Notas de Estadística Inferencial

Andrés Fabián Leal Archila

7 de julio de 2025

Índice general

Capítulo 1

Contenido del curso

Temas del curso de Estadística Inferencial

1. **Repaso de estadística descriptiva:** Capítulo 2.
   - Tipos de variables y escalas de medición
   - Medidas de tendencia central y dispersión
   - Tablas y gráficos
2. **Distribución normal y distribuciones muestrales:**
   - Distribución normal y su importancia
   - Estandarización y uso de tablas
   - Teorema Central del Límite
   - Distribución muestral de la media y la proporción
3. **Muestreo**
   - Población, muestra y parámetros
   - Tipos de muestreo: probabilístico y no probabilístico
   - Ejemplos y aplicaciones reales
   - Tamaño de muestra y error muestral
4. **Estimación**
   - Estimadores y estimación puntual
   - Intervalos de confianza para media, proporción y varianza
   - Interpretación y utilidad
5. **Pruebas de hipótesis para una población**
   - Hipótesis nula y alternativa
   - Prueba z y prueba t para la media
   - Prueba para proporciones
   - Valor-p y toma de decisiones
6. **Pruebas de hipótesis para dos poblaciones**
   - Pruebas z y t para dos medias (independientes y pareadas)

- Prueba para dos proporciones
- Comparación de varianzas (prueba F)

7. **Análisis de la varianza (ANOVA)**
   - ANOVA de un factor
   - Supuestos y procedimiento
   - Ejemplos y aplicaciones
8. **Regresión lineal simple y correlación**
   - Modelo de regresión lineal simple
   - Interpretación de coeficientes
   - Correlación de Pearson
   - Diagnóstico gráfico y supuestos
9. **Aplicaciones computacionales**
   - Uso de Python y R para inferencia estadística
   - Simulación y visualización de resultados
10. **Integración y casos de estudio**
    - Resolución de problemas integradores
    - Análisis de casos reales por carrera

Capítulo 2

Estadística Descriptiva

2.1. Introducción

La estadística descriptiva constituye el fundamento sobre el cual se construye todo el edificio de la ciencia estadística. Su importancia radica en su capacidad para transformar grandes volúmenes de datos brutos en información significativa y comprensible, facilitando así la toma de decisiones informadas en campos tan diversos como la ingeniería, la medicina, el marketing y la investigación científica.

Históricamente, la necesidad de organizar y resumir información ha acompañado a la humanidad desde las primeras civilizaciones. Los censos realizados en Estados Unidos desde 1790 ejemplifican esta tradición, mientras que el desarrollo teórico moderno debe mucho a la escuela matemática rusa, con figuras como P. L. Chebyshev (1821-1894) y sus discípulos A. A. Markov (1856-1922) y A. M. Lyapunov (1857-1918), quienes consolidaron la teoría de probabilidades como una ciencia matemática rigurosa.

Definición 2.0.1 (Estadística Descriptiva)**.** *La estadística descriptiva es la rama de la estadística que se encarga de la* ***recolección, organización, resumen y presentación de datos de manera informativa****, con el objetivo de facilitar su comprensión e interpretación.*

2.2. Conceptos Fundamentales

Definición 2.0.2 (Estadística)**.** *La estadística es el* ***arte y la ciencia de recolectar, analizar, presentar e interpretar datos*** *con el fin de facilitar la toma de decisiones más eficaces.*

Definición 2.0.3 (Dato)**.** *Un dato es un* ***hecho, información o cifra que se recolecta, analiza y resume*** *para su presentación e interpretación. Representa el valor específico que toma una variable en un individuo particular de la población o muestra bajo estudio.*

Definición 2.0.4 (Variable)**.** *Una variable es una* ***característica observable que puede tomar diferentes valores****. Las variables se clasifican en:*

- ***Variables Cualitativas (Categóricas):*** *Utilizan etiquetas o nombres para identificar atributos*
- ***Variables Cuantitativas (Numéricas):*** *Representan valores numéricos que indican cantidad*

Definición 2.0.5 (Escalas de Medición)**.** *Las escalas de medición determinan las operaciones matemáticas válidas que pueden realizarse con los datos:*

- ***Escala Nominal:*** *Etiquetas sin orden inherente (ej: tipo de sangre)*

- ***Escala Ordinal:*** *Datos con orden jerárquico (ej: calificación de servicio)*
- ***Escala de Intervalo:*** *Intervalos uniformes con punto cero arbitrario (ej: temperatura en Celsius)*
- ***Escala de Razón:*** *Intervalos uniformes con punto cero absoluto (ej: peso, altura)*

Definición 2.0.6 (Población y Muestra)**.**

- ***Población:*** *El conjunto completo de todos los elementos de interés en un estudio determinado*
- ***Muestra:*** *Un subconjunto representativo de la población*

Observación 2.0.1. *La distinción entre población y muestra es fundamental: las medidas calculadas a partir de una muestra se denominan* ***estadísticos muestrales****, mientras que las medidas poblacionales se conocen como* ***parámetros poblacionales****. Un estadístico muestral sirve como estimador puntual del parámetro poblacional correspondiente.*

2.3. Organización y Presentación de Datos

La organización efectiva de los datos es el primer paso hacia su comprensión. Los métodos más comunes incluyen:

Definición 2.0.7 (Tabla de Frecuencias)**.** *Una tabla de frecuencias es un* ***resumen tabular de los datos*** *que muestra las frecuencias absolutas, relativas, acumuladas y relativas acumuladas de cada valor o intervalo de valores.*

Definición 2.0.8 (Distribución de Frecuencias)**.** *Para datos agrupados en intervalos, la distribución de frecuencias organiza los datos en clases mutuamente excluyentes, facilitando la identificación de patrones y tendencias.*

2.4. Medidas de Tendencia Central

Las medidas de tendencia central describen el valor típico o representativo de un conjunto de datos, indicando hacia dónde se concentran las observaciones.

Definición 2.0.9 (Media Aritmética)**.** *La media aritmética es la* ***suma de todos los valores dividida por el número de observaciones****.*

Para una muestra:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{2.4.1}$$

Para una población:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{2.4.2}$$

Definición 2.0.10 (Mediana)**.** *La mediana es el* ***valor que ocupa la posición central*** *cuando los datos se ordenan de menor a mayor. Si el número de observaciones es par, es el promedio de los dos valores centrales.*

Definición 2.0.11 (Moda)**.** *La moda es el* ***valor que aparece con mayor frecuencia*** *en un conjunto de datos. Es la única medida de tendencia central aplicable a datos nominales.*

Observación 2.0.2. *La elección de la medida de tendencia central apropiada depende del tipo de datos y la distribución:*

- *La* ***media*** *es ideal para datos simétricos sin valores extremos*
- *La* ***mediana*** *es preferible para distribuciones asimétricas o con outliers*
- *La* ***moda*** *es útil para datos categóricos y para identificar valores más comunes*

2.5. Medidas de Dispersión

Las medidas de dispersión cuantifican la variabilidad o dispersión de los datos respecto a un valor central, proporcionando información crucial sobre la homogeneidad del conjunto de datos.

Definición 2.0.12 (Rango)**.** *El rango es la* ***diferencia entre el valor máximo y el valor mínimo*** *de un conjunto de datos:*

$$R = x_{(n)} - x_{(1)} \tag{2.5.1}$$

Definición 2.0.13 (Varianza)**.** *La varianza mide la* ***variabilidad promedio de los datos respecto a su media****, basándose en los cuadrados de las desviaciones.*

Para una muestra:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{2.5.2}$$

Para una población:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \tag{2.5.3}$$

Definición 2.0.14 (Desviación Estándar)**.** *La desviación estándar es la* ***raíz cuadrada positiva de la varianza****, expresada en las mismas unidades que los datos originales:*

$$s = \sqrt{s^2} \quad \textit{(muestra)} \qquad \sigma = \sqrt{\sigma^2} \quad \textit{(población)} \tag{2.5.4}$$

Definición 2.0.15 (Coeficiente de Variación)**.** *El coeficiente de variación es una* ***medida de variabilidad relativa*** *que permite comparar la dispersión entre conjuntos de datos con diferentes unidades o escalas:*

$$CV = \frac{s}{\bar{x}} \times 100\,\% \tag{2.5.5}$$

Observación 2.0.3. *Una desviación estándar menor indica mayor concentración de los datos alrededor de la media, lo que en contextos aplicados como control de calidad se interpreta como mayor consistencia y mejor calidad del proceso.*

2.6. Medidas de Posición y Forma

Definición 2.0.16 (Cuartiles)**.** *Los cuartiles son valores que* ***dividen los datos ordenados en cuatro partes iguales****:*

- Q_1 *(primer cuartil): 25 % de los datos son menores o iguales a este valor*
- Q_2 *(segundo cuartil): coincide con la mediana*
- Q_3 *(tercer cuartil): 75 % de los datos son menores o iguales a este valor*

Definición 2.0.17 (Percentiles)**.** *Los percentiles son valores que* ***dividen los datos en 100 partes iguales****. El percentil P_k es el valor por debajo del cual se encuentra el $k\,\%$ de las observaciones.*

Definición 2.0.18 (Asimetría (Sesgo))**.** *La asimetría mide la* ***falta de simetría*** *en la distribución de los datos:*

- *Sesgo positivo: cola más larga hacia la derecha*
- *Sesgo cero: distribución simétrica*
- *Sesgo negativo: cola más larga hacia la izquierda*

Definición 2.0.19 (Curtosis)**.** *La curtosis mide el* ***grado de concentración*** *de los datos alrededor de la media:*

- *Leptocúrtica: más concentrada que la normal (curtosis >3)*
- *Mesocúrtica: similar a la normal (curtosis = 3)*
- *Platicúrtica: menos concentrada que la normal (curtosis <3)*

2.7. Representación Gráfica

Definición 2.0.20 (Histograma)**.** *Un histograma es una* ***representación gráfica de la distribución de frecuencias*** *mediante barras rectangulares cuya área es proporcional a la frecuencia de cada intervalo.*

Definición 2.0.21 (Diagrama de Caja y Bigotes)**.** *El diagrama de caja resume la distribución mediante* ***cinco valores estadísticos clave****: valor mínimo, primer cuartil, mediana, tercer cuartil y valor máximo, facilitando la identificación de valores atípicos.*

Observación 2.0.4. *Los gráficos no solo facilitan la comprensión de los datos para el analista, sino que son herramientas fundamentales para comunicar hallazgos a audiencias técnicas y no técnicas.*

2.8. Teoremas Fundamentales

Teorema 2.1 (Desigualdad de Chebyshev)**.** *Para cualquier conjunto de datos, independientemente de su distribución, al menos* $1 - \frac{1}{k^2}$ *de los valores se encuentran dentro de* k *desviaciones estándar de la media, donde* $k > 1$*.*

Teorema 2.2 (Regla Empírica)**.** *Para datos que siguen aproximadamente una distribución normal:*

- *Aproximadamente 68 % de los datos está dentro de 1 desviación estándar de la media*
- *Aproximadamente 95 % de los datos está dentro de 2 desviaciones estándar de la media*
- *Aproximadamente 99.7 % de los datos está dentro de 3 desviaciones estándar de la media*

Teorema 2.3 (Teorema de Glivenko-Cantelli)**.** *La función de distribución empírica converge uniformemente a la función de distribución poblacional cuando el tamaño de la muestra tiende a infinito, estableciendo el fundamento teórico para la inferencia estadística.*

2.9. Aplicaciones Prácticas

2.9.1. Ingeniería y Control de Calidad

En ingeniería, la estadística descriptiva es fundamental para el Control Estadístico de Procesos (SPC). Por ejemplo, en el control de la profundidad de un chavetero, se calculan promedios y rangos de subgrupos para monitorear la estabilidad del proceso.

Observación 2.3.1. *En metodologías como Seis Sigma, la variación se cuantifica mediante la desviación estándar, buscando reducirla hasta alcanzar 3.4 defectos por millón de oportunidades.*

2.9.2. Medicina y Epidemiología

En investigación médica, las medidas descriptivas son esenciales para caracterizar poblaciones de estudio y resumir datos de biomarcadores. Los diagramas de caja son particularmente útiles para visualizar la distribución de respuestas a tratamientos y identificar valores atípicos.

2.9.3. Marketing y Análisis de Mercado

En marketing, los estadísticos descriptivos permiten estimar preferencias de consumidores y evaluar la efectividad de estrategias publicitarias. Los diagramas de Pareto ayudan a identificar los problemas más frecuentes o costosos.

2.10. Ejemplo Integrador

Consideremos el siguiente conjunto de datos que representa las edades de una muestra de 10 empleados: 25, 28, 30, 32, 35, 35, 40, 42, 45, 48.

Cálculos:

$$\bar{x} = \frac{25 + 28 + 30 + 32 + 35 + 35 + 40 + 42 + 45 + 48}{10} = 36 \text{ años} \tag{2.10.1}$$

$$\text{Mediana} = \frac{35 + 35}{2} = 35 \text{ años} \tag{2.10.2}$$

$$\text{Moda} = 35 \text{ años} \tag{2.10.3}$$

$$s^2 = \frac{1}{9}\sum_{i=1}^{10}(x_i - 36)^2 = 56{,}67 \text{ años}^2 \tag{2.10.4}$$

$$s = \sqrt{56{,}67} = 7{,}53 \text{ años} \tag{2.10.5}$$

Observación 2.3.2. *La proximidad entre la media y la mediana (36 vs 35) sugiere una distribución aproximadamente simétrica, lo que se confirma con el valor de sesgo cercano a cero.*

2.11. Herramientas Computacionales

2.12. Conclusiones

La estadística descriptiva proporciona las herramientas fundamentales para transformar datos brutos en información útil. Su dominio es esencial para cualquier análisis estadístico posterior y constituye la base sobre la cual se construyen técnicas más avanzadas de inferencia estadística.

Observación 2.3.3. *El análisis descriptivo no solo resume los datos, sino que también revela patrones, identifica anomalías y sugiere hipótesis para investigaciones futuras. Su valor radica tanto en su capacidad analítica como en su poder comunicativo.*

La transición natural de la estadística descriptiva hacia la inferencial marca el paso de la simple descripción de lo observado hacia la formulación de conclusiones sobre poblaciones no observadas completamente, estableciendo así el puente entre la certeza de los datos y la incertidumbre inherente a la toma de decisiones basada en evidencia parcial.

Capítulo 3

Distribución Normal y Distribuciones Muestrales

3.1. Fundamentos Teóricos

La distribución normal constituye el pilar fundamental de la estadística moderna y representa uno de los conceptos más importantes en la teoría de la probabilidad. Su desarrollo histórico ha sido influenciado significativamente por los trabajos de la llamada .Escuela Rusa", con contribuciones notables de figuras como P.L. Chebyshev y A.M. Liapunov, quienes marcaron un "nuevo período.en el desarrollo de la ciencia matemática.

Definición 3.0.1 (Distribución Normal)**.** *Una variable aleatoria continua X sigue una distribución normal con parámetros μ y σ^2, denotada como $X \sim N(\mu, \sigma^2)$, si su función de densidad de probabilidad está dada por:*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty \tag{3.1.1}$$

donde μ es la media poblacional y σ^2 es la varianza poblacional.

Observación 3.0.1. *La distribución normal, también conocida como campana de Gauss, posee propiedades únicas que la hacen especialmente útil en aplicaciones prácticas:*

- *Es simétrica alrededor de su media μ, siendo unimodal*
- *Su media, mediana y moda coinciden en el mismo valor*
- *Aproximadamente el 68 % del área bajo la curva se encuentra entre $\mu \pm \sigma$*
- *Alrededor del 95 % del área se encuentra entre $\mu \pm 2\sigma$*
- *Aproximadamente el 99.7 % del área se encuentra entre $\mu \pm 3\sigma$*
- *Queda completamente definida por sus dos parámetros: μ y σ*

Ejemplo 3.0.1. *Un ingeniero de control de calidad monitorea el peso de componentes electrónicos en una línea de producción. Si los pesos siguen una distribución normal con $\mu = 50$ gramos y $\sigma = 2$ gramos, podemos establecer que aproximadamente el 95 % de los componentes tendrán un peso entre 46 y 54 gramos ($\mu \pm 2\sigma$). Esta propiedad es fundamental para establecer límites de control en gráficas de control estadístico.*

Aunque esto se demostrará más adelante, es necesario saber que existen unos test para verificar si una variable numérica es normal, esto se muestra en el siguiente teorema:

Teorema 3.1 (Implementación de Test de Normalidad)**.** *Sea $X = \{x_1, x_2, \ldots, x_n\}$ una muestra aleatoria de una variable continua. Para verificar si X proviene de una distribución normal, se pueden aplicar los siguientes test estadísticos:*

- ***Test de Shapiro-Wilk:*** *Evalúa la hipótesis nula H_0: "La muestra proviene de una distribución normal". Se calcula el estadístico W y su p-valor.*
- ***Test de Kolmogorov-Smirnov:*** *Compara la función de distribución empírica de la muestra con la función de distribución acumulada de una normal estándar.*
- ***Test de Anderson-Darling:*** *Es una modificación del test de Kolmogorov-Smirnov, dando más peso a los extremos de la distribución.*
- ***Test de Jarque-Bera:*** *Evalúa la asimetría y curtosis de la muestra respecto a una normal.*

Implementación en Python:

```
from scipy import stats

# Supongamos que 'data' es la variable a analizar

# Shapiro-Wilk
shapiro_stat, shapiro_p = stats.shapiro(data)

# Kolmogorov-Smirnov (comparando con normal estándar)
ks_stat, ks_p = stats.kstest(data, 'norm', args=(data.mean(), data.std()))

# Anderson-Darling
anderson_result = stats.anderson(data, dist='norm')

# Jarque-Bera
jb_stat, jb_p = stats.jarque_bera(data)
```

Decisión: *Si el p-valor es menor que el nivel de significancia (α), se rechaza la hipótesis nula de normalidad. Para Anderson-Darling, se compara el estadístico con los valores críticos proporcionados por el test.*

Observación 3.1.1. ***Aplicación práctica y criterios de los métodos:***
Supongamos que un investigador recolecta los tiempos (en minutos) que tardan 50 estudiantes en resolver un examen. Para determinar si estos tiempos siguen una distribución normal, puede aplicar los siguientes test, considerando sus criterios y usos más comunes:

- ***Shapiro-Wilk:***
  - *Criterio: Recomendado para muestras pequeñas y medianas ($n \leq 2000$), sensible a desviaciones de normalidad.*
  - *Uso común: Psicología, medicina, ciencias sociales y educación, especialmente cuando el tamaño muestral es limitado.*
- ***Kolmogorov-Smirnov:***
  - *Criterio: Adecuado para muestras grandes, aunque menos potente que otros tests; requiere especificar los parámetros de la normal.*

- *Uso común: Comparación de distribuciones empíricas, validación en estudios de ingeniería y economía.*

- ***Anderson-Darling:***
  - *Criterio: Más sensible a los extremos (colas) de la distribución, útil para detectar valores atípicos.*
  - *Uso común: Control de calidad, análisis financiero y estudios donde los valores extremos son relevantes.*

- ***Jarque-Bera:***
  - *Criterio: Basado en asimetría y curtosis; apropiado para muestras grandes ($n > 30$).*
  - *Uso común: Series temporales, econometría y análisis financiero, especialmente en estudios de retornos de activos.*

Por ejemplo, si el p-valor del test de Shapiro-Wilk es 0.03 y el nivel de significancia es 0.05, se concluye que los datos no siguen una distribución normal. Este procedimiento es fundamental antes de aplicar pruebas paramétricas en diversas áreas científicas y profesionales.

3.2. El Teorema Central del Límite

El Teorema Central del Límite representa una de las herramientas más poderosas de la estadística, proporcionando la justificación teórica para el uso de la distribución normal en numerosas aplicaciones prácticas.

Teorema 3.2 (Teorema Central del Límite)**.** *Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una población con media μ y varianza finita σ^2. Cuando n es suficientemente grande, la distribución de la media muestral $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ se aproxima a una distribución normal:*

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \tag{3.2.1}$$

Esta aproximación mejora a medida que n aumenta, independientemente de la forma de la distribución original de la población.

Observación 3.2.1. *La importancia del Teorema Central del Límite radica en que permite hacer inferencias sobre la media poblacional sin necesidad de conocer la forma específica de la distribución de la población. En la práctica, se considera que $n \geq 30$ es suficiente para que la aproximación normal sea adecuada.*

El desarrollo teórico del TCL se fundamenta en teoremas de continuidad y convergencia. Los trabajos de Borovkov presentan el "segundo teorema de continuidadz el "tercer teorema de continuidad", que son fundamentales para entender cómo las propiedades de las funciones de distribución se preservan y cómo una secuencia de distribuciones converge a una distribución límite.

Ejemplo 3.2.1. *Un ingeniero químico estudia un proceso donde el tiempo de reacción tiene una distribución desconocida con media $\mu = 15$ minutos y desviación estándar $\sigma = 4$ minutos. Si toma muestras de $n = 64$ observaciones, el TCL garantiza que la media muestral $\bar{X}$ se distribuirá aproximadamente como $N(15, 4^2/64) = N(15, 0{,}25)$. Esto significa que la desviación estándar de la media muestral será $\sigma/\sqrt{n} = 4/8 = 0{,}5$ minutos.*

3.3. Distribuciones Muestrales

El concepto de distribución muestral es fundamental para comprender la inferencia estadística y la variabilidad inherente en los procesos de muestreo.

Definición 3.2.1 (Distribución Muestral)**.** *La distribución muestral de un estadístico es la distribución de probabilidad de ese estadístico calculado sobre todas las posibles muestras de un tamaño fijo extraídas de una población. Como distintas muestras pueden producir valores diferentes para el mismo estadístico, este se considera una variable aleatoria.*

3.3.1. Distribución Muestral de la Media

Observación 3.2.2. *Para la media muestral $\bar{X}$, se pueden establecer las siguientes propiedades importantes:*

- *Si la población original tiene distribución normal, $\bar{X}$ también tendrá distribución normal, independientemente del tamaño de muestra*
- *Si la población no es normal, el TCL asegura que $\bar{X}$ se aproxima a una distribución normal cuando n es grande*
- *La media de la distribución muestral de $\bar{X}$ es igual a la media poblacional: $E[\bar{X}] = \mu$*
- *La varianza de la distribución muestral de $\bar{X}$ es: $Var(\bar{X}) = \sigma^2/n$*

Ejemplo 3.2.2. *En un estudio de resistencia de cables eléctricos, la población tiene media $\mu = 150$ N y desviación estándar $\sigma = 10$ N. Si se toma una muestra de $n = 36$ cables, la distribución de la media muestral será:*

$$\bar{X} \sim N\left(150, \frac{10^2}{36}\right) = N(150, 2{,}78) \tag{3.3.1}$$

La probabilidad de que la media muestral exceda 153 N es:

$$P(\bar{X} > 153) = P\left(Z > \frac{153 - 150}{\sqrt{10^2/36}}\right) = P(Z > 1{,}80) \approx 0{,}0359 \tag{3.3.2}$$

3.3.2. Distribución Muestral de la Proporción

Definición 3.2.2 (Distribución Muestral de la Proporción)**.** *Si en una población la proporción de elementos con cierta característica es p, y en una muestra de tamaño n se observa la proporción muestral $\hat{p}$, entonces:*

- $E[\hat{p}] = p$
- $Var(\hat{p}) = \frac{p(1-p)}{n}$

Para n suficientemente grande, $\hat{p}$ se distribuye aproximadamente normal:

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right) \tag{3.3.3}$$

Ejemplo 3.2.3. *Una empresa de bebidas realiza una encuesta para evaluar la aceptación de un nuevo producto. Si el 40 % de la población prefiere el producto y se toma una muestra de $n = 100$ clientes, la probabilidad de que al menos 50 prefieran el producto es:*

$$\hat{p} \sim N\left(0{,}4, \frac{0{,}4 \times 0{,}6}{100}\right) = N(0{,}4, 0{,}0024) \tag{3.3.4}$$

$$P(\hat{p} \geq 0{,}5) = P\left(Z \geq \frac{0{,}5 - 0{,}4}{\sqrt{0{,}0024}}\right) = P(Z \geq 2{,}04) \approx 0{,}0207 \tag{3.3.5}$$

3.3.3. Otras Distribuciones Muestrales Importantes

Observación 3.2.3. *Además de la distribución normal, existen otras distribuciones fundamentales que surgen del muestreo:*

- ***Distribución ji-cuadrada*** *(χ^2): Para una muestra de tamaño n de una población normal, el estadístico $\frac{(n-1)S^2}{\sigma^2}$ sigue una distribución χ^2 con $n-1$ grados de libertad*
- ***Distribución t de Student****: Se utiliza para inferencias sobre la media cuando σ es desconocida y se estima con s*
- ***Distribución F****: Se aplica en comparación de varianzas y análisis de varianza (ANOVA)*

3.4. Aplicaciones Prácticas

3.4.1. Ingeniería y Control de Calidad

Ejemplo 3.2.4. *En el control estadístico de procesos (SPC), las gráficas de control se basan en las propiedades de la distribución normal. Una gráfica de control $\bar{X}$ monitorea las medias de subgrupos para detectar cambios en el promedio del proceso. Los límites de control se establecen típicamente en $\mu \pm 3\sigma/\sqrt{n}$, aprovechando que aproximadamente el 99.7 % de las observaciones deben estar dentro de estos límites cuando el proceso está bajo control.*

Si las medias de las muestras caen fuera de estos límites, esto indica que el proceso está fuera de control estadístico, señalando la necesidad de investigar y corregir causas asignables.

Ejemplo 3.2.5. *En el diseño de experimentos, un ingeniero químico optimiza la conversión porcentual de un proceso variando tiempo y temperatura de reacción. Utilizando un diseño factorial 2^2, el análisis de varianza (ANOVA) emplea la distribución F para comparar la influencia de diferentes factores. La metodología de superficie de respuesta (MSR) utiliza modelos de segundo orden para encontrar condiciones óptimas, donde las distribuciones muestrales son fundamentales para evaluar la significancia estadística de los efectos.*

3.4.2. Medicina y Investigación Clínica

Ejemplo 3.2.6. *En un ensayo clínico para evaluar la eficacia de un nuevo medicamento para reducir la presión arterial, se asignan aleatoriamente pacientes a grupos de tratamiento y control. Después del período de tratamiento, se utiliza la prueba t para dos muestras (basada en la distribución t de Student) para determinar si la reducción promedio de presión arterial en el grupo de tratamiento es significativamente diferente del grupo control.*

Si las reducciones promedio son $\bar{X}_1 = 15$ mmHg y $\bar{X}_2 = 8$ mmHg con desviaciones estándar $s_1 = 6$ y $s_2 = 5$ mmHg respectivamente, y tamaños de muestra $n_1 = n_2 = 30$, el estadístico de prueba sería:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{7}{\sqrt{\frac{36}{30} + \frac{25}{30}}} \approx 4{,}96 \tag{3.4.1}$$

3.4.3. Marketing y Estudios de Mercado

Ejemplo 3.2.7. *Una empresa analiza la satisfacción del cliente mediante encuestas. Si las calificaciones de satisfacción se distribuyen normalmente con media $\mu = 7{,}5$ y desviación estándar $\sigma = 1{,}2$ en una escala de 1 a 10, la empresa puede estimar que aproximadamente el 95 % de los clientes tendrán calificaciones entre 5.1 y 9.9 ($\mu \pm 2\sigma$).*

Para una muestra de $n = 100$ clientes, la distribución de la media muestral será $N(7{,}5, 1{,}2^2/100) = N(7{,}5, 0{,}0144)$, permitiendo hacer inferencias más precisas sobre la satisfacción promedio de toda la población de clientes.

3.5. Estandarización y Cálculo de Probabilidades

Observación 3.2.4. *La estandarización es un proceso fundamental que permite convertir cualquier variable normal en una variable normal estándar. Para una variable $X \sim N(\mu, \sigma^2)$, la transformación:*

$$Z = \frac{X - \mu}{\sigma} \tag{3.5.1}$$

produce una variable $Z \sim N(0, 1)$, lo que permite el uso de tablas estándar y facilita la comparación entre diferentes poblaciones.

Ejemplo 3.2.8. *La vida útil de un componente electrónico sigue una distribución $N(1200, 100^2)$ horas. Para encontrar la probabilidad de que un componente dure más de 1350 horas:*

$$P(X > 1350) = P\left(\frac{X - 1200}{100} > \frac{1350 - 1200}{100}\right) \tag{3.5.2}$$

$$= P(Z > 1{,}5) = 1 - \Phi(1{,}5) = 1 - 0{,}9332 = 0{,}0668 \tag{3.5.3}$$

donde Φ representa la función de distribución acumulativa de la normal estándar.

Ejemplo 3.2.9. *En una línea de producción, el peso de las botellas sigue $N(500, 25)$ gramos. Para cumplir normas de calidad, solo se aceptan botellas entre 490 g y 510 g. El porcentaje de producción aceptable es:*

$$P(490 < X < 510) = P\left(\frac{490 - 500}{5} < Z < \frac{510 - 500}{5}\right) \tag{3.5.4}$$

$$= P(-2 < Z < 2) = \Phi(2) - \Phi(-2) \tag{3.5.5}$$

$$= 0{,}9772 - 0{,}0228 = 0{,}9544 \tag{3.5.6}$$

Por tanto, el 95.44 % de las botellas cumplen la especificación.

3.6. Simulación y Verificación Computacional

Observación 3.2.5. *La simulación computacional permite verificar los conceptos teóricos y visualizar el comportamiento de las distribuciones muestrales. Para ilustrar el Teorema Central del Límite, se puede simular la distribución de medias muestrales a partir de poblaciones no normales y observar cómo la distribución resultante se aproxima a la normalidad.*

Ejemplo 3.2.10 (Simulación del TCL)**.** *Para demostrar el Teorema Central del Límite utilizando Python:*

```
import numpy as np
import matplotlib.pyplot as plt

# Población exponencial (no normal)
np.random.seed(42)
population = np.random.exponential(scale=1.0, size=100000)

# Generar múltiples medias muestrales
sample_means = []
for _ in range(1000):
    sample = np.random.choice(population, size=30)
    sample_means.append(np.mean(sample))
```

```
# Visualizar la distribución de medias muestrales
plt.hist(sample_means, bins=50, density=True, alpha=0.7)
plt.title('Distribución de Medias Muestrales (n=30)')
plt.xlabel('Media Muestral')
plt.ylabel('Densidad')
plt.show()
```

Este código demuestra cómo las medias muestrales se distribuyen normalmente incluso cuando la población original no es normal.

3.7. Ejercicios y Problemas

Ejemplo 3.2.11. *Un fabricante desea garantizar que el 90 % de sus productos duren al menos T horas. Si la vida útil es $N(2000, 150^2)$, encontrar el valor de T.*

Solución: *Se requiere $P(X > T) = 0{,}90$, lo que implica $P(X < T) = 0{,}10$. Para la distribución normal estándar, $P(Z < -1{,}2816) = 0{,}10$. Por tanto:*

$$T = \mu + Z\sigma = 2000 + (-1{,}2816)(150) = 1808 \text{ horas} \tag{3.7.1}$$

Ejemplo 3.2.12. *En un proceso industrial, las piezas tienen longitud normal $N(50, 0{,}25)$ mm. Determinar la proporción de piezas que miden entre 49.2 mm y 50.5 mm.*

Solución:

$$P(49{,}2 < X < 50{,}5) = P\left(\frac{49{,}2 - 50}{0{,}5} < Z < \frac{50{,}5 - 50}{0{,}5}\right) \tag{3.7.2}$$

$$= P(-1{,}6 < Z < 1{,}0) \tag{3.7.3}$$

$$= \Phi(1{,}0) - \Phi(-1{,}6) = 0{,}8413 - 0{,}0548 = 0{,}7865 \tag{3.7.4}$$

Aproximadamente el 78.65 % de las piezas cumple la especificación.

3.8. Conclusiones

La distribución normal y las distribuciones muestrales forman la base conceptual de la inferencia estadística moderna. El Teorema Central del Límite proporciona la justificación teórica para el uso generalizado de la distribución normal, mientras que las distribuciones muestrales permiten cuantificar la incertidumbre en las estimaciones estadísticas.

La comprensión de estos conceptos es fundamental para aplicaciones en ingeniería, medicina, marketing y otras disciplinas, donde la toma de decisiones basada en datos requiere un entendimiento sólido de la variabilidad y la incertidumbre inherentes en los procesos de medición y muestreo.

Observación 3.2.6. *Los conceptos presentados en este capítulo son prerrequisitos esenciales para el estudio de la estimación de parámetros, pruebas de hipótesis y análisis de regresión, temas que se desarrollarán en capítulos posteriores.*

Capítulo 4

Estimación

La estadística inferencial constituye uno de los pilares fundamentales de la ciencia estadística, permitiendo **utilizar los resultados derivados de las muestras para obtener conclusiones acerca de las características de una población**. A diferencia de la estadística descriptiva, que se enfoca en organizar y resumir datos observados, la inferencia estadística busca determinar propiedades de una población completa a partir de un subconjunto limitado de datos conocido como muestra.

Definición 4.0.1 (Parámetro y Estadístico)**.** *Un **parámetro** es un valor numérico que resume una característica de toda la población (por ejemplo, μ, p, σ^2). Un **estadístico** es una función calculada a partir de los datos de la muestra (por ejemplo, $\overline{X}$, $\hat{p}$, s^2).*

Dentro de la inferencia estadística, la **estimación** representa una de las dos áreas principales, siendo la otra las pruebas de hipótesis. La estimación busca proporcionar información sobre los parámetros poblacionales desconocidos mediante el uso de estadísticos muestrales.

Definición 4.0.2 (Estimador y Estimación)**.** *Un **estimador** es una función de los datos muestrales que se usa para aproximar un parámetro poblacional. El valor numérico que toma el estimador en una muestra concreta se denomina **estimación puntual**.*

*La **estimación por intervalos** proporciona un rango de valores (intervalo de confianza) dentro del cual es probable que se encuentre el parámetro poblacional, junto con un nivel de confianza asociado.*

Observación 4.0.1. *Un buen estimador debe cumplir con ciertas propiedades deseables: ser insesgado (su valor esperado es igual al parámetro), eficiente (mínima varianza), consistente (converge al parámetro al aumentar el tamaño muestral) y robusto (poco sensible a valores atípicos).*

Para comprender adecuadamente la estimación, es crucial dominar el concepto de **distribución muestral**, que representa la distribución de todos los posibles valores de un estadístico que podrían surgir al seleccionar todas las muestras posibles de un tamaño determinado.

4.1. Estimación para la Media Poblacional

La media poblacional (μ) representa el valor promedio de todos los elementos en una población. Es un parámetro constante que no se ve afectado por las observaciones de una muestra particular.

4.1.1. Estimación Puntual de la Media

Definición 4.0.3 (Media Muestral)**.** *El estimador puntual más importante para la media poblacional μ es la media muestral:*

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

donde $X_1, X_2, \ldots, X_n$ son las observaciones de la muestra.

Teorema 4.1 (Insesgadez de la Media Muestral)**.** *La media muestral $\overline{X}$ es un estimador insesgado de la media poblacional μ, es decir:*

$$\mathbb{E}[\overline{X}] = \mu$$

4.1.2. Estimación por Intervalos para la Media

Definición 4.1.1 (Intervalo de Confianza)**.** *Un intervalo de confianza para un parámetro poblacional es un rango de valores que, con un nivel de confianza dado $(1-\alpha)$, contiene el valor real del parámetro poblacional.*

La construcción de intervalos de confianza para la media depende de si la desviación estándar poblacional es conocida o desconocida.

Caso 1: Desviación estándar poblacional (σ) conocida

Cuando se dispone de información histórica abundante o datos previos que permiten conocer σ, el intervalo de confianza se basa en la distribución normal estándar:

$$\text{IC}_{1-\alpha} : \left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \right)$$

Caso 2: Desviación estándar poblacional (σ) desconocida

Esta es la situación más común en la práctica. Se utiliza la desviación estándar muestral (s) y la distribución t de Student:

$$\text{IC}_{1-\alpha} : \left(\overline{X} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}, \overline{X} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} \right)$$

Observación 4.1.1. *La distribución t de Student es similar a la normal estándar pero tiene çolas"más pesadas, reflejando la incertidumbre adicional de estimar σ. Los grados de libertad $(n-1)$ son fundamentales para determinar la forma exacta de la distribución.*

Ejemplo 4.1.1 (Intervalo de Confianza para la Media)**.** *Una muestra de $n = 25$ resistencias eléctricas tiene una media de $\overline{X} = 98$ ohmios y una desviación estándar de $s = 5$ ohmios. Calcule un intervalo de confianza del 95 % para la media poblacional.*

Solución:

1. *Nivel de confianza = 95 % $\Rightarrow \alpha = 0{,}05 \Rightarrow \alpha/2 = 0{,}025$*
2. *Grados de libertad $= n - 1 = 24$*
3. *Valor crítico: $t_{0{,}025,24} \approx 2{,}064$*
4. *Error estándar: $SE = \frac{s}{\sqrt{n}} = \frac{5}{\sqrt{25}} = 1$*
5. *Límites del intervalo:*

$$\textit{Límite inferior} = 98 - 2{,}064 \times 1 = 95{,}936 \tag{4.1.1}$$

$$\textit{Límite superior} = 98 + 2{,}064 \times 1 = 100{,}064 \tag{4.1.2}$$

Por tanto, el intervalo de confianza del 95 % para la media poblacional es $(95{,}936, 100{,}064)$ *ohmios.*

4.1.3. Supuestos y Condiciones de Aplicabilidad

Teorema 4.2 (Teorema del Límite Central)**.** *Si el tamaño de muestra n es suficientemente grande (generalmente $n \geq 30$), la distribución muestral de la media será aproximadamente normal, independientemente de la forma de la distribución poblacional original:*

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Observación 4.2.1. *Los supuestos fundamentales para la estimación de la media incluyen:*

- ***Normalidad:*** *Para muestras pequeñas, se asume que la población sigue una distribución normal.*
- ***Independencia:*** *Las observaciones deben ser independientes entre sí.*
- ***Robustez:*** *La distribución t es razonablemente robusta a desviaciones leves de la normalidad, especialmente con muestras grandes.*

4.2. Estimación para la Proporción Poblacional

La proporción poblacional (p) representa la fracción de elementos en una población que poseen una característica específica de interés. Es fundamental en estudios que involucran datos categóricos o binarios.

4.2.1. Estimación Puntual de la Proporción

Definición 4.2.1 (Proporción Muestral)**.** *El estimador puntual para la proporción poblacional p es la proporción muestral:*

$$\hat{p} = \frac{X}{n}$$

donde X es el número de elementos en la muestra que poseen la característica de interés y n es el tamaño de la muestra.

Teorema 4.3 (Insesgadez de la Proporción Muestral)**.** *La proporción muestral $\hat{p}$ es un estimador insesgado de la proporción poblacional p:*

$$\mathbb{E}[\hat{p}] = p$$

4.2.2. Estimación por Intervalos para la Proporción

Para muestras grandes, cuando se cumplen las condiciones $n\hat{p} \geq 5$ y $n(1-\hat{p}) \geq 5$, el intervalo de confianza se basa en la aproximación normal:

$$\text{IC}_{1-\alpha} : \left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

Ejemplo 4.3.1 (Intervalo de Confianza para la Proporción)**.** *En una encuesta realizada a 100 personas, 56 manifestaron preferir un producto específico. Calcule un intervalo de confianza del 95 % para la proporción poblacional.*

Solución:

1. *Proporción muestral:* $\hat{p} = \frac{56}{100} = 0{,}56$
2. *Nivel de confianza = 95 %* $\Rightarrow z_{0{,}025} = 1{,}96$

3. *Verificación de condiciones:* $n\hat{p} = 56 \geq 5$ *y* $n(1-\hat{p}) = 44 \geq 5$

4. *Error estándar:* $SE = \sqrt{\frac{0{,}56 \times 0{,}44}{100}} = \sqrt{0{,}002464} = 0{,}0496$

5. *Límites del intervalo:*

$$\text{Límite inferior} = 0{,}56 - 1{,}96 \times 0{,}0496 = 0{,}463 \tag{4.2.1}$$
$$\text{Límite superior} = 0{,}56 + 1{,}96 \times 0{,}0496 = 0{,}657 \tag{4.2.2}$$

El intervalo de confianza del 95 % para la proporción poblacional es $(0{,}463, 0{,}657)$.

Observación 4.3.1. *Los contextos típicos donde se aplica la estimación de proporciones incluyen:*

- *Encuestas de opinión pública*
- *Estudios clínicos y biomédicos*
- *Control de calidad en procesos de manufactura*
- *Investigación de mercado*

4.3. Estimación para la Varianza Poblacional

La varianza poblacional (σ^2) y su raíz cuadrada, la desviación estándar (σ), son medidas fundamentales de la dispersión o variabilidad de los datos.

4.3.1. Estimación Puntual de la Varianza

Definición 4.3.1 (Varianza Muestral)**.** *El estimador puntual insesgado para la varianza poblacional* σ^2 *es la varianza muestral:*

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

Observación 4.3.2. *La varianza muestral utiliza* $(n-1)$ *en el denominador en lugar de* n *para garantizar que sea un estimador insesgado. Esta corrección se conoce como corrección de Bessel.*

4.3.2. Estimación por Intervalos para la Varianza

Cuando la población sigue una distribución normal, el intervalo de confianza para la varianza se basa en la distribución chi-cuadrado (χ^2):

$$\text{IC}_{1-\alpha} : \left(\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}} \right)$$

Ejemplo 4.3.2 (Intervalo de Confianza para la Varianza)**.** *En una muestra de* $n = 10$ *mediciones se obtuvo* $s^2 = 2{,}5$*. Calcule un intervalo de confianza del 95 % para la varianza poblacional.*

Solución:

1. *Grados de libertad:* $n - 1 = 9$

2. *Valores críticos:* $\chi^2_{0{,}025,9} = 19{,}02$ *y* $\chi^2_{0{,}975,9} = 2{,}70$

3. Límites del intervalo:

$$\textit{Límite inferior} = \frac{9 \times 2{,}5}{19{,}02} = 1{,}18 \tag{4.3.1}$$

$$\textit{Límite superior} = \frac{9 \times 2{,}5}{2{,}70} = 8{,}33 \tag{4.3.2}$$

El intervalo de confianza del 95 % para la varianza poblacional es $(1{,}18, 8{,}33)$.

Observación 4.3.3. *La estimación de la varianza es crucial en diversas aplicaciones:*

- ***Control de calidad:*** *Garantizar que un proceso cumpla con especificaciones*
- ***Diseño experimental:*** *Reducir el error experimental para obtener inferencias más precisas*
- ***Finanzas:*** *Medir el riesgo asociado a inversiones*
- ***Estudios de capacidad:*** *Evaluar la capacidad de procesos industriales*

4.4. Interpretación y Aplicaciones de los Intervalos de Confianza

Definición 4.3.2 (Nivel de Confianza)**.** *El nivel de confianza* $(1 - \alpha)$ *es la probabilidad de que el intervalo calculado contenga el valor real del parámetro si el procedimiento de muestreo se repitiera muchas veces bajo las mismas condiciones.*

Observación 4.3.4. *Consideraciones importantes sobre los intervalos de confianza:*

- *Un intervalo más estrecho indica mayor precisión en la estimación*
- *La precisión se puede mejorar aumentando el tamaño de muestra o reduciendo la variabilidad*
- *El margen de error es la mitad de la amplitud del intervalo*
- *Un resultado estadísticamente significativo no implica necesariamente relevancia práctica*

4.5. Determinación del Tamaño de Muestra

Para estimar parámetros poblacionales con un margen de error específico y un nivel de confianza dado, es posible determinar el tamaño de muestra necesario:

Para la media (con σ conocida):

$$n = \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2$$

Para la proporción:

$$n = \left(\frac{z_{\alpha/2}}{E}\right)^2 \hat{p}(1 - \hat{p})$$

donde E es el margen de error deseado.

Observación 4.3.5. *La determinación del tamaño de muestra es fundamental en la planificación de estudios, especialmente cuando existen restricciones de presupuesto y tiempo. Para proporciones sin estimación previa, se puede usar* $\hat{p} = 0{,}5$ *para obtener el tamaño de muestra más conservador.*

4.6. Consideraciones Éticas y Prácticas

Observación 4.3.6. *Es fundamental que los resultados de estimación se presenten de manera ética e imparcial, incluyendo:*

- *Las estimaciones puntuales junto con sus intervalos de confianza*
- *El tamaño de muestra utilizado*
- *Las suposiciones realizadas*
- *Una interpretación clara del significado de los resultados*
- *Las limitaciones del estudio*

La estimación estadística proporciona las herramientas esenciales para transformar datos muestrales en información confiable que sustente la toma de decisiones en diversas disciplinas, desde la administración y economía hasta la ingeniería y medicina. La correcta aplicación de estas técnicas, junto con una interpretación adecuada de los resultados, constituye la base de la inferencia estadística moderna.

Capítulo 5

Fundamentos del Muestreo Estadístico

5.1. Introducción

El muestreo estadístico constituye una disciplina fundamental que permite obtener conclusiones fiables sobre grandes conjuntos de datos mediante el análisis de una porción representativa de ellos. Esta herramienta es esencial en la investigación científica, la toma de decisiones empresariales y el análisis de fenómenos sociales, económicos y naturales.

El presente capítulo explora los fundamentos teóricos del muestreo estadístico, sus diversas aplicaciones en múltiples campos y los métodos prácticos para su implementación, integrando perspectivas clásicas y modernas que han enriquecido esta disciplina a lo largo de su desarrollo histórico.

5.2. Conceptos Fundamentales

5.2.1. Población y Muestra

Definición 5.0.1. *Una **población** (o universo) es el conjunto completo de todos los elementos de interés en un estudio determinado. Puede ser finita o infinita.*

Definición 5.0.2. *Una **muestra** es un subconjunto de la población seleccionado para su análisis.*

Definición 5.0.3. *El **muestreo** es el proceso sistemático de seleccionar elementos de una población mediante un conjunto de reglas, procedimientos y criterios, con el objetivo de hacer inferencias válidas sobre la población completa.*

Observación 5.0.1. *El análisis de una población completa puede ser impracticable debido a limitaciones de tiempo, costo, accesibilidad o incluso por razones éticas. Por ejemplo, sería imposible e innecesario encuestar a todos los ciudadanos de un país para conocer la intención de voto, o analizar todos los componentes de un lote de productos para evaluar su calidad.*

Definición 5.0.4. *Un **censo** es un estudio que recolecta datos de toda la población, mientras que una **encuesta muestral** recolecta datos únicamente de una muestra.*

Definición 5.0.5. *La **población objetivo** es la población sobre la cual se desean hacer inferencias, mientras que la **población muestreada** es la población de la que efectivamente se toma la muestra.*

Definición 5.0.6. *El **marco de muestreo** es una lista completa de los componentes de la población objetivo de la cual se seleccionará la muestra.*

Observación 5.0.2. *La validez de las conclusiones de una encuesta muestral aplicadas a la población objetivo depende crucialmente de la semejanza entre la población objetivo y la población muestreada. Las discrepancias entre ambas pueden introducir sesgos sistemáticos en los resultados.*

5.2.2. Representatividad y Variabilidad

Definición 5.0.7. *Una muestra es* ***representativa*** *cuando encapsula toda la variabilidad posible de la población, permitiendo que los resultados obtenidos en la muestra sean extrapolables a la población completa.*

Observación 5.0.3. *La representatividad es el principio fundamental del muestreo estadístico. Una muestra no representativa puede conducir a conclusiones erróneas y decisiones inadecuadas, independientemente del tamaño de la muestra o la sofisticación de los métodos de análisis empleados.*

5.3. Ramas de la Estadística

El campo de la estadística se divide tradicionalmente en dos ramas complementarias:

Definición 5.0.8. *La* ***estadística descriptiva*** *se ocupa de la recolección, tabulación, análisis, interpretación y presentación de datos. Consiste en procedimientos para organizar y resumir datos, transformando información bruta en formas significativas mediante tablas, gráficas y medidas numéricas.*

Definición 5.0.9. *La* ***estadística inferencial*** *implica tomar una muestra de una población y realizar cálculos sobre esta muestra para determinar características de la población completa. Su propósito es hacer generalizaciones sobre una población basándose únicamente en una muestra.*

Observación 5.0.4. *Las inferencias estadísticas no pueden afirmarse con certidumbre absoluta debido a la variabilidad inherente en los procesos de muestreo. Por esta razón, la estadística inferencial emplea el lenguaje de la probabilidad para cuantificar la incertidumbre asociada con las conclusiones.*

5.3.1. Estadísticos y Parámetros

Definición 5.0.10. *Los* ***estadísticos muestrales*** *(o simplemente estadísticos) son valores numéricos calculados a partir de una muestra.*

Definición 5.0.11. *Los* ***parámetros poblacionales*** *(o parámetros) son medidas numéricas que caracterizan a una población completa.*

Definición 5.0.12. *Un* ***estimador puntual*** *es un estadístico muestral que se utiliza para estimar el valor de un parámetro poblacional correspondiente.*

Observación 5.0.5. *Las propiedades deseables de un buen estimador puntual incluyen:*

- ***Insesgadez****: El valor esperado del estimador es igual al parámetro que estima*
- ***Eficiencia****: Tiene la menor varianza posible entre todos los estimadores insesgados*
- ***Consistencia****: Converge al parámetro verdadero cuando el tamaño muestral tiende a infinito*

Ejemplo 5.0.1. *Si deseamos estimar la altura promedio de todos los estudiantes de una universidad (parámetro poblacional μ), podemos calcular la altura promedio de una muestra de 100 estudiantes (estadístico muestral $\bar{x}$). En este caso, $\bar{x}$ es el estimador puntual de μ.*

5.4. Variables Estadísticas y Escalas de Medición

La naturaleza de los datos determina las técnicas estadísticas apropiadas para su análisis. Es fundamental comprender los diferentes tipos de variables y sus escalas de medición.

Definición 5.0.13. *Las **variables cualitativas** (o categóricas) utilizan etiquetas o nombres para identificar atributos de los elementos estudiados.*

Las variables cualitativas se clasifican en:

Definición 5.0.14. ***Escala nominal**: Los datos son etiquetas que identifican atributos sin orden inherente.*

Definición 5.0.15. ***Escala ordinal**: Los datos poseen las propiedades de los datos nominales, pero además tienen un orden o jerarquía natural.*

Ejemplo 5.0.2. *Variables nominales: género (masculino, femenino), nacionalidad (mexicana, estadounidense, canadiense), color de ojos (azul, verde, café). Variables ordinales: nivel educativo (primaria, secundaria, universidad), calificación de servicio (excelente, bueno, regular, malo).*

Definición 5.0.16. *Las **variables cuantitativas** (o numéricas) son valores numéricos que indican cantidad o magnitud.*

Las variables cuantitativas se subdividen en:

Definición 5.0.17. ***Variables discretas**: Toman un número contable de valores, generalmente números enteros.*

Definición 5.0.18. ***Variables continuas**: Pueden tomar cualquier valor dentro de un intervalo determinado.*

Ejemplo 5.0.3. *Variables discretas: número de hijos en una familia, cantidad de errores en un documento, número de automóviles vendidos por día. Variables continuas: peso de una persona, temperatura ambiente, tiempo transcurrido en completar una tarea.*

5.5. Técnicas de Muestreo

La selección de una muestra apropiada es crucial para obtener resultados válidos y confiables. Las técnicas de muestreo se clasifican en dos categorías principales:

5.5.1. Muestreo Probabilístico

Definición 5.0.19. *El **muestreo probabilístico** es aquel en el cual cada elemento de la población tiene una probabilidad conocida y no nula de ser incluido en la muestra.*

Observación 5.0.6. *Los métodos de muestreo probabilístico son los más recomendables para investigación cuantitativa porque permiten calcular la precisión de las estimaciones y realizar inferencias estadísticas válidas.*

Muestreo Aleatorio Simple

Definición 5.0.20. *En el **muestreo aleatorio simple**, cada posible muestra de un tamaño específico tiene la misma probabilidad de ser seleccionada.*

Ejemplo 5.0.4. *Para seleccionar una muestra aleatoria simple de 50 estudiantes de una universidad con 5,000 estudiantes, podríamos asignar un número del 1 al 5,000 a cada estudiante y usar una tabla de números aleatorios o un generador de números aleatorios para seleccionar 50 números.*

Muestreo Sistemático

Definición 5.0.21. *En el **muestreo sistemático**, se selecciona el primer elemento al azar y luego cada k-ésimo elemento de la lista, donde $k = N/n$ (N = tamaño de la población, n = tamaño de la muestra).*

Observación 5.0.7. *El muestreo sistemático es menos costoso y requiere menos tiempo que el muestreo aleatorio simple, pero debe usarse con precaución cuando existe periodicidad en los datos que coincida con el intervalo de selección.*

Ejemplo 5.0.5. *La DIAN utiliza muestreo sistemático sobre listas de contribuyentes, seleccionando cada k-ésimo registro para auditoría fiscal, lo que permite una distribución uniforme de las auditorías a lo largo del tiempo.*

Muestreo Estratificado

Definición 5.0.22. *En el **muestreo estratificado**, la población se divide en grupos homogéneos llamados estratos, y se selecciona una muestra aleatoria simple de cada estrato.*

Observación 5.0.8. *Este método funciona mejor cuando la variabilidad dentro de cada estrato es pequeña comparada con la variabilidad entre estratos. Ayuda a evitar sesgos y generalmente produce estimaciones más precisas que el muestreo aleatorio simple.*

Ejemplo 5.0.6. *Las alcaldías de ciudades principales aplican muestreo estratificado en encuestas de percepción ciudadana, dividiendo la muestra por localidades y estratos socioeconómicos para asegurar representatividad proporcional en cada segmento poblacional.*

Muestreo por Conglomerados

Definición 5.0.23. *En el **muestreo por conglomerados**, la población se divide en grupos llamados conglomerados, se seleccionan aleatoriamente algunos conglomerados y se estudian todos o una submuestra de sus elementos.*

Observación 5.0.9. *Este método puede ser más eficiente económicamente, especialmente cuando los elementos están geográficamente dispersos. El muestreo bietápico es una variante donde se realiza un segundo muestreo dentro de los conglomerados seleccionados.*

Ejemplo 5.0.7. *El DANE utiliza muestreo por conglomerados en el Censo y la GEIH, seleccionando manzanas o veredas como conglomerados, y dentro de ellas los hogares, asegurando representatividad por regiones y estratos socioeconómicos mientras optimiza los recursos logísticos.*

5.5.2. Muestreo No Probabilístico

Definición 5.0.24. *En el **muestreo no probabilístico**, los elementos se seleccionan sin una probabilidad conocida de inclusión, lo que impide determinar estadísticamente la precisión de las estimaciones.*

Observación 5.0.10. *Aunque estos métodos son menos rigurosos desde el punto de vista estadístico, pueden ser útiles en investigación exploratoria o cuando las restricciones de tiempo y recursos impiden el uso de métodos probabilísticos.*

Los tipos principales incluyen:

- **Muestreo de conveniencia**: Selección basada en la accesibilidad
- **Muestreo subjetivo**: Selección basada en el juicio del investigador

- **Muestreo por cuotas**: Selección para cumplir ciertas proporciones predeterminadas
- **Muestreo bola de nieve**: Selección a través de referidos de participantes iniciales

Ejemplo 5.0.8. *Las Secretarías de Salud emplean muestreo por conveniencia para estudios rápidos durante brotes epidémicos, seleccionando hospitales accesibles que permitan recolectar información urgente para la toma de decisiones en salud pública.*

Ejemplo 5.0.9. *Las ONG utilizan muestreo bola de nieve para estudiar poblaciones difíciles de censar, como migrantes irregulares, donde los participantes iniciales refieren a otros miembros de la comunidad, creando una red de contactos que facilita el acceso a esta población oculta.*

5.6. Aplicaciones Multidisciplinarias

El muestreo estadístico encuentra aplicaciones en prácticamente todas las áreas del conocimiento y la actividad humana:

5.6.1. Ingeniería y Control de Calidad

Ejemplo 5.0.10. *En control de calidad, se utilizan gráficas de control (como $\bar{X}$, R, p y np) para monitorear la estabilidad de procesos productivos. Una empresa manufacturera puede tomar muestras de 5 productos cada hora para verificar que las dimensiones se mantengan dentro de especificaciones, detectando variaciones antes de que se produzcan defectos masivos.*

5.6.2. Ciencias de la Salud

Ejemplo 5.0.11. *En ensayos clínicos, se utilizan técnicas de muestreo para evaluar la eficacia de nuevos medicamentos. Un estudio puede comparar la recuperación de pacientes tratados con un nuevo fármaco versus un placebo, usando muestreo aleatorio estratificado por edad y sexo para controlar variables confusoras.*

5.6.3. Ciencias Sociales y Marketing

Ejemplo 5.0.12. *Las encuestas de opinión política utilizan muestreo probabilístico para estimar las preferencias electorales. Con una muestra representativa de 1,500 personas, se puede estimar la intención de voto de toda la población con un margen de error de aproximadamente $\pm 2{,}5\,\%$ al 95 % de confianza.*

5.6.4. Economía y Finanzas

Ejemplo 5.0.13. *En econometría, se utilizan muestras para estimar relaciones entre variables económicas. Un estudio puede examinar la relación entre educación e ingresos usando una muestra representativa de trabajadores, aplicando técnicas de regresión para controlar factores como experiencia laboral, sector económico y ubicación geográfica.*

5.7. Perspectiva Histórica y Desarrollo

Observación 5.0.11. *El desarrollo de la estadística moderna ha sido impulsado por figuras clave como Sir Ronald A. Fisher, quien estableció los principios fundamentales del diseño experimental y la importancia de la aleatorización. Charles Spearman contribuyó con el desarrollo de métodos de correlación y análisis factorial. Los trabajos pioneros de John Graunt en el análisis de datos de mortalidad marcaron el inicio de la interpretación estadística sistemática de datos poblacionales.*

Observación 5.0.12. *El pensamiento estadístico moderno se fundamenta en tres principios esenciales:*

1. *Todo trabajo ocurre en un sistema de procesos interconectados*
2. *La variación existe en todos los procesos*
3. *Entender y reducir la variación son claves para el éxito*

5.8. Ejemplos Prácticos de Implementación

5.8.1. Casos de Estudio en el Contexto Colombiano

Cuadro 5.1: Ejemplos de aplicación de técnicas de muestreo en Colombia

Tipo de muestreo	**Entidad/Contexto**	**Descripción y aplicación práctica**
Aleatorio simple	DANE (encuestas piloto)	Selección aleatoria de viviendas para pruebas metodológicas previas a encuestas nacionales
Estratificado	DANE, Alcaldías	División de la muestra por zonas geográficas, estratos socioeconómicos y características demográficas
Conglomerados	DANE (GEIH, Censo)	Selección de unidades primarias (manzanas, veredas) seguida de selección de hogares dentro de cada conglomerado
Sistemático	DIAN (auditorías fiscales)	Selección de cada k-ésimo registro de contribuyentes para procesos de fiscalización
Conveniencia	Secretarías de Salud	Selección de centros de salud accesibles durante emergencias epidemiológicas
Bola de nieve	ONG, estudios sociales	Referenciación entre participantes para acceder a poblaciones vulnerables o de difícil acceso

5.8.2. Ejercicios de Simulación

Ejemplo 5.0.14. ***Simulación de muestreo por conglomerados (DANE):*** *Considere una población de 10,000 viviendas agrupadas en 200 manzanas (50 viviendas por manzana). Seleccione aleatoriamente 20 manzanas y, dentro de cada una, tome una muestra aleatoria de 10 viviendas para estimar la proporción de hogares con acceso a internet, sabiendo que la proporción real es 0.75.*

Ejemplo 5.0.15. ***Simulación de muestreo sistemático (DIAN):*** *Genere una lista de 5,000 empresas donde 15 % son evasoras fiscales. Implemente un muestreo sistemático con tamaño de muestra 100 para estimar la proporción de evasores. Repita el experimento 500 veces para analizar la distribución de las estimaciones muestrales.*

Ejemplo 5.0.16. ***Simulación de muestreo estratificado (Encuesta ciudadana):*** *Simule una ciudad con 6 localidades distribuidas en tres estratos socioeconómicos: tres localidades de*

estrato bajo (2,000 hogares cada una), dos de estrato medio (3,000 hogares cada una) y una de estrato alto (1,000 hogares). Realice un muestreo estratificado proporcional para una muestra total de 600 hogares y estime el ingreso promedio considerando ingresos medios de $1,200,000, $2,500,000 y $5,000,000 respectivamente.

5.9. Herramientas Computacionales

5.9.1. Implementación en Python

```
# Ejemplo de muestreo sistemático
import numpy as np
np.random.seed(42)

# Parámetros de población
N = 5000  # Tamaño de población
n = 100   # Tamaño de muestra
k = N // n  # Intervalo de muestreo

# Selección del punto de inicio aleatorio
start = np.random.randint(0, k)

# Generación de índices sistemáticos
indices_sistematicos = np.arange(start, N, k)[:n]

print(f"Intervalo de muestreo: {k}")
print(f"Punto de inicio: {start}")
print(f"Primeros 10 índices: {indices_sistematicos[:10]}")
```

5.9.2. Implementación en R

```
# Ejemplo de muestreo estratificado
set.seed(123)

# Definición de estratos poblacionales
poblacion_total <- 13000
estratos <- c(
  rep("bajo", 6000),
  rep("medio", 6000),
  rep("alto", 1000)
)

# Muestreo estratificado proporcional
muestra_estratificada <- c(
  sample(which(estratos == "bajo"), 300),
  sample(which(estratos == "medio"), 300),
  sample(which(estratos == "alto"), 50)
)

# Verificación de la distribución muestral
distribucion_muestra <- table(estratos[muestra_estratificada])
print(distribucion_muestra)
```

5.10. Consideraciones Éticas y Metodológicas

Observación 5.0.13. *La implementación de técnicas de muestreo debe considerar aspectos éticos fundamentales, incluyendo el consentimiento informado de los participantes, la confidencialidad de los datos, y la minimización de sesgos que puedan afectar negativamente a grupos vulnerables. La transparencia metodológica es esencial para la validación y replicabilidad de los resultados.*

Observación 5.0.14. *En el contexto de los grandes datos y la inteligencia artificial, las técnicas tradicionales de muestreo se complementan con nuevos enfoques como el muestreo adaptativo, el muestreo por importancia y técnicas de remuestreo como bootstrap y jackknife, que permiten abordar problemas complejos en entornos de datos masivos.*

5.11. Conclusiones

El muestreo estadístico representa una herramienta fundamental para la investigación científica y la toma de decisiones basada en evidencia. Su correcta aplicación requiere un entendimiento profundo de los principios teóricos, una cuidadosa consideración del contexto específico de aplicación, y una implementación rigurosa que garantice la validez y confiabilidad de los resultados.

Los ejemplos presentados del contexto colombiano ilustran la versatilidad y aplicabilidad práctica de estas técnicas en diferentes sectores, desde la administración pública hasta la investigación social y económica. La evolución continua de las herramientas computacionales y metodológicas mantiene al muestreo estadístico como una disciplina dinámica, adaptándose constantemente a nuevos desafíos y oportunidades en el análisis de datos.

Observación 5.0.15. *El dominio de las técnicas de muestreo estadístico es esencial para cualquier profesional que trabaje con datos, proporcionando las bases metodológicas necesarias para extraer conocimiento significativo y tomar decisiones informadas en un mundo caracterizado por la variabilidad y la incertidumbre.*

Capítulo 6

Prueba de Hipótesis para una Población

La prueba de hipótesis constituye una de las herramientas fundamentales de la inferencia estadística, permitiendo tomar decisiones informadas sobre parámetros poblacionales a partir de evidencia muestral. Este capítulo desarrolla los conceptos teóricos y aplicaciones prácticas de las pruebas de hipótesis para una población.

6.1. Introducción y Conceptos Fundamentales

Definición 6.0.1 (Prueba de Hipótesis)**.** *Una **prueba de hipótesis** es un procedimiento estadístico que, basándose en la evidencia de una muestra y la teoría de la probabilidad, permite determinar si una afirmación sobre un parámetro poblacional es razonable.*

Observación 6.0.1. *La prueba de hipótesis forma parte de la inferencia estadística clásica junto con la estimación. Su objetivo principal es hacer inferencias sobre una población a partir de una cantidad limitada de observaciones muestrales. La disciplina de la probabilidad sirve como puente entre la estadística descriptiva y la estadística inferencial.*

Definición 6.0.2 (Hipótesis Estadística)**.** *Una **hipótesis estadística** es una afirmación sobre el valor de un parámetro poblacional. Se distinguen dos tipos:*

- ***Hipótesis nula** (H_0)**:** Afirmación que se asume tentativamente como verdadera para realizar la prueba. Frecuentemente expresa igualdad o ausencia de efecto.*
- ***Hipótesis alternativa** (H_1 **o** H_a)**:** Afirmación que se acepta si la evidencia muestral contradice a H_0. Representa la hipótesis de investigación.*

Observación 6.0.2. *La verdad o falsedad de una hipótesis estadística nunca se conoce con certeza absoluta, a menos que se examine toda la población, lo cual es impracticable en la mayoría de los casos. Por tanto, se utiliza una muestra aleatoria para proporcionar evidencia que respalde o contradiga la hipótesis.*

6.1.1. Errores en las Pruebas de Hipótesis

Definición 6.0.3 (Errores Tipo I y Tipo II)**.** *Al utilizar una muestra para tomar decisiones sobre la población, pueden ocurrir dos tipos de errores:*

- ***Error Tipo I** (α)**:** Rechazar H_0 cuando en realidad es verdadera. Su probabilidad se denomina **nivel de significancia**.*
- ***Error Tipo II** (β)**:** No rechazar H_0 cuando en realidad es falsa.*

- ***Potencia de la prueba:*** $1-\beta$, *probabilidad de rechazar correctamente* H_0 *cuando es falsa.*

Observación 6.0.3. *Existe una relación inversa entre los errores tipo I y tipo II: si se intenta reducir uno, el otro tiende a aumentar para un tamaño de muestra dado. El nivel de significancia* α *se especifica antes de la recopilación de datos según la importancia relativa de los riesgos.*

6.1.2. Metodología de la Prueba de Hipótesis

Teorema 6.1 (Pasos de una Prueba de Hipótesis)**.** *El procedimiento sistemático para realizar una prueba de hipótesis consta de los siguientes pasos:*

1. *Establecer las hipótesis nula y alternativa*
2. *Elegir un nivel de significancia* (α)
3. *Identificar el estadístico de prueba y la distribución muestral apropiados*
4. *Formular la regla de decisión*
5. *Recopilar los datos de la muestra y calcular el estadístico de prueba*
6. *Tomar la decisión estadística y establecer la conclusión*

Definición 6.1.1 (Valor p)**.** *El* ***valor p*** *es una probabilidad que mide la evidencia de la muestra contra la hipótesis nula. Cuanto menor sea el valor p, mayor será la evidencia contra* H_0*. La regla de decisión es: si valor* $p < \alpha$*, se rechaza* H_0*.*

6.2. Prueba de Hipótesis para la Media Poblacional

6.2.1. Caso 1: Desviación Estándar Poblacional Conocida

Teorema 6.2 (Prueba Z para la Media)**.** *Sea* $X_1, X_2, \ldots, X_n$ *una muestra aleatoria de una población con media* μ *y desviación estándar conocida* σ*. Para probar* $H_0 : \mu = \mu_0$*, el estadístico de prueba es:*

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

que sigue una distribución normal estándar bajo H_0*.*

Observación 6.2.1. *La prueba Z se aplica cuando:*

- *La población se distribuye normalmente, o*
- *El tamaño de la muestra es suficientemente grande* ($n \geq 30$) *por el Teorema Central del Límite*
- *Se dispone de datos históricos extensos o aplicaciones de control de calidad*

Ejemplo 6.2.1 (Aplicación en Control de Calidad)**.** *Una máquina llena bolsas de café con peso objetivo* $\mu_0 = 500$ *g y desviación estándar* $\sigma = 10$ *g. Se toma una muestra de* $n = 36$ *bolsas obteniéndose* $\overline{x} = 497$ *g. ¿Existe evidencia de que la máquina se está desviando del objetivo? Use* $\alpha = 0{,}05$*.*

Solución:

1. $H_0 : \mu = 500$, $H_1 : \mu \neq 500$ *(prueba bilateral)*
2. $\alpha = 0{,}05$, $z_{0{,}025} = 1{,}96$

3. $Z = \frac{497-500}{10/\sqrt{36}} = \frac{-3}{1{,}667} = -1{,}8$

4. $|Z| = 1{,}8 < 1{,}96$, *no se rechaza* H_0

5. *Valor-p* $= 2P(Z > 1{,}8) = 2 \times 0{,}0359 = 0{,}0718 > 0{,}05$

Conclusión: *No hay evidencia suficiente para afirmar que la máquina está desviándose del objetivo.*

6.2.2. Caso 2: Desviación Estándar Poblacional Desconocida

Teorema 6.3 (Prueba t para la Media)**.** *Sea* $X_1, X_2, \ldots, X_n$ *una muestra aleatoria de una población normal con media* μ *y desviación estándar desconocida. Para probar* $H_0 : \mu = \mu_0$*, el estadístico de prueba es:*

$$t = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

que sigue una distribución t de Student con $(n-1)$ *grados de libertad bajo* H_0*.*

Observación 6.3.1. *La prueba t es robusta a desviaciones moderadas de la normalidad, especialmente cuando el tamaño de la muestra es grande. Para muestras muy pequeñas* $(n < 30)$ *y poblaciones marcadamente asimétricas, es apropiado considerar procedimientos no paramétricos.*

Ejemplo 6.3.1 (Aplicación en Análisis Químico)**.** *Un laboratorio afirma que el contenido medio de vitamina C en un jugo es 50 mg por botella. Una muestra de* $n = 25$ *botellas da* $\overline{x} = 48{,}5$ *mg,* $s = 3$ *mg. £Es válida la afirmación al 5 % de significancia?*

Solución:

1. $H_0 : \mu = 50$, $H_1 : \mu \neq 50$

2. $\alpha = 0{,}05$, $t_{0{,}025,24} = 2{,}064$

3. $t = \frac{48{,}5-50}{3/\sqrt{25}} = \frac{-1{,}5}{0{,}6} = -2{,}5$

4. $|t| = 2{,}5 > 2{,}064$, *se rechaza* H_0

5. *Valor-p* $\approx 0{,}02 < 0{,}05$

Conclusión: *Hay evidencia suficiente para rechazar la afirmación del laboratorio.*

6.3. Prueba de Hipótesis para la Proporción Poblacional

Definición 6.3.1 (Proporción Poblacional)**.** *Una* ***proporción*** *es la razón entre el número de elementos con una característica específica y el número total de observaciones. La estimación puntual para la proporción poblacional* p *es la proporción muestral* $\hat{p} = x/n$*.*

Teorema 6.4 (Prueba Z para la Proporción)**.** *Para una muestra de tamaño* n *con* x *éxitos, el estadístico de prueba para* $H_0 : p = p_0$ *es:*

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

que sigue aproximadamente una distribución normal estándar cuando $np_0 \geq 5$ *y* $n(1-p_0) \geq 5$*.*

Observación 6.4.1. *Las pruebas de proporción son fundamentales en:*

- *Encuestas de opinión y estudios de mercado*

- *Estudios clínicos y epidemiológicos*
- *Control de calidad en procesos industriales*
- *Análisis de efectividad de tratamientos*

Ejemplo 6.4.1 (Aplicación en Investigación de Mercado)**.** *En una encuesta, 62 de 100 personas prefieren un nuevo producto. £Difiere significativamente la proporción de preferencia del 60 %? Use* $\alpha = 0{,}05$*.*

Solución:

1. $H_0 : p = 0{,}6$, $H_1 : p \neq 0{,}6$
2. $\hat{p} = 0{,}62$, $n = 100$
3. $Z = \frac{0{,}62-0{,}6}{\sqrt{0{,}6\times 0{,}4/100}} = \frac{0{,}02}{0{,}049} = 0{,}41$
4. $|Z| = 0{,}41 < 1{,}96$, *no se rechaza* H_0
5. *Valor-p* $\approx 0{,}68 > 0{,}05$

Conclusión: *No hay evidencia suficiente para afirmar que la proporción difiere del 60 %.*

6.4. Prueba de Hipótesis para la Varianza Poblacional

Definición 6.4.1 (Varianza Poblacional)**.** *La* ***varianza poblacional*** σ^2 *mide la dispersión de los valores alrededor de la media. Su estimador puntual es la varianza muestral* S^2*.*

Teorema 6.5 (Prueba Chi-cuadrado para la Varianza)**.** *Para una muestra aleatoria de tamaño* n *de una población normal, el estadístico de prueba para* $H_0 : \sigma^2 = \sigma_0^2$ *es:*

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

que sigue una distribución chi-cuadrado con $(n-1)$ *grados de libertad bajo* H_0*.*

Observación 6.5.1. *La prueba chi-cuadrado para la varianza no es robusta a desviaciones de la normalidad. Si la población no es normal, el valor p calculado puede ser significativamente diferente del valor p verdadero. Se recomienda evaluar la normalidad mediante gráficas de probabilidad normal y pruebas de bondad de ajuste.*

Observación 6.5.2. *Las pruebas de varianza son especialmente relevantes en:*

- ***Control de calidad:*** *Para asegurar consistencia en procesos manufactureros*
- ***Six Sigma:*** *La reducción de variabilidad es fundamental en esta metodología*
- ***Diseño experimental:*** *Para evaluar homogeneidad de unidades experimentales*
- ***Índices de capacidad:*** *Cp y Cpk se basan en la desviación estándar del proceso*

Ejemplo 6.5.1 (Aplicación en Control de Calidad)**.** *Un fabricante de componentes electrónicos requiere que la varianza en las medidas de resistencia no exceda* $\sigma_0^2 = 4$ *ohmios*^{2}*. Una muestra de* $n = 20$ *componentes da* $s^2 = 6{,}2$ *ohmios*^{2}*. £Cumple el proceso con la especificación? Use* $\alpha = 0{,}05$*.*

Solución:

1. $H_0 : \sigma^2 \leq 4$, $H_1 : \sigma^2 > 4$ *(prueba unilateral derecha)*

2. $\alpha = 0{,}05$, $\chi^2_{0{,}05,19} = 30{,}14$
3. $\chi^2 = \frac{(20-1)\times 6{,}2}{4} = \frac{19\times 6{,}2}{4} = 29{,}45$
4. $\chi^2 = 29{,}45 < 30{,}14$, *no se rechaza* H_0

Conclusión: *No hay evidencia suficiente para afirmar que el proceso no cumple la especificación.*

6.5. Consideraciones Prácticas y Recomendaciones

Observación 6.5.3. *Al interpretar resultados de pruebas de hipótesis, es importante considerar:*

- *Un valor p pequeño indica evidencia contra* H_0*, pero no prueba causalidad*
- *Un resultado "no significativo"no prueba que* H_0 *sea verdadera*
- *Siempre interpretar los resultados en el contexto del problema*
- *Considerar la significancia práctica además de la estadística*
- *Evaluar el tamaño del efecto y la potencia de la prueba*

Observación 6.5.4. *La elección del nivel de significancia* α *debe basarse en:*

- *La importancia relativa de los errores tipo I y tipo II*
- *Las consecuencias de tomar una decisión incorrecta*
- *Las convenciones del área de aplicación*
- *El costo de obtener más datos para reducir los errores*

6.6. Ejercicios Propuestos

1. Un fabricante afirma que la vida media de sus bombillas es de 1200 horas. Una muestra de 40 bombillas tiene media 1170 horas y $s = 80$ horas. £Se puede rechazar la afirmación al 5 % de significancia?
2. En una encuesta, 45 de 150 personas prefieren un nuevo diseño. £Difiere significativamente la proporción de la hipótesis de que el 30 % prefiere el nuevo diseño? Use $\alpha = 0{,}05$.
3. Un proceso de manufactura requiere que la varianza del peso de los productos no exceda 2.5 g^2. Una muestra de 25 productos da $s^2 = 3{,}8$ g^2. £Cumple el proceso con la especificación al 1 % de significancia?
4. Un laboratorio afirma que el contenido medio de sodio en una bebida es 100 mg. Una muestra de 16 bebidas da $\overline{x} = 104$ mg, $s = 6$ mg. £Es creíble la afirmación al 1 % de significancia?

Capítulo 7

Pruebas de Hipótesis para Dos Poblaciones

La comparación de dos poblaciones constituye una extensión natural de las pruebas de hipótesis para una población y representa una de las aplicaciones más frecuentes en la inferencia estadística. Este capítulo desarrolla los métodos estadísticos fundamentales para comparar parámetros de dos poblaciones independientes.

7.1. Fundamentos de la Inferencia Bipoblacional

Definición 7.0.1 (Inferencia Estadística Bipoblacional)**.** *La **inferencia estadística bipoblacional** es el conjunto de procedimientos estadísticos que permiten obtener conclusiones sobre la comparación entre dos poblaciones a partir de muestras independientes extraídas de cada una de ellas.*

Observación 7.0.1. *La inferencia bipoblacional extiende los conceptos de la estadística inferencial clásica al contexto de comparación. Mientras que en las pruebas de una población se contrasta un parámetro contra un valor hipotético, en las pruebas bipoblacionales se comparan parámetros de dos poblaciones diferentes. Esta extensión es fundamental en campos como la medicina (comparación de tratamientos), ingeniería (comparación de materiales o procesos), y ciencias sociales (comparación de grupos).*

Definición 7.0.2 (Muestras Independientes)**.** *Dos muestras se consideran **independientes** cuando la selección de elementos de una muestra no influye en la selección de elementos de la otra muestra. Esta independencia es crucial para la validez de las pruebas estadísticas bipoblacionales.*

7.2. Comparación de Medias Poblacionales

7.2.1. Prueba Z para Diferencia de Medias

Teorema 7.1 (Prueba Z para Dos Medias con Varianzas Conocidas)**.** *Sean $X_1, X_2, \ldots, X_{n_1}$ y $Y_1, Y_2, \ldots, Y_{n_2}$ muestras aleatorias independientes de poblaciones normales con medias μ_1, μ_2 y varianzas conocidas σ_1^2, σ_2^2 respectivamente. Para probar $H_0 : \mu_1 - \mu_2 = D_0$, el estadístico de prueba es:*

$$Z = \frac{(\overline{X} - \overline{Y}) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

que sigue una distribución normal estándar bajo H_0.

Observación 7.1.1. *Esta prueba se aplica cuando las desviaciones estándar poblacionales son conocidas, situación que puede presentarse en aplicaciones de control de calidad con datos históricos extensos o cuando se asume que los procesos están .en control". El Teorema Central del Límite garantiza la aproximación normal cuando los tamaños muestrales son grandes ($n_1, n_2 \geq 30$), incluso si las poblaciones no son normales.*

Ejemplo 7.1.1 (Comparación de Tiempos de Atención)**.** *Dos sucursales de un banco reportan tiempos medios de atención de 9.2 minutos ($n_1 = 20$, $\sigma_1 = 1{,}1$) y 8.7 minutos ($n_2 = 22$, $\sigma_2 = 1{,}3$). £Existe diferencia significativa al 5 % de significancia?*

Solución:

1. $H_0 : \mu_1 = \mu_2$, $H_1 : \mu_1 \neq \mu_2$
2. $Z = \frac{9{,}2-8{,}7}{\sqrt{1{,}1^2/20+1{,}3^2/22}} = \frac{0{,}5}{\sqrt{0{,}0605+0{,}0769}} = \frac{0{,}5}{0{,}371} = 1{,}35$
3. $|Z| = 1{,}35 < 1{,}96$, *no se rechaza* H_0
4. *Valor-p* $\approx 0{,}18 > 0{,}05$

Conclusión: *No hay evidencia suficiente para afirmar que los tiempos medios de atención difieren significativamente.*

7.2.2. Prueba t para Diferencia de Medias

Teorema 7.2 (Prueba t para Dos Medias con Varianzas Desconocidas Iguales)**.** *Sean $X_1, \ldots, X_{n_1}$ y $Y_1, \ldots, Y_{n_2}$ muestras independientes de poblaciones normales con medias μ_1, μ_2 y varianza común desconocida σ^2. Para probar $H_0 : \mu_1 - \mu_2 = D_0$, el estadístico de prueba es:*

$$t = \frac{(\overline{X} - \overline{Y}) - D_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

donde $S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$ es la varianza muestral combinada. Este estadístico sigue una distribución t de Student con $(n_1 + n_2 - 2)$ grados de libertad bajo H_0.

Teorema 7.3 (Prueba t de Welch para Varianzas Desiguales)**.** *Cuando no se puede asumir igualdad de varianzas, el estadístico de prueba es:*

$$t = \frac{(\overline{X} - \overline{Y}) - D_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Los grados de libertad se calculan mediante la aproximación de Satterthwaite.

Observación 7.3.1. *La prueba t es robusta a desviaciones moderadas de la normalidad, especialmente cuando los tamaños muestrales son grandes. La prueba t de Welch es más general y se recomienda cuando no se puede verificar la igualdad de varianzas mediante una prueba F previa.*

Ejemplo 7.3.1 (Comparación de Dietas)**.** *Se mide el nivel de colesterol en dos grupos: dieta A ($n_1 = 16$, $\overline{x}_1 = 195$ mg/dL, $s_1 = 15$) y dieta B ($n_2 = 14$, $\overline{x}_2 = 205$ mg/dL, $s_2 = 20$). £Difieren las medias al 5 % de significancia?*

Solución usando t de Welch:

1. $H_0 : \mu_1 = \mu_2$, $H_1 : \mu_1 \neq \mu_2$
2. $t = \frac{195-205}{\sqrt{15^2/16+20^2/14}} = \frac{-10}{\sqrt{14{,}06+28{,}57}} = \frac{-10}{6{,}53} = -1{,}53$
3. *Con gl aproximados* ≈ 24, $t_{0{,}025,24} = 2{,}064$
4. $|t| = 1{,}53 < 2{,}064$, *no se rechaza* H_0

Conclusión: *No hay evidencia suficiente para afirmar que las dietas producen niveles medios diferentes de colesterol.*

7.2.3. Prueba t para Muestras Pareadas

Definición 7.3.1 (Muestras Pareadas)**.** *Las **muestras pareadas** son aquellas en las que cada observación de una muestra está directamente relacionada con una observación específica de la otra muestra. Esta relación puede ser natural (mismo individuo medido en dos momentos) o artificial (individuos emparejados por características similares).*

Teorema 7.4 (Prueba t para Muestras Pareadas)**.** *Para n pares de observaciones $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n$ se definen las diferencias $D_i = X_i - Y_i$. Para probar $H_0 : \mu_D = D_0$, el estadístico de prueba es:*

$$t = \frac{\overline{D} - D_0}{S_D/\sqrt{n}}$$

que sigue una distribución t de Student con $(n-1)$ grados de libertad bajo H_0.

Observación 7.4.1. *El diseño pareado es especialmente útil para controlar fuentes de variabilidad externa y aumentar la potencia de la prueba. Al calcular diferencias dentro de cada par, se elimina la variabilidad entre unidades experimentales, focalizando la prueba en el efecto de interés.*

Ejemplo 7.4.1 (Efectividad de Tratamiento)**.** *Se mide la presión arterial de 12 pacientes antes y después de un tratamiento. La media de las diferencias es $\overline{d} = -5$ mmHg con $s_d = 4$ mmHg. £El tratamiento reduce significativamente la presión arterial al 5 %?*

Solución:

1. $H_0 : \mu_d = 0$, $H_1 : \mu_d < 0$ *(prueba unilateral)*
2. $t = \frac{-5}{4/\sqrt{12}} = \frac{-5}{1{,}155} = -4{,}33$
3. *Con gl = 11,* $t_{0{,}05,11} = -1{,}796$
4. $t = -4{,}33 < -1{,}796$, *se rechaza* H_0

Conclusión: *El tratamiento reduce significativamente la presión arterial.*

7.3. Comparación de Proporciones Poblacionales

Definición 7.4.1 (Proporción Poblacional)**.** *La **proporción poblacional** p es la fracción de elementos en una población que poseen una característica específica de interés. Su estimador puntual es la proporción muestral $\hat{p} = x/n$.*

Teorema 7.5 (Prueba Z para Diferencia de Proporciones)**.** *Para dos muestras independientes de tamaños n_1, n_2 con proporciones muestrales $\hat{p}_1, \hat{p}_2$, el estadístico de prueba para $H_0 : p_1 - p_2 = D_0$ es:*

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - D_0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

donde $\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$ es la proporción combinada.

Observación 7.5.1. *Esta prueba es válida cuando se cumplen las condiciones de muestra grande: $n_1\hat{p} \geq 5$, $n_1(1-\hat{p}) \geq 5$, $n_2\hat{p} \geq 5$, y $n_2(1-\hat{p}) \geq 5$. Esto asegura que la distribución muestral de la diferencia de proporciones se aproxime a la normal.*

Ejemplo 7.5.1 (Acceso a Internet)**.** *En dos ciudades, 60 de 120 hogares y 52 de 100 hogares tienen acceso a internet. £Difieren las proporciones al 1 % de significancia?*

Solución:

1. $H_0 : p_1 = p_2$, $H_1 : p_1 \neq p_2$
2. $\hat{p}_1 = 60/120 = 0{,}5$, $\hat{p}_2 = 52/100 = 0{,}52$
3. $\hat{p} = (60 + 52)/(120 + 100) = 112/220 = 0{,}509$
4. $Z = \frac{0{,}5-0{,}52}{\sqrt{0{,}509\times 0{,}491\times(1/120+1/100)}} = \frac{-0{,}02}{0{,}067} = -0{,}30$
5. $|Z| = 0{,}30 < 2{,}58$ *(para* $\alpha = 0{,}01$*)*

Conclusión: *No hay evidencia suficiente para afirmar que las proporciones difieren significativamente.*

7.3.1. Prueba Chi-cuadrado para Tablas de Contingencia

Teorema 7.6 (Prueba Chi-cuadrado de Independencia)**.** *Para una tabla de contingencia* 2×2*, el estadístico de prueba para la independencia es:*

$$\chi^2 = \sum_{i,j} \frac{(f_{ij}^{obs} - f_{ij}^{esp})^2}{f_{ij}^{esp}}$$

que sigue una distribución chi-cuadrado con $(r-1)(c-1)$ *grados de libertad.*

Observación 7.6.1. *La prueba chi-cuadrado de independencia aplicada a una tabla* 2×2 *es matemáticamente equivalente a la prueba Z para diferencia de proporciones. Se requiere que todas las frecuencias esperadas sean al menos 5 para garantizar la validez de la aproximación chi-cuadrado.*

7.4. Comparación de Varianzas Poblacionales

Definición 7.6.1 (Varianza Poblacional)**.** *La* ***varianza poblacional*** σ^2 *mide la dispersión de los valores alrededor de la media poblacional. Su estimador puntual es la varianza muestral* S^2*.*

Teorema 7.7 (Prueba F para Comparación de Varianzas)**.** *Para dos muestras independientes de poblaciones normales, el estadístico de prueba para* $H_0 : \sigma_1^2 = \sigma_2^2$ *es:*

$$F = \frac{S_1^2}{S_2^2}$$

donde $S_1^2 \geq S_2^2$*. Este estadístico sigue una distribución F con* $(n_1 - 1, n_2 - 1)$ *grados de libertad bajo* H_0*.*

Observación 7.7.1. *La prueba F es extremadamente sensible a desviaciones de la normalidad. Si las poblaciones no son normales, los valores p calculados pueden ser significativamente diferentes de los valores p verdaderos, llevando a conclusiones erróneas. Es crucial verificar la normalidad antes de aplicar esta prueba.*

Ejemplo 7.7.1 (Variabilidad en Especies)**.** *Se mide la longitud de hojas en dos especies: Especie 1 (*$n_1 = 8$*,* $s_1 = 1{,}2$ *cm) y Especie 2 (*$n_2 = 10$*,* $s_2 = 2{,}1$ *cm). £Difieren las varianzas al 5 % de significancia?*

Solución:

1. $H_0 : \sigma_1^2 = \sigma_2^2$, $H_1 : \sigma_1^2 \neq \sigma_2^2$
2. $F = \frac{s_2^2}{s_1^2} = \frac{2{,}1^2}{1{,}2^2} = \frac{4{,}41}{1{,}44} = 3{,}06$

3. Con gl = (9,7), $F_{0,025,9,7} \approx 4{,}82$ *y* $F_{0,975,9,7} \approx 0{,}21$

4. $0{,}21 < 3{,}06 < 4{,}82$*, no se rechaza* H_0

Conclusión: *No hay evidencia suficiente para afirmar que las varianzas difieren significativamente.*

7.5. Consideraciones Prácticas

Observación 7.7.2. *La selección del método apropiado depende de varios factores:*

- ***Tipo de parámetro:** media, proporción o varianza*
- ***Conocimiento de varianzas poblacionales:** conocidas o desconocidas*
- ***Supuestos de normalidad:** especialmente críticos para pruebas F*
- ***Independencia de muestras:** fundamental para todas las pruebas*
- ***Tamaños muestrales:** determinan la aplicabilidad de aproximaciones asintóticas*

Observación 7.7.3. *Cuando no se cumplen los supuestos de normalidad, especialmente para la comparación de varianzas, se pueden considerar:*

- ***Transformaciones de datos:** logarítmica, raíz cuadrada, etc.*
- ***Pruebas no paramétricas:** como la prueba de Levene para homogeneidad de varianzas*
- ***Métodos robustos:** que sean menos sensibles a desviaciones de los supuestos*

7.6. Ejercicios Propuestos

1. Dos algoritmos de búsqueda muestran tiempos medios de ejecución de 2.3 segundos ($n_1 = 15$, $s_1 = 0{,}4$) y 2.1 segundos ($n_2 = 18$, $s_2 = 0{,}3$). £Difieren significativamente al 5 %?

2. En una encuesta, 45 de 200 hombres y 38 de 150 mujeres prefieren un producto específico. £Difiere la proporción de preferencia entre géneros al 5 %?

3. Se mide el peso de 10 personas antes y después de una dieta. Las diferencias promedio son $\overline{d} = -3{,}2$ kg con $s_d = 2{,}1$ kg. £Es efectiva la dieta al 1 %?

4. Dos procesos de manufactura muestran varianzas muestrales de 0.25 ($n_1 = 12$) y 0.64 ($n_2 = 15$). £Difieren las varianzas poblacionales al 5 %?

Capítulo 8

Análisis de la Varianza (ANOVA)

El Análisis de Varianza constituye una extensión natural de las pruebas de hipótesis para comparar múltiples medias poblacionales simultáneamente. Este capítulo desarrolla los fundamentos teóricos y aplicaciones prácticas del ANOVA en sus diferentes modalidades, desde el diseño más simple hasta los experimentos factoriales complejos.

8.1. Fundamentos del Análisis de Varianza

Definición 8.0.1 (Análisis de Varianza)**.** *El **Análisis de Varianza (ANOVA)** es un conjunto de procedimientos estadísticos para el análisis de respuestas cuantitativas que permite determinar si las medias de dos o más poblaciones son iguales mediante la descomposición de la variación total en componentes atribuibles a diferentes fuentes.*

Observación 8.0.1. *El ANOVA tiende un puente entre la estadística descriptiva y la inferencia estadística. Mientras que los métodos descriptivos resumen las características de los datos muestrales, el ANOVA permite hacer inferencias sobre las características poblacionales basándose en esa información muestral. Su desarrollo ha revolucionado el análisis experimental en diversas disciplinas científicas.*

El concepto fundamental del ANOVA radica en la **dependencia estadística de una variable dependiente respecto de una o más variables explicativas**. Su objetivo es estimar o predecir el valor promedio de la variable dependiente basándose en los valores conocidos de las variables explicativas.

8.1.1. Fundamento Matemático del Método

Teorema 8.1 (Principio de Descomposición de la Varianza)**.** *El ANOVA se basa en la descomposición de la variación total de la variable dependiente en componentes significativos:*

$$\textit{Variación Total} = \textit{Variación entre grupos} + \textit{Variación dentro de grupos}$$

Esta descomposición permite obtener dos estimaciones independientes de la varianza poblacional común σ^2.

El procedimiento fundamental consiste en: 1. **Estimación basada en variabilidad entre medias muestrales** (variación entre grupos) 2. **Estimación basada en variabilidad dentro de cada muestra** (variación dentro de grupos o error)

Si la hipótesis nula (igualdad de medias poblacionales) es verdadera, ambas estimaciones son estimadores insesgados de σ^2. Sin embargo, si las medias poblacionales difieren, la estimación basada en la variabilidad entre medias tenderá a ser mayor.

8.1.2. Supuestos del Modelo ANOVA

Definición 8.1.1 (Supuestos del ANOVA Clásico)**.** *Para que las inferencias estadísticas del ANOVA sean válidas, se deben cumplir tres supuestos fundamentales:*

1. ***Normalidad:*** *La variable de respuesta en cada población sigue una distribución normal*
2. ***Homocedasticidad:*** *Las poblaciones tienen varianzas iguales (σ^2)*
3. ***Independencia:*** *Las observaciones dentro y entre grupos son independientes*

Observación 8.1.1. *La verificación de supuestos es crucial para la validez del análisis. La normalidad puede evaluarse mediante histogramas, gráficas de probabilidad normal, y pruebas como Jarque-Bera o Anderson-Darling. La homocedasticidad se diagnostica con gráficos de residuales y pruebas como Hartley o Brown-Forsythe. La independencia se garantiza mediante aleatorización en el diseño experimental.*

8.1.3. El Estadístico F y su Interpretación

Definición 8.1.2 (Estadístico F en ANOVA)**.** *El **estadístico F** es el cociente entre el Cuadrado Medio de los Tratamientos (CMTR) y el Cuadrado Medio del Error (CME):*

$$F = \frac{CMTR}{CME}$$

Teorema 8.2 (Interpretación del Estadístico F)**.**

- *Un **valor F cercano a 1** sugiere ausencia de diferencias significativas entre medias grupales*
- *Un **valor F grande** indica que la variación entre medias es considerablemente mayor que la variación dentro de grupos, evidenciando diferencias significativas*

Los **grados de libertad** representan el número de piezas de información independientes disponibles: - Para tratamientos: $k - 1$ (donde k es el número de grupos) - Para error: $N_T - k$ (donde N_T es el número total de observaciones) - Para total: $N_T - 1$

8.2. ANOVA de Un Factor

Definición 8.2.1 (ANOVA de Un Factor)**.** *El **ANOVA de un factor** (clasificación simple o unidireccional) analiza el efecto de un solo factor sobre una variable de respuesta cuantitativa, donde el factor tiene dos o más niveles o categorías.*

8.2.1. Diseño Completamente Aleatorizado

En un **diseño completamente aleatorizado**, los tratamientos se asignan aleatoriamente a las unidades experimentales, asegurando que la variación no explicada se distribuya uniformemente entre grupos, minimizando el sesgo.

8.2.2. Descomposición de la Varianza

Teorema 8.3 (Descomposición de Sumas de Cuadrados)**.** *Para el ANOVA de un factor, la variación total se descompone en:*

$$SSTotal = SCTR + SCE$$

donde:

- $SSTotal = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_{..})^2$ *(Suma de Cuadrados Total)*

- *SCTR = Suma de Cuadrados de Tratamientos (entre grupos)*
- *SCE = Suma de Cuadrados del Error (dentro de grupos)*

Los **Cuadrados Medios** se obtienen dividiendo las Sumas de Cuadrados por sus respectivos grados de libertad: - CMTR = SCTR/$(k-1)$ - CME = SCE/$(N_T - k)$

8.2.3. Prueba de Hipótesis en ANOVA

Teorema 8.4 (Hipótesis en ANOVA de Un Factor)**.** *Las hipótesis para el ANOVA de un factor son:*

- $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ *(todas las medias poblacionales son iguales)*
- H_1 : *No todas las medias poblacionales son iguales*

El estadístico de prueba es $F = CMTR/CME$.

Ejemplo 8.4.1 (Comparación de Resistencia entre Máquinas)**.** *Un ingeniero mide la resistencia (MPa) de piezas producidas por tres máquinas:*

Máquina 1	*45*	*47*	*44*
Máquina 2	*42*	*43*	*40*
Máquina 3	*48*	*50*	*47*

***Solución:** 1. H_0: $\mu_1 = \mu_2 = \mu_3$, H_1: No todas las medias son iguales 2. Medias: $\bar{x}_1 = 45{,}33$, $\bar{x}_2 = 41{,}67$, $\bar{x}_3 = 48{,}33$, $\bar{x}_{..} = 45{,}11$ 3. $SCTR = 3[(45{,}33 - 45{,}11)^2 + (41{,}67 - 45{,}11)^2 + (48{,}33 - 45{,}11)^2] = 66{,}81$ 4. $SCE = 13{,}01$ (suma de desviaciones cuadráticas dentro de grupos) 5. $CMTR = 66{,}81/2 = 33{,}41$, $CME = 13{,}01/6 = 2{,}17$ 6. $F = 33{,}41/2{,}17 = 15{,}39$ 7. Con $F_{0{,}05,2,6} = 5{,}14$, como $15{,}39 > 5{,}14$, se rechaza H_0*

***Conclusión:** Existen diferencias significativas en la resistencia media entre las máquinas.*

Teorema 8.5 (Prueba de comparaciones múltiples de Tukey)**.** *Sea un conjunto de k grupos independientes con medias poblacionales $\mu_1, \mu_2, \ldots, \mu_k$, y supóngase que se ha realizado un ANOVA de un factor que resulta significativo, es decir, se rechaza la hipótesis nula $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$. Entonces, el procedimiento de comparaciones múltiples de Tukey (Tukey HSD):*

1. *Permite comparar todas las posibles diferencias entre pares de medias $(\mu_i - \mu_j)$ mediante la construcción de intervalos de confianza simultáneos para cada par.*
2. *Utiliza la distribución del rango estudentizado para calcular un valor crítico que controla el error de tipo I familiar.*
3. *Si el intervalo de confianza para la diferencia entre dos medias no incluye el cero, se concluye que esas dos medias difieren significativamente.*
4. *Este método es aplicable, por ejemplo, para comparar la eficacia de diferentes tratamientos, mezclas o porcentajes de componentes en estudios experimentales donde se busca identificar cuáles grupos presentan diferencias estadísticamente significativas en sus medias.*

8.3. ANOVA de Dos Factores con Replicación

Definición 8.5.1 (Experimento Factorial)**.** *Un **experimento factorial** permite obtener conclusiones simultáneamente sobre dos o más factores. En el ANOVA de dos factores con replicación se estudian dos factores con múltiples observaciones (réplicas) para cada combinación de niveles.*

8.3.1. Diseño Factorial Completo

En un diseño factorial completo con factores A y B: - Factor A con a niveles - Factor B con b niveles - n réplicas por combinación - Total de observaciones: abn

Definición 8.5.2 (Interacción entre Factores)**.** *Se produce **interacción** cuando el efecto de un factor sobre la variable de respuesta depende del nivel del otro factor. Una interacción significativa implica que los efectos principales no pueden interpretarse aisladamente.*

8.3.2. Tabla ANOVA Extendida

Teorema 8.6 (Estructura ANOVA Bifactorial con Replicación)**.** *Para un diseño de dos factores con replicación, la tabla ANOVA incluye:*

Fuente	***SC***	***gl***	***CM***	***F***
Factor A	*SCA*	$a-1$	*CMA*	*CMA/CME*
Factor B	*SCB*	$b-1$	*CMB*	*CMB/CME*
Interacción AŒB	*SCAB*	$(a-1)(b-1)$	*CMAB*	*CMAB/CME*
Error	*SCE*	$ab(n-1)$	*CME*	
Total	*SCTotal*	$abn-1$		

Ejemplo 8.6.1 (Estudio GMAT)**.** *Un estudio evalúa puntuaciones GMAT considerando: - Factor A: Programa de preparación (3 niveles) - Factor B: Licenciatura de origen (2 niveles) - 5 estudiantes por combinación*

Este diseño permite examinar si existe interacción entre el programa de preparación y la licenciatura, es decir, si algún programa funciona mejor para estudiantes de ciertas licenciaturas.

8.4. ANOVA de Dos Factores sin Replicación

Definición 8.6.1 (Diseño de Bloques Aleatorizados)**.** *El **ANOVA de dos factores sin replicación** se aplica cuando se tienen dos factores pero solo una observación por combinación de niveles ($n = 1$). Un caso típico es el diseño de bloques aleatorizados.*

8.4.1. Limitaciones del Diseño

Observación 8.6.1. *La principal limitación de este diseño es la imposibilidad de estimar la interacción entre factores independientemente del error. La Suma de Cuadrados de Interacción (SCAB) y la del Error (SCE) están confundidas. Por tanto, se debe asumir ausencia de interacción significativa.*

Teorema 8.7 (Modelo Simplificado sin Interacción)**.** *Para un diseño de bloques aleatorizados con k tratamientos y b bloques:*

Fuente	***SC***	***gl***	***CM***	***F***
Tratamientos	*SCTR*	$k-1$	*CMTR*	*CMTR/CME*
Bloques	*SCBL*	$b-1$	*CMBL*	*CMBL/CME*
Error	*SCE*	$(k-1)(b-1)$	*CME*	
Total	*SCTotal*	$kb-1$		

8.5. Aplicaciones Prácticas del ANOVA

8.5.1. Aplicaciones por Disciplina

Ejemplo 8.7.1 (Ingeniería Industrial)**.** *Análisis de resistencia de piezas producidas en tres turnos: - Mañana: 42, 45, 41, 44, 43 MPa - Tarde: 40, 39, 41, 38, 40 MPa - Noche: 46, 47, 45, 48, 46 MPa*

***Objetivo:** Determinar si existen diferencias significativas en la resistencia media entre turnos.*

Ejemplo 8.7.2 (Medicina)**.** *Evaluación de reducción de presión arterial con tres tratamientos: - Tratamiento A: 8, 10, 7, 9, 8 mmHg - Tratamiento B: 12, 11, 13, 12, 14 mmHg - Tratamiento C: 9, 10, 8, 7, 9 mmHg*

***Objetivo:** Identificar el tratamiento más efectivo.*

Ejemplo 8.7.3 (Ingeniería de Sistemas)**.** *Comparación de tiempo de respuesta de algoritmos: - Algoritmo 1: 0.51, 0.49, 0.50, 0.52, 0.50 s - Algoritmo 2: 0.60, 0.62, 0.59, 0.61, 0.60 s - Algoritmo 3: 0.48, 0.47, 0.49, 0.48, 0.47 s*

***Objetivo:** Determinar cuál algoritmo es significativamente más rápido.*

8.5.2. Implementación Computacional

Observación 8.7.1. *La integración de software estadístico como Python, R, Minitab y Excel ha facilitado enormemente la aplicación del ANOVA. Esto permite enfocarse en el análisis e interpretación de resultados en lugar de cálculos manuales tediosos.*

8.6. Metodología del Análisis ANOVA

Teorema 8.8 (Procedimiento ANOVA de Un Factor)**.** *El análisis ANOVA sigue estos pasos sistemáticos:*

1. *Formular hipótesis H_0 y H_1*
2. *Calcular medias grupales y media global*
3. *Calcular sumas de cuadrados (entre grupos y dentro de grupos)*
4. *Calcular cuadrados medios*
5. *Calcular estadístico F*
6. *Determinar valor crítico o valor-p*
7. *Tomar decisión e interpretar resultados*

Observación 8.8.1. *La interpretación del ANOVA requiere considerar tanto la significancia estadística como la relevancia práctica. Un resultado estadísticamente significativo no siempre implica importancia práctica, especialmente con muestras muy grandes. Conversamente, diferencias prácticamente importantes pueden no ser estadísticamente significativas con muestras pequeñas.*

8.7. Consideraciones Adicionales

Observación 8.8.2. *El ANOVA es una herramienta poderosa que se extiende más allá de los diseños básicos presentados. Variantes avanzadas incluyen ANOVA de medidas repetidas, ANCOVA (análisis de covarianza), y diseños factoriales complejos con múltiples factores. Su flexibilidad lo convierte en fundamental para el análisis experimental en diversas disciplinas científicas y empresariales.*

Cuando los supuestos del ANOVA no se cumplen, existen alternativas: - **Transformaciones de datos** para normalizar distribuciones - **Pruebas no paramétricas** como Kruskal-Wallis - **Métodos robustos** menos sensibles a desviaciones de supuestos

8.8. Ejercicios Propuestos

1. Un laboratorio farmacéutico prueba un medicamento en tres dosis diferentes y mide la reducción de presión arterial en cada grupo. Realice un ANOVA para determinar si la dosis influye en la reducción promedio.

2. Compare los ingresos mensuales de familias en tres regiones diferentes. Simule datos apropiados y realice el análisis correspondiente.

3. Evalúe la satisfacción laboral en tres departamentos de una empresa utilizando una escala del 1 al 10. Determine si existen diferencias significativas entre departamentos.

4. Analice el tiempo de ejecución de cuatro algoritmos diferentes en condiciones similares. Implemente el análisis tanto en Python como en R.

Capítulo 9

Regresión y Correlación

El análisis de regresión constituye una de las herramientas estadísticas más fundamentales para el estudio de relaciones entre variables. Su importancia radica en la capacidad de modelar, explicar y predecir el comportamiento de fenómenos complejos a través de relaciones matemáticas. Este capítulo presenta los conceptos esenciales de la regresión lineal simple y múltiple, el análisis de correlación y sus extensiones más relevantes.

9.1. Fundamentos de la Regresión Lineal

Definición 9.0.1. *El **análisis de regresión** es una técnica estadística que estudia la dependencia de una variable respecto a otra u otras variables, con el objetivo de estimar o predecir el valor medio de la variable dependiente basándose en los valores conocidos de las variables explicativas.*

Definición 9.0.2. *La **regresión lineal simple** es el tipo más básico de análisis de regresión, que involucra una variable independiente (regresora) X y una variable dependiente (regresada) Y, donde la relación entre ambas se aproxima mediante una línea recta.*

9.1.1. Modelo Matemático

El modelo de regresión lineal simple se expresa como:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

donde:

- Y_i es la variable dependiente o de respuesta
- X_i es la variable independiente o regresora
- β_0 es el intercepto, que representa el valor medio de Y cuando $X = 0$
- β_1 es la pendiente, que indica el cambio en Y por cada unidad de cambio en X
- ϵ_i es el término de error estocástico

Observación 9.0.1. *El modelo de regresión lineal simple describe la relación promedio entre las variables, no la relación exacta. El término de error ϵ_i captura la variabilidad no explicada por el modelo lineal.*

9.1.2. Supuestos del Modelo Clásico

Para que las inferencias estadísticas sean válidas, el modelo debe satisfacer ciertos supuestos fundamentales:

Teorema 9.1 (Supuestos del Modelo Clásico de Regresión Lineal)**.** *Para una inferencia estadística válida, el modelo de regresión lineal debe cumplir:*

1. ***Linealidad en los parámetros**: El modelo es lineal en β_0 y β_1*
2. ***Valores fijos de las regresoras**: Los valores de X son fijos o independientes del error*
3. ***Media cero del error**: $E(\epsilon_i) = 0$*
4. ***Homoscedasticidad**: $Var(\epsilon_i) = \sigma^2$ (constante)*
5. ***No autocorrelación**: $Cov(\epsilon_i, \epsilon_j) = 0$ para $i \neq j$*
6. ***Normalidad**: $\epsilon_i \sim N(0, \sigma^2)$*

9.2. Estimación por Mínimos Cuadrados

Definición 9.1.1. *El **método de mínimos cuadrados ordinarios (MCO)** es el procedimiento que minimiza la suma de los cuadrados de los residuales para obtener los estimadores b_0 y b_1 de los parámetros poblacionales β_0 y β_1.*

La ecuación de regresión estimada se expresa como:

$$\hat{Y}_i = b_0 + b_1 X_i$$

Los estimadores de MCO se calculan mediante:

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

$$b_0 = \bar{Y} - b_1 \bar{X}$$

Teorema 9.2 (Propiedades de los Estimadores MCO)**.** *Bajo los supuestos del modelo clásico, los estimadores de MCO son los mejores estimadores lineales insesgados (MELI), lo que significa que:*

1. *Son insesgados: $E(b_0) = \beta_0$ y $E(b_1) = \beta_1$*
2. *Tienen varianza mínima entre todos los estimadores lineales insesgados*
3. *Son consistentes: convergen en probabilidad a los valores verdaderos*

Ejemplo 9.2.1. *Un investigador estudia la relación entre gasto en publicidad (X, en millones) y ventas (Y, en millones) con los siguientes datos:*

X	2	3	5	7	9
Y	4	5	7	10	15

Solución:

1. *Calcular medias: $\bar{X} = 5{,}2$, $\bar{Y} = 8{,}2$*
2. *Calcular la pendiente:*

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = 1{,}46$$

3. Calcular el intercepto:

$$b_0 = \bar{Y} - b_1\bar{X} = 8{,}2 - 1{,}46 \times 5{,}2 = 0{,}6$$

4. Ecuación estimada: $\hat{Y} = 0{,}6 + 1{,}46X$

Interpretación: *Por cada millón adicional invertido en publicidad, las ventas aumentan en promedio 1.46 millones.*

9.3. Análisis de Correlación

Definición 9.2.1. *El* ***coeficiente de correlación*** *(r) es una medida descriptiva de la fuerza y dirección de la relación lineal entre dos variables numéricas. Toma valores entre -1 y +1.*

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

Observación 9.2.1. ***Interpretación del coeficiente de correlación:***

- $r = +1$*: Relación lineal positiva perfecta*
- $r = -1$*: Relación lineal negativa perfecta*
- $r = 0$*: No hay relación lineal*
- $|r| > 0{,}8$*: Correlación fuerte*
- $0{,}5 < |r| < 0{,}8$*: Correlación moderada*
- $|r| < 0{,}5$*: Correlación débil*

Definición 9.2.2. *El* ***coeficiente de determinación*** *(R^2) representa la proporción de la variabilidad total de la variable dependiente que es explicada por la ecuación de regresión. Se relaciona con el coeficiente de correlación mediante* $R^2 = r^2$.

Observación 9.2.2. ***Correlación vs. Causalidad:*** *Es fundamental entender que la correlación no implica causalidad. El análisis de regresión y correlación indica cómo las variables están relacionadas, pero las conclusiones sobre causa y efecto deben basarse en el conocimiento experto del área de aplicación.*

9.4. Inferencia Estadística

9.4.1. Pruebas de Hipótesis

Teorema 9.3 (Prueba de Significancia de la Pendiente)**.** *Para evaluar si existe una relación lineal significativa, se prueba:*

$$H_0 : \beta_1 = 0 \quad \textit{(no hay relación lineal)} \tag{9.4.1}$$

$$H_a : \beta_1 \neq 0 \quad \textit{(existe relación lineal)} \tag{9.4.2}$$

El estadístico de prueba es: $t = \frac{b_1}{s_{b_1}}$*, donde* s_{b_1} *es el error estándar de* b_1.

9.5. Análisis de Residuos

Definición 9.3.1. *Los **residuales** son las diferencias entre los valores observados y los valores predichos por el modelo:* $e_i = Y_i - \hat{Y}_i$.

El análisis de residuos es crucial para:

- Evaluar la bondad del ajuste del modelo
- Verificar el cumplimiento de los supuestos
- Detectar observaciones atípicas e influyentes

Observación 9.3.1. ***Herramientas para el análisis de residuos:***

- *Gráficas de residuos vs. valores ajustados (evalúan linealidad y homoscedasticidad)*
- *Gráficas de probabilidad normal (verifican normalidad)*
- *Histogramas de residuos (evalúan distribución)*

9.6. Regresión Múltiple

Definición 9.3.2. *La **regresión múltiple** examina la relación entre una variable dependiente y dos o más variables independientes, extendiendo el concepto de regresión simple para considerar múltiples factores explicativos.*

El modelo de regresión múltiple se expresa como:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi} + \epsilon_i$$

Observación 9.3.2. *En regresión múltiple,* β_j *representa el **coeficiente de regresión parcial**, que indica el cambio promedio en* Y *por cada unidad de cambio en* X_j*, manteniendo constantes las demás variables explicativas.*

9.6.1. Multicolinealidad

Definición 9.3.3. *La **multicolinealidad** ocurre cuando dos o más variables explicativas están altamente correlacionadas entre sí, dificultando la estimación precisa de los coeficientes individuales.*

Observación 9.3.3. ***Remedios para la multicolinealidad:***

- *Utilizar información a priori*
- *Combinar datos de corte transversal y series temporales*
- *Omitir variables altamente colineales (con precaución)*
- *Transformar los datos*
- *Aplicar técnicas como análisis factorial*

9.7. Extensiones del Modelo Lineal

9.7.1. Modelos Polinomiales

Definición 9.3.4. *Los **modelos de regresión polinomial** incorporan términos de variables independientes elevadas a potencias para modelar relaciones curvilíneas, manteniendo la linealidad en los parámetros.*

Un modelo cuadrático se expresa como:

$$\hat{Y} = b_0 + b_1 X + b_2 X^2$$

9.7.2. Términos de Interacción

Definición 9.3.5. *La **interacción** ocurre cuando el efecto de una variable independiente sobre la dependiente cambia según el valor de otra variable independiente. Se modela incluyendo términos como $X_1 X_2$.*

9.7.3. Variables Indicadoras

Definición 9.3.6. *Las **variables indicadoras** (dummy) se utilizan para incorporar información cualitativa en modelos de regresión. Una variable con m categorías requiere $m-1$ variables indicadoras binarias.*

Observación 9.3.4. *Para evitar la trampa de la variable dicótoma (multicolinealidad perfecta), se debe usar $m-1$ variables indicadoras cuando se incluye un término constante, dejando una categoría como referencia.*

9.8. Regresión con Matrices

Definición 9.3.7. *La **formulación matricial** de la regresión expresada como $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ proporciona una forma compacta y eficiente de resolver modelos de regresión múltiple.*

Teorema 9.4 (Estimadores MCO en Notación Matricial)**.** *Los estimadores de mínimos cuadrados ordinarios se obtienen mediante:*

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

donde $\mathbf{X}'$ es la transpuesta de la matriz de diseño $\mathbf{X}$.

9.9. Regresión Logística

Definición 9.4.1. *La **regresión logística** se utiliza cuando la variable dependiente es binaria (toma solo dos valores), modelando la probabilidad de que ocurra un evento específico.*

La ecuación de regresión logística se formula como:

$$P(Y = 1|X_1, \ldots, X_p) = \frac{e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p}}$$

Observación 9.4.1. *En regresión logística, los parámetros se estiman mediante el método de máxima verosimilitud, no por mínimos cuadrados. La interpretación se realiza a través de odds ratios: e^{β_j} representa el cambio en las probabilidades relativas por cada unidad de cambio en X_j.*

9.10. Aplicaciones Disciplinarias

9.10.1. Ingeniería Industrial

Ejemplo 9.4.1. *Se estudia la relación entre tiempo de operación de una máquina (X) y número de piezas defectuosas (Y) para optimizar procesos de manufactura y predecir tasas de defectos.*

9.10.2. Medicina

Ejemplo 9.4.2. *Se analiza la relación entre dosis de medicamento (X) y reducción de presión arterial (Y) para determinar dosificaciones óptimas en tratamientos clínicos.*

9.10.3. Economía

Ejemplo 9.4.3. *Se modela el ingreso mensual (Y) en función de años de escolaridad (X) para estudiar el retorno económico de la educación.*

9.10.4. Administración

Ejemplo 9.4.4. *Una empresa estudia la relación entre gasto en publicidad (X) y ventas mensuales (Y) para optimizar la asignación de presupuesto publicitario.*

9.11. Consideraciones Finales

El análisis de regresión y correlación constituye una herramienta fundamental en el arsenal estadístico, con aplicaciones que abarcan desde la investigación científica hasta la toma de decisiones empresariales. La comprensión profunda de sus supuestos, limitaciones y extensiones es esencial para su aplicación correcta y la interpretación válida de resultados.

Observación 9.4.2. ***Principios clave para recordar:***

- *La regresión describe relaciones, no necesariamente causales*
- *Los supuestos del modelo deben verificarse siempre*
- *El análisis de residuos es fundamental para validar el modelo*
- *La selección del modelo apropiado depende del tipo de variable dependiente*
- *La interpretación debe realizarse en el contexto del problema específico*

Capítulo 10

Anexo: Funciones de Python y R para Estadística Inferencial

Este anexo presenta las funciones más utilizadas en Python y R para la estadística inferencial, acompañadas de ejemplos prácticos. El objetivo es proporcionar una referencia rápida para la implementación computacional de pruebas y procedimientos estadísticos habituales.

10.1. 1. Pruebas de hipótesis para una media

Python (`scipy.stats`)

- `stats.ttest_1samp(datos, popmean=valor)` Prueba t para una muestra
- `stats.norm.cdf(z)`, `stats.norm.ppf(p)` Funciones de la normal estándar (prueba z)

Ejemplo:

```
from scipy import stats
import numpy as np
x = np.array([10.2, 9.8, 10.5, 10.1, 9.9])
t_stat, p_val = stats.ttest_1samp(x, popmean=10)
print("t =", t_stat, "p =", p_val)
```

R

- `t.test(datos, mu=valor)` Prueba t para una muestra
- `pnorm(z)`, `qnorm(p)` Funciones de la normal estándar (prueba z)

Ejemplo:

```
x <- c(10.2, 9.8, 10.5, 10.1, 9.9)
t.test(x, mu=10)
```

2. Pruebas de hipótesis para dos medias

Python

- `stats.ttest_ind(a, b, equal_var=True/False)` Prueba t para dos muestras independientes
- `stats.ttest_rel(a, b)` Prueba t para muestras pareadas

Ejemplo:

```
a = np.random.normal(100, 10, 30)
b = np.random.normal(105, 10, 30)
t_stat, p_val = stats.ttest_ind(a, b, equal_var=False)
```

R

- `t.test(a, b, var.equal=TRUE/FALSE)` Prueba t para dos muestras independientes
- `t.test(a, b, paired=TRUE)` Prueba t para muestras pareadas

Ejemplo:

```
a <- rnorm(30, mean=100, sd=10)
b <- rnorm(30, mean=105, sd=10)
t.test(a, b, var.equal=FALSE)
```

3. Pruebas de proporciones

Python (`statsmodels`)

- `proportions_ztest([x1, x2], [n1, n2])` Prueba z para dos proporciones

Ejemplo:

```
from statsmodels.stats.proportion import proportions_ztest
count = [45, 55]  # éxitos en cada grupo
nobs = [100, 120] # tamaños de muestra
stat, pval = proportions_ztest(count, nobs)
```

R

- `prop.test(c(x1, x2), c(n1, n2), correct=FALSE)` Prueba para dos proporciones

Ejemplo:

```
prop.test(c(45, 55), c(100, 120), correct=FALSE)
```

4. Comparación de varianzas

Python

- `stats.levene(a, b)` Prueba de igualdad de varianzas (Levene)
- `stats.bartlett(a, b)` Prueba de Bartlett para varianzas

Ejemplo:

```
from scipy import stats
a = np.random.normal(0, 1, 30)
b = np.random.normal(0, 2, 30)
stat, p = stats.levene(a, b)
```

R

- `var.test(a, b)` Prueba F para dos varianzas

Ejemplo:

```
a <- rnorm(30, mean=0, sd=1)
b <- rnorm(30, mean=0, sd=2)
var.test(a, b)
```

5. Intervalos de confianza

Python

- `stats.t.interval(conf, df, loc=media, scale=sem)` IC para la media (t)
- `stats.norm.interval(conf, loc=media, scale=sem)` IC para la media (z)

Ejemplo:

```
media = np.mean(x)
sem = stats.sem(x)
ic = stats.t.interval(0.95, len(x)-1, loc=media, scale=sem)
```

R

- `t.test(x)$conf.int` IC para la media

Ejemplo:

```
t.test(x)$conf.int
```

6. ANOVA

Python

- `stats.f_oneway(grupo1, grupo2, grupo3, ...)` ANOVA de un factor

Ejemplo:

```
g1 = [10, 12, 11, 13]
g2 = [14, 15, 13, 16]
g3 = [11, 10, 12, 13]
f_stat, p_val = stats.f_oneway(g1, g2, g3)
```

R

- `aov(y  grupo, data=datos)` ANOVA de un factor
- `summary(anova_result)` Resumen del ANOVA

Ejemplo:

```
grupo <- factor(rep(1:3, each=4))
y <- c(10,12,11,13, 14,15,13,16, 11,10,12,13)
datos <- data.frame(y, grupo)
anova_result <- aov(y ~ grupo, data=datos)
summary(anova_result)
```

7. Regresión lineal y correlación

Python

- `stats.linregress(x, y)` — Regresión lineal simple y correlación
- `np.corrcoef(x, y)` — Matriz de correlación de Pearson

Ejemplo:

```
x = np.array([2, 3, 5, 7, 9])
y = np.array([4, 5, 7, 10, 15])
res = stats.linregress(x, y)
print("Pendiente:", res.slope, "Intercepto:", res.intercept, "r:", res.rvalue)
```

R

- `lm(y  x)` — Regresión lineal simple
- `cor(x, y)` — Correlación de Pearson

Ejemplo:

```
x <- c(2, 3, 5, 7, 9)
y <- c(4, 5, 7, 10, 15)
modelo <- lm(y ~ x)
summary(modelo)
cor(x, y)
```

8. Pruebas de bondad de ajuste y tablas

Python

- `stats.chisquare(f_obs, f_exp)` — Prueba chi-cuadrado de bondad de ajuste
- `stats.chi2_contingency(tabla)` — Prueba chi-cuadrado de independencia

Ejemplo:

```
from scipy.stats import chisquare, chi2_contingency
f_obs = [30, 14, 56]
f_exp = [33, 18, 49]
chi2, p = chisquare(f_obs, f_exp)
```

R

- `chisq.test(tabla)` — Prueba chi-cuadrado

Ejemplo:

```
tabla <- matrix(c(30,14,56,33,18,49), nrow=2, byrow=TRUE)
chisq.test(tabla)
```

9. Recursos y documentación

- https://docs.scipy.org/doc/scipy/reference/stats.html
- https://pandas.pydata.org/
- https://www.statsmodels.org/
- https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html
- https://cran.r-project.org/web/packages/e1071/

Bibliografía

[Academia de Estudios MIR, S.L., 2019] Academia de Estudios MIR, S.L. (2019). *Manual de estadística y epidemiología 2*. Academia de Estudios MIR, S.L., 12 edition.

[Anderson et al., 2008] Anderson, D. R., Sweeney, D. J., and Williams, T. A. (2008). *Estadística para administración y economía*. Thomson/Southwestern, 10 edition.

[Besterfield, 2009] Besterfield, H. (2009). *Control de Calidad*. Pearson Educación, 8 edition.

[Borovkov, 1988] Borovkov, A. A. (1988). *Estadística Matemática*. Editorial Mir Moscú.

[Gmurman, 1974] Gmurman, V. E. (1974). *Teoría de las Probabilidades y Estadística Matemática*. Editorial Mir.

[Gujarati, 2010] Gujarati, D. N. (2010). *Econometría*. McGraw-Hill Interamericana, 5 edition.

[Gutiérrez Pulido and De La Vara Salazar, 2009] Gutiérrez Pulido, H. and De La Vara Salazar, R. (2009). *Control Estadístico de Calidad y Seis Sigma*. McGraw-Hill Interamericana Editores, S.A. de C.V., 2 edition.

[Koroliuk, 1981] Koroliuk, V. S. (1981). *Manual de la Teoría de Probabilidades y Estadística Matemática*. Editorial Mir.

[Kuehl, 2001] Kuehl, R. O. (2001). *Diseño de Experimentos*. Thomson Learning, 2 edition.

[Levine et al., 2006] Levine, D. M., Krehbiel, T. C., and Berenson, M. L. (2006). *Estadística para Administración: Un Primer Curso*. Prentice Hall, 4 edition.

[Lind et al., 2015] Lind, D. A., Marchal, W. G., and Wathen, S. A. (2015). *Estadística aplicada a los negocios y la economía*. McGraw-Hill/Interamericana Editores, S.A. de C.V., 16 edition.

[Montgomery, 2004] Montgomery, D. C. (2004). *Diseño y Análisis de Experimentos*. Editorial Limusa, S.A. de C.V., Grupo Noriega Editores. Revisión técnica de Rodolfo Piña García y Griselda Zetina Vélez.

[Montgomery and Runger, 2018] Montgomery, D. C. and Runger, G. C. (2018). *Applied Statistics and Probability for Engineers*. Wiley.

[OpenStax, 2025] OpenStax (2025). 4.3 Correlation and Linear Regression Analysis. https://openstax.org/books/principles-data-science/pages/1-introduction. Extracto del libro "Principles of Data Science" de OpenStax. La fecha de publicación no se especifica en el extracto.

[Walpole et al., 2012] Walpole, R. E., Myers, R. H., and Myers, S. L. (2012). *Probabilidad y Estadística para Ingenieros*. Pearson, 9 edition.