Contenido

1. Estadística Descriptiva

- 1.1 Conceptos Básicos de Estadística:

- Definición de Estadística (Ampliación):
- **Tipos de Variables:**
- **Cualitativas:** Describen cualidades o atributos (ej. color, género).
- **Nominales:** No tienen un orden inherente (ej. color de ojos).
- **Ordinales:** Tienen un orden inherente (ej. nivel de satisfacción: bajo, medio, alto).
- **Cuantitativas:** Se miden numéricamente (ej. edad, altura).
- **Discretas:** Toman valores enteros (ej. número de hijos).
- **Continuas:** Toman cualquier valor dentro de un rango (ej. temperatura).
- **Escalas de Medición:**
- **Nominal:** Clasifica datos en categorías mutuamente excluyentes.
- **Ordinal:** Ordena datos en categorías con un orden significativo.
- **Intervalo:** Permite medir la diferencia entre valores, pero no tiene un cero absoluto (ej. temperatura en Celsius).
- **Razón:** Tiene un cero absoluto y permite comparar razones entre valores (ej. altura, peso).
- **Teoría de Decisión (Ampliación):**
- **Criterios de Decisión:**
- **Maximax:** Elegir la alternativa que ofrece el mejor resultado posible (optimista).
- **Maximin:** Elegir la alternativa que ofrece el mejor resultado en el peor escenario posible (pesimista).
- **Valor Esperado:** Elegir la alternativa con el valor esperado más alto.
- **Árboles de Decisión:** Diagramas que representan las posibles decisiones y sus resultados, ayudando a evaluar diferentes opciones.
- **Población (Ampliación):**
- **Parámetros Poblacionales Comunes:**
- **Media (μ):** Promedio de todos los valores en la población.
- **Desviación Estándar (σ):** Medida de la dispersión de los valores con respecto a la media.

- Proporción (p): Proporción de elementos en la población que tienen una característica específica.

- **Muestra Aleatoria (Ampliación):**

- **Tamaño de la Muestra:** Determinar el tamaño adecuado de la muestra es crucial para obtener resultados precisos.

- **Factores que Influyen:** Nivel de confianza deseado, margen de error aceptable, variabilidad de la población.

- **Sesgo de Muestreo:** Evitar sesgos en la selección de la muestra es fundamental para asegurar que sea representativa.

- Parámetros Aleatorios (Ampliación):

- **Estimadores:** Estadísticos muestrales que se utilizan para estimar los parámetros poblacionales.

- **Media Muestral ($\bar{x}$):** Estimador de la media poblacional (μ).

- **Desviación Estándar Muestral (s):** Estimador de la desviación estándar poblacional (σ).

- **Proporción Muestral ($\hat{p}$):** Estimador de la proporción poblacional (p).

- **Propiedades de los Estimadores:**

- **Insesgado:** El valor esperado del estimador es igual al parámetro poblacional.

- **Eficiente:** Tiene la menor varianza posible entre todos los estimadores insesgados.

- **Consistente:** Se acerca al parámetro poblacional a medida que aumenta el tamaño de la muestra.

- 1.2 Descripción de Datos (Aún Más Detallado):

- Datos Agrupados y No Agrupados (Ampliación):

- **Ventajas y Desventajas:**

- Datos No Agrupados:

- **Ventajas:** Proporcionan información detallada sobre cada observación.

- **Desventajas:** Pueden ser difíciles de manejar y analizar cuando hay una gran cantidad de datos.

- Datos Agrupados:

- **Ventajas:** Facilitan el resumen y la presentación de datos.
- **Desventajas:** Se pierde información detallada sobre cada observación.
- **Elección del Método:** La elección entre datos agrupados y no agrupados depende del tamaño del conjunto de datos y del objetivo del análisis.

- Frecuencia de Clase y Frecuencia Relativa (Ampliación):

- **Tabla de Frecuencias:** Organización tabular de los datos que muestra la frecuencia de cada clase o categoría.

- Frecuencia Acumulada: Suma de las frecuencias de todas las clases hasta una clase dada.

- **Frecuencia Relativa Acumulada:** Proporción de observaciones que caen dentro de una clase o categoría y todas las clases anteriores.
- **Punto Medio (Ampliación):**

- Uso: Se utiliza para representar todas las observaciones dentro de una clase al calcular medidas como la media y la varianza para datos agrupados.

- **Límites (Ampliación):**
- **Límites Reales:** Se ajustan los límites de clase para evitar huecos entre las clases adyacentes.
- **Límite Real Inferior:** Se resta la mitad de la unidad de medida al límite inferior.
- **Límite Real Superior:** Se suma la mitad de la unidad de medida al límite superior.

- 1.3 Medidas de Tendencia Central (Aún Más Detallado):

- Media Aritmética (Ampliación):

- **Sensibilidad a Valores Atípicos:** La media es sensible a valores atípicos (valores extremadamente altos o bajos).

- Uso: Se utiliza cuando los datos siguen una distribución simétrica y no hay valores atípicos.

- Media Geométrica (Ampliación):

- **Aplicaciones:** Cálculo de tasas de crecimiento promedio, rendimientos de inversiones.
- **Limitaciones:** No se puede calcular si hay valores negativos o cero.

- Media Ponderada (Ampliación):

- **Ejemplos:** Cálculo del promedio ponderado de calificaciones en un curso, cálculo del índice de precios al consumidor.
- **Mediana (Ampliación):**
- **Robustez:** La mediana es robusta a valores atípicos (no se ve afectada por valores extremos).
- **Uso:** Se utiliza cuando los datos tienen una distribución asimétrica o hay valores atípicos.
- Moda (Ampliación):
- **Distribuciones Multimodales:** Una distribución puede tener más de una moda.
- **Uso:** Se utiliza para identificar el valor más común en un conjunto de datos.

- 1.4 Medidas de Dispersión (Aún Más Detallado):

- **Varianza (Ampliación):**
- **Unidades:** Se mide en unidades al cuadrado, lo que dificulta su interpretación directa.
- **Uso:** Se utiliza como base para calcular la desviación estándar.
- Desviación Estándar (Ampliación):
- **Regla Empírica:** En una distribución normal, aproximadamente el 68% de los datos se encuentran dentro de una desviación estándar de la media, el 95% dentro de dos desviaciones estándar y el 99.7% dentro de tres desviaciones estándar.
- **Rango (Ampliación):**
- **Simplicidad:** Es fácil de calcular e interpretar.
- **Limitaciones:** No proporciona información sobre la distribución de los datos entre los valores máximo y mínimo.
- **Sensibilidad:** Es muy sensible a los valores atípicos.
- **Otras Medidas de Dispersión:**
- **Rango Intercuartílico (IQR):** Diferencia entre el tercer cuartil (Q3) y el primer cuartil (Q1). Robusto a valores atípicos.
- **Coeficiente de Variación (CV):** Desviación estándar dividida por la media. Permite comparar la dispersión entre conjuntos de datos con diferentes unidades de medida o diferentes medias.

- 1.5 Parámetros para Datos Agrupados (Aún Más Detallado):

- **Fórmulas Específicas:**

- Media para Datos Agrupados:

- $\bar{x} = \Sigma (f_i * m_i) / n$

- Donde:

- f_i: Frecuencia de la clase i.

- m_i: Punto medio de la clase i.

- n: Número total de observaciones.

- Varianza para Datos Agrupados:

- $s^2 = \Sigma [f_i * (m_i - \bar{x})^2] / (n - 1)$

- **Corrección de Sheppard:** Ajuste que se aplica a la varianza calculada para datos agrupados para compensar la pérdida de información debido al agrupamiento.

- 1.6 Distribución de Frecuencias (Aún Más Detallado):

- Tipos de Gráficos:

- **Histograma:** Gráfico de barras que muestra la distribución de frecuencias de datos agrupados.

- **Polígono de Frecuencias**: Línea que conecta los puntos medios de las barras de un histograma.

- **Ojiva:** Gráfico de línea que muestra la frecuencia acumulada o la frecuencia relativa acumulada.

- **Interpretación:** La forma de la distribución de frecuencias puede revelar información importante sobre los datos, como la simetría, la presencia de valores atípicos y la concentración de los valores.

- 1.7 Técnicas de Agrupación de Datos (Aún Más Detallado):

- **Regla de Sturges (Ampliación):**

- $k = 1 + 3.322 * \log(n)$

- Donde:

- k: Número de clases.

- n: Número de observaciones.

- **Consideraciones:** La regla de Sturges es una guía, pero el número óptimo de clases puede depender de la naturaleza de los datos.

- **Criterios de Amplitud (Ampliación):**

- **Amplitud Constante:** Todas las clases tienen la misma amplitud.

- Amplitud Variable: Las clases tienen diferentes amplitudes, lo que puede ser útil cuando los datos tienen una distribución muy asimétrica.

- Consideraciones Adicionales:

- **Evitar Clases Vacías:** Asegurarse de que cada clase contenga al menos algunas observaciones.

- **Evitar Clases Abiertas:** Evitar clases que no tienen un límite superior o inferior definido (ej. "más de 100").

- 1.8 Técnicas de Muestreo:

- Muestreo Aleatorio Simple (Ampliación):

- **Métodos de Selección:** Tabla de números aleatorios, generador de números aleatorios.

- **Ventajas:** Fácil de implementar.

- **Desventajas:** Puede no ser representativo si la población es heterogénea.

- **Muestreo Estratificado (Ampliación):**

- **Estratificación Proporcional:** El tamaño de la muestra en cada estrato es proporcional al tamaño del estrato en la población.

- **Estratificación Óptima:** El tamaño de la muestra en cada estrato se determina para minimizar la varianza del estimador.

- **Muestreo por Conglomerados (Ampliación):**

- Muestreo por Conglomerados en una Etapa: Se seleccionan algunos conglomerados al azar y se estudian todos los elementos dentro de esos conglomerados.

- Muestreo por Conglomerados en Dos Etapas: Se seleccionan algunos conglomerados al azar y luego se seleccionan algunos elementos al azar dentro de esos conglomerados.

- **Muestreo Sistemático (Ampliación):**

- Intervalo de Muestreo: Se calcula dividiendo el tamaño de la población entre el tamaño de la muestra deseado.

- 1.9 Histogramas (Aún Más Detallado):

- Construcción:

- Eje Horizontal: Representa las clases o categorías.

- **Eje Vertical:** Representa las frecuencias o las frecuencias relativas.

- **Barras:** La altura de cada barra representa la frecuencia de la clase correspondiente.

- **Interpretación:**

- **Forma:** Simétrica, asimétrica (positiva o negativa), unimodal, bimodal, multimodal.

- **Centro:** Ubicación de la media o la mediana.

- **Dispersión:** Rango o desviación estándar.

- **Valores Atípicos:** Observaciones que se encuentran muy alejadas del resto de los datos.

2. Fundamentos de la Teoría de Probabilidad

- 2.1 Técnicas de Conteo (Aún Más Detallado):

- Principio Aditivo (Ampliación):

- Ejemplo: Si tienes 3 camisas y 2 pantalones, tienes 3 + 2 = 5 opciones para vestirte si solo puedes elegir una prenda.

- **Principio Multiplicativo (Ampliación):**

- Ejemplo: Si tienes 3 camisas y 2 pantalones, tienes 3 * 2 = 6 combinaciones posibles para vestirte.

- Notación Factorial (Ampliación):

- Aplicaciones: Cálculo de permutaciones y combinaciones.

- **Convención**: 0! = 1

- **Permutaciones (Ampliación):**

- **Permutaciones con Repetición:** Si hay n objetos, donde n1 son iguales, n2 son iguales, ..., nk son iguales, el número de permutaciones es:

- n! / (n1! * n2! * ... * nk!)

- **Combinaciones (Ampliación):**

- Propiedades:

- C(n,0) = 1

- C(n,n) = 1

- C(n,r) = C(n, n-r)

- **Diagrama de Árbol (Ampliación):**

- Aplicaciones: Resolución de problemas de probabilidad, análisis de decisiones.

- **Teorema del Binomio (Ampliación):**

- (a + b)^n = Σ [C(n,k) * a^(n-k) * b^k] (desde k = 0 hasta n)

- Donde C(n,k) es el coeficiente binomial.

- 2.2 Teoría Elemental de Probabilidad :

- Espacio Muestral (S) (Ampliación):

- Espacio Muestral Discreto: Contiene un número finito o contable de resultados.

- **Espacio Muestral Continuo:** Contiene un número infinito no contable de resultados.

- Evento (A) (Ampliación):

- **Eventos Simples:** Contienen un solo resultado.

- **Eventos Compuestos:** Contienen más de un resultado.

- **Probabilidad de un Evento (P(A)) (Ampliación):**

- **Interpretaciones de la Probabilidad:**

- **Clásica:** Si todos los resultados son igualmente probables, P(A) = (número de resultados favorables) / (número total de resultados).

- **Frecuentista:** P(A) es la frecuencia relativa a largo plazo de la ocurrencia de A.

- **Subjetiva:** P(A) es una medida de la creencia personal en la ocurrencia de A.

- 2.3 Probabilidad de Eventos :

- **Unión (A ∪ B) (Ampliación):**

- Eventos Mutuamente Excluyentes: Si A y B son mutuamente excluyentes, $P(A \cup B) = P(A) + P(B)$.

- **Intersección (A ∩ B) (Ampliación):**

- **Eventos Independientes:** Si A y B son independientes, $P(A \cap B) = P(A) * P(B)$.

- **Complemento (A') (Ampliación):**

- Relación con P(A): $P(A) + P(A') = 1$

- **Diagramas de Venn (Ampliación):**

- Aplicaciones: Visualización de relaciones entre conjuntos, resolución de problemas de probabilidad.

- 2.4 Probabilidad con Técnicas de Conteo :

- Axiomas de Probabilidad (Ampliación):

- Consistencia: Los axiomas aseguran que las probabilidades sean consistentes y lógicas.

- **Teoremas (Ampliación):**

- Teorema de la Suma (General): $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

- **Teorema del Complemento:** $P(A') = 1 - P(A)$

2. Fundamentos de la Teoría de Probabilidad (Continuación Aún Más Detallado)

- 2.4 Probabilidad con Técnicas de Conteo (Ampliación Adicional):

- Teorema de la Suma para Más de Dos Eventos:

- Para tres eventos A, B y C:

- $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$

- Esta fórmula se extiende a n eventos con el principio de inclusión-exclusión.

- Ejemplos Prácticos:

- **Problema de las Bolas**: Una urna tiene 5 bolas rojas, 3 azules y 2 verdes. ¿Cuál es la probabilidad de sacar una roja o una azul?

- $P(R \cup A) = P(R) + P(A) = 5/10 + 3/10 = 8/10 = 0.8$ (ya que son mutuamente excluyentes)

- **Problema de los Cursos:** Un estudiante toma matemáticas (M) con probabilidad 0.7, física (F) con 0.6 y ambas con 0.5. ¿Cuál es la probabilidad de tomar al menos una de las dos?

- $P(M \cup F) = 0.7 + 0.6 - 0.5 = 0.8$

- 2.5 Probabilidad Condicional :

- **Interpretación Intuitiva:** Refleja cómo la información sobre la ocurrencia de un evento cambia nuestra creencia sobre la ocurrencia de otro.

- Propiedades:

- $P(A|B) \geq 0$

- $P(S|B) = 1$

- Si A y C son mutuamente excluyentes, entonces $P(A \cup C|B) = P(A|B) + P(C|B)$

- **Ejemplos Prácticos:**

- Problema de las Cartas: De una baraja de 52 cartas, se saca una al azar. Si sabemos que es una figura (J, Q, K), ¿cuál es la probabilidad de que sea de tréboles?

- $P(\text{Tréboles}|\text{Figura}) = P(\text{Tréboles} \cap \text{Figura}) / P(\text{Figura}) = (3/52) / (12/52) = 3/12 = 0.25$

- **Problema de la Enfermedad:** Un test de enfermedad tiene un 95% de sensibilidad (detecta la enfermedad si está presente) y un 90% de especificidad (no detecta la enfermedad si no está presente). Si el 1% de la población tiene la enfermedad, ¿cuál es la probabilidad de tenerla si el test es positivo? (Este es un caso clásico para la regla de Bayes, pero se puede resolver con condicionalidad básica).

- 2.6 Ley Multiplicativa (Ampliación Adicional):

- Extensión a Más de Dos Eventos:

- Para tres eventos A, B y C:

- $P(A \cap B \cap C) = P(A|B \cap C) * P(B|C) * P(C)$

- Esto es útil para calcular la probabilidad de secuencias de eventos dependientes.

- **Ejemplo Práctico:**

- **Problema de las Bolas sin Reemplazo**: Una urna tiene 4 bolas rojas y 6 azules. Se sacan dos bolas sin reemplazar la primera. ¿Cuál es la probabilidad de sacar dos rojas?

- $P(R1 \cap R2) = P(R1) * P(R2|R1) = (4/10) * (3/9) = 12/90 = 2/15 \approx 0.133$

- 2.7 Eventos Independientes: Regla de Bayes :

- Eventos Independientes :

- Dos eventos son independientes si $P(A|B) = P(A)$ o, equivalentemente, $P(B|A) = P(B)$ o $P(A \cap B) = P(A) * P(B)$.

- **Ejemplo:** Lanzar una moneda y un dado son eventos independientes, ya que el resultado de uno no afecta al otro.

- **Regla de Bayes (Ampliación):**

- **Contexto:** Es fundamental en la inferencia estadística, ya que permite actualizar nuestras creencias "a priori" con nueva evidencia para obtener creencias "a posteriori".

- **Forma General para Particiones:** Si la población se divide en k eventos mutuamente excluyentes y exhaustivos $A_1, A_2, ..., A_k$, entonces para cualquier evento B:

- $P(A_i|B) = [P(B|A_i) * P(A_i)] / \Sigma [P(B|A_j) * P(A_j)]$ (para j = 1 a k)

- Ejemplo Práctico (Continuación del Problema de la Enfermedad):

- Sea E = tener la enfermedad, P = test positivo.

- $P(E) = 0.01$, $P(\neg E) = 0.99$

- $P(P|E) = 0.95$ (sensibilidad), $P(P|\neg E) = 0.10$ (falso positivo)

- $P(E|P) = [0.95 * 0.01] / [(0.95 * 0.01) + (0.10 * 0.99)] = 0.0095 / (0.0095 + 0.099) = 0.0095 / 0.1085 \approx 0.0876$ (solo un 8.76% de probabilidad de tener la enfermedad a pesar de un test positivo)

3. Variables Aleatorias

- Definición: Una variable aleatoria es una función que asigna un valor numérico al resultado de un experimento aleatorio. En otras palabras, es una variable cuyo valor es un resultado numérico de un fenómeno aleatorio.

- Importancia: Las variables aleatorias permiten analizar y modelar fenómenos inciertos mediante herramientas matemáticas y estadísticas.

- Tipos: Se dividen principalmente en dos tipos: discretas y continuas.

3.1 Variables Aleatorias Discretas

- Definición: Una variable aleatoria discreta es aquella cuyo conjunto de posibles valores es finito o numerable (es decir, se pueden contar).

- Ejemplos: Número de caras al lanzar una moneda 5 veces, número de clientes que entran a una tienda en una hora.

3.1.1 Distribución de Probabilidad en Forma General

- **Definición:** La distribución de probabilidad de una variable aleatoria discreta describe cómo se distribuyen las probabilidades entre los diferentes valores que puede tomar la variable.

- **Representación:** Se puede representar mediante una función de masa de probabilidad (PMF), que asigna una probabilidad a cada valor posible de la variable.

- **Características:**

- La suma de todas las probabilidades debe ser igual a 1.

- Cada probabilidad debe estar entre 0 y 1.

- Ejemplo: Si X es el número de caras al lanzar una moneda dos veces, la PMF podría ser: P(X=0) = 0.25, P(X=1) = 0.5, P(X=2) = 0.25.

3.1.2 Valor Esperado (Esperanza Matemática)

- **Definición:** El valor esperado (E[X] o μ) de una variable aleatoria discreta es el promedio ponderado de los posibles valores de la variable, donde los pesos son las probabilidades de cada valor.

- Fórmula: $E[X] = \Sigma [x * P(x)]$, donde la suma se toma sobre todos los posibles valores x de la variable aleatoria.

- Interpretación: Representa el valor promedio que se esperaría obtener si se repitiera el experimento muchas veces.

- Ejemplo: Si X es el número de caras al lanzar una moneda dos veces (con la PMF anterior), entonces E[X] = 00.25 + 10.5 + 2*0.25 = 1.

3.1.3 Varianza y Desviación Estándar

- **Definición:**

- La varianza (Var[X] o σ^2) mide la dispersión de los valores de la variable aleatoria alrededor de su valor esperado.

- La desviación estándar (σ) es la raíz cuadrada de la varianza y proporciona una medida de dispersión en las mismas unidades que la variable aleatoria.

- **Fórmulas:**

- $Var[X] = E[(X - E[X])^2] = \Sigma [(x - E[X])^2 * P(x)]$

- $\sigma = \sqrt{Var[X]}$

- **Interpretación:** Una varianza o desviación estándar alta indica que los valores de la variable aleatoria están más dispersos alrededor del valor esperado, mientras que un valor bajo indica que están más concentrados.

- **Ejemplo:** Usando el ejemplo anterior, $Var[X] = (0-1)^2*0.25 + (1-1)^2*0.5 + (2-1)^2*0.25 = 0.5$, y $\sigma = \sqrt{0.5} \approx 0.707$.

3.1.4 Función Acumulada

- **Definición:** La función de distribución acumulativa (CDF), F(x), de una variable aleatoria discreta da la probabilidad de que la variable tome un valor menor o igual a x.

- **Fórmula:** $F(x) = P(X \leq x) = \Sigma P(t)$, donde la suma se toma sobre todos los valores $t \leq x$.

- **Características:**
- Es una función no decreciente.
- $F(-\infty) = 0$ y $F(\infty) = 1$.
- **Ejemplo:** Para el ejemplo de la moneda, $F(0) = P(X \leq 0) = 0.25$, $F(1) = P(X \leq 1) = 0.25 + 0.5 = 0.75$, $F(2) = P(X \leq 2) = 1$.

3.2 Variables Aleatorias Continuas

- Definición: Una variable aleatoria continua es aquella que puede tomar cualquier valor dentro de un rango dado.
- **Ejemplos:** Altura de una persona, temperatura de una habitación, tiempo que tarda en completarse una tarea.

3.2.1 Distribución de Probabilidad en Forma General

- Definición: La distribución de probabilidad de una variable aleatoria continua se describe mediante una función de densidad de probabilidad (PDF), $f(x)$.
- Características:
- $f(x) \geq 0$ para todo x.
- El área bajo la curva de $f(x)$ es igual a 1 (es decir, $\int f(x)\,dx = 1$, donde la integral se toma sobre todo el rango posible de x).
- **Interpretación:** La PDF no da la probabilidad directamente, sino que la probabilidad de que la variable tome un valor en un intervalo [a, b] es el área bajo la curva de $f(x)$ entre a y b (es decir, $P(a \leq X \leq b) = \int[a,b]\, f(x)\,dx$).

3.2.2 Valor Esperado (Esperanza Matemática)

- Definición: El valor esperado ($E[X]$ o μ) de una variable aleatoria continua es el promedio ponderado de los posibles valores de la variable, donde los pesos son dados por la PDF.
- **Fórmula:** $E[X] = \int [x * f(x)]\,dx$, donde la integral se toma sobre todo el rango posible de x.

- Interpretación: Similar a las variables discretas, representa el valor promedio que se esperaría obtener si se repitiera el experimento muchas veces.

3.2.3 Varianza y Desviación Estándar

- **Definición:**
- La varianza (Var[X] o σ^2) mide la dispersión de los valores de la variable aleatoria alrededor de su valor esperado.
- La desviación estándar (σ) es la raíz cuadrada de la varianza y proporciona una medida de dispersión en las mismas unidades que la variable aleatoria.
- Fórmulas:
- $Var[X] = E[(X - E[X])^2] = \int [(x - E[X])^2 * f(x)]\, dx$
- $\sigma = \sqrt{Var[X]}$
- **Interpretación:** Similar a las variables discretas, una varianza o desviación estándar alta indica mayor dispersión.

3.2.4 Función Acumulada

- **Definición:** La función de distribución acumulativa (CDF), F(x), de una variable aleatoria continua da la probabilidad de que la variable tome un valor menor o igual a x.
- **Fórmula:** $F(x) = P(X \leq x) = \int[-\infty,x] f(t)\, dt$
- **Características:**
- Es una función no decreciente.
- $F(-\infty) = 0$ y $F(\infty) = 1$.
- Relación con la PDF: La CDF es la integral de la PDF, y la PDF es la derivada de la CDF.

3.2.5 Cálculos de Probabilidad

- **Métodos:** Para calcular probabilidades con variables aleatorias continuas, se utilizan integrales de la PDF sobre intervalos específicos.

- Ejemplo: La probabilidad de que una variable aleatoria continua X tome un valor entre a y b es $P(a \leq X \leq b) = \int_{[a,b]} f(x)\,dx = F(b) - F(a)$.

4. Distribuciones de Probabilidad

- 4.1 Función de Probabilidad:

- Solo para Variables Discretas: Se conoce también como función de masa de probabilidad (PMF).

- **Características Clave:**

- Asigna una probabilidad positiva a cada valor posible de la variable discreta.

- La suma de todas las probabilidades es 1.

- **Ejemplo:** Para una variable aleatoria X que representa el número de caras al lanzar una moneda dos veces, la PMF es:

- $P(X=0) = 1/4$

- $P(X=1) = 2/4$

- $P(X=2) = 1/4$

- **Diferencia con la Función de Densidad :** La PMF da probabilidades directas para valores específicos; la PDF da densidades (se necesita integrar para obtener probabilidades de intervalos).

- 4.2 Distribución Binomial (Ampliación Adicional):

- Condiciones para Aplicarla:

1. Número Finito de Ensayos (n): Se realizan n experimentos idénticos.

2. Dos Resultados Posibles: Cada ensayo tiene solo dos salidas: "éxito" (S) o "fracaso" (F).

3. Probabilidad Constante de Éxito (p): La probabilidad de éxito es la misma en cada ensayo.

4. Ensayos Independientes: El resultado de un ensayo no afecta al de los demás.

- Función de Masa de Probabilidad (PMF):

- $P(X = k) = C(n, k) * p^k * (1-p)^{(n-k)}$

- Donde:
- $C(n, k)$ = combinación de n elementos tomados k a k.
- k = número de éxitos ($k = 0, 1, 2, ..., n$).
- **Valor Esperado y Varianza:**
- $E(X) = n * p$
- $Var(X) = n * p * (1-p)$
- Desv. Estándar(X) = $\sqrt{[n * p * (1-p)]}$
- **Forma de la Distribución:**
- Simétrica si $p = 0.5$.
- Asimétrica positiva si $p < 0.5$.
- Asimétrica negativa si $p > 0.5$.
- Se aproxima a una distribución normal cuando n es grande y p no está demasiado cerca de 0 o 1 (regla general: $np \geq 5$ y $n(1-p) \geq 5$).
- **Ejemplo Práctico:** Un lanzador de dardos tiene una probabilidad de 0.7 de acertar el centro. ¿Cuál es la probabilidad de acertar 3 veces en 5 lanzamientos?
- $P(X=3) = C(5,3) * (0.7)^3 * (0.3)^2 = 10 * 0.343 * 0.09 = 0.3087$.

- 4.3 Distribución Hipergeométrica :

- **Condiciones para Aplicarla:**

1. Población Finita (N): Hay N elementos en total.

2. Elementos Clasificados: K elementos son "éxitos" y (N-K) son "fracasos".

3. Muestra Sin Reemplazo: Se selecciona una muestra de tamaño n sin reemplazar los elementos.

4. Variable de Interés: Número de éxitos (k) en la muestra.

- **Función de Masa de Probabilidad (PMF):**
- $P(X = k) = [C(K, k) * C(N-K, n-k)] / C(N, n)$
- Donde:
- k = número de éxitos en la muestra (máximo entre 0 y min(K, n); mínimo entre K y n).
- **Valor Esperado y Varianza:**
- $E(X) = n * (K/N) = n * p$ (donde $p = K/N$ es la proporción de éxitos en la población)

- $Var(X) = n * (K/N) * [(N-K)/N] * [(N-n)/(N-1)]$
- El término $[(N-n)/(N-1)]$ es el factor de corrección por muestreo sin reemplazo. Si N es mucho más grande que n (ej. $N \geq 10n$), este factor se acerca a 1 y la distribución hipergeométrica se aproxima a la binomial.
- **Ejemplo Práctico:** Una caja tiene 20 productos, de los cuales 5 son defectuosos. Se toman 4 productos sin reemplazo. ¿Cuál es la probabilidad de que 2 sean defectuosos?
- $P(X=2) = [C(5,2) * C(15,2)] / C(20,4) = [10 * 105] / 4845 \approx 1050 / 4845 \approx 0.2167$.

- 4.4 Distribución de Poisson :

- **Condiciones para Aplicarla:**

1. Eventos Independientes: Los eventos ocurren de manera independiente entre sí.

2. Tasa Constante: La probabilidad de que ocurra un evento en un intervalo de tiempo o espacio es constante.

3. No Superposición: No pueden ocurrir dos eventos al mismo tiempo.

- Función de Masa de Probabilidad (PMF):

- $P(X = k) = (e^{(-\lambda)} * \lambda^k) / k!$
- Donde:
- λ = tasa promedio de eventos por intervalo ($\lambda > 0$).
- k = número de eventos (k = 0, 1, 2, ...).
- e = número de Euler (aproximadamente 2.71828).

- Valor Esperado y Varianza:

- $E(X) = \lambda$
- $Var(X) = \lambda$
- **Característica Única:** La media y la varianza son iguales.
- **Forma de la Distribución:**
- Asimétrica positiva cuando λ es pequeño.
- Se vuelve más simétrica a medida que λ aumenta (generalmente $\lambda \geq 20$ para una buena aproximación normal).
- **Aproximación a la Binomial:** Se puede usar como aproximación a la binomial cuando n es grande, p es pequeño y $np = \lambda$ (generalmente $n \geq 20$ y $p \leq 0.05$).

- Ejemplo Práctico: Un restaurante recibe un promedio de 3 pedidos por minuto. ¿Cuál es la probabilidad de recibir 5 pedidos en un minuto?

- $P(X=5) = (e^{-3} * 3^5) / 5! = (0.0498 * 243) / 120 \approx 12.1014 / 120 \approx 0.1008$.

- 4.5 Distribución Normal :

- **Características Principales:**

- Simetría: Tiene una forma de campana simétrica alrededor de la media μ.

- Colas Infinitas: Las colas se extienden hasta el infinito en ambos lados, pero nunca tocan el eje horizontal.

- Parámetros: Se define completamente por su media μ y su desviación estándar σ. Se denota como $X \sim N(\mu, \sigma^2)$.

- **Función de Densidad de Probabilidad (PDF):**

- $f(x) = (1 / (\sigma * \sqrt{2\pi})) * e^{(-(x-\mu)^2 / (2\sigma^2))}$

- **Valor Esperado y Varianza**:

- $E(X) = \mu$

- $Var(X) = \sigma^2$

- Regla Empírica (Ampliación):

- ~68.27% de los datos están entre μ - σ y μ + σ.

- ~95.45% de los datos están entre μ - 2σ y μ + 2σ.

- ~99.73% de los datos están entre μ - 3σ y μ + 3σ.

- Distribución Normal Estándar (Z):

- Es una distribución normal con media μ = 0 y desviación estándar σ = 1 ($Z \sim N(0, 1)$).

- **Transformación Z:** Cualquier variable normal $X \sim N(\mu, \sigma^2)$ se puede transformar a Z mediante:

- $Z = (X - \mu) / \sigma$

- Uso: Se utilizan tablas de distribución Z o calculadoras para encontrar probabilidades.

- **Aplicaciones:** Es la distribución más importante en estadística, ya que muchos fenómenos naturales, sociales y científicos siguen o se aproximan a ella. También es fundamental en la inferencia estadística debido al teorema del límite central.

- **Ejemplo Práctico:** La altura de los hombres en una población sigue una distribución normal con $\mu = 175$ cm y $\sigma = 6$ cm. ¿Cuál es la probabilidad de que un hombre tenga una altura mayor de 185 cm?
- $Z = (185 - 175) / 6 \approx 1.67$
- $P(X > 185) = P(Z > 1.67) = 1 - P(Z \leq 1.67) \approx 1 - 0.9525 = 0.0475$.

- 4.6 Distribución T-Student :

- Historia: Fue desarrollada por William Sealy Gosset en 1908 bajo el pseudónimo "Student".

- Características:

- Similar en forma a la distribución normal (simétrica, de campana), pero con colas más pesadas (es decir, hay más probabilidad de valores extremos).

- Parámetro: Grados de libertad (v), que dependen del tamaño de la muestra (generalmente $v = n - 1$).

- A medida que los grados de libertad aumentan, la distribución t se aproxima a la distribución normal estándar.

- **Uso:**

- Pruebas de hipótesis para la media de una población cuando la desviación estándar poblacional σ es desconocida y el tamaño de la muestra es pequeño ($n \leq 30$).
- Intervalos de confianza para la media de una población en las mismas condiciones.
- Pruebas de hipótesis para diferencias de medias de dos poblaciones.

- Función de Densidad de Probabilidad (PDF): Es más compleja que la normal, pero depende solo de los grados de libertad.

- Ejemplo Práctico: Se toma una muestra de 10 estudiantes y se mide su IQ. La media muestral es 105 y la desviación estándar muestral es 8. Se desea construir un intervalo de confianza del 95% para la media poblacional del IQ. Dado que n es pequeño y σ es desconocida, se usa la distribución t con $v = 9$. El valor crítico t es aproximadamente 2.262. El intervalo es: $105 \pm 2.262 * (8/\sqrt{10}) \approx 105 \pm 5.78$, es decir, (99.22, 110.78).

- 4.7 Distribución Chi Cuadrada (χ^2) :

- **Definición:** Se define como la suma de los cuadrados de v variables aleatorias independientes $Z_1, Z_2, ..., Zv$, cada una con distribución normal estándar. Es decir, $\chi^2 = Z_1^2 + Z_2^2 + ... + Zv^2$.

- **Parámetro:** Grados de libertad (v).

- **Características:**

- No es simétrica; es asimétrica positiva.

- Los valores son siempre no negativos (ya que son sumas de cuadrados).

- A medida que los grados de libertad aumentan, la distribución chi cuadrada se vuelve más simétrica y se aproxima a una distribución normal.

- Uso:

- Prueba de Bondad de Ajuste: Evaluar si un conjunto de datos observados se ajusta a una distribución teórica esperada.

- Prueba de Independencia: Evaluar si dos variables categóricas son independientes entre sí.

- **Prueba de Homogeneidad:** Evaluar si las proporciones de una variable categórica son iguales en varias poblaciones.

- Intervalos de confianza para la varianza de una población.

- Función de Densidad de Probabilidad (PDF): Depende solo de los grados de libertad.

- Ejemplo Práctico: Se quiere verificar si un dado está equilibrado. Se lanza 60 veces y se obtienen las siguientes frecuencias: 9 (1), 11 (2), 8 (3), 12 (4), 10 (5), 10 (6). La frecuencia esperada por cara es 10. El estadístico chi cuadrada es $\Sigma[(O - E)^2 / E] = [(9-10)^2/10] + ... + [(10-10)^2/10] = 0.1 + 0.1 + 0.4 + 0.4 + 0 + 0 = 1.0$. Con v = 5 grados de libertad, el valor p es mayor que 0.05, por lo que no hay evidencia para rechazar que el dado está equilibrado.

- 4.8 Distribución Fisher (F) :

- **Definición:** Se define como el cociente de dos variables aleatorias chi cuadrada independientes, cada una dividida por sus respectivos grados de libertad. Es decir, $F = (\chi_1^2 / v_1) / (\chi_2^2 / v_2)$, donde χ_1^2 y χ_2^2 son independientes.

- **Parámetros:** Grados de libertad del numerador (v_1) y grados de libertad del denominador (v_2).

- Características:

- No es simétrica; es asimétrica positiva.

- Los valores son siempre no negativos.

- La forma depende de ambos parámetros de grados de libertad.

- Uso:

- Prueba de Comparación de Varianzas: Evaluar si las varianzas de dos poblaciones son iguales.

- Análisis de Varianza (ANOVA): Evaluar si las medias de tres o más poblaciones son iguales.

- Pruebas de significancia en modelos de regresión múltiple.

- **Función de Densidad de Probabilidad (PDF):** Depende de ambos grados de libertad.

- Relación con la Distribución T: Si Z sigue una distribución t con v grados de libertad, entonces Z^2 sigue una distribución F con 1 y v grados de libertad.

- Ejemplo Práctico: Se quieren comparar las varianzas de dos grupos de estudiantes en un examen. Grupo 1: $n_1 = 15$, $s_1^2 = 25$. Grupo 2: $n_2 = 12$, $s_2^2 = 16$. El estadístico F es $s_1^2 / s_2^2 = 25 / 16 = 1.5625$. Con $v_1 = 14$ y $v_2 = 11$ grados de libertad, el valor p para una prueba bilateral es mayor que 0.05, por lo que no hay evidencia para rechazar que las varianzas son iguales.

5. Regresión Lineal

- ¿Qué es (Ampliación General)?

- **Es una técnica de análisis estadístico** que busca modelar la relación lineal entre una variable dependiente (también llamada variable de respuesta o Y) y una o más variables independientes (también llamadas variables explicativas, predictoras o X).
- **El objetivo principal** es predecir el valor de la variable dependiente a partir de los valores de las variables independientes, o bien explicar cómo cambia la variable dependiente cuando cambian las variables independientes.
- **Se divide en regresión lineal simple** (una sola variable independiente) y regresión lineal múltiple (varias variables independientes) — el temario se centra en la simple, pero incluiremos contexto general.

5.1 Regresión y Correlación

- Diferencia Clave Entre Regresión y Correlación:

- **Correlación:** Medida de la fuerza y dirección de la relación lineal entre dos variables. No implica causalidad, y las variables son tratadas de manera simétrica (no hay variable dependiente ni independiente).
- **Regresión:** Modelo que describe la relación funcional lineal entre una variable dependiente y una o más independientes. Implica una dirección en la relación (se predice Y a partir de X) y puede sugerir causalidad si el diseño del estudio lo permite.
- **Ejemplo**: La correlación entre el número de horas de estudio y la calificación en un examen indica qué tan fuerte está la relación lineal entre ambas. La regresión permite predecir la calificación esperada para un número dado de horas de estudio.

- Limitaciones de la Correlación Lineal:

- Solo mide relaciones lineales — no detecta relaciones no lineales (ej. una relación curva).
- Se ve afectada por valores atípicos.
- No implica causalidad (correlación no es causa).

5.1.1 Diagrama de Dispersión

- ¿Qué es con Más Detalle?

- Es un gráfico bidimensional donde cada punto representa una observación, con el valor de la variable independiente (X) en el eje horizontal y el valor de la variable dependiente (Y) en el eje vertical.

- Objetivos del Diagrama de Dispersión:

1. Detectar la Forma de la Relación: ¿Es lineal, curva, no existe relación?

2. Detectar la Dirección: ¿Es positiva (a medida que X aumenta, Y aumenta) o negativa (a medida que X aumenta, Y disminuye)?

3. Detectar la Fuerza: ¿Los puntos están muy agrupados (relación fuerte) o muy dispersos (relación débil)?

4. Identificar Valores Atípicos: Puntos que se alejan mucho del resto, lo que podría afectar el modelo de regresión.

5. Identificar Valores de Influencia: Puntos que, aunque no sean atípicos en valor, tienen un impacto significativo en la pendiente o el intercepto de la recta de regresión.

- Ejemplo: Un diagrama de dispersión con puntos que siguen una línea ascendente sugieren una correlación positiva; si siguen una línea descendente, una correlación negativa; si están dispersos sin orden, no hay correlación lineal.

- Herramientas Adicionales en el Diagrama de Dispersión:

- Recta de Regresión: Se puede superponer la recta ajustada para visualizar el modelo.

- Burbujas: Si hay una tercera variable, el tamaño de las burbujas puede representarla (diagrama de dispersión de burbujas).

5.1.2 Regresión Lineal Simple

- Modelo Teórico Poblacional:

- La relación verdadera en la población se expresa como:

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

- Donde:

- Y = Variable dependiente.

- X = Variable independiente.

- β_0 = Intercepto poblacional (valor de Y cuando X = 0).
- β_1 = Pendiente poblacional (cambio en Y por cada unidad de cambio en X).
- ε = Término de error (representa la variabilidad en Y que no es explicada por X; se asume que $\varepsilon \sim N(0, \sigma^2)$).

- Modelo Estimado Muestral:

- Dado que no conocemos β_0 y β_1, los estimamos a partir de la muestra con b_0 y b_1, obteniendo la recta de regresión estimada:
- $\hat{Y} = b_0 + b_1X$
- Donde:
- $\hat{Y}$ (Y gorro) = Valor predicho de Y para un dado X.
- b_0 = Estimador del intercepto poblacional β_0.
- b_1 = Estimador de la pendiente poblacional β_1.

- Método de los Mínimos Cuadrados Ordinarios (MCO):

- Es el método más común para estimar b_0 y b_1. Consiste en minimizar la suma de los cuadrados de los residuos (SCR):
- $SCR = \Sigma (Y_i - \hat{Y}_i)^2 = \Sigma (Y_i - b_0 - b_1X_i)^2$
- Donde Y_i es el valor observado de Y para la i-ésima observación, y $\hat{Y}_i$ es el valor predicho.

- Fórmulas para los Estimadores MCO:

- Pendiente (b_1):
- $b_1 = [n\Sigma(X_iY_i) - (\Sigma X_i)(\Sigma Y_i)] / [n\Sigma(X_i^2) - (\Sigma X_i)^2]$
- O, de manera más compacta:
- $b_1 = Cov(X, Y) / Var(X)$
- Donde Cov(X, Y) es la covarianza entre X y Y, y Var(X) es la varianza de X.
- Intercepto (b_0):
- $b_0 = \bar{Y} - b_1\bar{X}$
- Donde $\bar{Y}$ es la media de Y y $\bar{X}$ es la media de X.

- Interpretación de los Coeficientes:

- b_1: Por cada unidad que aumenta X, se espera que Y aumente (si b_1 es positivo) o disminuya (si b_1 es negativo) en b_1 unidades, manteniéndose constantes otras variables (en el caso simple, no hay otras variables).

- b_0: Es el valor predicho de Y cuando X = 0. Importante: Solo tiene interpretación práctica si X = 0 es un valor plausible en el contexto del problema.

- Ejemplo Práctico:

- Se quiere modelar la relación entre el número de horas de estudio (X) y la calificación en un examen (Y). Los datos son:

Horas (X) Calificación (Y)

2 60

4 70

5 75

7 85

8 90

- Cálculos:

- $\Sigma X = 26$, $\Sigma Y = 380$, $\Sigma XY = 2090$, $\Sigma X^2 = 158$, $n = 5$
- $b_1 = (52090 - 26380) / (5*158 - 26^2) = (10450 - 9880) / (790 - 676) = 570 / 114 = 5$
- $\bar{Y} = 380/5 = 76$, $\bar{X} = 26/5 = 5.2$
- $b_0 = 76 - 5*5.2 = 76 - 26 = 50$
- Modelo estimado: $\hat{Y} = 50 + 5X$

- Interpretación: Por cada hora adicional de estudio, se espera que la calificación aumente en 5 puntos. Si alguien no estudia (X=0), se espera una calificación de 50 puntos (interpretación plausible en este contexto).

5.1.3 Correlación

- Coeficiente de Correlación de Pearson (r):

- Es la medida más común de correlación lineal. Se calcula como:

- $r = [n\Sigma(X_iY_i) - (\Sigma X_i)(\Sigma Y_i)] / \sqrt{[n\Sigma(X_i^2) - (\Sigma X_i)^2]} * \sqrt{[n\Sigma(Y_i^2) - (\Sigma Y_i)^2]}$

- O, de manera más compacta:

- $r = Cov(X, Y) / (\sigma_x * \sigma_Y)$
- Donde σ_x es la desviación estándar de X y σ_Y es la desviación estándar de Y.
- **Rango de Valores:** $-1 \leq r \leq 1$.

- r = 1: Correlación positiva perfecta (todos los puntos caen en una recta ascendente).

- r = -1: Correlación negativa perfecta (todos los puntos caen en una recta descendente).

- r = 0: No hay correlación lineal (los puntos no siguen una tendencia lineal).

- Interpretación de la Magnitud:

- 0.00 - 0.19: Correlación muy débil.
- 0.20 - 0.39: Correlación débil.
- 0.40 - 0.59: Correlación moderada.
- 0.60 - 0.79: Correlación fuerte.
- 0.80 - 1.00: Correlación muy fuerte.

- Relación con la Pendiente de la Regresión:

- La signo de r es la misma que la de b_1 (ambos positivos o ambos negativos).
- Pero su magnitud no es la misma — r mide la fuerza de la relación, mientras que b_1 mide el cambio en Y por unidad de X.

- Coeficiente de Correlación de Spearman (ρ):

- Medida de la relación monotónica (no necesariamente lineal) entre dos variables. Se basa en los rangos de los datos, por lo que es robusto a valores atípicos.

- Ejemplo Práctico (continuación del caso de horas y calificaciones):

- $\Sigma Y^2 = 60^2 + 70^2 + 75^2 + 85^2 + 90^2 = 3600 + 4900 + 5625 + 7225 + 8100 = 29450$
- $r = (52090 - 26380) / \sqrt{[(5158 - 26^2)(529450 - 380^2)]} = 570 / \sqrt{[(114)(147250 - 144400)]} = 570 / \sqrt{[114*2850]} = 570 / \sqrt{[324900]} = 570 / 570 = 1$
- Esto indica una correlación positiva perfecta (lo cual es esperado en este ejemplo artificial, ya que todos los puntos caen en la recta de regresión).

5.1.4 Determinación y Análisis de los Coeficientes de Correlación y de Determinación

- Coeficiente de Determinación (R^2):

- Definición: Es el cuadrado del coeficiente de correlación de Pearson ($R^2 = r^2$) en el caso de la regresión lineal simple. En regresión múltiple, es más complejo pero tiene una interpretación similar.

- **Interpretación Clave:** Representa la proporción de la variabilidad total en la variable dependiente (Y) que es explicada por la variable independiente (X) en el modelo de regresión.
- **Rango de Valores:** $0 \leq R^2 \leq 1$.
- $R^2 = 0$: El modelo no explica ninguna variabilidad en Y (la recta de regresión es horizontal, igual a la media de Y).
- $R^2 = 1$: El modelo explica toda la variabilidad en Y (todos los puntos caen en la recta de regresión).
- **Cálculo Alternativo:**
- $R^2 = 1 - (SCR / SCT)$
- Donde:
- SCR = Suma de los cuadrados de los residuos (variabilidad no explicada).
- SCT = Suma de los cuadrados totales (variabilidad total en Y: $SCT = \Sigma(Y_i - \bar{Y})^2$).
- SCE = Suma de los cuadrados de la explicación (variabilidad explicada por el modelo: $SCE = \Sigma(\hat{Y}_i - \bar{Y})^2 = SCT - SCR$).
- **Análisis de los Coeficientes:**
- **Significancia Estadística de los Coeficientes:**
- Se realizan pruebas de hipótesis para determinar si los coeficientes (especialmente b_1) son estadísticamente diferentes de cero.
- **Hipótesis para la pendiente:**
- H_0: $\beta_1 = 0$ (no hay relación lineal entre X y Y)
- H_1: $\beta_1 \neq 0$ (hay una relación lineal entre X y Y)
- **Estadístico de Prueba:** Se usa la distribución t con (n - 2) grados de libertad:
- $t = b_1 / SE(b_1)$
- Donde $SE(b_1)$ es el error estándar de la pendiente.
- Si el valor p es menor que el nivel de significancia (α, generalmente 0.05), se rechaza H_0 y se concluye que hay una relación lineal significativa entre X y Y.
- **Intervalos de Confianza para los Coeficientes**: Se pueden calcular intervalos de confianza para β_0 y β_1 para conocer el rango de valores en el que se espera que se encuentren los parámetros poblacionales.

- Ejemplo Práctico :

- $R^2 = r^2 = (1)^2 = 1$. Esto significa que el 100% de la variabilidad en las calificaciones es explicada por el número de horas de estudio (lo cual es normal en un ejemplo artificial).
- La prueba de hipótesis para β_1 daría un valor p muy pequeño, indicando que la relación es significativa.

5.1.5 Distribución Normal Bidimensional (Ampliación Adicional)

- ¿Qué es?
- Es la generalización de la distribución normal unidimensional a dos variables aleatorias X y Y. Se define completamente por:

1. La media de X (μ_X) y la media de Y (μ_Y).
2. La varianza de X (σ_X^2) y la varianza de Y (σ_Y^2).
3. El coeficiente de correlación entre X y Y (ρ).

- Se denota como $(X, Y) \sim N_2(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$.

- Función de Densidad de Probabilidad (PDF):

- Es una función tridimensional en forma de "campana" que describe la probabilidad conjunta de X y Y.

- Importancia en la Regresión Lineal:

- Los supuestos del modelo de regresión lineal simple (especialmente que los errores ε son normales, independientes y con varianza constante) implican que Y, dado X, sigue una distribución normal unidimensional con media $\beta_0 + \beta_1 X$ y varianza σ^2. Esto está relacionado con la distribución normal bidimensional

- Importancia en la Regresión Lineal :

- Si (X, Y) siguen una distribución normal bidimensional, entonces la relación entre X y Y es exclusivamente lineal — no hay relaciones no lineales ocultas.
- En este caso, el coeficiente de correlación de Pearson es la medida perfecta de la fuerza de la relación entre ambas variables.
- Las pruebas de significancia para el coeficiente de correlación y para la pendiente de la regresión son equivalentes (dan el mismo valor p).

- Propiedades Adicionales:

- Las marginales de una distribución normal bidimensional son normales: $X \sim N(\mu_X, \sigma_X^2)$ y $Y \sim N(\mu_Y, \sigma_Y^2)$.
- Las condicionales también son normales: $Y|X=x \sim N(\mu_Y + \rho(\sigma_Y/\sigma_X)(x - \mu_X), \sigma_Y^2(1 - \rho^2))$ y $X|Y=y \sim N(\mu_X + \rho(\sigma_X/\sigma_Y)(y - \mu_Y), \sigma_X^2(1 - \rho^2))$.
- Si $\rho = 0$, entonces X y Y son independientes (lo cual no es cierto para distribuciones no normales).

5.1.6 Intervalos de Confianza y Pruebas para el Coeficiente de Correlación

- Prueba de Hipótesis para el Coeficiente de Correlación Poblacional (ρ):

- El objetivo es determinar si hay una relación lineal significativa entre X y Y en la población.

- Hipótesis Comunes:

- H_0: $\rho = 0$ (no hay correlación lineal en la población)
- H_1: $\rho \neq 0$ (hay correlación lineal en la población) — prueba bilateral
- También se pueden hacer pruebas unilaterales (H_1: $\rho > 0$ o H_1: $\rho < 0$)

- Estadístico de Prueba:

- Si H_0 es verdadera y los datos siguen una distribución normal bidimensional, se usa la distribución t con (n - 2) grados de libertad:
- $t = r * \sqrt{[(n - 2) / (1 - r^2)]}$
- **Alternativa (transformación de Fisher):** Para intervalos de confianza o pruebas con $\rho \neq 0$, se usa la transformación z de Fisher:
- $z' = (1/2) * \ln[(1 + r) / (1 - r)]$
- Esta transformación aproxima a una distribución normal con media $z'_\rho = (1/2) * \ln[(1 + \rho) / (1 - \rho)]$ y varianza $1/(n - 3)$.

- Intervalo de Confianza para ρ:

- Se calcula usando la transformación de Fisher:

1. Calcular z' a partir de r.

2. Calcular el intervalo de confianza para z'ρ: $z' \pm z(\alpha/2) * \sqrt{[1/(n - 3)]}$

3. Invertir la transformación para obtener el intervalo de confianza para ρ:

- ρ = [(e^(2z'_inf) - 1) / (e^(2z'_inf) + 1), (e^(2z'_sup) - 1) / (e^(2z'_sup) + 1)]

- Ejemplo Práctico:

- Supongamos que en una muestra de n = 20 observaciones, el coeficiente de correlación es r = 0.5. Queremos probar si ρ ≠ 0 con α = 0.05.
- t = 0.5 * √[(20 - 2) / (1 - 0.25)] = 0.5 * √[18 / 0.75] = 0.5 * √24 ≈ 0.5 * 4.899 ≈ 2.449
- Grados de libertad = 18. El valor crítico t es ±2.101.
- Como 2.449 > 2.101, se rechaza H_0: hay evidencia de una correlación lineal significativa en la población.

- Intervalo de Confianza para ρ:

- z' = 0.5 * ln[(1 + 0.5)/(1 - 0.5)] = 0.5 * ln(3) ≈ 0.5 * 1.0986 ≈ 0.5493
- Intervalo para z': 0.5493 ± 1.96 * √[1/(20 - 3)] ≈ 0.5493 ± 1.96 * 0.2425 ≈ 0.5493 ± 0.4753 → (0.074, 1.0246)
- Intervalo para ρ: [(e^(20.074) - 1)/(e^(20.074) + 1), (e^(21.0246) - 1)/(e^(21.0246) + 1)] ≈ (0.074, 0.773)
- Interpretación: Estamos seguros en un 95% de que el coeficiente de correlación poblacional está entre 0.074 y 0.773.

5.1.7 Errores de Medición

- ¿Qué son con Más Detalle?

- Son desviaciones entre el valor medido de una variable y su valor verdadero (que generalmente no conocemos). Pueden ser:

- Errores Aleatorios: Varían de una medición a otra en forma impredecible, pero se cancelan en promedio. Afectan la precisión de las mediciones.

- Errores Sistemáticos: Son desviaciones constantes o con un patrón, que no se cancelan. Afectan la exactitud de las mediciones.

- Impacto en la Regresión Lineal:

- Errores en la Variable Independiente (X):

- Si X tiene errores de medición, los estimadores MCO de b_0 y b_1 son insesgados y consistentes solo si los errores en X son independientes de Y y de los errores en Y.

- Si no es así, se produce el sesgo de errores de medición (también llamado sesgo de regresión atenuada), donde la pendiente b_1 se estima con un valor menor en magnitud que el verdadero β_1.

- Errores en la Variable Dependiente (Y):

- Estos errores son absorbidos por el término de error ε del modelo de regresión. Siempre y cuando se cumplan los supuestos del modelo ($\varepsilon \sim N(0, \sigma^2)$, independientes, etc.), los estimadores MCO siguen siendo insesgados y eficientes.

- Métodos para Manejar Errores de Medición:

- Modelos de Variables Instrumentales: Se usa una variable "instrumental" que está correlacionada con X verdadero pero no con los errores en X ni con ε.

- Modelos de Regresión con Errores en Variables (EIV): Métodos específicos que tienen en cuenta los errores en X.

Supuestos del Modelo de Regresión Lineal Simple (Ampliación Adicional — tema implícito pero crucial)

- Para que los estimadores MCO sean insesgados, eficientes y consistentes, y para que las pruebas de hipótesis y los intervalos de confianza sean válidos, se deben cumplir los siguientes supuestos:

1. Linealidad: La relación entre X y Y en la población es lineal ($Y = \beta_0 + \beta_1 X + \varepsilon$).

2. Media Cero de los Errores: $E(\varepsilon_i) = 0$ para todas las i.

3. Homocedasticidad: La varianza de los errores es constante para todos los valores de X: $Var(\varepsilon_i) = \sigma^2$.

4. Independencia de los Errores: Los errores son independientes entre sí: $Cov(\varepsilon_i, \varepsilon_j) = 0$ para $i \neq j$.

5. No Correlación entre X y los Errores: $Cov(X_i, \varepsilon_i) = 0$.

6. Normalidad de los Errores (para inferencias pequeñas): $\varepsilon_i \sim N(0, \sigma^2)$.

- Prueba de los Supuestos:

- Linealidad y Homocedasticidad: Se analizan los gráficos de residuos (residuos vs. valores predichos $\hat{Y}$, o residuos vs. X).

- Si los residuos se dispersan aleatoriamente alrededor de cero sin patrón, se cumple la linealidad y la homocedasticidad.
- Si hay un patrón curvo, la linealidad no se cumple.
- Si la dispersión de los residuos cambia con $\hat{Y}$ o X, hay heterocedasticidad.

- Independencia de los Errores: Se usa el test de Durbin-Watson (para datos de serie temporal) o análisis de patrones en los residuos.

- Normalidad de los Errores: Se usan gráficos de histograma de residuos, diagrama de probabilidad normal (Q-Q plot) o pruebas de normalidad (Shapiro-Wilk, Kolmogorov-Smirnov).

- Consecuencias de Violar los Supuestos:

- Insesgadez perdida: Los estimadores no se acercan al parámetro poblacional en promedio.

- Eficacia perdida: Hay estimadores más precisos.

- Errores estándar incorrectos: Las pruebas de hipótesis y los intervalos de confianza no son válidos.

6. Estadística Aplicada

- ¿Qué es (Ampliación General)?

- Es la rama de la estadística que se encarga de aplicar métodos y técnicas estadísticos para resolver problemas prácticos en el mundo real. A diferencia de la estadística matemática (que se centra en la teoría y la deducción), la estadística aplicada se basa en la colección, análisis, interpretación y presentación de datos para tomar decisiones informadas en campos como la ciencia, la medicina, los negocios, la economía, la sociología, entre otros.
- El temario se centra en el muestreo, la estimación y la prueba de hipótesis — herramientas fundamentales para la inferencia estadística (pasar de los datos de una muestra a conclusiones sobre una población).

6.1 Muestreo

- ¿Por Qué Muestreamos?

- Costo: Estudiar toda la población es demasiado caro.

- Tiempo: Estudiar toda la población toma demasiado tiempo.

- Imposibilidad: En algunos casos, es imposible estudiar toda la población (ej. probar productos destructivos, como bombillas).

- Precisión: Un muestreo bien diseñado puede proporcionar resultados tan precisos como un censo.

6.1.1 Tipos de Muestreo

Se clasifican en dos grandes categorías:

- 1. Muestreo Probabilístico (Aleatorio): Cada elemento de la población tiene una probabilidad conocida y positiva de ser seleccionado. Garantiza que la muestra sea representativa (en promedio) y permite calcular el error de muestreo.

- Muestreo Aleatorio Simple (MAS):

- Descripción: Cada conjunto de n elementos de la población tiene la misma probabilidad de ser seleccionado como muestra.

- Métodos de Selección: Tablas de números aleatorios, generadores de números aleatorios, selección aleatoria manual (ej. sacar bolas de una urna).

- Ventajas: Fácil de entender y aplicar cuando la población está bien definida y accesible.

- Desventajas: Puede ser ineficiente si la población es grande o heterogénea. Requiere una lista completa de la población (marco muestral).

- Muestreo Estratificado:

- Descripción: La población se divide en subconjuntos homogéneos llamados estratos (ej. por género, edad, región). Luego se selecciona una muestra aleatoria simple de cada estrato.

- Tipos de Estratificación:

- Proporcional: El tamaño de la muestra en cada estrato es proporcional al tamaño del estrato en la población.

- **Óptima (Neyman):** El tamaño de la muestra en cada estrato se determina para minimizar el error de muestreo, teniendo en cuenta la variabilidad dentro de cada estrato y el costo de muestrear en cada uno.
- **Ventajas:** Mayor precisión que el MAS cuando la población es heterogénea. Permite analizar cada estrato por separado.
- **Desventajas:** Requiere conocer la estructura de la población para definir los estratos. Es más complejo de implementar.
- **Muestreo por Conglomerados:**
- **Descripción:** La población se divide en subconjuntos llamados conglomerados (ej. escuelas, barrios, ciudades) que son heterogéneos (similares a la población en su conjunto). Luego se seleccionan algunos conglomerados al azar y se estudian todos los elementos dentro de ellos (muestreo en una etapa) o se selecciona una muestra de elementos dentro de los conglomerados seleccionados (muestreo en dos etapas).
- **Diferencia con el Estratificado:** Los estratos son homogéneos entre sí; los conglomerados son heterogéneos.
- **Ventajas:** No requiere una lista completa de la población, solo una lista de conglomerados. Es económico y fácil de implementar en áreas geográficas amplias.
- **Desventajas:** Menos preciso que el MAS o el estratificado si los conglomerados son pequeños.
- **Muestreo Sistemático:**
- **Descripción:** Se selecciona un elemento al azar de la población y luego se seleccionan los elementos subsiguientes a intervalos regulares.
- **Pasos:**

1. **Calcular el intervalo de muestreo**: k = N / n (donde N es el tamaño de la población y n es el tamaño de la muestra).
2. Seleccionar un número aleatorio r entre 1 y k.
3. Los elementos seleccionados son: r, r + k, r + 2k, ..., hasta completar la muestra.

- **Ventajas:** Fácil de implementar. No requiere una lista completa de la población si el intervalo se puede aplicar en forma secuencial.
- **Desventajas**: Puede ser sesgado si hay un patrón en la población que coincide con el intervalo de muestreo (ej. una lista donde cada k-ésimo elemento es diferente).

- **2. Muestreo No Probabilístico (No Aleatorio**): No se asignan probabilidades conocidas a los elementos de la población. La muestra puede no ser representativa, y no se puede calcular el error de muestreo. Se usa cuando el muestreo probabilístico es imposible o demasiado costoso, pero los resultados deben interpretarse con precaución.
- **Muestreo por Conveniencia**: Se selecciona la muestra de elementos que están fácilmente accesibles (ej. encuestar a personas en la calle).
- **Muestreo por Juicio (Purposive):** El investigador selecciona la muestra basándose en su juicio sobre qué elementos son representativos.
- **Muestreo por Snowball (Nieve):** Se selecciona un pequeño grupo de elementos, y estos a su vez recomiendan otros elementos para la muestra (útil para poblaciones difíciles de alcanzar).
- **Muestreo por Cuotas:** Se define un número de cuotas para diferentes grupos de la población, y se selecciona la muestra para cumplir con las cuotas (similar al estratificado, pero sin selección aleatoria).
- **Marco Muestral:**
- Es la lista o fuente que contiene todos los elementos de la población que pueden ser seleccionados en la muestra. Debe ser lo más completo y preciso posible para evitar el sesgo de marco.

6.1.2 Teorema del Límite Central (TLC)

- **¿Qué Es con Más Detalle?**
- Es uno de los teoremas más importantes en estadística. Establece que, independientemente de la distribución de la población, la distribución de las medias muestrales se aproxima a una distribución normal a medida que el tamaño de la muestra n aumenta.
- **Condiciones Clave:**
- El tamaño de la muestra debe ser suficientemente grande (generalmente $n \geq 30$, aunque si la población es normal, el TLC se cumple para cualquier n).
- Las observaciones deben ser independientes entre sí.
- **Características de la Distribución de Medias Muestrales (según el TLC):**

- **Media**: La media de las medias muestrales es igual a la media de la población: $\mu_{\bar{Y}} = \mu$.
- **Varianza**: La varianza de las medias muestrales es igual a la varianza de la población dividida por el tamaño de la muestra: $Var(\bar{Y}) = \sigma^2 / n$.
- Desviación Estándar: La desviación estándar de las medias muestrales se conoce como error estándar de la media (EEM) o error estándar muestral (ESM): $\sigma_{\bar{Y}} = \sigma / \sqrt{n}$.
- Importancia del TLC:
- Permite usar métodos de inferencia estadística basados en la distribución normal, incluso cuando la distribución de la población es desconocida o no normal, siempre que el tamaño de la muestra sea suficientemente grande.
- Ejemplo Práctico: Si la población de ingresos de una ciudad tiene una distribución asimétrica positiva, la distribución de las medias de muestras de tamaño 50 tomadas de esa población será aproximadamente normal.

6.1.3 Distribución Muestral de la Media

- ¿Qué Es?
- Es la distribución de todas las posibles medias muestrales $\bar{Y}$ que se pueden obtener de todas las posibles muestras de tamaño n tomadas de una población.
- Casos Específicos:
1. La población es normal ($X \sim N(\mu, \sigma^2)$):
- Entonces, la distribución de la media muestral es también normal, independientemente del tamaño de la muestra: $\bar{Y} \sim N(\mu, \sigma^2 / n)$.
- Si σ es desconocida y n es pequeño ($n \leq 30$), se usa la distribución t en lugar de la normal: $\bar{Y} \sim t(\mu, s^2 / n)$ con $v = n - 1$ grados de libertad.
2. La población no es normal, pero n es grande ($n \geq 30$) (TLC):
- La distribución de la media muestral se aproxima a la normal: $\bar{Y} \sim N(\mu, \sigma^2 / n)$ (aproximación).
- Si σ es desconocida, se reemplaza por la desviación estándar muestral s, y la aproximación sigue siendo válida: $\bar{Y} \sim N(\mu, s^2 / n)$ (aproximación).
- Transformación Z para la Media Muestral:

- Si $\bar{Y} \sim N(\mu, \sigma^2 / n)$, entonces:
- $Z = (\bar{Y} - \mu) / (\sigma / \sqrt{n}) \sim N(0, 1)$

- Si σ es desconocida y n es grande, se usa:

- $Z = (\bar{Y} - \mu) / (s / \sqrt{n}) \sim N(0, 1)$ (aproximación)

- Ejemplo Práctico: La altura de los hombres en una población es normal con $\mu = 175$ cm y $\sigma = 6$ cm. Se toma una muestra de $n = 25$ hombres. ¿Cuál es la probabilidad de que la media muestral sea mayor de 177 cm?

- $\bar{Y} \sim N(175, 6^2 / 25) = N(175, 1.44)$
- $Z = (177 - 175) / (6 / \sqrt{25}) = 2 / 1.2 \approx 1.67$
- $P(\bar{Y} > 177) = P(Z > 1.67) \approx 1 - 0.9525 = 0.0475$.

6.1.4 Distribución Muestral de una Proporción

- ¿Qué Es?

- Es la distribución de todas las posibles proporciones muestrales $\hat{p}$ que se pueden obtener de todas las posibles muestras de tamaño n tomadas de una población con proporción p de elementos que tienen una característica específica.

- Definición de Proporción Muestral:

- $\hat{p} = x / n$, donde x es el número de elementos en la muestra que tienen la característica.

- Aproximación Normal (TLC para Proporciones):

- A medida que el tamaño de la muestra n aumenta, la distribución de $\hat{p}$ se aproxima a una distribución normal, siempre que se cumplan las condiciones de muestra suficientemente grande:
- $np \geq 5$ y $n(1 - p) \geq 5$ (o $np \geq 10$ y $n(1 - p) \geq 10$, según algunas fuentes).

- Características de la Distribución Aproximada:

- **Media:** La media de las proporciones muestrales es igual a la proporción de la población: $\mu_{\hat{p}} = p$.
- **Varianza:** La varianza de las proporciones muestrales es: $Var(\hat{p}) = p(1 - p) / n$.
- **Desviación Estándar:** La desviación estándar de las proporciones muestrales se conoce como error estándar de la proporción: $\sigma_{\hat{p}} = \sqrt{[p(1 - p) / n]}$.

- Transformación Z para la Proporción Muestral:

- Si $\hat{p} \sim N(p, p(1 - p) / n)$ (aproximación), entonces:

- $Z = (\hat{p} - p) / \sqrt{[p(1 - p) / n]} \sim N(0, 1)$

- Si p es desconocido (lo cual es usualmente el caso), se reemplaza por $\hat{p}$ en el error estándar:

- $Z = (\hat{p} - p) / \sqrt{[\hat{p}(1 - \hat{p}) / n]} \sim N(0, 1)$ (aproximación)

- Corrección de Continuidad:

- Dado que $\hat{p}$ es una variable discreta y la normal es continua, se puede aplicar una corrección de continuidad para mejorar la aproximación (ej. si se calcula $P(\hat{p} \geq 0.5)$, se usa $P(\hat{p} \geq 0.495)$ en la distribución normal).

- Ejemplo Práctico: El 30% de los estudiantes de una universidad son de origen rural. Se toma una muestra de n = 100 estudiantes. ¿Cuál es la probabilidad de que la proporción muestral de estudiantes de origen rural sea mayor de 0.35?

- $np = 100*0.3 = 30 \geq 5$, $n(1-p) = 70 \geq 5 \rightarrow$ aproximación normal es válida.

- $\mu_{\hat{p}} = 0.3$, $\sigma_{\hat{p}} = \sqrt{[0.3*0.7 / 100]} = \sqrt{[0.21 / 100]} \approx 0.0458$

- $Z = (0.35 - 0.3) / 0.0458 \approx 0.05 / 0.0458 \approx 1.09$

- $P(\hat{p} > 0.35) = P(Z > 1.09) \approx 1 - 0.8621 = 0.1379$.

6.2 Estimación

- ¿Qué Es?

- Es el proceso de utilizar datos de una muestra para obtener información sobre un parámetro poblacional (ej. media, proporción, varianza). Se divide en estimación puntual y estimación por intervalo.

6.2.1 Estimación Puntual

- ¿Qué Es?

- Consiste en obtener un solo valor numérico que se considera el mejor estimador del parámetro poblacional. Este valor se llama estimador puntual.

- **Características Deseables de un Estimador:**

1. Insesgado: El valor esperado del estimador es igual al parámetro poblacional: $E(\hat{\theta}) = \theta$. Si no es así, se dice que tiene sesgo: $Sesgo = E(\hat{\theta}) - \theta$.

2. Eficiente: Entre todos los estimadores insesgados, es el que tiene la menor varianza (es decir, es el más preciso).

3. Consistente: A medida que el tamaño de la muestra n aumenta, el estimador se acerca al parámetro poblacional (en probabilidad).

4. Suficiente: Contiene toda la información de la muestra que es relevante para estimar el parámetro.

- Ejemplos de Estimadores Puntuales Comunes:

- Media Poblacional (μ): El estimador puntual es la media muestral ($\bar{Y}$). Es insesgado, eficiente y consistente.

- $\bar{Y} = (\Sigma Y_i) / n$

- Proporción Poblacional (p): El estimador puntual es la proporción muestral ($\hat{p}$). Es insesgado, eficiente y consistente.

- $\hat{p} = x / n$ (donde x es el número de elementos con la característica)

- Varianza Poblacional (σ^2): El estimador puntual es la varianza muestral (s^2) (con corrección de grados de libertad). Es insesgado y consistente.

- $s^2 = [\Sigma(Y_i - \bar{Y})^2] / (n - 1)$

- Nota: Si se usa n en el denominador ($s^2_n = \Sigma(Y_i - \bar{Y})^2 / n$), el estimador es sesgado hacia abajo.

- Ejemplo Práctico: Se toma una muestra de 50 estudiantes y se mide su peso, obteniendo una media muestral de 65 kg. El estimador puntual de la media poblacional de peso de los estudiantes es 65 kg.

6.2.2 Estimación por Intervalo

- ¿Qué Es?

- Consiste en obtener un rango de valores dentro del cual se espera que se encuentre el parámetro poblacional con un nivel de confianza dado. Este rango se llama intervalo de confianza (IC).

- Conceptos Clave:

- Nivel de Confianza (1 - α): Probabilidad de que el intervalo de confianza contenga el parámetro poblacional verdadero. Comúnmente se usa 90%, 95% o 99% (es decir, $\alpha = 0.10, 0.05$ o 0.01).

- Márgen de Error (E): Mitad de la longitud del intervalo de confianza. Representa la precisión de la estimación.

- Forma General de un Intervalo de Confianza:

- Estimador puntual ± (Valor crítico * Error estándar del estimador)

- Interpretación Correcta del IC:

- "Estamos seguros en un (1 - α)*100% de que el parámetro poblacional verdadero se encuentra entre [límite inferior, límite superior]."

- No se debe decir: "La probabilidad de que el parámetro esté en el intervalo es (1 - α)*100%", ya que el parámetro es un valor constante (no aleatorio) y el intervalo es el que varía entre muestras.

- Relación Entre Nivel de Confianza, Tamaño de Muestra y Márgen de Error:

- A mayor nivel de confianza, mayor es el margen de error (menos precisión).
- A mayor tamaño de muestra, menor es el margen de error (mayor precisión).
- Se puede calcular el tamaño de muestra necesario para obtener un margen de error deseado (ver sección adicional).

6.2.3 Intervalo de Confianza para una Media

Se distinguen dos casos:

- Caso 1: Desviación Estándar Poblacional (σ) Conocida o Tamaño de Muestra Grande ($n \geq 30$) (TLC):

- Se usa la distribución normal estándar (Z).

- Fórmula del IC:

- $\bar{Y} \pm Z_{(\alpha/2)} * (\sigma / \sqrt{n})$
- Si σ es desconocida y $n \geq 30$, se reemplaza por la desviación estándar muestral (s):
- $\bar{Y} \pm Z_{(\alpha/2)} * (s / \sqrt{n})$
- Valor Crítico $Z_{(\alpha/2)}$: Es el valor de Z que deja una probabilidad de α/2 en cada cola de la distribución normal.

- Para 95% de confianza: $Z_{(0.025)} = 1.96$

- Para 90% de confianza: $Z_{(0.05)} = 1.645$

- Para 99% de confianza: $Z_{(0.005)} = 2.576$

- **Caso 2:** Desviación Estándar Poblacional (σ) Desconocida y Tamaño de Muestra Pequeño ($n \leq 30$):

- Se usa la distribución T-Student (t) con $v = n - 1$ grados de libertad.

- **Fórmula del IC:**

- $\bar{Y} \pm t_{(\alpha/2, v)} * (s / \sqrt{n})$

- Valor Crítico $t_{(\alpha/2, v)}$: Es el valor de t que deja una probabilidad de $\alpha/2$ en cada cola de la distribución t con v grados de libertad. Es mayor que $Z_{(\alpha/2)}$ para el mismo nivel de confianza, lo que hace que el intervalo sea más amplio (compensando la incertidumbre de no conocer σ).

- **Ejemplo Práctico (Caso 1):** Se mide la altura de 100 estudiantes, obteniendo $\bar{Y} = 165$ cm y $s = 8$ cm. Construir un IC del 95% para la media poblacional.

- $Z_{(0.025)} = 1.96$

- Error estándar = $8 / \sqrt{100} = 0.8$

- Márgen de error = $1.96 * 0.8 = 1.568$

- IC = $165 \pm 1.568 \rightarrow (163.432, 166.568)$

- **Interpretación:** Estamos seguros en un 95% de que la media de altura de todos los estudiantes está entre 163.43 cm y 166.57 cm.

- **Ejemplo Práctico (Caso 2):** Se toma una muestra de 15 estudiantes y se mide su IQ, obteniendo $\bar{Y} = 105$ y $s = 8$. Construir un IC del 95% para la media poblacional.

- $v = 15 - 1 = 14$, $t_{(0.025, 14)} = 2.145$

- Error estándar = $8 / \sqrt{15} \approx 2.066$

- Márgen de error = $2.145 * 2.066 \approx 4.43$

- IC = $105 \pm 4.43 \rightarrow (100.57, 109.43)$

6.2.4 Intervalo de Confianza para una Proporción

- **Condición:** La muestra debe ser suficientemente grande para que la aproximación normal sea válida: $n\hat{p} \geq 5$ y $n(1 - \hat{p}) \geq 5$ (o $n\hat{p} \geq 10$ y $n(1 - \hat{p}) \geq 10$).

- Fórmula del IC:

- $\hat{p} \pm Z_{(\alpha/2)} * \sqrt{[\hat{p}(1 - \hat{p}) / n]}$
- Valor Crítico $Z_{(\alpha/2)}$: Mismo que para la media (1.96 para 95% de confianza, etc.).

- Intervalo de Wilson (Mejorado): Para muestras pequeñas o proporciones cercanas a 0 o 1, se usa el intervalo de Wilson, que es más preciso:

- $[(\hat{p} + Z_{(\alpha/2)}^2/(2n) \pm Z_{(\alpha/2)}\sqrt{[\hat{p}(1 - \hat{p})/n + Z_{(\alpha/2)}^2/(4n^2)]}) / (1 + Z_{(\alpha/2)}^2/n)]$

- Ejemplo Práctico: En una encuesta a 200 personas, 120 afirman apoyar a un candidato. Construir un IC del 90% para la proporción poblacional de apoyo.

- $\hat{p} = 120 / 200 = 0.6$
- $n\hat{p} = 120 \geq 5$, $n(1-\hat{p}) = 80 \geq 5 \rightarrow$ válido.
- $Z_{(0.05)} = 1.645$
- Error estándar = $\sqrt{[0.6*0.4 / 200]} = \sqrt{[0.24 / 200]} \approx 0.0346$
- Márgen de error = $1.645 * 0.0346 \approx 0.057$
- IC = $0.6 \pm 0.057 \rightarrow (0.543, 0.657)$

- Interpretación: Estamos seguros en un 90% de que la proporción de personas que apoyan al candidato está entre 54.3% y 65.7%.

6.2.5 Tamaño de Muestra Necesario

- Para la Media:

- Se desea un margen de error E. Despejando n de la fórmula del margen de error:
- $n = (Z_{(\alpha/2)} * \sigma / E)^2$
- Si σ es desconocida, se puede usar una estimación previa de σ (de una muestra anterior o de la literatura).

- Para la Proporción:

- Se desea un margen de error E. Despejando n de la fórmula del margen de error:
- $n = p(1 - p) * (Z_{(\alpha/2)} / E)^2$
- Si p es desconocido, se usa p = 0.5 (lo que maximiza n, asegurando que el margen de error se cumpla independientemente de p).

- Ejemplo Práctico (Para la Proporción): Se quiere encuestar para estimar la proporción de personas que usan internet, con un margen de error de 3% y un nivel de confianza de 95%. ¿Cuál es el tamaño de muestra necesario?

- $Z_{(0.025)} = 1.96$, $E = 0.03$, $p = 0.5$
- $n = 0.5 \cdot 0.5 \cdot (1.96 / 0.03)^2 = 0.25 \cdot (65.333)^2 \approx 0.25 \cdot 4268.44 \approx 1067.11$
- Se redondea hacia arriba: n = 1068 personas.

6.3 Prueba de Hipótesis

- ¿Qué Es con Más Detalle?

- Es un proceso sistemático para evaluar si la evidencia de una muestra es suficiente para rechazar una afirmación sobre un parámetro poblacional. Se basa en el principio de falsación: se intenta refutar la hipótesis nula, ya que es más fácil demostrar que algo es falso que que es verdadero.

6.3.1 Errores Tipo I y II

- Error Tipo I (α):

- Definición: Rechazar la hipótesis nula (H_0) cuando en realidad es verdadera.

- Probabilidad: Se denota como α (nivel de significancia del test).

- **Ejemplo:** Un test de enfermedad da positivo a una persona que no la tiene (falso positivo).

- **Error Tipo II (β):**

- Definición: No rechazar la hipótesis nula (H_0) cuando en realidad es falsa.

- Probabilidad: Se denota como β.

- **Ejemplo:** Un test de enfermedad da negativo a una persona que sí la tiene (falso negativo).

- Potencia del Test (1 - β):

- Definición: Probabilidad de rechazar H_0 cuando es falsa (es decir, la probabilidad de tomar la decisión correcta cuando H_1 es verdadera).

- Importancia: Se desea maximizar la potencia del test (generalmente se busca $1 - \beta \geq 0.80$ o 0.90).

- Relación Entre α y β:

- Para un tamaño de muestra dado, reducir α aumenta β (y viceversa).

- La única forma de reducir ambos errores al mismo tiempo es aumentar el tamaño de la muestra.

6.3.2 Pasos para Realizar una Prueba de Hipótesis

1. Establecer las Hipótesis:

- Hipótesis Nula (H_0): Afirmación que se intenta refutar. Es una afirmación de "no efecto" o "no diferencia". Siempre contiene un signo de igualdad (=, ≤, ≥).

- Hipótesis Alternativa (H_1 o H_a): Afirmación que se acepta si se rechaza H_0. Es una afirmación de "efecto" o "diferencia". Contiene un signo de desigualdad (≠, >, <).

- Tipos de Pruebas:

- Bilateral (o de dos colas): H_1: $\theta \neq \theta_0$ (busca cualquier diferencia, positiva o negativa).

- **Unilateral Derecha (o de cola derecha):** H_1: $\theta > \theta_0$ (busca una diferencia positiva).

- Unilateral Izquierda (o de cola izquierda): H_1: $\theta < \theta_0$ (busca una diferencia negativa).

2. Elegir el Nivel de Significancia (α):

- Probabilidad máxima de cometer un error Tipo I. Comúnmente $\alpha = 0.05$ (5%), pero se puede usar $\alpha = 0.10$ o $\alpha = 0.01$ según la gravedad del error Tipo I.

3. Seleccionar el Estadístico de Prueba y su Distribución:

- El estadístico de prueba es una función de los datos de la muestra que se compara con una distribución teórica (normal, t, chi cuadrada, F, etc.) bajo la suposición de que H_0 es verdadera.

- Ejemplos:

- Para la media (σ conocida o $n \geq 30$): Estadístico Z.
- Para la media (σ desconocida y $n \leq 30$): Estadístico t.
- Para la proporción: Estadístico Z.

4. Determinar la Región de Rechazo o Calcular el Valor p:

- Región de Rechazo: Conjunto de valores del estadístico de prueba que llevan a rechazar H_0. Se define por los valores críticos.

6.3.3 Prueba de Hipótesis para una Media

Se distinguen los mismos casos que en el intervalo de confianza:

- Caso 1: σ Conocida o $n \geq 30$ (TLC) — Estadístico Z

- Hipótesis Comunes:

- Bilateral: H_0: $\mu = \mu_0$; H_1: $\mu \neq \mu_0$
- Unilateral Derecha: H_0: $\mu \leq \mu_0$; H_1: $\mu > \mu_0$
- Unilateral Izquierda: H_0: $\mu \geq \mu_0$; H_1: $\mu < \mu_0$

- Estadístico de Prueba:

- $Z = (\bar{Y} - \mu_0) / (\sigma / \sqrt{n})$
- Si σ es desconocida y $n \geq 30$: $Z = (\bar{Y} - \mu_0) / (s / \sqrt{n})$

- **Ejemplo Práctico:** Se afirma que la media de ingresos de los trabajadores de una empresa es $50,000 al año. Se toma una muestra de 100 trabajadores, obteniendo $\bar{Y}$ = $52,000 y s = $8,000. Probar si la media es diferente de $50,000 con $\alpha = 0.05$.

- H_0: $\mu = 50{,}000$; H_1: $\mu \neq 50{,}000$ (bilateral)
- $Z = (52{,}000 - 50{,}000) / (8{,}000 / \sqrt{100}) = 2{,}000 / 800 = 2.5$
- Valores críticos: ± 1.96
- Valor p: $2 * P(Z \geq 2.5) = 2 * (1 - 0.9938) = 0.0124$
- Decisión: $2.5 > 1.96$ o $0.0124 < 0.05$ → Rechazar H_0.

- Conclusión: Hay evidencia estadísticamente significativa para afirmar que la media de ingresos es diferente de $50,000.

- Caso 2: σ Desconocida y $n \leq 30$ — Estadístico t

- Hipótesis Comunes: Mismas que en el caso 1.

- Estadístico de Prueba:

- $t = (\bar{Y} - \mu_0) / (s / \sqrt{n})$
- Grados de libertad: $v = n - 1$

- Ejemplo Práctico: Se afirma que la media de IQ de los estudiantes de un colegio es 110. Se toma una muestra de 15 estudiantes, obteniendo $\bar{Y} = 105$ y $s = 8$. Probar si la media es menor de 110 con $\alpha = 0.05$.

- H_0: $\mu \geq 110$; H_1: $\mu < 110$ (unilateral izquierda)
- $t = (105 - 110) / (8 / \sqrt{15}) = -5 / 2.066 \approx -2.42$
- Grados de libertad = 14; valor crítico = -1.761

- Valor p: $P(t \le -2.42) \approx 0.014$

- **Decisión:** $-2.42 < -1.761$ o $0.014 < 0.05$ → Rechazar H_0.

- Conclusión: Hay evidencia estadísticamente significativa para afirmar que la media de IQ es menor de 110.

6.3.4 Prueba de Hipótesis para una Proporción

- Condición: Muestra suficientemente grande: $np_0 \ge 5$ y $n(1 - p_0) \ge 5$ (donde p_0 es la proporción de H_0).

- Hipótesis Comunes:

- Bilateral: H_0: $p = p_0$; H_1: $p \ne p_0$

- Unilateral Derecha: H_0: $p \le p_0$; H_1: $p > p_0$

- Unilateral Izquierda: H_0: $p \ge p_0$; H_1: $p < p_0$

- Estadístico de Prueba:

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

- Ejemplo Práctico: Se afirma que el 60% de los votantes apoyan a un candidato. Se encuesta a 200 votantes, y 110 lo apoyan. Probar si la proporción es menor de 60% con $\alpha = 0.05$.

- H_0: $p \ge 0.60$; H_1: $p < 0.60$ (unilateral izquierda)

- $\hat{p} = 110 / 200 = 0.55$

- $np_0 = 200*0.60 = 120 \ge 5$; $n(1-p_0) = 80 \ge 5$ → válido.

- $Z = (0.55 - 0.60) / \sqrt{[0.60*0.40 / 200]} = -0.05 / 0.0346 \approx -1.445$

- Valor crítico = -1.645

- Valor p: $P(Z \le -1.445) \approx 0.074$

- **Decisión:** $-1.445 > -1.645$ o $0.074 \ge 0.05$ → No rechazar H_0.

- Conclusión: No hay evidencia estadísticamente significativa para afirmar que la proporción de apoyo es menor de 60%.

6.3.5 Pruebas de Hipótesis Adicionales

- Prueba de Hipótesis para la Varianza de una Población:

- Usa la distribución Chi Cuadrada (χ^2).

- Estadístico de prueba: $\chi^2 = (n - 1)s^2 / \sigma_0^2$ (con $v = n - 1$ grados de libertad).

- Prueba de Hipótesis para la Diferencia de Dos Medias:

- Casos: Varianzas poblacionales conocidas o desconocidas, muestras independientes o apareadas.

- Usa la distribución Z (si varianzas conocidas o n grandes) o la distribución t (si varianzas desconocidas y n pequeñas).

- Prueba de Hipótesis para la Diferencia de Dos Proporciones:

- Usa la distribución Z.

- Estadístico de prueba: $Z = (\hat{p}_1 - \hat{p}_2) / \sqrt{[\bar{p}(1 - \bar{p})(1/n_1 + 1/n_2)]}$, donde $\bar{p}$ es la proporción combinada.

6.3.6 Potencia del Test y Tamaño de Muestra Necesario para Pruebas de Hipótesis

- Potencia del Test:

- Depende de:

1. Nivel de significancia α (a mayor α, mayor potencia).
2. Tamaño del efecto (a mayor diferencia entre el parámetro verdadero y el valor de H_0, mayor potencia).
3. Tamaño de muestra n (a mayor n, mayor potencia).

- Se puede calcular la potencia para un tamaño de muestra dado o el tamaño de muestra necesario para obtener una potencia deseada.

- Tamaño de Muestra Necesario para la Media:

- Para una prueba unilateral: $n = (Z_{\alpha} + Z_{\beta})^2 * \sigma^2 / \Delta^2$

- Para una prueba bilateral: $n = (Z_{\alpha/2} + Z_{\beta})^2 * \sigma^2 / \Delta^2$

- Donde Δ es el tamaño de efecto deseado (diferencia entre μ verdadero y μ_0).

- Tamaño de Muestra Necesario para la Proporción:

- Para una prueba unilateral: $n = (Z_{\alpha}\sqrt{[p_0(1-p_0)]} + Z_{\beta}\sqrt{[p_1(1-p_1)]})^2 / (p_1 - p_0)^2$

- Para una prueba bilateral: $n = (Z_{\alpha/2}\sqrt{[p_0(1-p_0)]} + Z_{\beta}\sqrt{[p_1(1-p_1)]})^2 / (p_1 - p_0)^2$

- Donde p_1 es la proporción verdadera que se quiere detectar.