de objetividad tal como fue concebida en la modernidad, socavando la idea de que es posible acceder a una verdad universal, imparcial y libre de influencias culturales o subjetivas. En lugar de considerar la objetividad como un estándar absoluto, la postmodernidad la percibe como una construcción social y cultural, moldeada por discursos, relaciones de poder y contextos históricos.

1.3.2 Críticas postmodernas a la objetividad

1. **La verdad como construcción social:** Filósofos postmodernos como Michel Foucault y Jacques Derrida argumentaron que lo que llamamos “verdad” es el resultado de prácticas discursivas y estructuras de poder. La objetividad, desde esta perspectiva, no es neutral, sino una herramienta utilizada para legitimar ciertos conocimientos y excluir otros. Por ejemplo, Foucault mostró cómo las instituciones científicas han sido históricamente influidas por intereses políticos y económicos, moldeando lo que se considera verdadero.
2. **El rechazo a las narrativas universales:** Jean-François Lyotard, en su obra *La condición postmoderna*, criticó las “metanarrativas” de la modernidad, como el progreso y la ciencia, que pretendían ofrecer explicaciones globales y objetivas de la realidad. Para Lyotard, la postmodernidad celebra las “micronarrativas”, perspectivas locales y situadas que reflejan la diversidad y la pluralidad del conocimiento humano.
3. **El conocimiento situado:** La feminista Donna Haraway introdujo el concepto de “conocimientos situados”, que sostiene que todo conocimiento está condicionado por la posición del sujeto que lo produce. Según Haraway, la objetividad debe entenderse no como una ausencia de perspectiva, sino como un reconocimiento crítico de las propias limitaciones y el contexto desde el cual se genera el conocimiento.
4. **Relativismo y pluralismo:** En la postmodernidad, la objetividad es reemplazada por una aceptación del relativismo y el pluralismo epistemológico. Esto no implica que todas las verdades sean iguales, sino que ninguna puede reclamar supremacía universal sin considerar su contexto. Por ejemplo, una investigación científica puede ser rigurosa dentro de su marco metodológico, pero su validez depende de los valores y presupuestos que subyacen a ese marco.

1.3.3 La objetividad como diálogo y negociación

En lugar de desechar por completo la idea de objetividad, la postmodernidad sugiere replantearla como un proceso dinámico de diálogo y negociación entre perspectivas diversas. Este enfoque reconoce que:

- No existe un punto de vista totalmente neutro, pero es posible alcanzar consensos provisionales basados en la crítica y el intercambio.
- La objetividad no elimina los sesgos, pero puede gestionarlos mediante la transparencia y la reflexividad.
- El conocimiento es siempre parcial, pero puede ser útil y significativo si se reconoce su contexto y sus límites.

1.3.4 Tensiones y desafíos

La crítica postmoderna a la objetividad ha abierto nuevas posibilidades para incluir voces marginadas y valorar diferentes formas de conocimiento. Sin embargo, también plantea desafíos:

- ¿Cómo evaluamos la validez de diferentes perspectivas sin recurrir a criterios universales?
- ¿Cómo evitamos el riesgo del relativismo extremo, en el que todo conocimiento tiene el mismo peso, independientemente de su rigor o evidencia?
- ¿Cómo reconciliamos la diversidad de narrativas con la necesidad de acciones colectivas basadas en hechos comunes, como en el caso del cambio climático?

La objetividad en la postmodernidad no es un ideal fijo, sino una práctica crítica y reflexiva que valora la pluralidad de perspectivas sin renunciar a la búsqueda de consensos. En lugar de perseguir una verdad absoluta e independiente, la postmodernidad nos invita a reconocer la complejidad, contingencia y contextualidad del conocimiento. Este enfoque no rechaza la objetividad, sino que la transforma en una herramienta más humana y adaptable para comprender y actuar en un mundo diverso y en constante cambio.

1.4 La información

La información puede definirse como el conjunto de datos organizados y contextualizados que transmiten significado. No es solo un registro de hechos aislados, sino una estructura que conecta elementos para generar conocimiento. Claude Shannon, en su teoría de la información, la describió como una medida de reducción de incertidumbre en un sistema, enfatizando su dimensión cuantitativa.

Sin embargo, la información también posee una dimensión cualitativa: depende de quién la interpreta, del contexto en el que se utiliza y del propósito que sirve. Es materia prima para la construcción del conocimiento, el puente entre lo desconocido y lo entendible, un recurso que define no solo cómo comprendemos el mundo, sino cómo lo transformamos.

1.4.1 Las ciencias son políticas

Las ciencias no existen en un vacío neutral; están inmersas en sistemas sociales, económicos y culturales que las moldean y que, a su vez, son influenciados por ellas. Esta relación las convierte en intrínsecamente políticas. Por “políticas”, no solo se entiende la intervención de intereses gubernamentales o ideológicos, sino también la dinámica de poder que define qué se investiga, quién financia dicha investigación y cómo se aplica su conocimiento.

Por ejemplo, los avances en biotecnología pueden abordar problemas de hambre global, pero también generar desigualdades si los beneficios se concentran en sectores privilegiados. El cambio climático, otro ejemplo, es un campo donde la ciencia enfrenta intereses corporativos y políticos que intentan influir en sus hallazgos y narrativas. En este sentido, las ciencias reflejan no solo la búsqueda de la verdad, sino también los valores y prioridades de la sociedad que las produce.

1.4.2 Las ciencias son herramientas

Si bien las ciencias están impregnadas de políticas y valores, también son herramientas esenciales para entender y actuar sobre el mundo. Su capacidad de generar conocimiento riguroso permite desarrollar tecnologías, formular políticas públicas efectivas y abordar desafíos globales. En este sentido, las ciencias son instrumentos al servicio de la humanidad, aunque la dirección de su uso depende de las intenciones humanas.

Por ejemplo, las ciencias médicas han permitido erradicar enfermedades y mejorar la calidad de vida, mientras que la física y la ingeniería han desarrollado tecnologías que revolucionan las comunicaciones, la energía y el transporte. Sin embargo, como herramientas, las ciencias son éticamente neutras: su impacto depende de cómo las aplicamos y con qué fines.

1.4.3 ¿Qué es un modelo científico?

Un modelo científico es una representación simplificada de un fenómeno complejo que permite comprender, predecir y manipular su comportamiento. Los modelos son esenciales en la ciencia porque nos ofrecen una manera de explorar aspectos de la realidad que no pueden observarse directamente o que son demasiado complicados para abordarse en su totalidad.

Por ejemplo, en la meteorología, los modelos climáticos simulan sistemas atmosféricos para prever patrones de lluvia y temperatura. En la física, los modelos cuánticos explican el comportamiento de partículas subatómicas, mientras que en las ciencias sociales, los modelos económicos intentan captar las dinámicas de mercado.

Sin embargo, todo modelo científico es una aproximación, no una réplica exacta de la realidad. Su valor radica en su utilidad, no en su perfección. Como dijo el estadístico George Box, “todos los modelos son erróneos, pero algunos son útiles”. Esta afirmación subraya que

los modelos no son verdades absolutas, sino herramientas conceptuales diseñadas para guiar nuestro entendimiento y acción.

1.4.4 Información y ciencia

La información, las ciencias y los modelos científicos son elementos interdependientes que estructuran nuestra forma de entender y actuar sobre el mundo. La información es la base del conocimiento; las ciencias son tanto herramientas como reflejo de las dinámicas políticas que las atraviesan; y los modelos científicos son construcciones que hacen posible desentrañar la complejidad de la realidad. Juntos, estos conceptos no solo iluminan el mundo, sino también nuestras propias limitaciones y posibilidades como seres humanos en busca de sentido.

La replicabilidad es un principio fundamental en la ciencia, que garantiza la validez y la fiabilidad de los hallazgos científicos. Consiste en la capacidad de repetir un experimento o estudio bajo condiciones similares y obtener resultados consistentes. Este proceso actúa como un mecanismo de control que permite verificar si los descubrimientos están libres de errores, sesgos o influencias externas. En disciplinas empíricas, la replicabilidad refuerza la confianza en el conocimiento generado, estableciendo estándares que distinguen a la ciencia de otras formas de comprensión, como las creencias o la opinión. Es, en esencia, la piedra angular que sostiene la credibilidad del método científico.

Sin embargo, en la práctica, la replicabilidad enfrenta desafíos significativos. Factores como la falta de transparencia en los métodos, la presión por publicar resultados novedosos y las limitaciones en la disponibilidad de recursos dificultan la reproducción de estudios. En años recientes, varias disciplinas, desde la psicología hasta la biomedicina, han experimentado lo que se denomina la “crisis de replicabilidad”, donde numerosos resultados publicados no pueden ser replicados por otros investigadores. Esta situación ha llevado a un llamado global para adoptar prácticas más rigurosas, como el uso de pre-registros, la publicación de datos abiertos y la implementación de análisis reproducibles, que refuercen la integridad de la ciencia y aseguren que el conocimiento generado sea verdaderamente confiable.

Lo objetivo no existe. Tenemos conocimiento más subjetivo y menos subjetivo. El conocimiento menos subjetivo es el conocimiento científico, porque se basa en lo replicable.

1.5 Definiciones

La estadística, como disciplina, se basa en el análisis sistemático de información cuantitativa o cualitativa para describir, interpretar y predecir fenómenos. Tres conceptos clave estructuran este análisis: los datos, las variables y los individuos. Cada uno cumple un rol específico, interconectado y esencial en el proceso de generación de conocimiento.

1.5.1 ¿Qué son los datos?

Los datos son el registro de observaciones, medidas o respuestas recopiladas sobre un fenómeno o población de interés. Constituyen la materia prima del análisis estadístico y pueden ser numéricos (cuantitativos) o descriptivos (cualitativos). Los datos por sí mismos son neutros; adquieren significado solo cuando se contextualizan y se analizan en relación con una pregunta o hipótesis.

Por ejemplo, el registro de temperaturas diarias en una ciudad o las respuestas a una encuesta sobre hábitos de consumo son datos. Estos sirven como base para identificar patrones, realizar comparaciones y construir modelos que representen aspectos de la realidad.

1.5.2 ¿Qué son las variables?

Las variables son características o atributos que pueden tomar diferentes valores entre los individuos de un conjunto o población. Representan las dimensiones específicas que se desean observar o analizar en un estudio.

Las variables se clasifican según su naturaleza:

1. **Cuantitativas:** Representan cantidades medibles, como la altura, el peso o los ingresos.
 - **Discretas:** Toman valores específicos, como el número de hijos en una familia.
 - **Continuas:** Pueden tomar cualquier valor dentro de un rango, como la temperatura o el tiempo.
2. **Cualitativas:** Representan categorías o cualidades, como el género, la nacionalidad o el color de los ojos.
 - **Nominales:** No tienen un orden intrínseco, como el estado civil.
 - **Ordinales:** Poseen un orden, como los niveles educativos.

Las variables permiten organizar y estructurar los datos para realizar análisis significativos. Por ejemplo, en un estudio sobre la educación, “nivel educativo” podría ser una variable cualitativa ordinal, mientras que “puntaje en una prueba” sería cuantitativa continua.

1.5.3 ¿Qué son los individuos?

Los individuos son las unidades básicas de observación o análisis dentro de una población. Pueden ser personas, objetos, organismos, eventos o entidades, dependiendo del contexto del estudio.

Por ejemplo:

- En un censo poblacional, los individuos son las personas encuestadas.
- En un experimento agrícola, los individuos podrían ser plantas o parcelas de tierra.
- En un análisis de mercado, los individuos son los productos o consumidores.

Cada individuo se caracteriza por tener valores específicos para las variables consideradas en el estudio. Por ejemplo, en una encuesta sobre hábitos alimenticios, un individuo (una persona) podría tener valores como edad = 30 años, género = femenino, y frecuencia de consumo de frutas = 5 veces por semana.

1.6 Primer acercamiento: ¿Cómo son los datos?

En el panorama actual, donde la información fluye en volúmenes masivos y de manera constante, el análisis de datos se ha convertido en un componente esencial para comprender fenómenos sociales. Sin embargo, en el ámbito de las ciencias sociales, persiste una brecha significativa entre las competencias cuantitativas necesarias para el análisis de datos y las herramientas empleadas por los investigadores. Esta brecha no solo es tecnológica, sino también epistemológica y cultural, lo que genera una resistencia hacia el desarrollo de habilidades cuantitativas.

El uso de programación en el análisis de datos es frecuentemente percibido como un campo exclusivo de disciplinas STEM (ciencia, tecnología, ingeniería y matemáticas). Esta percepción puede desmotivar a los científicos sociales, quienes a menudo ven estas competencias como ajenas a sus intereses y tradiciones. Sin embargo, esta distancia es un obstáculo tanto para la investigación académica como para el desarrollo profesional en un mundo donde las decisiones basadas en datos son cada vez más frecuentes.

A medida que las metodologías de análisis avanzan, la integración de enfoques cuantitativos se vuelve inevitable. La combinación de grandes volúmenes de datos (*big data*), tecnologías accesibles y herramientas estadísticas ha revolucionado la forma en que se abordan preguntas de investigación. Desde el análisis de redes sociales hasta el estudio de comportamientos colectivos, las ciencias sociales dependen cada vez más de métodos que requieren un manejo competente de datos y software especializado.

Sin embargo, el uso de herramientas básicas, como hojas de cálculo, limita el alcance y la profundidad de los análisis. Aunque son útiles en los primeros pasos, carecen de la capacidad para manejar datasets complejos, realizar análisis avanzados o automatizar procesos, lo que resulta en una dependencia de enfoques manuales y un posible sesgo en los resultados.

La resistencia al desarrollo de competencias cuantitativas no es solo un problema de acceso a tecnología, sino también de cultura académica. Durante décadas, las ciencias sociales han privilegiado enfoques cualitativos y narrativos, subestimando las posibilidades de las herramientas cuantitativas para enriquecer la comprensión de los fenómenos sociales. Esta resistencia cultural se traduce en una falta de programas educativos orientados al análisis de datos en ciencias sociales, perpetuando la brecha tecnológica.

Reconstruir esta relación requiere un cambio de perspectiva en varios niveles:

1. **Formación interdisciplinaria:** Incorporar competencias de programación, análisis de datos y estadística en la formación de los científicos sociales. Esto no implica abandonar los enfoques cualitativos, sino complementarlos con habilidades cuantitativas que amplíen su capacidad de análisis.
2. **Acceso a herramientas accesibles y avanzadas:** Herramientas como R, Python y software estadístico especializado permiten a los científicos sociales realizar análisis más complejos sin depender de costosos recursos. Estas herramientas deben integrarse en los currículos de las ciencias sociales, acompañadas de una pedagogía que enfatice su utilidad.
3. **Cambio de mentalidad hacia los datos:** Los datos no deben ser vistos solo como números, sino como representaciones de fenómenos humanos y sociales. Este enfoque humanista puede ayudar a reducir la distancia entre el científico social y los métodos cuantitativos, reconociendo que los datos también cuentan historias y reflejan dinámicas sociales significativas.
4. **Enfoque en la relevancia profesional:** Más allá de la academia, el análisis de datos es una habilidad esencial en sectores como el marketing, la política, la planificación urbana y el desarrollo comunitario. Los científicos sociales con competencias cuantitativas tienen una ventaja competitiva en un mercado laboral que valora cada vez más las decisiones basadas en datos.

Un primer acercamiento a los datos puede verse de esta manera. Para nuestro ejemplo vamos a pensar en un conjunto de 100 personas que recién ingresan a una universidad en Bogotá en 1990, indagamos el género y la edad de cada una de ellas. Estos datos se pueden ver de esta forma.

18.4	17.5	20.4	16.4	18.1	16.6	16.1	16.8	19	17.6
19	15.2	18.6	18.4	17.9	19	18	18	17.4	18
16.6	17.4	17.1	19.4	17.9	18.7	16.9	16.6	14.8	16.3
17.9	18.4	15.7	18.6	13.8	19.3	17.4	15.5	16.3	18
20.5	17.8	17.7	17.5	18.9	18.9	18.2	18.5	15.2	16.9
16.3	17	18.6	19.1	17.2	17.4	17.1	18	17.4	18.2
17.8	17.4	18.6	18.4	18	18.2	18.5	19.8	16.7	18.1
19	19.3	15.8	19.1	18.3	17.9	16.6	17	15.8	18.2
15.3	15.1	17.9	18.7	18.3	14.6	16.7	17.4	15.4	16.9
15.9	18.8	19	17.7	18.2	17.1	18.3	17.7	19.2	19.6

Figure 1.1: Figura 1: Edades de 100 personas coloreadas por género, siendo femeninos los rojos y masculinos los azules. Fuente: elaboración propia a partir de simulación.

Esta forma de ver los datos no nos entrega una información fácil de comprender. Por esta razón, la **visualización de datos** es importante en el análisis estadístico. Para comprender mejor, abordemos primero la edad. Podemos ver los datos de la edad como puntos en el eje horizontal.

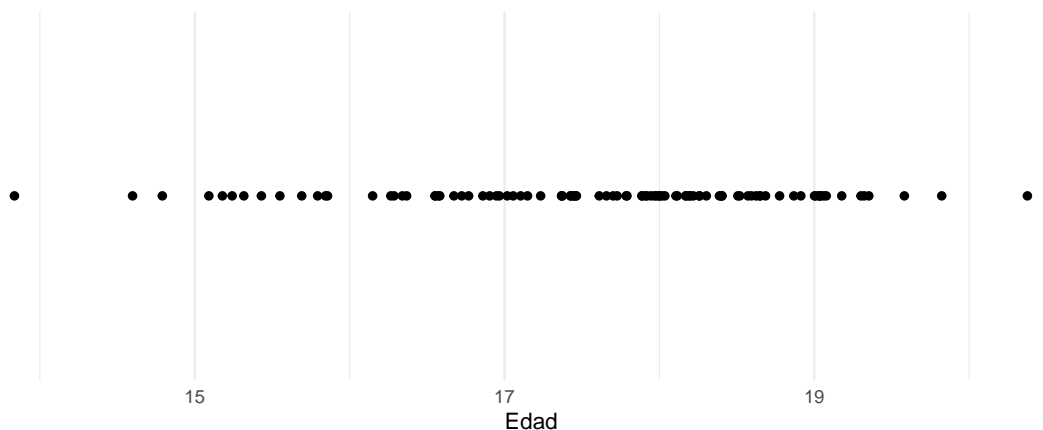


Figure 1.2: Figura 2: Edades de 100 personas en el eje horizontal. Fuente: elaboración propia a partir de simulación.

Esta visualización nos entrega un poco más de información, podemos entender el valor más alto, el más bajo, y en general el espacio que ocupan los datos en el eje horizontal. Los datos que ocupan mucho espacio se llaman **dispersos**, si el espacio es poco, se llaman **concentrados**. Más adelante veremos medidas para esta característica y profundizaremos al respecto. Para verlos mejor, podemos diseminarlos verticalmente. Este es un truco que ayuda a comprender

mejor los datos, impidiendo que se superpongan los puntos. Para esto se agrega ruido en el eje vertical, pero este ruido no tiene significado.

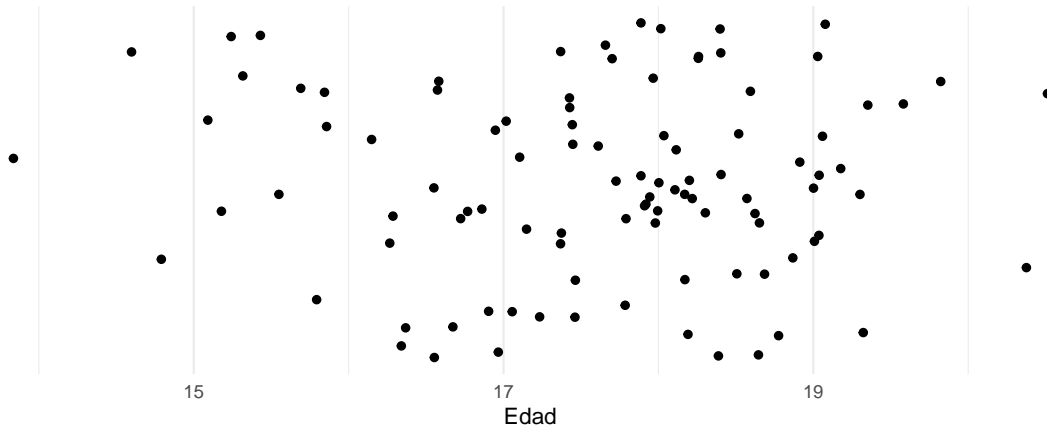


Figure 1.3: Figura 3: Edades de 100 personas en el eje horizontal con ruido vertical. Fuente: elaboración propia a partir de simulación.

Podemos trazar líneas imaginarias para entender mejor los datos. Las primeras líneas imaginarias son el máximo y el mínimo. Al trazar estas líneas podemos contener el 100% de los datos. Es muy fácil.

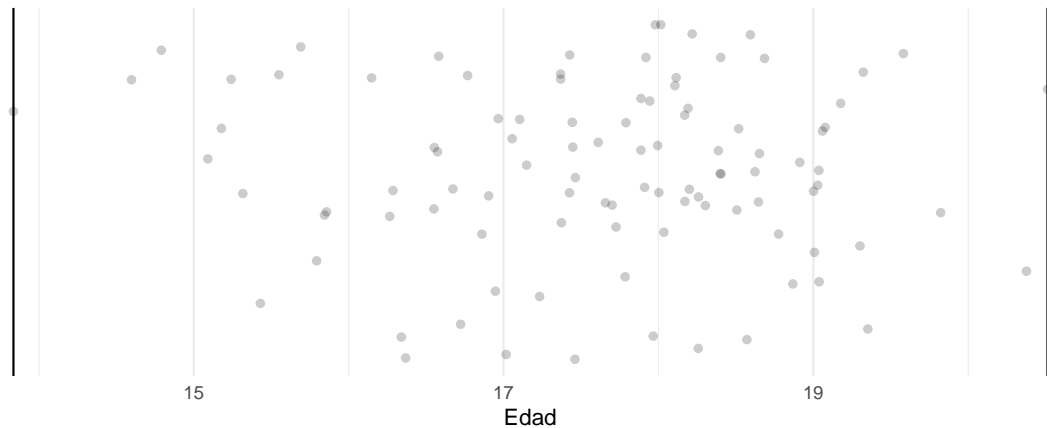


Figure 1.4: Figura 4: Edades de 100 personas en el eje horizontal con máximo y mínimo. Fuente: elaboración propia a partir de simulación.

divide los datos en dos conjuntos de igual magnitud. A la derecha de la línea se encuentra la misma cantidad de datos que a la izquierda. Esta línea se encuentra en un punto muy

importante del eje horizontal, este valor se denomina la mediana. La mediana de un conjunto de datos es el valor que divide a los datos en dos conjuntos de igual magnitud.

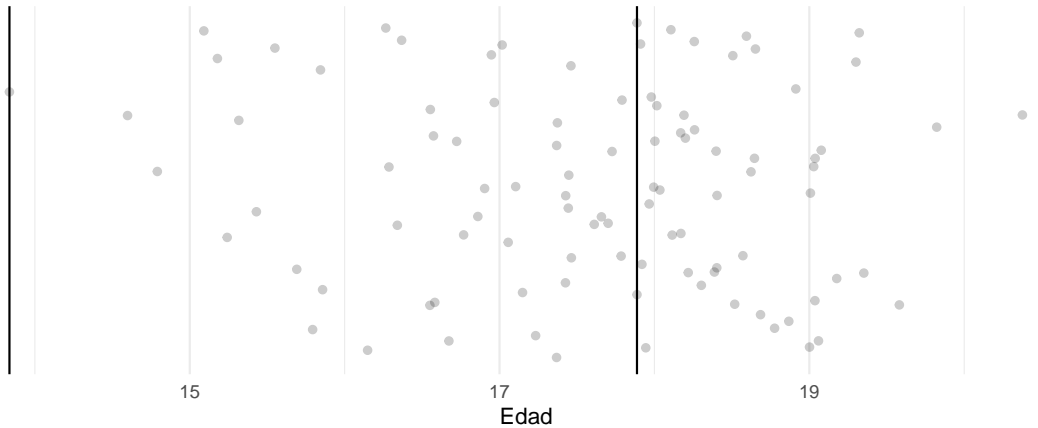


Figure 1.5: Figura 5: Edades de 100 personas en el eje horizontal con su mediana. Fuente: elaboración propia a partir de simulación.

Usando más líneas imaginarias podemos dividirlos en cuatro partes iguales. Estas líneas imaginarias distribuyen los datos de la edad en cuatro conjuntos de igual magnitud. Al igual que la mediana, estos valores son importantes. Se denominan cuartiles. En cada uno de los conjuntos resultantes, se encuentra el 25% de los datos.

Entonces, funciona de la siguiente forma: el cuartil cero (Q_0) corresponde al valor mínimo; el primer cuartil (Q_1) separa el 25% de los datos; el segundo cuartil (Q_2) coincide con la mediana, porque separa el 50% de los datos; el tercer cuartil (Q_3) separa el 75% de los datos; y el cuarto cuartil (Q_4) coincide con el máximo de los datos.

Estos cuartiles en general no tienen la misma distancia entre ellos. Lo usual es que se ubiquen en distancias diferentes según los datos.

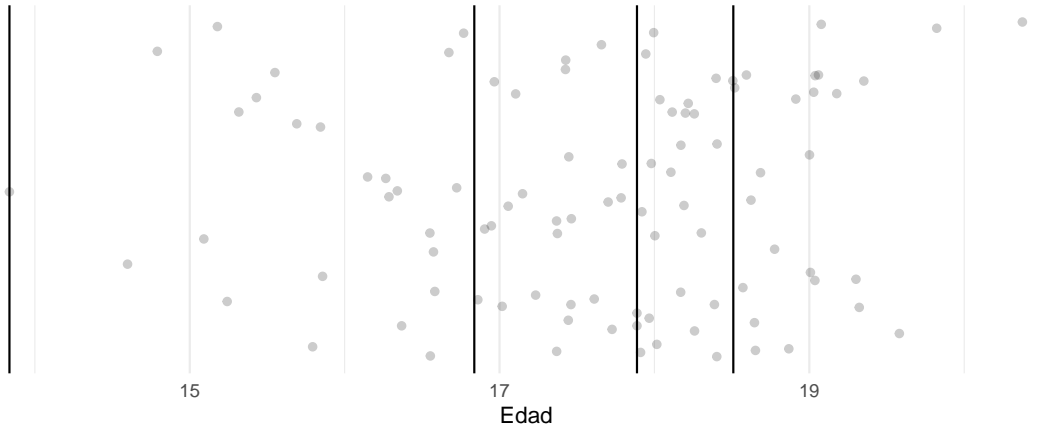


Figure 1.6: Figura 6: Edades de 100 personas en el eje horizontal con cuartiles. Fuente: elaboración propia a partir de simulación.

Estas líneas imaginarias que son importantes, se pueden consolidar en un solo gráfico, que se denomina gráfico de caja y bigotes. Este gráfico está conformado por la mediana y los cuartiles. Este es un gráfico esencial en el análisis de datos y lo vamos a ver en muchas investigaciones.

Este gráfico tiene un cambio con respecto a la construcción anterior: aquí se utilizan un máximo teórico y un mínimo teórico. Esto se realiza con el fin de identificar visualmente los datos de los extremos.

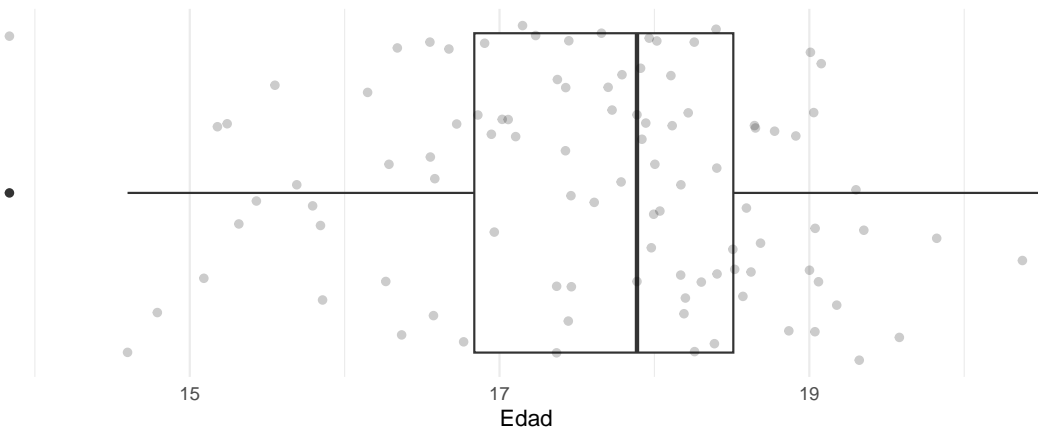
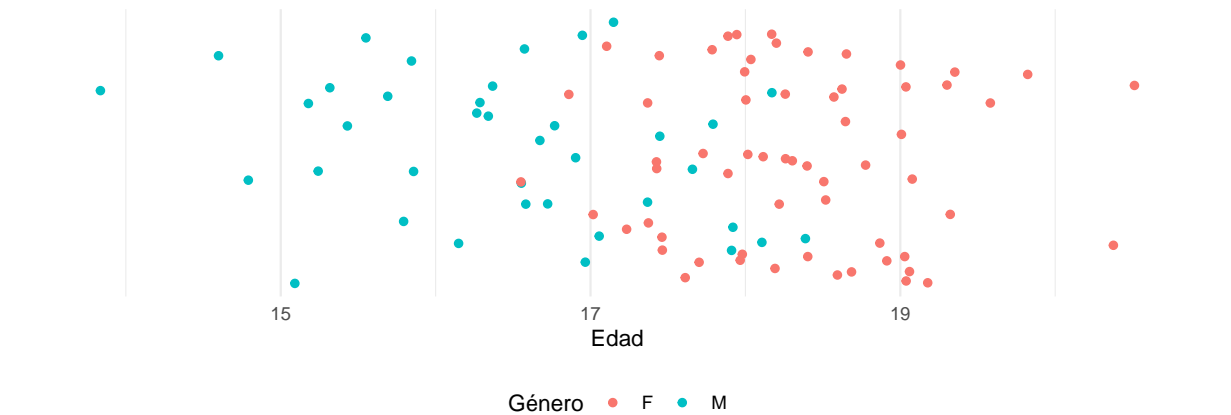


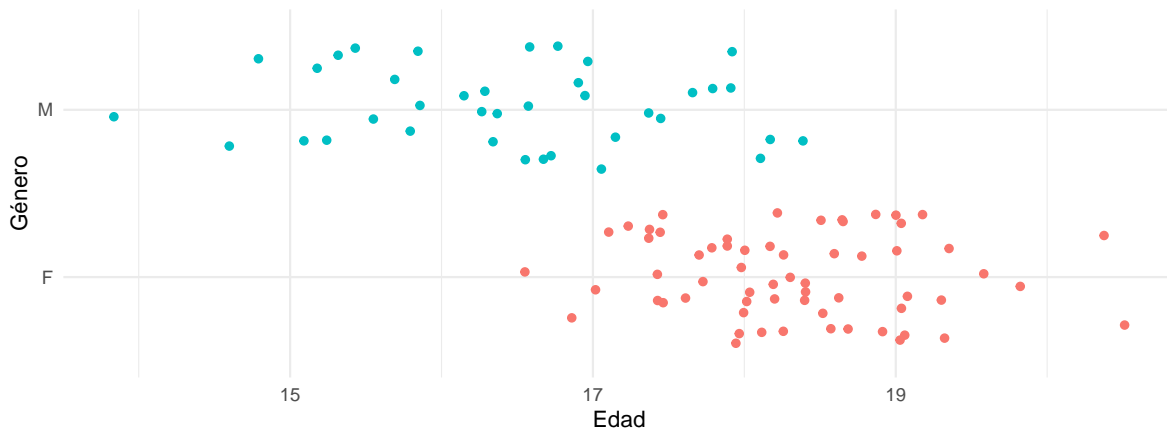
Figure 1.7: Figura 7: Edades de 100 personas, gráfico de caja y bigotes. Fuente: elaboración propia a partir de simulación.

Abordemos el género ahora. Podemos usar el color para identificar el género en los datos. Realizamos el mismo procedimiento añadiendo el color del género. En este caso, ya podemos

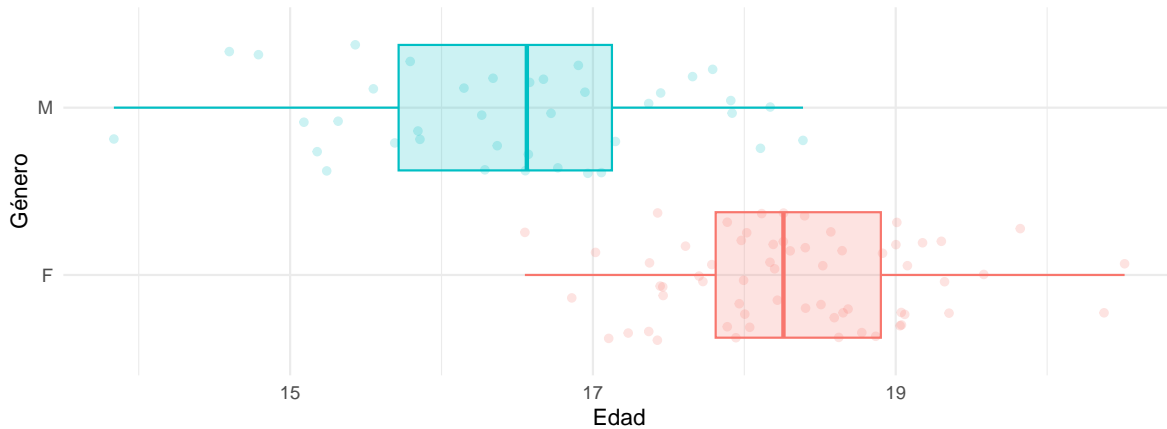
identificar la tendencia, de puntos rojos más a la derecha y azules a la izquierda.



Organizamos los datos verticalmente por género. Esto os permite tener dos nubes de puntos y facilita la interpretación, ahora es más notoria la tendencia hacia a la derecha y hacia la izquierda de los puntos rojos y azules respectivamente.



Al elaborar una gráfico de caja para cada género es posible ver la tendencia. Esta característica que no resultaba fácil de identificar en el primer gráfico, ahora es muy notoria. Las medianas y los cuartiles de los datos agrupados por género difieren. Esto nos permite obtener hipótesis que podremos comprobar más adelante.



1.7 Tipos de datos

El análisis de datos es la columna vertebral de la investigación científica y social, donde se busca extraer patrones, relaciones y tendencias a partir de observaciones sistemáticas. Los datos son el insumo principal de este análisis, y comprender sus diferentes tipos y propiedades es esencial para elegir las metodologías y enfoques más adecuados. A continuación, se profundiza en los conceptos clave presentados, estructurándolos en torno a su relevancia en la investigación.

1.7.1 Tendencias y patrones globales

Las tendencias representan comportamientos agregados que emergen de las características individuales de una población. A diferencia de una afirmación puntual o específica sobre un individuo, las tendencias buscan capturar patrones generalizables que describen cómo un grupo, en promedio, se comporta respecto a una variable o conjunto de variables. Este enfoque es crucial en disciplinas como la economía, la sociología y la epidemiología, donde los fenómenos globales informan decisiones de política pública o estrategias organizacionales.

Por ejemplo, al analizar los hábitos de consumo en un país, una tendencia podría revelar que el promedio de gasto en entretenimiento es más alto en grupos jóvenes que en mayores de 50 años. Sin embargo, esta tendencia no asegura que todos los jóvenes gasten más en entretenimiento, sino que describe un comportamiento predominante dentro de la población estudiada. Este desajuste entre las tendencias y los comportamientos individuales resalta la importancia de interpretar los datos globales como representativos, pero no absolutos.

1.7.2 Unidades de análisis

Los individuos son las unidades básicas de observación en cualquier estudio. Representan los elementos de la población objeto de análisis, que pueden ser personas, empresas, organismos o eventos, dependiendo del contexto de la investigación. En términos analíticos, los individuos son la fuente de datos a partir de los cuales se construyen modelos, hipótesis y conclusiones.

La elección de los individuos y su caracterización es clave, ya que determina la validez y generalizabilidad de los resultados. Por ejemplo, en un estudio sobre hábitos alimenticios, seleccionar individuos representativos de diferentes regiones y estratos socioeconómicos garantizará que las conclusiones puedan extrapolarse al resto de la población. Este concepto está íntimamente ligado al diseño del muestreo y la operacionalización de las variables.

1.8 Variables: Características observables y medibles

Las variables son los rasgos o atributos que describen a los individuos y sobre los cuales se recopila información. Funcionan como el puente entre las observaciones individuales y los análisis que buscan generalizar comportamientos.

1.8.1 Clasificación de las variables

1. **Cuantitativas:** Estas variables representan magnitudes numéricas que se pueden medir. Por otro lado, las variables numéricas o cuantitativas permiten medir y realizar cálculos matemáticos. La distinción entre discretas y continuas tiene implicaciones metodológicas: mientras que las discretas suelen analizarse con conteos y proporciones, las continuas requieren herramientas que consideren distribuciones más complejas. Estas variables son esenciales en áreas como la economía y la biología, donde los fenómenos físicos y financieros se representan mediante medidas precisas.
 - **Discretas:** Toman valores finitos o contables, como el número de hijos en una familia.
 - **Continuas:** Admiten infinitos valores dentro de un rango, como la temperatura medida en grados Celsius.
2. **Categóricas:** Representan atributos o categorías que no necesariamente poseen un valor numérico. Las variables categóricas agrupan a los individuos en categorías o clases. Su análisis permite identificar frecuencias y distribuciones dentro de la población. Por ejemplo, analizar la afiliación política de un grupo puede mostrar que el 40% prefiere un partido A, mientras que el 30% opta por un partido B. Este tipo de variable es fundamental en estudios demográficos y de percepción social, donde las características subjetivas y de identidad son el foco principal.

- **Nominales:** Categorías sin orden intrínseco, como el color de los ojos o el estado civil.
- **Ordinales:** Categorías con un orden lógico, como los niveles educativos o el nivel de satisfacción.

1.8.2 Medición: Asignación de valores a las observaciones

La medición es el proceso mediante el cual las características de los individuos se transforman en datos observables y cuantificables. Este proceso debe ser riguroso y consistente, basado en reglas preestablecidas que garanticen la comparabilidad y la reproducibilidad de los resultados.

Por ejemplo, en un estudio sobre bienestar subjetivo, es esencial que las escalas utilizadas para medir “felicidad” o “satisfacción” estén claramente definidas, estandarizadas y validadas. Una medición precisa y confiable no solo facilita el análisis estadístico, sino que también asegura que las conclusiones se basen en datos sólidos y significativos.

Los tipos de datos, las variables y las técnicas de medición constituyen los elementos centrales del análisis en cualquier disciplina. Entender cómo se relacionan las tendencias globales con las características individuales, cómo se seleccionan y clasifican las variables, y cómo se mide con precisión cada atributo son habilidades fundamentales para el análisis riguroso. Este marco integrado no solo es esencial para la investigación académica, sino también para la toma de decisiones informadas en contextos profesionales y sociales. La reconceptualización del trabajo con datos requiere, por tanto, no solo conocimiento técnico, sino también una reflexión crítica sobre las implicaciones de cómo estructuramos y utilizamos la información.

2 Preparación de los datos

2.1 Imputación de datos

2.2 Identificación de atípicos

2.3 De categórico a numérico

2.4 De numérico a categórico

2.5 Trabajo con fechas

3 Estadística descriptiva para una variable

3.1 Medidas de frecuencia

En el análisis de datos dentro de las ciencias sociales, las medidas de frecuencia permiten describir de manera precisa la distribución de variables **categorías** y ayudan a comprender patrones generales en poblaciones o muestras. Estas medidas son esenciales para resumir y comunicar la información de forma efectiva, aportando una base sólida para el análisis estadístico posterior.

3.1.1 Frecuencia

La **frecuencia absoluta** es el conteo del número de veces que aparece una categoría específica en un conjunto de datos. Es la medida más básica de estadística descriptiva y es crucial para comprender la distribución de variables cualitativas. Las categorías en un conjunto de datos representan respuestas a preguntas de encuestas, atributos de los participantes o clasificaciones en estudios sociales.

Tip

Ejemplo aplicado a las ciencias sociales: En un estudio sobre preferencias políticas, los encuestados pueden expresar su afinidad por diferentes partidos. Si en una muestra de 500 personas, 150 apoyan el partido A, 200 el partido B y 150 no apoyan a ningún partido, la frecuencia de cada categoría sería:

- **Partido A:** 150
- **Partido B:** 200
- **Ninguno:** 150

Este conteo ayuda a determinar cuál es el partido con mayor número de seguidores y, por ende, cuál tiene una posición predominante en la muestra.

3.1.2 Proporción

La **proporción** se utiliza para expresar la frecuencia relativa de una categoría con respecto al total de observaciones, facilitando la comparación entre diferentes grupos de datos. Se calcula dividiendo la frecuencia de una categoría por el total de observaciones.

Warning

Fórmula:

$$p_i = \frac{\#freq_i}{n}$$

Donde:

- p_i es la proporción.
- $\#freq_i$ es la frecuencia de la categoría específica (i -ésima).
- n es el número total de observaciones.

Tip

Ejemplo contextualizado: Supongamos que en una encuesta sobre el uso de redes sociales, se encontró que 320 de 800 personas utilizan redes sociales diariamente. La proporción de usuarios diarios es:

$$p_{rs} = \frac{320}{800} = 0.40 \text{ o } 40\%$$

Este resultado indica que el 40% de los encuestados son usuarios diarios de redes sociales, un hallazgo importante para estudios sobre comportamiento digital y hábitos de consumo en la sociedad.

Las frecuencias relativas suelen presentarse en porcentajes. Un **porcentaje** es una forma de expresar una fracción o proporción sobre un total de 100, facilitando la comparación y comprensión de datos en contextos variados. Se utiliza ampliamente en estadísticas, finanzas y análisis cuantitativo para representar partes de un todo. La palabra “porcentaje” proviene del latín “per centum,” que significa “por cada cien.” Su uso se remonta a la antigua Roma, donde los comerciantes y cobradores de impuestos empleaban fracciones y cálculos semejantes al porcentaje moderno para facilitar transacciones y registros contables. A lo largo de los siglos, el concepto se perfeccionó, especialmente durante el Renacimiento, con la expansión del comercio y la necesidad de representar proporciones de forma estandarizada, llevando al símbolo “%” que hoy es universal.

3.1.3 Moda

La **moda** es la categoría con mayor frecuencia en un conjunto de datos y representa el valor más común o frecuente. Esta medida es particularmente útil en estudios sociales cuando se analizan características como la ocupación, el nivel de estudios o las opiniones sobre políticas públicas. La moda permite identificar tendencias dominantes o respuestas mayoritarias en la población estudiada.

Tip

Ejemplo aplicado: Imaginemos un estudio que investiga la ocupación principal de los trabajadores en una ciudad. Si los datos muestran que, de 1,000 encuestados, 450 trabajan en el sector servicios, 300 en el sector manufacturero y 250 en el sector tecnológico. En este caso la moda es “Sector servicios”

Este dato resalta que la ocupación predominante en la muestra es el sector servicios, una observación relevante para estudios de desarrollo económico y planificación urbana.

3.1.4 Nota sobre variables continuas

La moda tiene aplicaciones limitadas en variables continuas, como el ingreso o la altura, ya que estas tienden a no repetir valores exactos con frecuencia significativa. Sin embargo, en ciertas investigaciones sociales, agrupar datos en rangos puede hacer que la moda sea más útil. Por ejemplo, si se clasifica el ingreso en intervalos como “menos de \$10,000”, “\$10,001-\$20,000”, ..., etc., puede ser posible identificar una moda representativa del grupo con mayor frecuencia.

Tip

Ejemplo adaptado: Si se analizan los ingresos de 500 hogares y el intervalo “\$10,001-\$20,000” tiene la mayor cantidad de hogares (150), entonces la moda del ingreso agrupado es ese intervalo específico. Esta información puede ser crucial para entender el nivel socioeconómico predominante en un área geográfica y para la formulación de políticas de asistencia económica.

3.2 Medidas de tendencia central

Las **medidas de tendencia central** son estadísticas descriptivas que representan el valor típico o central de un conjunto de datos. Estas medidas permiten comprender dónde se encuentra el “centro” de una distribución y son fundamentales en la investigación cuantitativa en ciencias sociales, donde se estudian fenómenos como ingresos, opiniones o resultados de encuestas.

3.2.1 Media aritmética

La **media aritmética**, comúnmente conocida como promedio, es la medida de tendencia central más utilizada. Se obtiene sumando todos los valores de un conjunto de datos y dividiendo el resultado por el número total de observaciones. La media es ideal para describir conjuntos de datos simétricos y es especialmente útil en estudios de ciencias sociales cuando se necesita resumir características cuantitativas como la edad, el ingreso o la puntuación de una encuesta.

Warning

Fórmula de la media:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Donde:

- \bar{x} es la media,
- n es el número de observaciones,
- x_i representa cada valor individual del conjunto de datos.

Tip

Ejemplo en ciencias sociales: Supongamos que estamos estudiando los ingresos mensuales en una comunidad y los datos obtenidos en dólares son: 1,200, 1,500, 1,800, 2,000 y 20,000. La media de estos ingresos es:

$$\bar{x} = \frac{1,200 + 1,500 + 1,800 + 2,000 + 20,000}{5} = \frac{26,500}{5} = 5,300$$

Aunque la media es 5,300, este valor puede no representar adecuadamente la distribución, ya que un ingreso atípico de 20,000 distorsiona el promedio.

3.2.2 Mediana

La **mediana** es la medida de tendencia central que divide un conjunto de datos ordenados en dos partes iguales, es decir, la mitad de los datos está por debajo y la otra mitad por encima de la mediana. Esta medida es más resistente a los valores extremos que la media, por lo que es preferida en distribuciones sesgadas o con outliers.



Tip

Procedimiento para encontrar la mediana:

- Si el número de observaciones es impar, la mediana es el valor central.
- Si es par, la mediana es el promedio de los dos valores centrales.



Tip

Ejemplo aplicado: Si analizamos el número de hijos en familias de una comunidad y los datos ordenados son: 1, 2, 2, 3 y 10, la mediana es 2, ya que es el valor del medio.

3.2.3 Otras medidas de tendencia central

3.2.3.1 Media armónica

La **media armónica** es útil para conjuntos de datos que representan tasas o razones, ya que pondera los valores de forma que las observaciones más pequeñas tengan un mayor impacto. Se define como el inverso del promedio de los inversos de los valores.



Warning

Fórmula de la media armónica:

$$\bar{x}_{\text{arm}} = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}$$

Donde:



Tip

Ejemplo en ciencias sociales: Supongamos que estudiamos la eficiencia de diferentes métodos de transporte urbano en términos de tiempo por viaje. Si los tiempos por viaje (en minutos) son 10, 15 y 20, la media armónica es:

$$\bar{x}_{\text{arm}} = \frac{3}{\frac{1}{10} + \frac{1}{15} + \frac{1}{20}} \approx 13.04$$

Esta media pondera más los viajes cortos y es útil en análisis donde las tasas individuales son significativas.

3.2.3.2 Media geométrica

La **media geométrica** es apropiada para datos que representan crecimiento proporcional o tasas de cambio, como el crecimiento de la población o el rendimiento económico. Es el (n)-ésimo raíz del producto de todos los valores.

Warning

Fórmula de la media geométrica:

$$\bar{x}_{\text{geo}} = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}$$

Tip

Ejemplo en ciencias sociales: Si una población crece anualmente a tasas de 1.05, 1.10 y 1.08, la media geométrica del crecimiento es:

$$\bar{x}_{\text{geo}} = \sqrt[3]{1.05 \times 1.10 \times 1.08} \approx 1.076$$

Lo que implica un crecimiento promedio anual del 7.6%.

3.2.3.3 Comparación de las medidas

En resumen, la **media aritmética** es ideal para datos simétricos y sin valores extremos, la **mediana** es preferible en distribuciones sesgadas o con valores atípicos, y las **medias armónica y geométrica** se utilizan en contextos específicos relacionados con tasas o datos multiplicativos. Estas medidas permiten a los investigadores de las ciencias sociales interpretar adecuadamente los datos y tomar decisiones basadas en análisis cuantitativos sólidos.

3.3 Medidas de localización

3.3.1 Mínimo y máximo

El mínimo es el valor más bajo en el conjunto de datos y el máximo es el valor más alto. En estudios sociales, pueden utilizarse para detectar posibles extremos en las variables, como el ingreso o la edad.

3.3.2 Cuantiles

Los cuantiles dividen los datos en partes iguales. En estudios sociales, los cuantiles son útiles para evaluar la distribución de ingresos o educación. Por ejemplo, el cuartil más bajo (25%) representa el grupo con los ingresos más bajos.

 Warning

$$Q_{\alpha}(x) \sim X_{(\alpha)}$$

3.3.3 Percentiles

Los percentiles son un caso especial de cuantiles y se usan para ver la posición de un valor dentro de una distribución. En educación, el percentil 90 indica que un estudiante superó al 90% de sus compañeros en una prueba estandarizada.

3.3.4 Cuartiles

Los cuartiles dividen los datos en cuatro partes. El primer cuartil (Q1) es el 25% más bajo, y el tercer cuartil (Q3) es el 25% más alto. Estos son útiles para evaluar la dispersión de los ingresos dentro de una población.

3.4 Medidas de dispersión

3.4.1 Varianza

La varianza mide qué tan dispersos están los datos respecto a la media. En estudios de desigualdad de ingresos, una alta varianza indicaría grandes disparidades entre los ingresos de las personas.

 Warning

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



Tip

Ejemplo: En una muestra de ingresos de 2000, 2500, 3000, 3500 y 4000, la varianza nos indica cuánto se alejan estos valores de la media.

3.4.2 Desviación estándar

La desviación estándar es la raíz cuadrada de la varianza y nos proporciona una medida de dispersión en las mismas unidades que los datos originales. Por ejemplo, si la desviación estándar de los ingresos en una población es alta, indica que hay una gran variabilidad en los ingresos.



Warning

$$S_x = \sqrt{S_x^2}$$

3.4.3 Rango intercuartílico (IQR)

El rango intercuartílico es la diferencia entre el tercer cuartil (Q3) y el primer cuartil (Q1). Mide la dispersión central de los datos y es útil para evitar que los valores extremos influyan en la interpretación.



Warning

$$ran(x) = Q_3(x) - Q_1(x)$$

Ejemplo: En una encuesta de satisfacción con el gobierno, el IQR podría mostrar la variación en las respuestas del 50% central, ignorando los valores más extremos de descontento o satisfacción.

3.4.4 Rango

El rango es la diferencia entre el valor máximo y el mínimo. Aunque es fácil de calcular, puede verse afectado por valores atípicos. Por ejemplo, en el análisis de ingresos, el rango puede ser muy amplio si hay pocos individuos con ingresos extremadamente altos.



Warning

$$Rango = \max(x) - \min(x)$$

3.5 Ejercicios y actividades

3.5.1 Hablemos bien

No diga: la gente votó **en promedio** por el candidato X

Diga: la gente votó en **mayor proporción** por el candidato X

3.5.2 Actividad

No diga:

- Esta alternativa es muy buena, nos ahorra menos dolores de cabeza
- La edad promedio de los estudiantes está entre 15 y 20 años.
- En total, uno de cada tres estudiantes no sabe estadística.

Analice

- Más de la mitad de los estudiantes que presentaron la prueba saben leer por debajo del promedio del puntaje de inglés.
- El 70% de los colombianos tienen ingresos por debajo de la media.

4 Correlación

Estadística descriptiva

```
## Reescribir bien
```

5 Dos variables

Es posible medir la relación entre dos variables, pero esto depende de qué tipo de variables son.

- Dos cualitativas
- Dos cuantitativas
- Una cuantitativa y una cualitativa

5.1 Dos variables cualitativas

Para poder medir la relación entre dos variables de tipo cualitativo usamos tablas de contingencia.

5.2 Dos variables cuantitativas

La relación entre dos variables cuantitativas puede ser medida con los siguientes estadísticos.

- Covarianza
- Correlación de Pearson
- Correlación de Spearman

5.3 Covarianza

Covarianza entre dos variables

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

5.4 Correlación de Pearson

$$\rho_{xy} = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{(n-1) S_x S_y}$$

5.5 Correlación de Spearman

$$\rho = 1 - \frac{6 \sum D^2}{n(n^2 - 1)}$$

donde D es la diferencia entre los correspondientes estadísticos de orden de x - y. N es el número de parejas de datos.

5.6 Una variable cualitativa y otra cuantitativa

Para medir asociación entre una variable cuantitativa y otra cualitativa se puede desagregar la media.

5.7 Correlación

¿Qué es la causalidad?

6 Distribuciones

7 Visualización

Estadística descriptiva

8 Visualización

Iniciamos el tema de visualización de datos con una lectura refrescante: [Gráficas para la ciencia y ciencia para las gráficas](#)

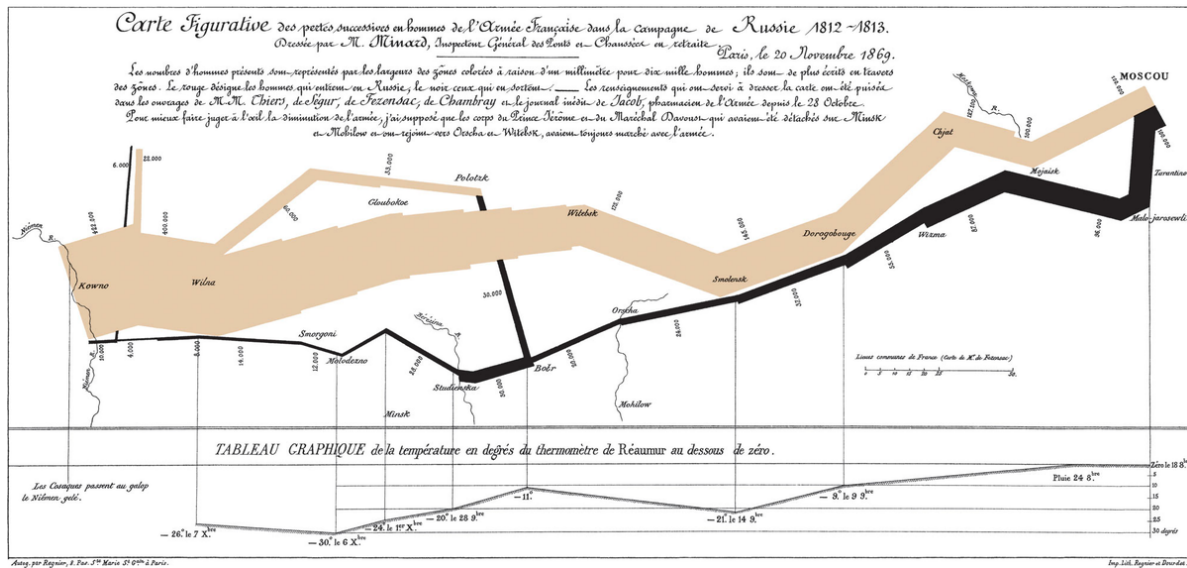
8.1 ¿Por qué visualización?

- [El cuarteto de Anscombe](#)
- [Seeing theory](#)

8.2 Veamos ejemplos

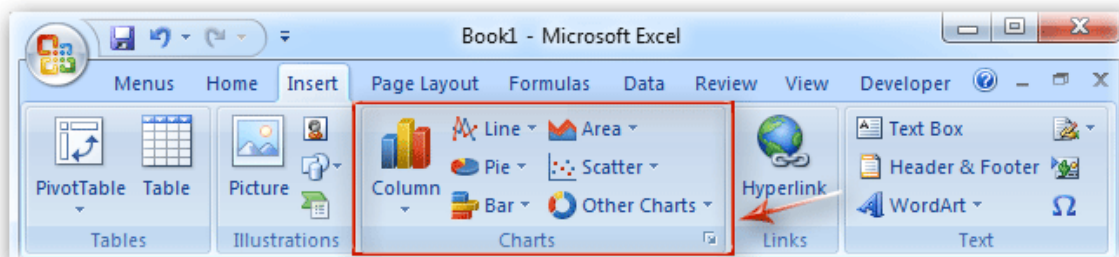
- [Information is beautiful](#)
- [gapmminder](#)
- [shiny:](#)
- [plotly](#)
- [Beter Life Index OECD](#)
- [Top 10 Countries by Inflation Rate \(1980-2018\)](#)

8.3 El ejemplo de Napoleón



Fuente: [Data visualization education using the storytelling with Minard's figurative map](#)

8.4 Gráficos tradicionales



8.5 Escala de Cleveland y McGill



8.6 Buenas prácticas de visualización

Buenas prácticas de la visualización - Víctor Pascual

8.7 Gráficos

Los gráficos más comunes utilizados en datos son los siguientes.

1. **Gráfico de barras**: útil para comparar categorías. Variación: **gráfico de lollipop**, donde se usan puntos conectados por líneas en lugar de barras.
2. **Gráfico de torta (pastel)**: muestra proporciones. Variación: **gráfico de dona**, que es similar pero con un espacio vacío en el centro.
3. **Histograma**: representa la distribución de frecuencias de una variable cuantitativa. Variación: **gráfico de densidad**, que suaviza las frecuencias en una curva continua para mostrar la distribución de los datos.
4. **Gráfico de dispersión**: muestra la relación entre dos variables cuantitativas. Variación: **gráfico de jitter**, que separa los puntos amontonados para revelar la densidad de los datos.

5. **Boxplot (diagrama de caja y bigotes)**: resume la distribución de una variable. Variación: **gráfico de violín**, que añade una visualización de la densidad en ambos lados del gráfico.
6. **Gráfico de líneas**: útil para visualizar tendencias a lo largo del tiempo. Variación: **gráfico de áreas**, donde el área bajo la línea está sombreada, destacando la magnitud.
7. **Mapa de calor (heatmap)**: visualiza patrones de datos a través de variaciones de color. Variación: **clustered heatmap**, que agrupa los datos por similitud, facilitando la interpretación de patrones.
8. **Gráfico de burbujas**: similar al gráfico de dispersión, pero con el tamaño de las burbujas que representa una tercera variable.
9. **Gráfico de radar (o de araña)**: muestra múltiples variables radiales. Variación: **gráfico de radar de área**, que sombrea el área debajo de los valores para enfatizar la comparación entre categorías.

Part II

Estadística inferencial

Part III

Contenido

En esta primera sección se examina la estadística descriptiva. El contenido ha sido seleccionado cuidadosamente con el fin de agregar valor a los análisis cuantitativos que se proponen al interior de las investigaciones en ciencias sociales.

Introducción

Este primer tema introduce a los estudiantes en los conceptos básicos de la estadística inferencial..

Intervalos de confianza

Pruebas de hipótesis paramétricas

Pruebas de hipótesis no paramétricas

9 Construyendo un marco epistemológico para la inferencia estadística

Estadística inferencial

Para comprender cómo surgió la inferencia estadística y en qué se fundamenta, es necesario entender primero las bases epistemológicas de la ciencia. Los métodos que hoy empleamos en estadística, especialmente en inferencia, tienen raíces profundas en la filosofía clásica y el desarrollo del método científico. A través de la deducción, la inducción y otras formas de razonamiento, los científicos han perfeccionado métodos para obtener conocimientos que sean precisos y replicables.

10 Los antiguos griegos y la deducción

10.1 Sócrates y la mayéutica

Sócrates fue uno de los primeros filósofos en enfatizar la importancia del cuestionamiento como herramienta para alcanzar la verdad. Su método, la **mayéutica**, consistía en formular preguntas para ayudar a su interlocutor a descubrir conocimientos por sí mismo, partiendo de sus propias creencias y explorando las inconsistencias en sus respuestas. A través del diálogo y la introspección, Sócrates buscaba llevar a los demás hacia una mejor comprensión de conceptos abstractos como la justicia, la verdad y el bien. Esta metodología sentó las bases para el pensamiento crítico, un pilar fundamental en la ciencia moderna.

10.2 Platón y la dialéctica

Platón, discípulo de Sócrates, amplió la mayéutica y formuló la **dialéctica** como un método para alcanzar conocimientos más profundos mediante la confrontación de ideas opuestas. A través del diálogo y la tensión entre las tesis y antítesis, Platón creía que se podía llegar a la síntesis, es decir, a una comprensión superior y más completa de la realidad. Este método dialéctico influyó en el desarrollo de sistemas de lógica y pensamiento analítico que aún sustentan la base epistemológica de la ciencia.

10.3 Aristóteles y la lógica

Aristóteles sistematizó la lógica como un método de razonamiento para llegar a conclusiones válidas a partir de premisas establecidas. En sus obras, como el **Organon**, formalizó el uso de la lógica deductiva y desarrolló una metodología para analizar y entender los principios subyacentes de los fenómenos. La lógica aristotélica no solo sentó las bases para el razonamiento científico, sino que también proporcionó las herramientas para la creación de sistemas de clasificación y el desarrollo de conceptos abstractos en la ciencia y las matemáticas.

10.4 El silogismo

Uno de los aportes más significativos de Aristóteles a la epistemología es el **silogismo**, una forma de razonamiento deductivo que permite derivar conclusiones a partir de dos o más premisas. El silogismo establece que si las premisas son verdaderas, la conclusión necesariamente debe serlo. Este tipo de razonamiento deductivo es un modelo de inferencia lógica que ha servido de base para el desarrollo de sistemas matemáticos y estadísticos. Un ejemplo clásico sería:

- Todos los hombres son mortales.
- Sócrates es un hombre.
- Por lo tanto, Sócrates es mortal.

Un ejemplo del uso del mecanismo deductivo es el segundo libro más editado de la historia, *Los Elementos* de Euclides, que organiza el conocimiento geométrico mediante un sistema axiomático. En este texto, Euclides parte de unos pocos postulados y axiomas fundamentales, a partir de los cuales deduce rigurosamente una serie de teoremas y proposiciones. Este enfoque deductivo no solo demostró la efectividad de la lógica en las matemáticas, sino que también influyó profundamente en la metodología científica, sirviendo de modelo para estructurar el conocimiento de manera lógica y coherente.

Otros ejemplos son *Ética demostrada según el orden geométrico* de Spinoza y *Philosophiæ Naturalis Principia Mathematica* de Isaac Newton. En *Ética*, Spinoza estructura su filosofía siguiendo el estilo geométrico de Euclides, utilizando definiciones, axiomas y proposiciones para desarrollar sus ideas sobre la naturaleza de Dios, la mente y la moralidad. Por su parte, en *Principia Mathematica*, Newton aplica un razonamiento deductivo para establecer las leyes del movimiento y la gravitación universal, partiendo de principios fundamentales y llegando a conclusiones que explican fenómenos físicos observables. Estos textos muestran cómo el método deductivo ha sido un pilar para avanzar en diversas disciplinas, desde la filosofía hasta la física.

11 El *Novum Organum* de Bacon

11.1 La deducción contra la inducción

En el siglo XVII, Francis Bacon introdujo un enfoque revolucionario en su obra *Novum Organum*, en la que defendía la **inducción** como método para el conocimiento científico. Este enfoque rompió con la tradición aristotélica de deducción estricta, proponiendo que, en lugar de solo partir de premisas generales, los científicos deberían observar y analizar los fenómenos específicos para, a partir de ellos, generalizar leyes y principios.

- **La deducción:** es un proceso de razonamiento que va de lo general a lo particular. Parte de leyes o teorías ya establecidas y aplica esas premisas para llegar a conclusiones específicas. La deducción asegura conclusiones válidas si las premisas son verdaderas, pero no permite descubrir nuevas leyes o principios.
- **La inducción:** es el proceso de observación de casos particulares para generar conclusiones generales o teorías. En la inducción, el conocimiento se construye a partir de patrones observados en la realidad, permitiendo la creación de nuevas hipótesis y teorías. Sin embargo, este método no garantiza la certeza absoluta de sus conclusiones, ya que estas son probabilísticas y dependen de la representatividad de los datos.

El trabajo de Bacon es fundamental porque sentó las bases para una ciencia basada en la observación empírica, un enfoque que siglos más tarde sería crucial en la inferencia estadística.

11.2 Actividad: ¿Qué es un cisne negro?

El concepto de “cisne negro” se refiere a eventos altamente improbables e impredecibles, pero con un gran impacto cuando ocurren. La expresión fue popularizada por el filósofo Nassim Nicholas Taleb y subraya la limitación de los métodos inductivos, ya que una amplia observación de cisnes blancos no garantiza que no existan cisnes negros. Este concepto es clave para entender los límites de la inferencia estadística y la probabilidad, pues resalta la posibilidad de eventos fuera de nuestras expectativas basadas en observaciones pasadas.

12 Fisher, Neyman, Pearson

12.1 Las reglas para hacer inducción

Ronald A. Fisher, Jerzy Neyman y Egon Pearson fueron fundamentales para estructurar las **reglas para hacer inducción** en el contexto de la estadística moderna. Su trabajo permitió la formalización de métodos inferenciales que ayudan a generalizar conclusiones a partir de muestras. Estas reglas establecen la estructura de las pruebas de hipótesis y la generación de intervalos de confianza, permitiendo a los científicos tomar decisiones con base en evidencia empírica.

12.2 Generación de conocimiento a partir de datos

Fisher, Neyman y Pearson desarrollaron metodologías para derivar conocimiento a partir de datos de manera rigurosa, incorporando conceptos como la probabilidad y el error estadístico. A través de la estadística inferencial, lograron definir un proceso sistemático para probar hipótesis, medir la incertidumbre y proporcionar intervalos de confianza, contribuyendo significativamente a las ciencias experimentales y sociales.

12.3 Inferencia estadística

La inferencia estadística surgió en el siglo XX como una disciplina clave en la estadística, impulsada por la necesidad de tomar decisiones informadas a partir de datos. Su desarrollo fue influenciado por figuras como Ronald A. Fisher, Jerzy Neyman y Egon Pearson, quienes sentaron las bases de los métodos inferenciales que permiten generalizar conclusiones de una muestra a una población más amplia. Fisher introdujo conceptos fundamentales como el “p-valor” y la prueba de hipótesis, mientras que Neyman y Pearson formalizaron la teoría de pruebas con su trabajo sobre errores tipo I y II y la formulación de intervalos de confianza. La inferencia estadística se consolidó rápidamente en diversas áreas científicas, desde la biología y la medicina hasta las ciencias sociales y económicas, transformando la manera en que los investigadores validan teorías y estiman parámetros poblacionales. A lo largo del tiempo, esta área ha evolucionado, incorporando herramientas computacionales y métodos bayesianos que amplían las posibilidades de análisis en contextos de datos complejos y grandes volúmenes de información.

13 Muestreo

Introducción al mundo cuantitativo

14 Contexto del muestreo

14.1 Tendencias

El análisis de datos no genera afirmaciones individuales. Se identifican comportamientos globales en torno a un fenómeno, que no corresponden al comportamiento de los individuos de manera puntual. Las tendencias son comportamientos globales, que los individuos acatan probablemente.

14.2 Individuos

Son unidades de análisis sobre las cuales vamos a generar un modelo. Son el sujeto de nuestra teoría.

14.3 Población

Para cualquier pregunta que interese responder, primero es necesario dirigir la atención a un grupo particular de individuos: personas, ciudades, animales, televisores, discos rígidos, tornillos o lamparitas.

14.4 Muestra

Es un subconjunto de la población.

14.5 Representatividad

Una muestra es representativa de la población cuando todas las características importantes de la población tienen que estar en la muestra en la misma proporción que en la población.

15 Muestreo

¿Qué hacemos para probar la sopa? Revolvemos la olla con una cuchara, sacamos una porción -una muestra- la saboreamos y sacamos una conclusión sobre toda la sopa de la olla sin haber en realidad probado toda. Si la muestra ha sido tomada adecuadamente - sin elegir tramposamente la parte buena - tendremos una buena idea del sabor de la totalidad de la sopa. Esto se hace en estadística, más específicamente en inferencia estadística.

Los investigadores quieren averiguar algo sobre una población, pero no tienen tiempo o dinero para estudiar a todos los individuos que la conforman. Por lo tanto, ¿qué hacen? Seleccionan una cantidad pequeña de unidades muestrales de la población (esto se llama una muestra), estudian esas unidades, generalmente individuos, y utilizan esa información para sacar conclusiones sobre toda de la población.

15.1 Actividad: Investiguemos

¿Qué es un cisne negro y qué historia esconde?

15.2 Conceptos de muestreo

- Marco
- Diseño
- Error
- Tamaño

15.3 Marco muestral

El marco muestral es el conjunto de todos los elementos o unidades de la población que son accesibles para ser seleccionados en la muestra.

Este marco debe ser representativo de (preferiblemente contener toda) la población objetivo para garantizar la validez de los resultados. Un marco muestral bien definido es crucial para evitar sesgos en la selección de la muestra.

15.4 Diseño muestral

El diseño muestral es el plan que describe cómo se selecciona la muestra a partir del marco muestral. Puede incluir diferentes técnicas de muestreo, como el muestreo aleatorio simple, el muestreo estratificado, el muestreo por conglomerados, entre otros.

La elección del diseño depende de los objetivos del estudio, las características de la población y los recursos disponibles.

Es el diseño muestral lo que le da representatividad a la muestra.

15.5 Error muestral

El error muestral es la diferencia entre el valor estimado a partir de la muestra y el valor real en la población. Este error surge debido a que la muestra es solo una parte de la población y no refleja completamente su variabilidad.

El tamaño de la muestra, el diseño muestral y el método de estimación influyen en la magnitud del error muestral.

15.6 Tamaño muestral

El tamaño muestral es la cantidad de unidades que se seleccionarán del marco muestral para ser incluidas en el estudio. Un tamaño muestral adecuado es fundamental para asegurar la precisión y confiabilidad de los resultados.

La determinación del tamaño muestral depende del nivel de confianza deseado, el margen de error aceptable y la variabilidad esperada en la población.

Es el tamaño muestral lo que le da significancia a las estimaciones. No hay muestras significativas.

15.7 Distintos abordajes del muestreo

15.7.1 Muestreo no probabilístico

No todos los elementos tienen probabilidad de ser seleccionados

- La muestra no es representativa (rigurosamente)
- No es posible calcular el error muestral
- Requiere menos recursos

15.7.2 Muestreo probabilístico

Todos los individuos en la población tienen una probabilidad específica de ser seleccionados para la muestra.

16 Muestreo no probabilístico

- Por cuotas
- Bola de nieve
- Discrecional
- Conveniencia
- Accidental

16.1 Muestreo por cuotas

Consiste en dividir la población en segmentos y obtener una cuota de cada segmento.

Se utiliza cuando se tienen segmentos relevantes pero no se tiene acceso al marco muestral.

En un estudio sobre preferencias de compra de automóviles, se decide utilizar el muestreo por cuotas para asegurar que la muestra sea representativa en términos de género y edad. El investigador establece cuotas basadas en la distribución de la población:

- **Género:** 50% hombres, 50% mujeres.
- **Edad:** 25% de 18-29 años, 25% de 30-39 años, 25% de 40-49 años, 25% de 50 años o más.

El investigador luego selecciona participantes hasta que se cumplan las cuotas establecidas, por ejemplo, 100 hombres y 100 mujeres, distribuidos equitativamente entre los diferentes grupos de edad. Este método asegura que todas las categorías importantes estén adecuadamente representadas en la muestra.

16.2 Bola de nieve

Cada individuo refiere nuevos individuos.

Se utiliza con poblaciones sensibles.

En un estudio sobre los hábitos de ahorro entre migrantes de un país específico, se utiliza el muestreo bola de nieve debido a la dificultad de acceder a esta población. El investigador comienza con un pequeño grupo de migrantes conocidos que participan en el estudio. Luego,

estos participantes refieren a otros migrantes que también podrían estar interesados en participar.

A medida que más personas son entrevistadas, se continúa pidiendo referencias, lo que permite que la muestra “crezca como una bola de nieve”. Este método es especialmente útil para poblaciones difíciles de alcanzar o cuando no existe un marco muestral claro.

16.3 Muestreo discrecional

Consiste en tomar de la población los individuos que resulten representativos bajo el juicio de un experto.

Se utiliza cuando hay un experto que ha realizado estudios previos.

En un estudio exploratorio sobre las opiniones de expertos en inteligencia artificial, se decide utilizar el muestreo discrecional. El investigador selecciona intencionadamente a un grupo de 10 expertos reconocidos en el campo, basándose en su conocimiento y reputación en la industria. La selección no es aleatoria, sino que se realiza según el criterio del investigador, quien elige a estos expertos por considerarlos los más apropiados para proporcionar información valiosa y relevante para el estudio.

16.4 Muestreo por conveniencia

Consiste en estudiar los individuos más cercanos o voluntarios, ya que estos son más accesibles.

Se utiliza en estudios cuyas poblaciones no son accesibles.

En un estudio piloto sobre las preferencias de snacks saludables entre estudiantes universitarios, el investigador utiliza el muestreo por conveniencia debido a limitaciones de tiempo y recursos.

El investigador selecciona a los estudiantes del curso de muestreo para que respondan a una breve encuesta. Este método se elige porque los participantes son fácilmente accesibles, aunque no necesariamente representan a toda la población estudiantil.

16.5 Muestreo accidental

Se selecciona un mecanismo de selección o acceso a los individuos encontrándolos por casualidad en un espacio definido sin ningún juicio previo.

Se utiliza en poblaciones particularmente grandes con fines descriptivos y prácticos.

En un estudio sobre las tendencias de la moda un investigador utiliza el muestreo accidental. Se aborda a los primeros 30 clientes que ingresan en una tienda de ropa durante una mañana para que respondan a una encuesta de satisfacción.

Este método es accidental porque se seleccionan a los participantes en el lugar y momento de recolección.

17 Muestreo probabilístico

- Aleatorio simple
- Por etapas
- Por conglomerados
- Por estratos

17.1 Muestreo aleatorio simple

- Definir la población
- Obtener el marco muestral
- Definir un tamaño muestral n
- Elegir aleatoriamente n individuos de la población

En un estudio sobre los hábitos de lectura de la población adulta en una ciudad, se decide utilizar el muestreo aleatorio simple. A partir de un listado completo de los 10,000 residentes adultos de la ciudad, el investigador selecciona 500 personas utilizando un generador de números aleatorios.

Cada individuo tiene la misma probabilidad de ser seleccionado, lo que asegura que la muestra sea representativa de toda la población, minimizando sesgos en la selección.

17.2 Muestreo por etapas

- Bietápico (2 etapas)
- Polietápico (3 etapas o más)
- Es necesario calcular tamaño muestral en cada etapa

En un estudio nacional sobre el nivel educativo en zonas rurales, se utiliza el muestreo por etapas para facilitar la selección de la muestra.

1. **Primera etapa:** Se seleccionan al azar 10 estados de un total de 32 en el país.
2. **Segunda etapa:** Dentro de cada estado seleccionado, se eligen aleatoriamente 3 municipios rurales.

3. **Tercera etapa:** En cada municipio, se seleccionan aleatoriamente 5 escuelas primarias.
4. **Cuarta etapa:** Dentro de cada escuela seleccionada, se eligen al azar 30 estudiantes para participar en el estudio.

Este método permite una selección eficiente y representativa en poblaciones grandes y dispersas, utilizando un enfoque jerárquico y secuencial.

17.3 Muestreo por conglomerados

- Son heterogéneos en su interior.
- Son homogéneos entre sí
- Se selecciona una muestra primaria
- Se selecciona una muestra secundaria
- Ejemplo: ciudades, salones, centros comerciales

En un estudio sobre el impacto de programas educativos en escuelas primarias, se utiliza el muestreo por conglomerados para simplificar el proceso de recolección de datos.

1. **Primera etapa:** Se seleccionan al azar 20 escuelas primarias de un listado completo de 200 en una región determinada.
2. **Segunda etapa:** Dentro de cada escuela seleccionada, se eligen al azar 3 clases para participar en el estudio.
3. **Tercera etapa:** Se recopilan datos de todos los estudiantes de las clases seleccionadas.

Este método es útil cuando la población está agrupada en unidades naturales, como escuelas o barrios, y facilita la logística y los costos al reducir el número de unidades primarias que se deben tratar.

17.4 Muestreo estratificado

- Son homogéneos en su interior
- Son heterogéneos entre sí
- Se selecciona una censo primario
- Se selecciona una muestra secundaria

En un estudio sobre la satisfacción laboral en una empresa multinacional, se utiliza el muestreo estratificado para asegurar que todas las divisiones y niveles jerárquicos de la empresa estén representados.

1. **Primera etapa:** Se divide a los empleados en diferentes estratos según su departamento (por ejemplo, Finanzas, Recursos Humanos, Marketing, etc.).

2. **Segunda etapa:** Se selecciona una muestra aleatoria de empleados de cada estrato para participar en la encuesta de satisfacción.

Este método garantiza que cada subgrupo relevante dentro de la población esté representado en la muestra, permitiendo comparaciones más precisas entre diferentes departamentos y niveles jerárquicos.

17.5 Técnicas de muestreo

[Técnicas de muestreo](#)

17.6 Factores de expansión

Cuando tenemos un diseño muestral, es posible saber cuántos individuos de la población son representados por un individuo de la muestra. Esta cantidad se denomina *Factores de expansión*.

Los factores de expansión son particularmente útiles para calcular totales.

17.7 Actividad

En esta actividad vamos a analizar una encuesta.

1. Elija una encuesta de la lista-
2. Escudriñe la documentación. Busque el muestreo.
3. Lea el muestreo cuidadosamente.
4. Escriba cómo imagina los detalles de la ejecución: Obtención del marco muestral

Encuestas

- [Encuesta Nacional de Demografía y Salud \(ENDS\)/Ministerio de Salud y Protección Social](#)
- [Encuesta TIC Hogares](#)
- [Índice de brecha Digital](#)
- [Encuesta Nacional Agropecuaria](#)
- [Encuesta Multipropósito](#)

18 Estimación

Estadística inferencial

19 Conceptos iniciales

Iniciamos por los conceptos que son fundamentales para entender la estimación en inferencia estadística. Estos conceptos ayudan a diferenciar entre la población y la muestra, así como a identificar las características numéricas que se desea estimar.

19.1 Parámetro

Un **parámetro** es un valor numérico que describe una característica específica de una población completa. Ejemplos incluyen la media poblacional (μ), la varianza poblacional (σ^2) o la proporción poblacional (P). Dado que los parámetros se refieren a poblaciones enteras, son constantes, pero, en la práctica, generalmente son desconocidos. Por lo tanto, se necesita recurrir a métodos estadísticos para estimarlos a partir de muestras.

19.2 Estimador

Un **estimador** es una regla o fórmula utilizada para aproximar un parámetro poblacional a partir de los datos muestrales. Un estimador es una variable aleatoria y, por lo tanto, puede variar de una muestra a otra. Los estimadores se utilizan para hacer inferencias sobre la población, y algunos ejemplos comunes incluyen la media muestral \bar{x} y la proporción muestral \hat{p} . La calidad de un estimador se evalúa en función de sus propiedades, como la consistencia y la unbiasedness (inexistencia de sesgo).

19.3 Muestra

Una **muestra** es un subconjunto de elementos extraídos de una población. La muestra se utiliza para realizar estimaciones sobre la población completa sin necesidad de medir a todos sus miembros. El tamaño y la calidad de la muestra son fundamentales para asegurar que las inferencias realizadas sean precisas y representativas. Una muestra puede ser aleatoria, estratificada o por conveniencia, dependiendo del método utilizado para su selección.

19.4 Población

La **población** se refiere al conjunto completo de individuos, elementos o unidades que comparten una característica particular que se está estudiando. Por ejemplo, si estamos interesados en el ingreso anual de todos los empleados de una empresa, la población sería todos los empleados. En la inferencia estadística, es crucial definir claramente la población para poder aplicar métodos de muestreo adecuados y realizar estimaciones precisas.

20 Estimación puntual

La **estimación puntual** se refiere al uso de un solo valor, calculado a partir de la muestra, para estimar un parámetro desconocido de la población. Este enfoque proporciona una respuesta única a la pregunta de interés, pero no informa sobre la precisión de la estimación. Por ejemplo, la media muestral \bar{x} es una estimación puntual de la media poblacional (μ).

- **Media muestral:** Se toma una muestra de los ingresos anuales de 100 empleados de una empresa. La media de estos ingresos en la muestra es de \$50,000. Este valor se utiliza como una **estimación puntual** de la media de los ingresos de toda la población de empleados.
- **Proporción muestral:** En una encuesta realizada a 500 personas, 320 afirman que prefieren trabajar desde casa. La proporción muestral es $\hat{p} = \frac{320}{500} = 0.64$, lo cual es una **estimación puntual** de la proporción real de personas que prefieren el trabajo remoto en la población general.
- **Varianza muestral:** Una muestra de 50 estudiantes tiene una varianza de puntajes en un examen de 16 puntos cuadrados. Este valor se utiliza como una **estimación puntual** de la varianza de los puntajes en la población completa de estudiantes.

20.1 Estimadores y parámetros

Los estimadores son utilizados para estimar parámetros poblacionales. Algunos ejemplos incluyen:

- **Media muestral** \bar{x} para estimar la media poblacional (μ).
- **Varianza muestral** s^2 para estimar la varianza poblacional (σ^2).
- **Proporción muestral** \hat{p} para estimar la proporción poblacional (P).
- **Correlación de Pearson** $\hat{\rho}$ para estimar la correlación poblacional ρ .

20.2 Aplicación interactiva

Para ilustrar la idea de estimación puntual, podemos estimar el número utilizando simulaciones. Aquí tienes un recurso que permite interactuar con este concepto:

[Estimación puntual](#)

20.3 Ley de los Grandes Números

La **Ley de los Grandes Números** es un principio fundamental en la teoría de la probabilidad que establece que, a medida que aumenta el tamaño de una muestra, la media muestral \bar{x} se acercará a la media poblacional (μ) de la población de la que se extrajo la muestra. Esta ley se basa en la idea de que las fluctuaciones aleatorias tienden a cancelarse entre sí en muestras más grandes, lo que resulta en una estimación más precisa del parámetro poblacional.

Existen dos versiones de la Ley de los Grandes Números:

1. **Ley débil de los grandes números:** Establece que, para cualquier valor ϵ positivo, la probabilidad de que la media muestral se desvíe de la media poblacional en más de ϵ tiende a cero a medida que el tamaño de la muestra (n) aumenta. Es decir, la media muestral se convierte en un estimador consistente de la media poblacional.
2. **Ley fuerte de los grandes números:** Afirmación más fuerte que la versión débil, establece que la media muestral converge casi seguramente a la media poblacional a medida que n tiende al infinito. Esto significa que, con una probabilidad de 1, la media muestral se aproximará a la media poblacional a medida que se tomen más y más muestras.

La Ley de los Grandes Números es fundamental en estadística y asegura que los resultados obtenidos de las muestras se volverán más representativos de la población a medida que se aumente el tamaño de la muestra, lo que permite realizar inferencias más confiables sobre la población completa.

20.4 Consistencia

Un **estimador puntual** es **consistente** si, a medida que el tamaño de la muestra aumenta, la estimación se aproxima cada vez más al verdadero parámetro poblacional. En otras palabras, un estimador consistente converge en probabilidad al parámetro que se está estimando. Esta propiedad es crucial porque asegura que con muestras más grandes, nuestras estimaciones se vuelven más precisas y confiables, reduciendo la variabilidad y el error de estimación.

21 Estimación por intervalo

La **estimación por intervalo** proporciona un rango de valores dentro del cual se espera que se encuentre el parámetro poblacional con un cierto nivel de confianza. A diferencia de una estimación puntual, que ofrece un único valor, el intervalo de confianza incluye una medida de la incertidumbre asociada con la estimación. Esto permite a los investigadores entender no solo qué valor se estima, sino también la precisión y confiabilidad de dicha estimación.

21.1 Precisión y Exactitud

Precisión y exactitud son dos conceptos fundamentales en la estadística y la investigación que se utilizan para evaluar la calidad de las estimaciones y mediciones. Aunque a menudo se utilizan de manera intercambiable en el lenguaje cotidiano, tienen significados distintos en el contexto estadístico.

21.1.1 Precisión

La **precisión** se refiere a la consistencia y reproducibilidad de las mediciones. Un conjunto de datos es preciso si las mediciones son cercanas entre sí, independientemente de si son correctas o no. En otras palabras, la precisión indica cuán dispersos están los valores en relación con la media. Un alto grado de precisión significa que las mediciones tienden a agruparse en torno a un valor central.

- **Ejemplo de Precisión:** Supongamos que un grupo de científicos mide la temperatura de un líquido en tres ocasiones y obtiene los siguientes valores: 22.1°C, 22.0°C y 22.2°C. Aunque la medición puede no ser la temperatura real del líquido, los valores son consistentes entre sí, lo que indica alta precisión.

21.1.2 Exactitud

La **exactitud**, por otro lado, se refiere a cuán cerca está una medición del valor verdadero o del objetivo. Una medición es exacta si se aproxima al valor real. En este caso, la exactitud evalúa la validez de los datos en relación con la realidad.

- **Ejemplo de Exactitud:** Siguiendo el mismo ejemplo anterior, si la temperatura real del líquido es de 23.0°C y las mediciones obtenidas fueron 22.1°C, 22.0°C y 22.2°C, podemos decir que las mediciones son imprecisas a pesar de ser consistentes, ya que están lejos del valor verdadero.

21.1.3 Diferencias entre Precisión y Exactitud

Concepto	Definición	Enfoque	Ejemplo
Precisión	Consistencia y reproducibilidad de mediciones	Distribución de datos	Mediciones: 10.1, 10.0, 10.2 (altamente precisas pero no exactas si el valor verdadero es 11.0)
Exactitud	Cercanía de las mediciones al valor verdadero	Proximidad al valor verdadero	Mediciones: 11.0, 11.1, 11.2 (altamente exactas si el valor verdadero es 11.0, aunque pueden no ser precisas si los valores no son consistentes)

La precisión y exactitud son cruciales en la estadística y en el diseño de investigaciones. Una alta precisión en los resultados es deseable, pero no es suficiente por sí sola. Si los datos son precisos pero no exactos, las conclusiones extraídas pueden ser engañosas y llevar a decisiones erróneas.

1. **Diseño de Estudios:** Durante el diseño de estudios, los investigadores deben asegurarse de que las herramientas de medición sean tanto precisas como exactas para obtener datos confiables.
2. **Interpretación de Resultados:** La evaluación de la precisión y exactitud de los datos ayuda a los investigadores a entender la confiabilidad de sus estimaciones y a interpretar correctamente los resultados.
3. **Mejora de Métodos:** Comprender la diferencia entre estos conceptos permite a los investigadores identificar áreas de mejora en sus métodos de recopilación de datos, ajustando su enfoque para aumentar tanto la precisión como la exactitud.

Los conceptos de precisión y exactitud están intrínsecamente relacionados con los intervalos de confianza, ya que estos intervalos son una herramienta estadística diseñada para expresar la incertidumbre en torno a una estimación puntual. Un intervalo de confianza proporciona un rango de valores dentro del cual se espera que se encuentre el parámetro poblacional verdadero, lo que refleja la **exactitud** de la estimación. Si el intervalo es estrecho, indica una alta precisión en las mediciones, lo que sugiere que repetidas mediciones generarían resultados consistentes. Sin embargo, si el intervalo de confianza incluye valores muy alejados del verdadero parámetro, sugiere que la estimación puede no ser precisa. Por lo tanto, un intervalo de confianza bien construido no solo comunica la variabilidad de los datos, sino que también integra las nociones de precisión y exactitud en la evaluación de la validez de los resultados estadísticos.

21.2 Ejemplo introductorio

21.2.1 Intervalo de confianza para la media

Para calcular un intervalo de confianza para la media de una población, se puede utilizar la distribución t de Student o la distribución normal, dependiendo del tamaño de la muestra y de si se conoce la varianza poblacional.

- **Ejemplo:** Supongamos que un investigador quiere estimar la media de la altura de los estudiantes de una universidad. Toma una muestra aleatoria de 30 estudiantes y calcula que la media muestral es de 1.70 metros con una desviación estándar de 0.10 metros. Dado que el tamaño de la muestra es pequeño, se utiliza la distribución t de Student. Si se quiere un nivel de confianza del 95%, el intervalo de confianza se calcula de la siguiente manera:

$$IC = \bar{x} \pm t_{\alpha/2} \left(\frac{s}{\sqrt{n}} \right)$$

Donde $t_{\alpha/2}$ es el valor crítico de t correspondiente al nivel de confianza del 95% y 29 grados de libertad (n-1), (s) es la desviación estándar muestral y (n) es el tamaño de la muestra. Supongamos que $t_{0.025} \approx 2.045$ para este caso:

$$IC = 1.70 \pm 2.045 \left(\frac{0.10}{\sqrt{30}} \right) \approx 1.70 \pm 0.374$$

El intervalo de confianza sería aproximadamente (1.63, 1.77). Esto significa que el investigador puede estar 95% seguro de que la media de altura de todos los estudiantes está entre 1.63 y 1.77 metros.

21.2.2 Intervalo de confianza para una proporción

El **intervalo de confianza para una proporción** se calcula utilizando la proporción muestral \hat{p} y el error estándar de la proporción. Se utiliza la distribución normal, ya que se asume que la proporción muestral sigue una distribución normal cuando el tamaño de muestra es suficientemente grande.

- **Ejemplo:** Imaginemos que en una encuesta a 500 personas, 320 afirman que prefieren trabajar desde casa. La proporción muestral es:

$$\hat{p} = \frac{320}{500} = 0.64$$

El intervalo de confianza se puede calcular de la siguiente manera:

$$IC = \hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Si elegimos un nivel de confianza del 95%, $z_{0.025} \approx 1.96$:

$$IC = 0.64 \pm 1.96 \sqrt{\frac{0.64 \times (1 - 0.64)}{500}} \approx 0.64 \pm 0.045$$

Esto resulta en un intervalo de confianza de (0.595, 0.685). Por lo tanto, el investigador puede estar 95% seguro de que la proporción real de personas que prefieren trabajar desde casa en la población general está entre el 59.5% y el 68.5%.

21.2.3 Intervalo de confianza para la varianza

El **intervalo de confianza para la varianza** se calcula utilizando la distribución chi-cuadrado. La varianza muestral (s^2) es el estimador puntual, y el intervalo se ajusta de acuerdo con la distribución de esta varianza bajo el supuesto de normalidad.

- **Ejemplo:** Supongamos que un investigador mide los tiempos de espera en una fila y obtiene una varianza muestral de ($\hat{s}^2 = 16$) minutos² con una muestra de 15 observaciones. Para calcular un intervalo de confianza del 95% para la varianza poblacional, se utiliza la fórmula:

$$IC = \left(\frac{(n-1)s^2}{\chi_{\alpha/2}^2}, \frac{(n-1)s^2}{\chi_{1-\alpha/2}^2} \right)$$

Donde (n) es el tamaño de la muestra y (χ^2) son los valores críticos de la distribución chi-cuadrado. Si (n=15) y con un nivel de confianza del 95%, se pueden encontrar los valores críticos:

$$IC = \left(\frac{(15-1) \times 16}{\chi_{0.025,14}^2}, \frac{(15-1) \times 16}{\chi_{0.975,14}^2} \right)$$

Si $\chi_{0.025,14}^2 \approx 27.688$ y $\chi_{0.975,14}^2 \approx 5.629$:

$$IC \approx \left(\frac{14 \times 16}{27.688}, \frac{14 \times 16}{5.629} \right) \approx (8.520, 39.658)$$

Esto significa que el investigador puede estar 95% seguro de que la varianza del tiempo de espera en la población está entre 8.520 y 39.658 minutos².

21.3 Aplicación interactiva

Para ilustrar la idea de estimación por intervalo, puedes utilizar simulaciones interactivas que permiten explorar estos conceptos. Aquí tienes un recurso que permite interactuar con este concepto:

[Estimación por intervalo](#)

21.4 Consistencia

Un **intervalo de confianza** es considerado **consistente** si, al aumentar el tamaño de la muestra, la longitud del intervalo tiende a reducirse. Esto indica que la estimación se vuelve más precisa y refleja mejor el parámetro poblacional.

21.5 Construcción de los intervalos de confianza

La construcción de intervalos de confianza puede realizarse de dos maneras:

1. **Método analítico:** Utiliza fórmulas y propiedades matemáticas para calcular los intervalos de confianza. Este método es eficiente cuando se cumplen las condiciones necesarias, como la normalidad de los datos.
2. **Método computacional:** Utiliza técnicas de remuestreo, como el bootstrapping, para estimar la distribución del estimador y generar intervalos de confianza sin necesidad de asumir una distribución específica.

21.6 Ventajas y desventajas de la construcción analítica

La construcción analítica de los intervalos de confianza presenta pros y contras

- **Ventajas:** - Menor costo computacional, ya que implica cálculos directos a partir de los datos. - Resultados más rápidos si se cumplen los supuestos de normalidad.
- **Desventajas:** - Dependencia de supuestos que pueden no cumplirse en datos reales. - Puede no ser adecuado para distribuciones no estándar o en situaciones de muestras pequeñas.

21.7 Ventajas y desventajas de la construcción computacional

La construcción computacional de los intervalos de confianza presenta pros y contras

- **Ventajas:** - Flexible, ya que no depende de supuestos sobre la distribución de los datos. - Adecuado para cualquier tipo de muestra y variabilidad en los datos.
- **Desventajas:** - Puede ser intensivo en recursos computacionales, especialmente para grandes volúmenes de datos. - Mayor tiempo de procesamiento y necesidad de software especializado.

21.8 Fórmulas para algunos intervalos

- Intervalo de confianza para la media:

$$IC = \bar{x} \pm z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right) \quad (\text{si } \sigma \text{ es conocido})$$

$$IC = \bar{x} \pm t_{\alpha/2} \left(\frac{s}{\sqrt{n}} \right) \quad (\text{si } \sigma \text{ es desconocido})$$

- Intervalo de confianza para la proporción:

$$IC = \hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

- Intervalo de confianza para la varianza:

$$IC = \left(\frac{(n-1)s^2}{\chi_{\alpha/2}^2}, \frac{(n-1)s^2}{\chi_{1-\alpha/2}^2} \right)$$

22 Bootstrapping

El **bootstrapping** es un método no paramétrico que se utiliza para estimar la distribución de un estimador y, por ende, sus intervalos de confianza. Este método genera múltiples muestras simuladas (con reemplazo) de la muestra original, permitiendo construir un intervalo de confianza sin depender de la distribución subyacente de los datos.

22.1 Ejemplo: Diferencia de medias

Para evaluar la **diferencia de medias** entre dos poblaciones, se construye un intervalo de confianza alrededor de la diferencia de medias muestrales. Dependiendo de si se asumen varianzas iguales o diferentes, se pueden aplicar diferentes métodos, como el uso de la distribución *t* de Student o el método de bootstrapping.

- **Ejemplo:** Supongamos que queremos comparar las alturas promedio de hombres y mujeres en una población. Recogemos datos de 50 hombres y 50 mujeres y encontramos que la media de los hombres es 1.80 metros y la media de las mujeres es 1.65 metros. La diferencia de medias es 0.15 metros. Aplicando el bootstrapping, podemos crear múltiples muestras de nuestras muestras originales para estimar la distribución de la diferencia de medias y, así, construir un intervalo de confianza.

22.2 Ejemplo: Diferencia de proporciones

El **intervalo de confianza para la diferencia de proporciones** entre dos poblaciones se construye de manera similar al de una proporción, pero utilizando la diferencia entre las proporciones muestrales. Dependiendo de si se asume o no la homogeneidad de varianzas entre las dos proporciones, se pueden aplicar diferentes métodos.

- **Ejemplo:** Si en una encuesta a 300 hombres, el 70% responde que prefiere el trabajo remoto, y en una encuesta a 250 mujeres, el 60% responde lo mismo, la diferencia de proporciones es 0.10. Para construir un intervalo de confianza, podemos aplicar el bootstrapping para simular la diferencia de proporciones y calcular el intervalo.

22.3 Bootstrapping ejemplo

Para explorar el método de bootstrapping, puedes interactuar con simulaciones que permiten visualizar cómo funciona este proceso en la práctica. Aquí tienes un recurso que permite interactuar con este concepto:

[Bootstrapping](#)

23 Tamaño muestral

Estadística inferencial

El **tamaño muestral** es un aspecto crucial en el diseño de estudios estadísticos, ya que influye directamente en la precisión y confiabilidad de las estimaciones. Comprender cómo determinar el tamaño de muestra adecuado es fundamental para obtener resultados válidos y aplicables a una población más amplia.

24 Error muestral

El **margen de error** es una medida que refleja la cantidad de incertidumbre asociada a una estimación puntual. Se define como la mitad de la longitud de un intervalo de confianza y representa la variabilidad esperada en una estimación debido al muestreo. Cuanto mayor sea el margen de error, menos confiable será la estimación.

Cálculo del margen de error:

El margen de error se puede calcular utilizando la siguiente fórmula:

$$\text{Margen de error} = Z \times \left(\frac{\sigma}{\sqrt{n}} \right)$$

Donde: - Z es el valor crítico de la distribución normal estándar (por ejemplo, 1.96 para un nivel de confianza del 95%). - σ es la desviación estándar de la población (o de la muestra si la desviación estándar de la población no está disponible). - n es el tamaño de la muestra.

24.1 Ejemplo:

- **Encuesta electoral:** En una encuesta a 1,000 votantes, el 48% afirma que votará por el candidato A. Si el margen de error es de $\pm 3\%$, esto indica que el porcentaje real de votantes que apoyan al candidato A se estima que está entre el 45% y el 51%. Esto se puede expresar como un intervalo de confianza de (0.45, 0.51).
- **Estimación de ingresos:** Se estima que el ingreso promedio de una población es de \$50,000 con un margen de error de $\pm \$2,000$. Esto implica que el ingreso promedio real en la población se espera que esté entre \$48,000 y \$52,000. En este caso, el margen de error nos proporciona una indicación clara de la posible variabilidad en los ingresos de la población.

24.2 Empate técnico

El **empate técnico** ocurre cuando las diferencias entre dos estimaciones puntuales, como medias o proporciones, no son estadísticamente significativas debido al margen de error. En este caso, las diferencias observadas podrían ser el resultado de la variabilidad muestral y no reflejan una verdadera diferencia en la población.

24.3 Ejemplo:

- **Encuesta electoral:** En una encuesta a 1,200 votantes, el candidato A obtiene el 46% de las intenciones de voto y el candidato B el 44%, con un margen de error de $\pm 3\%$. Dado que el margen de error abarca los porcentajes de ambos candidatos (43% a 49% para A y 41% a 47% para B), se considera un empate técnico. Esto sugiere que no hay suficiente evidencia para afirmar que uno de los candidatos es preferido sobre el otro.
- **Competencia de ventas entre productos:** En un análisis de ventas, el producto X alcanza el 35% de participación en el mercado y el producto Y el 33%, con un margen de error de $\pm 2\%$. Aquí, el margen de error (33% a 37% para X y 31% a 35% para Y) se superpone, lo que lleva a declarar un empate técnico entre los productos. Este tipo de análisis es crucial en la investigación de mercados, donde decisiones de marketing deben basarse en resultados estadísticamente significativos.

25 Tamaño de muestra

El **tamaño de muestra** afecta directamente la precisión de las estimaciones. Un mayor tamaño de muestra generalmente reduce el error estándar, disminuye el margen de error y mejora la confiabilidad de las estimaciones. Sin embargo, también implica un aumento en los costos y el tiempo de recolección de datos.

25.1 Tamaño de muestra óptimo

El tamaño de muestra óptimo se calcula bajo el supuesto de que cada registro genera un costo. En este sentido, la optimización se realiza disminuyendo el tamaño muestral con la restricción de que las estimaciones sean estadísticamente distintas de cero.

Consideraciones para determinar el tamaño óptimo:

1. **Costos:** Es fundamental considerar los costos asociados a la recolección de datos. Un tamaño muestral mayor a lo necesario puede generar costos innecesarios, mientras que un tamaño menor puede resultar en una estimación menos útil.
2. **Variabilidad:** La variabilidad dentro de la población impacta la determinación del tamaño de la muestra. Si la población es altamente variable, se necesitará un tamaño de muestra mayor para capturar esta variabilidad.
3. **Nivel de confianza y margen de error deseados:** Un mayor nivel de confianza o un margen de error más pequeño requerirá un tamaño de muestra más grande. Por ejemplo, si deseamos un nivel de confianza del 99% en lugar del 95%, el tamaño de la muestra tendrá que aumentar.
4. **Tamaño poblacional:** En poblaciones muy grandes, la relación entre el tamaño de la muestra y el tamaño poblacional es menos crítica. Sin embargo, en poblaciones pequeñas, se debe considerar el efecto de la muestra sobre el total.

26 Fórmulas para Estimación

26.1 1. Estimación de una Proporción con Tamaño Poblacional Conocido

Cuando se conoce el tamaño de la población y se desea estimar una proporción, se puede utilizar la siguiente fórmula:

$$n = \frac{N \cdot Z^2 \cdot p(1-p)}{(N-1) \cdot E^2 + Z^2 \cdot p(1-p)}$$

Donde: - n = tamaño de la muestra - N = tamaño de la población - Z = valor crítico de la distribución normal (por ejemplo, 1.96 para un nivel de confianza del 95%) - p = proporción estimada de la población (por ejemplo, 0.5 si no se conoce) - E = margen de error deseado

26.2 2. Estimación de una Proporción con Tamaño Poblacional Desconocido

Cuando el tamaño de la población es desconocido, se utiliza la siguiente fórmula simplificada:

$$n = \frac{Z^2 \cdot p(1-p)}{E^2}$$

Donde: - n = tamaño de la muestra - Z = valor crítico de la distribución normal (por ejemplo, 1.96 para un nivel de confianza del 95%) - p = proporción estimada de la población (puede ser 0.5 si se desea máxima variabilidad) - E = margen de error deseado

26.3 3. Estimación de una Media con Tamaño Poblacional Conocido

Si el tamaño de la población es conocido y se desea estimar la media, se utiliza la siguiente fórmula:

$$n = \frac{N \cdot Z^2 \cdot \sigma^2}{(N - 1) \cdot E^2 + Z^2 \cdot \sigma^2}$$

Donde: - n = tamaño de la muestra - N = tamaño de la población - Z = valor crítico de la distribución normal (por ejemplo, 1.96 para un nivel de confianza del 95%) - σ = desviación estándar de la población - E = margen de error deseado

26.4 4. Estimación de una Media con Tamaño Poblacional Desconocido

Cuando el tamaño de la población es desconocido, se utiliza la siguiente fórmula:

$$n = \frac{Z^2 \cdot \sigma^2}{E^2}$$

Donde:

- n = tamaño de la muestra
- Z = valor crítico de la distribución normal (por ejemplo, 1.96 para un nivel de confianza del 95%)
- σ = desviación estándar de la población (o de la muestra si es necesario)
- E = margen de error deseado

Es importante considerar que, al utilizar estas fórmulas, se deben tener en cuenta supuestos como la normalidad de la población y la aleatoriedad de la muestra. Además, los valores de p y σ deben ser estimados de forma adecuada para obtener resultados confiables. Para poblaciones pequeñas, es recomendable aplicar una corrección de población finita si se utiliza la fórmula con el tamaño poblacional conocido.

Estas fórmulas son herramientas esenciales en la investigación estadística y permiten a los investigadores diseñar estudios que produzcan estimaciones precisas y confiables.

26.5 Ejemplo:

- En un estudio sobre hábitos de consumo, se determina que el costo de encuestar a un individuo es de \$10. Si se quiere estimar el gasto promedio mensual con un margen de error de $\pm \$5$ y un nivel de confianza del 95%, se puede calcular el tamaño de muestra necesario utilizando la fórmula del tamaño de muestra:

$$n = \left(\frac{Z^2 \cdot \sigma^2}{E^2} \right)$$

Donde:

- E es el margen de error deseado.

Al determinar que la desviación estándar de los gastos mensuales es \$50, se puede calcular el tamaño muestral óptimo para cumplir con los criterios de estudio.

- **Costo-beneficio:** Supongamos que al calcular el tamaño muestral óptimo se determina que es de 100 encuestas. Si se decide llevar a cabo 150 encuestas, aunque se logra una mayor precisión, el costo adicional debe justificarse por los beneficios esperados del estudio.

En resumen, un tamaño muestral mayor al óptimo produce estimaciones que son distintas de cero, pero incurre en costos innecesarios. Por otro lado, un tamaño muestral menor al óptimo produce estimaciones que son estadísticamente iguales a cero, disminuyendo la utilidad del ejercicio. Por lo tanto, encontrar un equilibrio entre precisión y costo es esencial en el diseño de estudios estadísticos.

26.6 Aplicación Interactiva

Para ayudar a comprender cómo calcular el tamaño de muestra, se puede utilizar la siguiente herramienta interactiva:

[Cálculo del tamaño de muestra](#)

Esta aplicación te permite ingresar parámetros como el nivel de confianza, el margen de error, y la proporción esperada o la desviación estándar, dependiendo de si deseas estimar una proporción o una media. La herramienta calculará automáticamente el tamaño de muestra necesario para tu estudio, facilitando así la planificación y diseño de investigaciones.

Instrucciones:

1. Accede al enlace proporcionado.

2. Selecciona el tipo de estimación que deseas realizar (proporción o media) y (si lo conoces) el tamaño poblacional.
3. Ingresa el margen de error que necesitas.
4. Haz clic en “Calcular” para obtener el tamaño de muestra recomendado.

Esta aplicación es especialmente útil para investigadores y profesionales que desean asegurarse de que sus estimaciones sean precisas y confiables, optimizando así los recursos destinados a la recolección de datos.

27 Hipótesis

27.1 La hipótesis en el contexto del desarrollo científico

Epistemológicamente, una **hipótesis** es una proposición que establece una relación entre variables, formulada para ser sometida a prueba a través de la observación y la experimentación. Este proceso es esencial en el desarrollo científico, ya que las hipótesis funcionan como un puente entre la teoría y los datos empíricos. Una hipótesis permite a los investigadores establecer un marco para la investigación, formulando preguntas que pueden ser respondidas mediante métodos científicos. Al plantear una hipótesis, el investigador no solo está formulando una afirmación, sino que también está invitando a la crítica y la validación a través de experimentos y observaciones.

Las pruebas de hipótesis son fundamentales en este proceso, ya que permiten evaluar la validez de las afirmaciones mediante un enfoque sistemático y basado en datos. Este enfoque no solo fomenta la acumulación de conocimiento, sino que también contribuye a la evolución de teorías dentro de un campo determinado. La capacidad de una hipótesis para ser refutada o confirmada es lo que la hace esencial para el avance del conocimiento científico, manteniendo siempre un carácter provisional y sujeto a revisión.

27.1.1 Qué es la reproducibilidad y por qué es importante

La **reproducibilidad** es un principio fundamental en la ciencia, que implica que los resultados de un experimento deben poder ser replicados por otros investigadores utilizando los mismos métodos. Esto no solo fortalece la validez de los hallazgos, sino que también asegura que el conocimiento científico se basa en evidencias sólidas y verificables.

- **Importancia:** La reproducibilidad es esencial para la credibilidad de la ciencia y se convierte en una medida del rigor y la fiabilidad de los estudios realizados. Sin reproducibilidad, los hallazgos pueden ser considerados anecdóticos y su impacto en la ciencia puede ser severamente cuestionado.

27.1.2 Reproducibilidad

La **crisis actual de la reproducibilidad** en la ciencia se refiere a la creciente preocupación sobre la capacidad de replicar resultados científicos. Estudios han demostrado que un número significativo de resultados

en diversas disciplinas no puede ser replicado, lo que plantea interrogantes sobre la validez de esos hallazgos.

- **Los escándalos en la ciencia:** Casos de fraude, mala conducta científica y prácticas poco éticas han salido a la luz, socavando la confianza en la investigación. Esto ha llevado a un llamado a la transparencia en los métodos y datos utilizados, así como a la implementación de mejores prácticas para asegurar la reproducibilidad.

```
# arreglar todo esto
```

27.1.3 Probabilidad

Es posible ver la probabilidad como el estudio del comportamiento de los datos generados a partir de una distribución conocida.

27.1.4 Estadística

Análogamente, la estadística es el estudio del comportamiento de las distribuciones asociadas a un conjunto de datos dado.

27.1.5 Modelo estadístico

Un modelo estadístico es un conjunto de supuestos matemáticos que se realizan sobre la distribución asociada a un conjunto de datos.

27.1.6 Objetivo

Establecer si existe **suficiente evidencia** en una muestra aleatoria para **rechazar** o **no rechazar** la **hipótesis nula** a nivel poblacional. El objetivo es determinar cuál hipótesis explica mejor los datos observados en la población.

27.2 Arquitectura de las pruebas de hipótesis

En el ámbito de la estadística y la investigación científica, la comprensión y aplicación de las pruebas de hipótesis es fundamental para validar teorías y conclusiones basadas en datos. Estas pruebas se estructuran de manera que permiten a los investigadores tomar decisiones informadas acerca de la relación entre variables y la existencia de efectos en estudios empíricos. Un buen punto de partida para ilustrar la lógica subyacente en las pruebas de hipótesis es la famosa analogía de la *tetera de Russell*, la cual ejemplifica la importancia de la carga de la prueba y su relación con la validación de afirmaciones complejas.

27.2.1 La tetera de Russell y la carga de la prueba

Bertrand Russell, filósofo y matemático británico, introdujo la analogía de una pequeña tetera de porcelana que orbita el Sol entre la Tierra y Marte, tan diminuta que ningún telescopio podría detectarla. La premisa de Russell es que, si alguien afirmara la existencia de esta tetera, la responsabilidad de demostrar su existencia recaería sobre esa persona y no sobre los escépticos, quienes no tendrían que probar la inexistencia de la misma. Este argumento sirve para destacar un principio esencial en la lógica y la ciencia: la carga de la prueba recae en quienes hacen afirmaciones poco usuales.

Russell empleó esta metáfora en el contexto de la religión, pero su aplicación es igualmente poderosa en la ciencia y la estadística. En el campo de las pruebas de hipótesis, se establece que la evidencia es la herramienta clave para sostener o rechazar afirmaciones. Esta carga de la prueba se traduce en un proceso metódico y riguroso que permite a los investigadores evaluar si los datos apoyan una hipótesis específica.

27.2.2 La arquitectura fundamental de una prueba de hipótesis

En estadística, la arquitectura de las pruebas de hipótesis se sostiene en dos componentes principales: la **hipótesis nula (H_0)** y la **hipótesis alternativa (H_a)**.

- **Hipótesis nula (H_0):** Es la suposición inicial que niega la existencia de un efecto o relación en la población de estudio. Es un punto de partida que establece que cualquier efecto observado es producto de la variabilidad natural o de un azar inherente. La hipótesis nula se formula de manera que, para rechazarla, se debe contar con evidencia suficiente que contradiga su planteamiento. No obstante, debido a su enunciado, su verificación resulta muy costosa.
- **Hipótesis alternativa (H_a):** Esta hipótesis propone la existencia de un efecto o relación y es la contraparte de la hipótesis nula. Si los datos recolectados ofrecen pruebas suficientes, se puede rechazar la hipótesis nula en favor de la hipótesis alternativa.

Este enfoque dual asegura que los investigadores se acerquen al análisis de manera objetiva, comenzando con la suposición de que no hay un efecto y buscando evidencia que justifique su rechazo.

Tip

Mejorar el ejemplo, lo demostrable y lo indemostrable

Ejemplo aplicado a las ciencias sociales: Considere el siguiente problema de investigación: **¿Los videojuegos violentos causan comportamientos violentos?**. La prueba de hipótesis en este caso se estructuraría de la siguiente manera:

- **Hipótesis nula (H_0):** Los videojuegos violentos no causan un aumento en los comportamientos violentos.
- **Hipótesis alternativa (H_a):** Los videojuegos violentos causan un aumento en los comportamientos violentos.

El propósito de la prueba de hipótesis es analizar los datos recolectados en estudios experimentales u observacionales para determinar si hay suficiente evidencia empírica para rechazar la hipótesis nula y aceptar la hipótesis alternativa. Aquí, la prueba recae en la demostración de H_a : si los datos muestran una relación estadísticamente significativa entre la exposición a videojuegos violentos y comportamientos agresivos, entonces la hipótesis nula puede ser rechazada razonablemente.

27.2.3 Reflexiones sobre la evidencia y la carga de la prueba

La analogía de la tetera de Russell es particularmente relevante al considerar la **evidencia necesaria para sostener o rechazar una hipótesis**. Así como Russell enfatizaba que es quien afirma la existencia de la tetera quien debe probar su existencia, en las pruebas de hipótesis, los investigadores deben demostrar que los datos apoyan la hipótesis alternativa y no asumir la verdad de H_0 sin evidencia adecuada. De lo contrario, el rigor científico y la validez de las conclusiones estarían en entredicho.

Este enfoque protege a la ciencia de afirmaciones infundadas y promueve una práctica investigativa basada en pruebas y datos sólidos. Al entender la arquitectura de las pruebas de hipótesis, los investigadores pueden abordar preguntas complejas con un marco metodológico que garantiza que sus conclusiones estén respaldadas por evidencia empírica y análisis crítico.

27.3 Terminología

La comprensión profunda de la terminología asociada con las pruebas de hipótesis es esencial para el análisis y la interpretación de datos en la estadística. A continuación, se presentan definiciones y explicaciones detalladas de los conceptos clave que constituyen la base de cualquier prueba de hipótesis.

27.3.1 Hipótesis estadística

Una **hipótesis estadística** es una afirmación sobre un parámetro poblacional que se somete a prueba mediante datos muestrales. Este tipo de hipótesis se formula con el fin de validar o refutar una suposición sobre la población a partir de la evidencia obtenida en la muestra. Dependiendo de la naturaleza de la afirmación, una hipótesis puede clasificarse en:

- **Hipótesis simple:** Es aquella que especifica un valor único y preciso para el parámetro de la población. Por ejemplo, si se formula la hipótesis $H_0 : \mu = \mu_0$, se está afirmando que la media poblacional es exactamente igual a un valor conocido μ_0 .
- **Hipótesis compuesta:** Esta hipótesis especifica un rango de valores para el parámetro poblacional en lugar de un único valor. Ejemplos de hipótesis compuestas incluyen $H : \mu \geq \mu_0$, $H : \mu \leq \mu_0$, $H : \mu > \mu_0$, $H : \mu < \mu_0$, y $H : \mu \neq \mu_0$.

En estas hipótesis, el valor μ_0 se conoce como el **valor hipotético**, que es el punto de referencia contra el cual se compara la estimación muestral para determinar si la hipótesis debe ser aceptada o rechazada.

27.3.2 Sistema de hipótesis

El **sistema de hipótesis** es el conjunto de hipótesis contrapuestas que se comparan en el proceso de la prueba de hipótesis. Este sistema incluye dos componentes principales:

- **Hipótesis nula (H_0):** Es la hipótesis de partida que se presume cierta y que no se rechaza a menos que los datos muestren una evidencia fuerte en su contra. La hipótesis nula se formula generalmente como una afirmación de igualdad o de ausencia de efecto. Ejemplos de hipótesis nulas pueden ser:
 - $H_0 : \mu \leq \mu_0$, que implica que la media poblacional es menor o igual al valor hipotético.
 - $H_0 : \mu \geq \mu_0$, que sugiere que la media es mayor o igual al valor hipotético.
 - $H_0 : \mu = \mu_0$, que postula que la media es exactamente igual al valor hipotético.

- **Hipótesis alternativa (H_1):** Representa la afirmación que se acepta si los datos muestrales proporcionan suficiente evidencia en contra de H_0 . Es la hipótesis que postula la existencia de un efecto, una diferencia o una relación significativa. Algunos ejemplos incluyen:
 - $H_1 : \mu > \mu_0$, que sostiene que la media poblacional es mayor que el valor hipotético.
 - $H_1 : \mu < \mu_0$, que indica que la media es menor que el valor hipotético.
 - $H_1 : \mu \neq \mu_0$, que señala que la media es diferente al valor hipotético, sin especificar si es mayor o menor.

Este sistema de hipótesis es el núcleo de la prueba estadística y establece un marco para decidir si los datos muestrales respaldan o refutan la hipótesis nula.

27.3.3 Tipos de error

En el proceso de tomar decisiones basadas en pruebas de hipótesis, es posible cometer errores. Estos errores se clasifican en dos tipos principales:

- **Error tipo I:** Ocurre cuando se rechaza la hipótesis nula siendo esta verdadera. Es decir, se concluye incorrectamente que hay un efecto o diferencia cuando en realidad no lo hay. La probabilidad de cometer un error tipo I se denota por α , que también se conoce como el **nivel de significancia** de la prueba. Este es un parámetro que se establece de antemano y refleja el nivel de tolerancia al riesgo de rechazar una hipótesis verdadera.
- **Error tipo II:** Se produce cuando no se rechaza la hipótesis nula cuando esta es falsa. En otras palabras, se falla en detectar un efecto o diferencia que realmente existe. La probabilidad de cometer un error tipo II se denota por β . El complemento de β se llama **potencia de la prueba** y representa la probabilidad de rechazar correctamente una hipótesis nula falsa.

La relación entre los tipos de error y la potencia de la prueba es fundamental para comprender la eficacia y precisión de una prueba estadística. Un investigador debe equilibrar estos errores al diseñar una prueba, ya que reducir la probabilidad de un error tipo I puede incrementar la probabilidad de un error tipo II, y viceversa.

27.3.4 Nivel de significancia

El **nivel de significancia** (α) es un umbral predefinido que establece la probabilidad máxima aceptable de cometer un error tipo I. Comúnmente, los niveles de significancia utilizados son 0.05, 0.01, o 0.10. Este valor define el criterio para decidir si los resultados de una prueba son estadísticamente significativos. Si el valor p obtenido de la prueba es menor o igual a α ,

se rechaza la hipótesis nula en favor de la hipótesis alternativa, indicando que los resultados observados son lo suficientemente raros como para no atribuirse al azar.

La elección del nivel de significancia depende del contexto y las consecuencias de cometer un error tipo I. En estudios donde el costo de un error tipo I es alto, se elige un α más bajo (e.g., 0.01). En otros casos, donde la tolerancia al riesgo es mayor, se podría optar por un α más alto (e.g., 0.10).

27.3.5 Integrando los conceptos en la práctica

La correcta formulación de hipótesis y la comprensión de los tipos de error y del nivel de significancia permiten a los investigadores tomar decisiones informadas basadas en los datos muestrales. En un estudio, el análisis de hipótesis proporciona un camino claro para evaluar si una afirmación sobre la población es plausible o si debe ser rechazada en función de la evidencia empírica. Este enfoque estructurado es la base para el análisis crítico y la validación de resultados en investigación científica y en aplicaciones prácticas en diversos campos.

27.4 AQUÍ VAMOS

27.5 Tipos de error

En una prueba de hipótesis, pueden ocurrir dos tipos de errores.

- **error tipo 1:** Rechazar la hipótesis nula cuando esta es verdadera. La probabilidad de cometer este error es el nivel de significancia α .
- **error tipo 2:** No rechazar la hipótesis nula cuando esta es falsa. La probabilidad de cometer este error se denota como β .

27.6 Tipos de error

En una prueba de hipótesis, pueden ocurrir dos tipos de errores.

27.7 Nivel de significancia

El **nivel de significancia** (α) es la probabilidad de cometer un error tipo 1, es decir, rechazar la hipótesis nula cuando en realidad es cierta. Este nivel se establece antes de realizar la prueba y comúnmente se fija en 0.05 o 0.01.

•

$$\alpha = P(\text{error tipo 1}) = P(\text{rechazar } H_0 \mid H_0)$$

27.8 Rechazar la hipótesis nula

Rechazar H_0 implica que se ha encontrado algo en la muestra tan improbable bajo la hipótesis nula, que lleva al investigador a favorecer la hipótesis alternativa H_1 . Sin embargo, siempre existe la posibilidad de cometer un error tipo 1 al hacer esta decisión.

En estadística, se prefiere no rechazar H_0 erróneamente a rechazarla sin suficiente evidencia. Por esto, H_0 se mantiene a menos que haya **evidencia contundente** que obligue a revocarla. Este enfoque refleja un principio conservador en la ciencia, donde se requiere una alta carga de prueba para cambiar el estado actual del conocimiento.

27.9 Valor p

El **valor p** es la probabilidad de observar un estadístico de prueba tan extremo o más extremo que el observado, bajo la suposición de que la hipótesis nula es cierta.

- Si $p < \alpha$, se rechaza H_0 , lo que sugiere que la evidencia muestral no es compatible con H_0 .
- Si $p \geq \alpha$, no se rechaza H_0 , indicando que los datos son consistentes con H_0 .

Es importante recordar que el valor p no mide la probabilidad de que H_0 sea verdadera, sino la probabilidad de los datos observados bajo H_0 .

27.10 Decisión

La decisión en una prueba de hipótesis se basa en comparar el valor p con el nivel de significancia α :

- **Si** $p < \alpha$, se **rechaza** H_0 .
- **Si** $p \geq \alpha$, **no se rechaza** H_0 .

Esta decisión refleja si la evidencia contra H_0 es lo suficientemente fuerte para considerarla improbable bajo su supuesta veracidad.

27.11 Prueba

El proceso para realizar una prueba de hipótesis incluye los siguientes pasos:

1. **Establecer las hipótesis:** Definir H_0 y H_1 basándose en el problema de investigación.
2. **Formular el sistema de hipótesis** y seleccionar el nivel de significancia α .
3. **Calcular el valor p :** Utilizar los datos muestrales para calcular el estadístico de prueba y el correspondiente valor p .
4. **Tomar la decisión:** Rechazar H_0 si $p < \alpha$ o no rechazar H_0 si $p \geq \alpha$.
5. **Interpretar los resultados:** Explicar el resultado en el contexto del problema de investigación, indicando si hay evidencia suficiente para apoyar H_1 .

27.11.1 El papel de la significancia estadística y el error

El proceso de las pruebas de hipótesis incluye la definición de un nivel de significancia (), que es la probabilidad de rechazar la hipótesis nula cuando es verdadera (error tipo I). Comúnmente, se utiliza un nivel de significancia de 0.05, lo que implica que hay un 5% de probabilidad de cometer un error al rechazar H_0 incorrectamente.

El resultado de una prueba estadística produce un valor p , que representa la probabilidad de observar los datos obtenidos, o más extremos, bajo la suposición de que H_0 es cierta. Si el valor p es menor que el nivel de significancia, se rechaza H_0 en favor de H_1 . Sin embargo, si el valor p es mayor, no se rechaza H_0 , lo que sugiere que no hay suficiente evidencia para afirmar un efecto significativo.

28 Pruebas y tipos de pruebas

Para probar una hipótesis los procedimientos se dividen en cuatro tipos

- ...

28.1 Pruebas paramétricas

Las **pruebas paramétricas** estudian un parámetro de dimensión finita (una colección finita de parámetros), por esta razón presentan supuestos específicos sobre la distribución de los datos.

Las **pruebas paramétricas** estudian un parámetro de dimensión infinita (funciones de densidad).

Analíticas: Hacen uso de propiedades generales de las distribuciones.

Simulación estocástica: Hacen uso de simulación computacional para obtener el valor-p.

Asumen que los datos siguen una distribución conocida, usualmente normal. Estas pruebas requieren que se cumplan ciertos supuestos, como la homogeneidad de varianzas y la linealidad.

Estas pruebas son más potentes cuando los datos cumplen estos supuestos, ya que utilizan toda la información disponible en los datos.

Cuando los datos no cumplen los supuestos, las pruebas pierden validez y los estudios rigor científico.

- Prueba t de Student (para la comparación de medias)
- ANOVA (para la comparación de medias entre múltiples grupos)

28.2 Pruebas no paramétricas analíticas

Las **pruebas no paramétricas analíticas** hacen uso de propiedades matemáticas generales de las distribuciones continuas. Por lo tanto, no requieren supuestos tan fuertes sobre la distribución de los datos.

Son útiles cuando los datos no cumplen los supuestos necesarios para las pruebas paramétricas, como en casos de datos con distribuciones no normales o escalas de medición ordinales.

Son menos potentes que las pruebas paramétricas en los casos paramétricos. No obstante, son más flexibles y aplicables a una mayor variedad de situaciones.

- Prueba de Mann-Whitney (para la comparación de dos grupos independientes)
- Prueba de Wilcoxon (para la comparación de dos grupos pareados)
- Prueba de Kruskal-Wallis (para la comparación de múltiples grupos)

28.3 Pruebas no paramétricas de simulación estocástica

Las **pruebas de permutaciones** son métodos estadísticos no paramétricos que evalúan la significancia de un estadístico de prueba al permutar, simular o remuestrear los datos observados con el fin de obtener una muestra de estadísticos bajo H_0 .

Al comparar el estadístico observado con la muestra de los estadísticos obtenidos, se puede determinar si el efecto observado es significativo.

- **Permutación:** Se basa en la idea de que si no hay efecto, todas las permutaciones de los datos son igualmente probables.
- **Simulación:** Se basa en la idea de que es posible simular la hipótesis nula y encontrar una muestra para el estimador bajo estas condiciones.
- **Remuestreo:** Se basa en la idea de que una muestra aleatoria de una muestra aleatoria es una muestra aleatoria.

Al comparar el estadístico observado con la distribución de los estadísticos obtenidos de las permutaciones, se puede determinar si el efecto observado es significativo.

Calculo del valor p: Se calcula como la proporción de permutaciones en las que el estadístico de prueba es al menos tan extremo como el observado.

Estas pruebas no dependen de supuestos sobre la distribución de los datos, lo que las hace muy flexibles y robustas frente a diversas condiciones de los datos.