Unidad: Estadística Básica y Aplicada

Fundamentos estadísticos para Ciencia de Datos

Nicolás Sidicaro
Marzo 2025

Agenda para Hoy

1. **Medidas de tendencia central y dispersión**
2. **Distribuciones de probabilidad**
3. **Inferencia estadística**
4. **Correlación y regresión**
5. **Tests de hipótesis**

1. Medidas de tendencia central

Media aritmética

La **media aritmética** o promedio es la suma de todos los valores dividida por el número de observaciones:

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

Propiedades importantes:

- Es sensible a valores extremos (outliers)
- Minimiza la suma de los cuadrados de las desviaciones
- Es el "centro de gravedad" de la distribución

Aplicación en economía: Usado para calcular ingreso promedio, rendimiento medio de inversiones, productividad promedio, etc.

Mediana y moda

Mediana

La **mediana** es el valor que ocupa la posición central cuando los datos están ordenados:

- Si n es impar, es el valor en la posición $(n+1)/2$
- Si n es par, es el promedio de los valores en las posiciones $n/2$ y $(n/2)+1$

Propiedades importantes:

- No se ve afectada por valores extremos (robusta)
- Divide la distribución en dos partes iguales
- Minimiza la suma de las desviaciones absolutas

Aplicación: Salario mediano, precio mediano de viviendas, tiempo mediano de espera

Mediana y moda

Moda

La **moda** es el valor que aparece con mayor frecuencia.

Propiedades:

- Puede no ser única (distribuciones multimodales)
- No requiere que los datos sean numéricos
- Útil para variables categóricas

Aplicación: Producto más vendido, opción más elegida en encuestas, talla de ropa más demandada

Medidas de posición: Cuartiles y

Los **cuartiles** dividen un conjunto de datos ordenados en cuatro partes iguales:

- **Primer cuartil (Q_1)**: 25% de los datos están por debajo
- **Segundo cuartil (Q_2)**: Equivalente a la mediana
- **Tercer cuartil (Q_3)**: 75% de los datos están por debajo

Los **percentiles** generalizan este concepto, dividiendo los datos en 100 partes iguales.

Aplicaciones:

- Evaluación de desempeño ("en el 10% superior")
- Crecimiento infantil ("peso en el percentil 75")
- Determinación de puntos de corte para análisis de crédito

Caso de estudio: En análisis de ingresos de una población, los deciles (percentiles 10, 20, ..., 90) son fundamentales para estudiar desigualdad y estructurar políticas fiscales.

Medidas de dispersión

Rango

Es la diferencia entre el valor máximo y mínimo:

$$\text{Rango} = x_{max} - x_{min}$$

Limitación: Muy sensible a valores extremos

Rango intercuartílico (IQR)

Es la diferencia entre el tercer y primer cuartil:

$$\text{IQR} = Q_3 - Q_1$$

Ventaja: Robusto frente a valores extremos

Aplicación: Identificación de valores atípicos (outliers):

- Valores atípicos leves: < Q_1 - 1.5×IQR o > Q_3 + 1.5×IQR
- Valores atípicos extremos: < Q_1 - 3×IQR o > Q_3 + 3×IQR

Varianza y desviación estándar

Varianza

La **varianza** mide la dispersión promedio respecto a la media:

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

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

Nota: En la varianza muestral dividimos por $n-1$ (grados de libertad) para obtener un estimador insesgado.

Varianza y desviación estándar

Desviación estándar

La **desviación estándar** es la raíz cuadrada de la varianza:

$$\sigma = \sqrt{\sigma^2} \text{ o } s = \sqrt{s^2}$$

Ventaja: Se expresa en las mismas unidades que los datos originales

Aplicaciones:

- Medición de riesgo en inversiones financieras (volatilidad)
- Control de calidad en procesos industriales
- Evaluación de consistencia en desempeño empresarial

Coeficiente de variación

El **coeficiente de variación** es una medida de dispersión relativa:

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

Ventajas:

- Permite comparar dispersión entre conjuntos de datos con diferentes escalas o unidades. Es adimensional (expresado como porcentaje)

Aplicaciones:

- Comparar volatilidad entre diferentes activos financieros
- Comparar variabilidad en indicadores económicos entre países

Interpretación:

- CV < 10%: Dispersión baja
- 10% ≤ CV < 20%: Dispersión moderada
- 20% ≤ CV < 30%: Dispersión alta
- CV ≥ 30%: Dispersión muy alta

Asimetría y curtosis

Asimetría

Mide la falta de simetría en una distribución:

$$\text{Asimetría} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3$$

Interpretación:

- **Asimetría = 0:** Distribución simétrica
- **Asimetría > 0:** Asimetría positiva (cola derecha más larga)
- **Asimetría < 0:** Asimetría negativa (cola izquierda más larga)

Asimetría y curtosis

Curtosis

Mide el grado de concentración de los valores alrededor del centro:

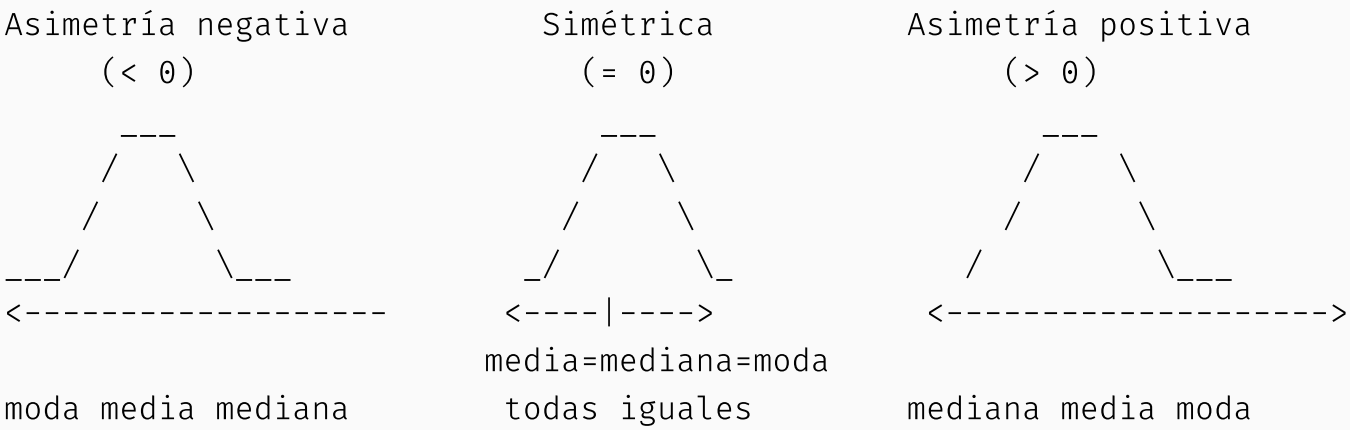
$$\text{Curtosis} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^4 - 3$$

Interpretación:

- **Curtosis = 0:** Mesocúrtica (como la distribución normal)
- **Curtosis > 0:** Leptocúrtica (más puntiaguda que la normal)
- **Curtosis < 0:** Platicúrtica (más plana que la normal)

Nota: El -3 es para que la normal tenga curtosis 0 (forma "estandarizada")

Visualización de la asimetría



Ejemplos reales:

- **Asimetría positiva:** Distribución de ingresos en la mayoría de países (muchas personas con ingresos bajos, pocas con ingresos muy altos)
- **Asimetría negativa:** Calificaciones en exámenes muy fáciles (muchas notas altas, pocas bajas)
- **Simétrica:** Medidas antropométricas como altura en una población homogénea

2. Distribuciones de probabilidad

Una **distribución de probabilidad** es un modelo matemático que describe la probabilidad de ocurrencia de cada valor posible de una variable aleatoria.

Variables aleatorias

- **Variable aleatoria discreta:** Toma valores específicos y contables
  - Ejemplos: Número de clientes, cantidad de productos defectuosos
- **Variable aleatoria continua:** Puede tomar cualquier valor dentro de un intervalo
  - Ejemplos: Tiempo de espera, precio de acciones, altura de personas

Distribuciones de probabilidad

Funciones asociadas

- **Función de masa de probabilidad (PMF):** Para variables discretas
  - $P(X = x)$ = probabilidad de que X tome exactamente el valor x
- **Función de densidad de probabilidad (PDF):** Para variables continuas
  - $f(x)dx$ = probabilidad de que X esté en un intervalo infinitesimal dx alrededor de x
- **Función de distribución acumulada (CDF):** Para ambos tipos
  - $F(x) = P(X \leq x)$ = probabilidad de que X tome un valor menor o igual a x

Distribuciones discretas importantes

Distribución Bernoulli

Modela un experimento con solo dos resultados posibles ("éxito" o "fracaso"):

$$P(X = x) = p^x(1-p)^{1-x} \quad \text{para } x \in \{0, 1\}$$

Parámetros: p = probabilidad de éxito **Media:** $\mu = p$ **Varianza:** $\sigma^2 = p(1-p)$

Aplicaciones:

- Resultado de un único lanzamiento de moneda
- Compra/no compra en una oportunidad de venta
- Aprobación/no aprobación de un crédito

Distribuciones discretas importantes

Distribución Binomial

Modela el número de éxitos en n ensayos independientes de Bernoulli:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{para } k = 0, 1, 2, \ldots, n$$

Parámetros: n = número de ensayos, p = probabilidad de éxito **Media:** $\mu = np$ **Varianza:** $\sigma^2 = np(1-p)$

Distribución de Poisson

Modela el número de eventos que ocurren en un intervalo fijo de tiempo o espacio:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \quad \text{para } k = 0, 1, 2, \ldots$$

Parámetro: λ = tasa media de ocurrencia **Media:** $\mu = \lambda$ **Varianza:** $\sigma^2 = \lambda$

Condiciones:

1. Los eventos ocurren de forma independiente
2. La probabilidad de que ocurra exactamente un evento en un intervalo pequeño es proporcional a la longitud del intervalo
3. La probabilidad de que ocurran dos o más eventos en un intervalo muy pequeño es despreciable

Aplicaciones:

- Número de clientes que llegan a un banco por hora
- Número de llamadas recibidas en un call center
- Cantidad de transacciones fraudulentas detectadas por día

Distribuciones continuas: La Normal

La **distribución normal** o gaussiana es la más importante en estadística:

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

Parámetros: μ = media, σ = desviación estándar

Forma: Simétrica, en forma de campana

Propiedades:

- La media, mediana y moda coinciden
- Aproximadamente el 68% de los datos están a ±1 desviación estándar de la media
- Aproximadamente el 95% de los datos están a ±2 desviaciones estándar
- Aproximadamente el 99.7% de los datos están a ±3 desviaciones estándar

Aplicaciones:

- Distribución de errores de medición
- ¿Todas? Por lo general, suponemos distribuciones normales en Economía

Distribución normal estándar

La **distribución normal estándar** tiene media 0 y desviación estándar 1:

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

Estandarización: Cualquier variable normal $X \sim N(\mu, \sigma^2)$ puede transformarse a una normal estándar mediante:

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

Ventaja: Permite usar tablas estandarizadas para calcular probabilidades.

Ejemplo: Si los salarios en una empresa siguen una distribución normal con media $\mu = 3000$ y desviación estándar $\sigma = 500$:

1. ¿Qué porcentaje de empleados gana más de $3500?
   - Estandarizar: $z = \frac{3500 - 3000}{500} = 1$
   - $P(X > 3500) = P(Z > 1) = 1 - P(Z < 1) = 1 - 0.8413 = 0.1587$ o 15.87%

Otras distribuciones continuas

Distribución t de Student

Similar a la normal pero con colas más pesadas. Útil cuando estimamos la media con varianza desconocida y muestra pequeña.

Parámetro: ν = grados de libertad **Propiedades:**

- Simétrica alrededor de 0
- A medida que ν aumenta, se aproxima a la normal estándar
- Para $\nu > 30$, prácticamente indistinguible de la normal estándar

Aplicación: Intervalos de confianza y pruebas de hipótesis para medias muestrales con n pequeño

Otras distribuciones continuas

Distribución Chi-cuadrado (χ^2)

Surge de la suma de cuadrados de variables normales estándar independientes.

Parámetro: k = grados de libertad **Propiedades:**

- Siempre toma valores positivos
- Asimétrica positiva (se vuelve más simétrica cuando k aumenta)

Aplicaciones:

- Pruebas de bondad de ajuste
- Pruebas de independencia en tablas de contingencia
- Intervalos de confianza para varianzas

Distribución F y Exponencial

Distribución F

Surge del cociente de dos variables chi-cuadrado independientes, cada una dividida por sus grados de libertad.

Parámetros: d_1 y d_2 = grados de libertad del numerador y denominador **Propiedades:**

- Siempre toma valores positivos
- Asimétrica positiva

Aplicación: Análisis de varianza (ANOVA) y pruebas de igualdad de varianzas

Distribución F y Exponencial

Distribución Exponencial

Modela el tiempo de espera hasta la ocurrencia del próximo evento en un proceso de Poisson.

$$f(x) = \lambda e^{-\lambda x} \quad \text{para } x \geq 0$$

Parámetro: λ = tasa (eventos por unidad de tiempo) **Media:** $\mu = 1/\lambda$ **Varianza:** $\sigma^2 = 1/\lambda^2$
Propiedad de falta de memoria: $P(X > s + t | X > s) = P(X > t)$

Aplicaciones:

- Tiempo entre llegadas de clientes
- Vida útil de componentes (sin deterioro)
- Duración de llamadas telefónicas

3. Inferencia estadística: Fundamentos

La **inferencia estadística** nos permite extraer conclusiones sobre una población a partir de una muestra.

Conceptos básicos

- **Población:** Conjunto completo de elementos que queremos estudiar
- **Muestra:** Subconjunto de la población
- **Parámetro:** Cantidad que describe una característica de la población (ej. μ, σ)
- **Estadístico:** Cantidad calculada a partir de los datos muestrales (ej. $\bar{x}$, s)

Muestreo aleatorio simple

Cada elemento de la población tiene la misma probabilidad de ser seleccionado, y las selecciones son independientes.

Propiedades:

- Produce muestras representativas
- Base de muchos métodos estadísticos
- Los estadísticos calculados son variables aleatorias con sus propias distribuciones

Otros métodos de muestreo:

- Muestreo estratificado
- Muestreo por conglomerados
- Muestreo sistemático

Distribuciones muestrales

Una **distribución muestral** es la distribución de probabilidad de un estadístico obtenido de muestras aleatorias de igual tamaño.

Distribución muestral de la media

Si tomamos todas las posibles muestras de tamaño n y calculamos sus medias, obtenemos la distribución muestral de $\bar{X}$.

Propiedades:

- Media: $E(\bar{X}) = \mu$ (insesgado)
- Varianza: $\mathbf{Var}(\bar{X}) = \frac{\sigma^2}{n}$
- Error estándar: $SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}$

Nota: El error estándar disminuye a medida que aumenta el tamaño de la muestra (n)

Teorema del Límite Central (TLC)

Uno de los teoremas más importantes en estadística:

"Si una muestra es lo suficientemente grande, la distribución muestral de la media sigue aproximadamente una distribución normal, independientemente de la distribución de la población original."

$$\bar{X} \overset{\text{aprox}}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$

Implicaciones: Permite construir intervalos de confianza y realizar pruebas de hipótesis incluso cuando la población no sigue una distribución normal.

Estimación puntual

La **estimación puntual** utiliza un único valor (estadístico) para estimar un parámetro poblacional desconocido.

Propiedades deseables de un estimador

1. **Insesgamiento:** El valor esperado del estimador iguala al parámetro

$$E(\hat{\theta}) = \theta$$

2. **Eficiencia:** Tiene la menor varianza entre todos los estimadores insesgados

3. **Consistencia:** El estimador converge al parámetro cuando el tamaño de la muestra crece

$$\lim_{n \to \infty} P(|\hat{\theta}_n - \theta| < \varepsilon) = 1 \text{ para todo } \varepsilon > 0$$

4. **Suficiencia:** Contiene toda la información relevante en la muestra sobre el parámetro

Estimación puntual

Estimadores comunes

- Media muestral $\bar{x}$ como estimador de μ
- Varianza muestral s^2 como estimador de σ^2
- Proporción muestral $\hat{p}$ como estimador de p

Ejemplo: Si en una muestra de 500 compras online, 85 resultaron en devolución, nuestro estimador puntual de la tasa de devolución poblacional es $\hat{p} = 85/500 = 0.17$ o 17%.

Estimación por intervalos

La **estimación por intervalos** proporciona un rango de valores que probablemente contiene el parámetro poblacional desconocido.

Intervalo de confianza

Un **intervalo de confianza (IC)** del (1-α)×100% para un parámetro θ es un intervalo [L,U] calculado a partir de la muestra, tal que:

$$P(L \leq \theta \leq U) = 1 - \alpha$$

Interpretación: Si tomamos muchas muestras y construimos un IC del 95% para cada una, aproximadamente el 95% de estos intervalos contendrán el verdadero valor del parámetro.

Nota: El parámetro θ es fijo (no aleatorio), son los límites L y U los que varían de muestra a muestra.

Estimación por intervalos

IC para la media poblacional (σ conocida)

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Donde $z_{\alpha/2}$ es el valor crítico de la distribución normal estándar (ej. para IC del 95%, $z_{0.025} = 1.96$)

IC para la media poblacional (σ desconocida)

$$\bar{x} \pm t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}$$

Donde $t_{\alpha/2,n-1}$ es el valor crítico de la distribución t con n-1 grados de libertad

Factores que afectan la amplitud del IC

La amplitud de un intervalo de confianza está determinada por:

1. **Nivel de confianza (1-α):**
    - Mayor confianza → Intervalo más amplio
    - Menor confianza → Intervalo más estrecho
2. **Tamaño de la muestra (n):**
    - Mayor tamaño → Intervalo más estrecho
    - Menor tamaño → Intervalo más amplio
3. **Variabilidad de los datos (σ o s):**
    - Mayor variabilidad → Intervalo más amplio
    - Menor variabilidad → Intervalo más estrecho

Factores que afectan la amplitud del IC

Implicaciones prácticas:

- Para aumentar la precisión (reducir amplitud) manteniendo el mismo nivel de confianza, debemos aumentar el tamaño de la muestra
- La relación entre la amplitud y n es proporcional a $1/\sqrt{n}$
- Para reducir la amplitud a la mitad, necesitamos cuadruplicar el tamaño de la muestra

Ejemplo: Si un IC del 95% para el gasto medio por cliente es [145, 155] dólares basado en una muestra de 100 clientes, ¿qué tamaño de muestra necesitaríamos para reducir la amplitud a $5 dólares mientras mantenemos el mismo nivel de confianza?

4. Correlación: relaciones lineales

La **correlación** mide la fuerza y dirección de la relación lineal entre dos variables.

Coeficiente de correlación de Pearson

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

Propiedades:

- Oscila entre -1 y 1
- r = 1: correlación positiva perfecta
- r = -1: correlación negativa perfecta
- r = 0: no hay correlación lineal

4. Correlación: relaciones lineales

La **correlación** mide la fuerza y dirección de la relación lineal entre dos variables.

Coeficiente de correlación de Pearson

Interpretación aproximada:

- $|r| < 0.3$: correlación débil
- $0.3 \leq |r| < 0.7$: correlación moderada
- $|r| \geq 0.7$: correlación fuerte

Advertencias:

- Correlación no implica causalidad
- Mide solo relaciones lineales
- Es sensible a valores extremos
- Una correlación de cero no significa independencia (pueden existir relaciones no lineales)

Correlaciones

Interpretación en diferentes contextos:

Finanzas:

- Correlación entre activos: fundamental para diversificación de portafolios
- r = 0.8 entre dos acciones: se mueven fuertemente en la misma dirección
- r = -0.7 entre oro y ciertos índices bursátiles: cobertura en crisis

Marketing:

- r = 0.65 entre gasto publicitario y ventas: relación positiva moderada-fuerte
- r = -0.4 entre precio y volumen de ventas: relación negativa moderada

Macroeconomía:

- r = 0.75 entre crecimiento del PIB y empleo: relación positiva fuerte
- r = -0.6 entre tasa de interés y inversión privada: relación negativa moderada

Correlaciones

Correlación vs causalidad:

Correlaciones espurias: Relaciones estadísticas sin conexión causal

- Ejemplo: Correlación entre consumo de helados y ahogamientos (ambos aumentan en verano → variable confusora: temperatura)

Dirección causal: La correlación no indica qué variable influye en la otra

- ¿El aumento de publicidad causa más ventas o las empresas con más ventas gastan más en publicidad?

Regresión lineal simple

La **regresión lineal simple** modela la relación entre una variable dependiente (Y) y una variable independiente (X):

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Componentes:

- β_0: intercepto (valor de Y cuando X=0)
- β_1: pendiente (cambio en Y por unidad de cambio en X)
- ε: término de error (residuos) con $\varepsilon \sim N(0, \sigma^2)$

Regresión lineal simple

Método de mínimos cuadrados ordinarios (MCO)

Encuentra los valores de β_0 y β_1 que minimizan la suma de los cuadrados de los residuos:

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Fórmulas:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Ecuación de la recta estimada:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Interpretación de coeficientes

Ejemplo: Modelo de ventas

Supongamos un modelo estimado:

$$\text{Ventas} = 15000 + 2.5 \times \text{Publicidad} + \varepsilon$$

Interpretación:

- **Intercepto** $\hat{\beta}_0 = 15000$: Ventas esperadas cuando no hay gasto en publicidad
- **Pendiente** $\hat{\beta}_1 = 2.5$: Por cada dólar adicional invertido en publicidad, las ventas aumentan en promedio 2.5 dólares

Interpretación de coeficientes

Ejemplo: Modelo de precios de vivienda

$$\text{Precio} = 80000 + 1200 \times \text{Tamaño} - 5000 \times \text{Distancia} + \varepsilon$$

Interpretación:

- **$\hat{\beta}_1 = 1200$**: Por cada metro cuadrado adicional, el precio aumenta en promedio 1200 dólares, manteniendo la distancia constante
- **$\hat{\beta}_2 = -5000$**: Por cada kilómetro adicional de distancia al centro, el precio disminuye en promedio 5000 dólares, manteniendo el tamaño constante

Nota: La interpretación "manteniendo las demás variables constantes" es crucial en regresión múltiple (efecto parcial).

Coeficiente de determinación (R^2)

El **coeficiente de determinación** (R^2) mide la proporción de la varianza en la variable dependiente que es predecible a partir de las variables independientes:

$$R^2 = \frac{\text{Varianza explicada}}{\text{Varianza total}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

También se puede calcular como:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Propiedades:

- Oscila entre 0 y 1 (o 0% y 100%)
- $R^2 = 0$: el modelo no explica nada de la variabilidad
- $R^2 = 1$: el modelo explica toda la variabilidad

Coeficiente de determinación (R^2)

Interpretación:

- $R^2 = 0.65$: el 65% de la variación en Y es explicada por el modelo
- El 35% restante se debe a factores no incluidos o aleatorios

Advertencia: R^2 siempre aumenta al añadir variables, incluso si no son relevantes. Por eso en regresión múltiple se prefiere el R^2 ajustado.

Regresión lineal múltiple

La **regresión lineal múltiple** extiende el modelo simple a múltiples predictores:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

Ventajas:

- Permite modelar relaciones más complejas
- Reduce el sesgo por variables omitidas
- Evalúa el efecto de cada variable controlando por las demás

Regresión lineal múltiple

R^2 ajustado

Corrige el R^2 para tener en cuenta el número de predictores:

$$R^2_{adj} = 1 - \frac{(1 - R^2)(n - 1)}{n - p - 1}$$

Donde n es el tamaño de la muestra y p el número de predictores.

Regresión lineal múltiple

Multicolinealidad

Problema cuando existe alta correlación entre variables independientes:

- Coeficientes inestables y difíciles de interpretar
- Errores estándar inflados
- Dificultad para determinar la importancia relativa de cada variable

Detección: Factor de inflación de la varianza (VIF) **Solución:** Eliminar variables redundantes, usar técnicas de regularización

Supuestos del modelo de regresión

Para que los estimadores MCO sean BLUE (Best Linear Unbiased Estimators):

1. **Linealidad:** La relación entre X e Y es lineal
2. **Independencia:** Las observaciones son independientes entre sí
3. **Homocedasticidad:** La varianza del error es constante para todos los valores de X
4. **Normalidad:** Los errores siguen una distribución normal
5. **No multicolinealidad perfecta:** No existe una relación lineal perfecta entre variables independientes

Supuestos del modelo de regresión

Consecuencias de violación de supuestos:

- **Violación de linealidad:** Estimaciones sesgadas
- **Heteroscedasticidad:** Estimadores ineficientes, errores estándar incorrectos
- **Autocorrelación:** Estimadores ineficientes, inferencia inválida
- **No normalidad:** Afecta la inferencia en muestras pequeñas

Supuestos del modelo de regresión

Diagnóstico:

- Análisis de residuos
- Gráficos de dispersión
- Pruebas estadísticas específicas (Breusch-Pagan, Durbin-Watson, etc.)

5. Tests de hipótesis: Fundamentos

Un **test de hipótesis** es un procedimiento formal para evaluar la evidencia proporcionada por los datos contra una afirmación específica.

Componentes básicos:

1. **Hipótesis nula (H_0):** Afirmación que se asume como verdadera inicialmente
    - Típicamente establece "no hay efecto" o "no hay diferencia"
    - Se formula como una igualdad: $\mu = \mu_0$, $p = p_0$, etc.
2. **Hipótesis alternativa (H_1 o H_a):** Afirmación contraria a H_0
    - Puede ser unilateral (>) o (<) o bilateral (≠)
    - Lo que el investigador generalmente quiere demostrar
3. **Estadístico de prueba:** Valor calculado a partir de los datos muestrales
    - Cuantifica la evidencia contra H_0

5. Tests de hipótesis: Fundamentos

Un **test de hipótesis** es un procedimiento formal para evaluar la evidencia proporcionada por los datos contra una afirmación específica.

Componentes básicos:

1. **Distribución del estadístico:** Distribución de probabilidad del estadístico bajo H_0
2. **Región crítica/de rechazo:** Conjunto de valores que llevan al rechazo de H_0
3. **Valor p:** Probabilidad de obtener un resultado tan o más extremo que el observado, asumiendo que H_0 es verdadera

Procedimiento de prueba de hipótesis

Pasos a seguir:

1. **Formular las hipótesis** H_0 y H_1

2. **Seleccionar el nivel de significancia (α)**

   - Típicamente 0.05 (5%) o 0.01 (1%)
   - Representa la probabilidad máxima aceptable de rechazar H_0 cuando es verdadera (Error Tipo I)

3. **Seleccionar y calcular el estadístico de prueba apropiado**

Procedimiento de prueba de hipótesis

Pasos a seguir:

1. **Determinar el valor crítico o valor p**
   - Valor crítico: depende de α y la distribución del estadístico
   - Valor p: probabilidad calculada a partir del estadístico observado
2. **Tomar una decisión estadística**
   - Si valor $p \leq \alpha$: rechazar H_0
   - Si valor $p > \alpha$: no rechazar H_0
3. **Formular la conclusión en el contexto del problema**
   - Interpretar el resultado en términos prácticos
   - Considerar la significancia estadística vs. práctica

Nota: No rechazar H_0 no significa "aceptar" H_0 o "probar" que H_0 es verdadera; simplemente significa que no hay suficiente evidencia para rechazarla.

Errores en pruebas de hipótesis

Existen dos tipos principales de errores:

Error Tipo I (α)

- Rechazar H_0 cuando es verdadera (falso positivo)
- Probabilidad = α (nivel de significancia)

Error Tipo II (β)

- No rechazar H_0 cuando es falsa (falso negativo)
- Probabilidad = β

Potencia del test (1-β)

- Probabilidad de rechazar H_0 cuando es falsa (verdadero positivo)
- Capacidad del test para detectar un efecto cuando realmente existe

Errores en pruebas de hipótesis

	H_o es verdadera	H_o es falsa
Rechazar H_o	Error Tipo I (α)	Decisión correcta ($1-\beta$ = Potencia)
No rechazar H_o	Decisión correcta ($1-\alpha$)	Error Tipo II (β)

Errores en pruebas de hipótesis

Factores que afectan la potencia:

- Tamaño de la muestra (n): mayor n → mayor potencia
- Nivel de significancia (α): mayor α → mayor potencia
- Tamaño del efecto: efecto mayor → mayor potencia
- Variabilidad de los datos: menor variabilidad → mayor potencia

Prueba t para una media

Evalúa si la media de una población es igual a un valor específico.

Hipótesis:

- H_0: $\mu = \mu_0$
- H_1: $\mu \neq \mu_0$ (bilateral), $\mu > \mu_0$ o $\mu < \mu_0$ (unilateral)

Estadístico de prueba:

- **Caso σ conocida:** $z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$
- **Caso σ desconocida:** $t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$

Distribución bajo H_0:

- z ~ N(0,1) si σ es conocida
- t ~ t(n-1) si σ es desconocida

Prueba t para una media

Regla de decisión (α = 0.05):

- Bilateral: Rechazar H_0 si |t| > t(0.025, n-1)
- Unilateral derecha: Rechazar H_0 si t > t(0.05, n-1)
- Unilateral izquierda: Rechazar H_0 si t < -t(0.05, n-1)

Ejemplo: ¿Es el tiempo medio de espera en un banco diferente de 10 minutos?

- Una muestra de 25 clientes da $\bar{x} = 11.5$ minutos y $s = 3.2$ minutos
- $t = \frac{11.5 - 10}{3.2/\sqrt{25}} = 2.34$
- Con 24 g.l. y α = 0.05, el valor crítico es t(0.025, 24) = 2.064
- Como |t| > 2.064, rechazamos H_0

Prueba t para dos muestras

Compara las medias de dos poblaciones independientes.

Hipótesis:

- H_0: $\mu_1 = \mu_2$ (o $\mu_1 - \mu_2 = 0$)
- H_1: $\mu_1 \neq \mu_2$ (bilateral), $\mu_1 > \mu_2$ o $\mu_1 < \mu_2$ (unilateral)

Estadístico de prueba (varianzas iguales):

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

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 combinada

Grados de libertad: $n_1 + n_2 - 2$

Prueba t para dos muestras

Estadístico (varianzas desiguales - Welch):

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Grados de libertad aproximados: fórmula de Welch-Satterthwaite

Ejemplo: Comparar el retorno de dos estrategias de inversión:

- Estrategia A: $\bar{x}_A = 8.2\%$, $s_A = 2.1\%$, $n_A = 20$
- Estrategia B: $\bar{x}_B = 6.9\%$, $s_B = 1.9\%$, $n_B = 20$
- ¿Es el retorno medio de A significativamente mayor que el de B?

ANOVA: Análisis de varianza

Compara las medias de más de dos grupos simultáneamente.

Hipótesis:

- H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ (todas las medias son iguales)
- H_1: Al menos una media es diferente de las demás

Lógica del ANOVA:

Descompone la variabilidad total en:

- **Variabilidad entre grupos** (explicada por el factor)
- **Variabilidad dentro de grupos** (error residual)

Estadístico F:

$$F = \frac{\text{Varianza entre grupos}}{\text{Varianza dentro de grupos}}$$

ANOVA: Análisis de varianza

Distribución bajo H_0: $F \sim F(k-1, n-k)$

Regla de decisión: Rechazar H_0 si $F > F(\alpha, k-1, n-k)$

Ventaja sobre múltiples pruebas t:

- Control del error Tipo I
- Mayor eficiencia

Limitación:

- Si rechazamos H_0, sabemos que existe diferencia pero no entre qué grupos específicamente
- Se requieren pruebas post-hoc (Tukey, Bonferroni, etc.) para identificar los grupos que difieren

Prueba chi-cuadrado

Prueba de bondad de ajuste

Evalúa si una distribución de frecuencias observada se ajusta a una distribución teórica.

Hipótesis:

- H_0: La distribución observada se ajusta a la distribución esperada
- H_1: La distribución observada difiere de la esperada

Estadístico:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

donde O_i = frecuencia observada, E_i = frecuencia esperada

Grados de libertad: k - 1 - m (k = categorías, m = parámetros estimados)

Prueba chi-cuadrado

Prueba de independencia

Evalúa si dos variables categóricas están asociadas.

Hipótesis:

- H_0: Las variables son independientes
- H_1: Las variables están asociadas

Tabla de contingencia: Organiza frecuencias por categorías cruzadas

Frecuencias esperadas bajo independencia:

$$E_{ij} = \frac{n_{i.} \times n_{.j}}{n}$$

Grados de libertad: (filas-1) × (columnas-1)

Test paramétricas vs no param.

Pruebas paramétricas

- Asumen una distribución específica de la población (típicamente normal)
- Utilizan parámetros como media y varianza
- Mayor potencia cuando se cumplen los supuestos
- Ejemplos: prueba t, ANOVA, regresión lineal

Pruebas no paramétricas

- No asumen una distribución específica ("libres de distribución")
- Típicamente basadas en rangos u órdenes, no en valores brutos
- Más robustas frente a valores atípicos y desviaciones de normalidad
- Menor potencia que las pruebas paramétricas cuando los supuestos de estas se cumplen
- Útiles con datos ordinales o cuando las muestras son pequeñas
- Ejemplos: prueba de Wilcoxon, prueba de Mann-Whitney, prueba de Kruskal-Wallis

Test paramétricas vs no param.

¿Cuándo usar pruebas no paramétricas?

- Datos que violan claramente la normalidad
- Muestras muy pequeñas (n < 30)
- Datos en escala ordinal
- Presencia de valores atípicos extremos
- Cuando se requiere mayor robustez

Resumen y conceptos clave

1. Estadística descriptiva

- Las medidas de tendencia central (media, mediana, moda) describen el centro de los datos
- Las medidas de dispersión (varianza, desviación estándar, IQR) cuantifican la variabilidad
- La asimetría y curtosis caracterizan la forma de la distribución

2. Distribuciones de probabilidad

- Las distribuciones discretas modelan variables con valores contables (binomial, Poisson)
- Las distribuciones continuas modelan variables que toman cualquier valor en un intervalo (normal, t, chi-cuadrado)
- La distribución normal es fundamental por el Teorema del Límite Central

Resumen y conceptos clave

3. Inferencia estadística

- Permite generalizar de muestras a poblaciones
- La estimación puntual proporciona un único valor
- Los intervalos de confianza cuantifican la incertidumbre
- El error estándar disminuye con la raíz cuadrada del tamaño muestral

4. Correlación y regresión

- La correlación mide la fuerza y dirección de relaciones lineales
- La regresión modela la relación funcional entre variables
- El R^2 mide la proporción de varianza explicada por el modelo

Resumen y conceptos clave

5. Tests de hipótesis

- Procedimiento formal para evaluar afirmaciones estadísticas
- Los errores tipo I y II representan decisiones incorrectas
- El valor p cuantifica la evidencia contra la hipótesis nula
- La significancia estadística no implica necesariamente relevancia práctica

Aplicaciones en economía y negocios

Finanzas

- Estimación de retornos esperados y riesgo (volatilidad)
- Pruebas de eficiencia de mercado
- Valoración de activos y opciones

Marketing

- Análisis de efectividad publicitaria
- Segmentación de mercados mediante análisis de clúster
- Modelado de comportamiento del consumidor

Gestión

- Control de calidad y mejora de procesos
- Pronóstico de ventas y demanda
- Optimización de inventarios

Aplicaciones en economía y negocios

Macroeconomía

- Análisis de series temporales para variables económicas
- Pruebas de teorías económicas
- Evaluación de impacto de políticas públicas

Caso de estudio: A/B Testing

Las empresas tecnológicas y de e-commerce utilizan pruebas de hipótesis para comparar diferentes versiones de sitios web, aplicaciones o estrategias de marketing:

- H_0: No hay diferencia en la tasa de conversión entre las versiones A y B
- H_1: Existe diferencia significativa entre las tasas de conversión
- Los datos de interacciones de usuarios se analizan para tomar decisiones basadas en evidencia

Recursos adicionales

Libros recomendados:

- "Estadística para Administración y Economía" (Anderson, Sweeney & Williams)
- "Introducción a la Econometría" (Wooldridge)
- "Econometría" (Gujarati)
- "Estadística para los Negocios y la Economía" (Newbold, Carlson & Thorne)

Software estadístico:

- R y RStudio
- Python con bibliotecas como pandas, statsmodels, scikit-learn
- SPSS
- Stata
- Excel (para análisis básicos)

Próxima clase: Aplicaciones en R y análisis exploratorio avanzado