UNIVERSIDAD AUTÓNOMA DE SAN LUIS POTOSÍ

Facultad de Ciencias

Apuntes, Ejercicios y Soluciones Estadística Aplicada

Material Didáctico

Autor:
Paul Hernández Herrera

May 19, 2026

-

Contents

INTRODUCCION

Este material didáctico ha sido desarrollado como una guía de apoyo para los estudiantes que cursan la materia de Estadística Aplicada. Contiene un resumen breve de cada unidad, describiendo los temas, definiciones y fórmulas más importantes de cada sección. Está diseñado para facilitar el aprendizaje y la resolución de problemas en esta materia, impartida en la Facultad de Ciencias para diversas carreras, como Licenciatura en Biología, Licenciatura en Matemáticas Aplicadas, Ingeniería Biomédica, LAEC e Ingeniería Electrónica.

El documento está dividido en siete unidades que corresponden al programa del curso de Estadística Aplicada. Cada unidad incluye:

- Una descripción breve y clara de los temas.
- Las definiciones básicas y las fórmulas más representativas.
- Ejercicios relacionados diseñados para que el alumno los resuelva de manera autónoma, apoyándose en el material y la teoría impartida en clase.

Además, se proporciona la solución detallada de cada ejercicio, explicando paso a paso el procedimiento necesario. Este enfoque permite al estudiante desarrollar una base sólida para resolver los ejercicios propuestos y comprender el proceso de resolución de problemas.

El contenido de las unidades es el siguiente:

1. Conceptos básicos de estadística
2. Medidas descriptivas y técnicas de visualización.
3. Teoría para construir intervalos de confianza.
4. Inferencia estadística con una muestra.
5. Inferencia estadística con dos muestras.
6. Análisis de varianza.
7. Regresión lineal.

Este material está diseñado para ser una herramienta complementaria en el aprendizaje de la estadística aplicada y fortalecer las habilidades del estudiante en esta área.

1 UNIDAD 1: CONCEPTOS BASICOS DE ESTADISTICA

OBJETIVOS: Al finalizar esta unidad, el estudiante será capaz de:

1. Identificar y comprender los conceptos fundamentales de la estadística, incluyendo población, muestra, variables y tipos de datos.
2. Aplicar correctamente los pasos para construir una tabla de distribución de frecuencias a partir de un conjunto de datos.
3. Representar datos estadísticos utilizando herramientas gráficas como histogramas, gráficas de pastel y otras visualizaciones adecuadas.
4. Resolver problemas prácticos que involucren la organización, representación y análisis básico de datos utilizando las técnicas aprendidas.
5. Integrar los conceptos y herramientas básicos de estadística para interpretar datos y comunicar resultados de manera clara y efectiva.

1.1 Definiciones básicas:

Datos: Son hechos o información recopilados de diversas fuentes, como mediciones, respuestas a preguntas, detalles sobre objetos u observaciones, y que pueden ser utilizados para obtener conocimiento o tomar decisiones.

Estadística: La estadística es un conjunto de técnicas y métodos utilizados para diseñar experimentos, recopilar información, organizar datos, resumirlos, presentarlos, analizarlos, interpretarlos y, finalmente, llegar a conclusiones basadas en la información recopilada. En esencia, la estadística ayuda a comprender y sacar sentido de los datos, permitiendo tomar decisiones informadas y obtener conocimientos significativos a partir de la información recopilada.

Población: Conjunto completo que incluye a todos los elementos que se desean examinar o estudiar.

Censo: Proceso mediante el cual se recopila información de todos y cada uno de los miembros de una población.

Muestra: Grupo o subconjunto representativo de elementos tomados de una población. En lugar de realizar un censo, se selecciona un grupo más pequeño pero representativo para realizar análisis y conclusiones que se aplicarán a toda la población.

1.2 Tipos de datos:

Parámetro: es una medida numérica que describe una característica o propiedad específica de una población.

Existen dos tipos de datos. **Datos cuantitativos:** son valores numéricos que se utilizan para medir cantidades o magnitudes. Ejemplos de datos cuantitativos incluyen la edad de una persona, la temperatura en grados Celsius o el número de productos vendidos. **Datos cualitativos o categóricos:** En lugar de utilizar números, se utilizan etiquetas o nombres para describir las categorías. Ejemplos de datos cualitativos incluyen el color de los ojos (por ejemplo, azul, verde o marrón), la marca de un automóvil (por ejemplo, Toyota, Ford o Honda) o el estado civil de una persona (por ejemplo, soltero, casado o divorciado).

1.2.1 Datos Nominales:

Los datos nominales son aquellos que consisten exclusivamente en nombres, etiquetas o categorías que no pueden ordenarse de acuerdo a un esquema de valor numérico. Ejemplos de datos nominales incluyen colores (como rojo, verde o azul) o tipos de frutas (como manzana, plátano o naranja).

1.2.2 Datos ordinales:

Los datos ordinales son datos que tienen un orden predefinido, pero las diferencias entre los valores no tienen un significado numérico específico. En otras palabras, puedes establecer un orden relativo, pero no puedes decir cuánto mayor es un valor en comparación con otro. Ejemplos de datos ordinales son las categorías como "alto", "mediano" y "bajo", donde se establece un orden relativo, pero no se puede medir la diferencia exacta entre ellos.

1.2.3 Datos de intervalo:

Los datos de intervalo son similares a los datos ordinales, pero tienen la característica adicional de que las diferencias entre sus elementos tienen un significado numérico. Sin embargo, estos datos no tienen un punto de partida natural en cero, lo que significa que cero no indica la ausencia de la cantidad (no se puede decir que un valor es el "doble" de otro). Ejemplos de datos de intervalo incluyen la temperatura en grados Celsius o años calendario, donde las diferencias entre los valores tienen un significado, pero cero grados Celsius o el año cero no representan la ausencia de temperatura o tiempo.

1.2.4 Datos de razón:

Los datos de razón son similares a los datos de intervalo, pero tienen la propiedad adicional de tener un punto de partida o cero inherente que indica la ausencia completa de la cantidad medida. Por lo tanto, cero en datos de razón significa que no hay cantidad presente. Ejemplos de datos de razón incluyen el peso en kilogramos, la velocidad en kilómetros por hora, la altura en metros, etc. En estos casos, cero kilogramos, cero kilómetros por hora o cero metros indican la ausencia total de la cantidad medida.

1.3 Diseño de Experimentos

En el diseño de experimentos, es fundamental asegurarse que los datos muestrales se recojan adecuadamente para evitar conclusiones estadísticas erróneas. La aleatoriedad desempeña un papel crucial en la determinación de los datos que se deben recopilar.

Los datos se pueden obtener de dos formas diferentes: a través de estudios observacionales y experimentos.

Estudio Observacional: En este enfoque, observamos y medimos características específicas, pero no intentamos manipular a los sujetos que están siendo estudiados.

Experimento: En un experimento, aplicamos un tratamiento específico y observamos su efecto en los sujetos de estudio.

1.3.1 Estudios Observacionales

Por lo general, los estudios observacionales se dividen en tres tipos:

1. **Estudio Transversal:** En este tipo de estudio, se observan, miden y recopilan los datos en un solo momento. Por ejemplo, encuestas sobre preferencias alimenticias.

2. **Estudio Retrospectivo:** Estos estudios se basan en datos del pasado. Por ejemplo, analizar registros médicos históricos para comprender la prevalencia de una enfermedad.

3. **Estudio Prospectivo, Longitudinal o Cohorte:** En este caso, los datos se recopilan en el futuro y se obtienen de grupos (llamados cohortes) que comparten factores comunes. Por ejemplo, seguir a un grupo de personas a lo largo del tiempo para estudiar cómo ciertos factores de salud afectan su calidad de vida en el futuro.

1.4 Experimentos:

1.4.1 Estudio ciego:

Placebo: El placebo es una sustancia que no tiene propiedades curativas, pero puede tener un efecto terapéutico si el paciente cree que está tomando un medicamento efectivo.

Efecto placebo: Este efecto ocurre cuando un paciente no tratado informa mejorías en sus síntomas.

Estudio ciego: Es un tipo de estudio diseñado para reducir el efecto placebo. En este enfoque, los pacientes no saben si están recibiendo un placebo o un tratamiento real. El estudio ciego permite determinar si el tratamiento es significativamente más efectivo que el efecto placebo mediante comparaciones estadísticas entre los grupos que reciben tratamiento y los que reciben placebo. Por ejemplo: Imagina un estudio en el que se prueba la eficacia de un nuevo suplemento vitamínico. Los participantes se dividen en dos grupos: uno recibe el suplemento real y el otro un placebo. Ninguno de los participantes sabe si está tomando el suplemento real o el placebo, ya que las píldoras se ven y saben idénticas. El estudio se realiza de esta manera para eliminar cualquier influencia de las expectativas de los participantes sobre los resultados.

1.4.2 Bloques:

Cuando realizas un experimento para probar diferentes tratamientos, agrupa a los sujetos en bloques o grupos que tienen características similares.

Diseño Experimental Completamente Aleatorizado: En este enfoque, los sujetos se asignan a diferentes bloques de manera completamente aleatoria.

Diseño rigurosamente controlado: En este caso, los sujetos se seleccionan cuidadosamente para que los miembros de cada bloque sean similares en las características que son importantes para el experimento.

1.4.3 Réplica y tamaño de muestra:

Además de controlar los efectos de las variables, el tamaño de la muestra es otro elemento crucial. Las muestras deben ser lo suficientemente grandes para evitar que la variabilidad aleatoria, que es común en muestras muy pequeñas, enmascare los efectos reales de los diferentes tratamientos. Cuando repetimos un experimento, esto se denomina "réplica", y es efectivo cuando contamos con suficientes participantes para identificar las diferencias que surgen entre los diferentes tratamientos. En otro contexto, la "réplica" se refiere a la repetición o duplicación de un experimento para confirmar o verificar los resultados. Con la réplica, aumentamos la probabilidad de detectar efectos significativos del tratamiento en muestras grandes.

Seleccionar un tamaño de muestra lo suficientemente grande para poder identificar con precisión los efectos reales entre las diversas opciones. Además, es importante obtener la muestra de manera apropiada, preferiblemente utilizando un método que se base en la aleatoriedad

1.4.4 Estrategias de muestreo

Si los datos muestrales no se reúnen de forma adecuada, resultarían tan inútiles que ninguna cantidad de tortura estadística podrá salvarlos.

Muestra aleatoria: se compone de miembros de una población que son seleccionados de tal manera que cada individuo tiene igual probabilidad de ser elegido. Por ejemplo: Seleccionar al azar 100 nombres de una guía telefónica para realizar una encuesta sobre preferencias de restaurantes en una ciudad

Muestra Aleatoria Simple: Es un conjunto de n sujetos seleccionados de una población de manera que cada posible combinación de n sujetos tenga la misma probabilidad de ser elegida.

1.5 Técnicas de conteo:

Cuando tienes un conjunto grande de datos, a menudo es útil organizarlos y resumirlos de una manera más manejable. Una forma común de hacerlo es construir una tabla que enumere todos los valores posibles de los datos, ya sea individualmente o agrupados en intervalos, junto con el número de veces que cada valor aparece en los datos. Esta tabla se llama "distribución de frecuencias".

Distribución de frecuencias: Esta tabla muestra qué valores son más comunes y cuántas veces se repiten en tu conjunto de datos, lo que facilita la comprensión de las tendencias y patrones en los datos.

1.5.1 Construcción de distribución de frecuencia:

1. **Decide cuántas clases o grupos (n) vas a crear:** Para resumir tus datos, primero debes decidir en cuántas clases o grupos (n) quieres dividir tu conjunto de datos. Estos grupos te permitirán organizar los datos de manera más comprensible.

2. **Calcula el tamaño de cada intervalo (ΔI):** El tamaño de cada intervalo se obtiene dividiendo la diferencia entre el valor máximo ($\max\_val$) y el valor mínimo ($\min\_val$) de tus datos por el número de clases (n). Este cálculo define el ancho de cada intervalo:

$$\Delta I = \frac{\max\_val - \min\_val}{n}.$$

3. **Define los límites de los intervalos:** Los límites de los intervalos se calculan comenzando con el valor mínimo ($\min\_val$) y sumando múltiplos del tamaño del intervalo (ΔI) para obtener cada límite:

$$\begin{aligned} l_0 &= \min\_val + 0 * \Delta I \\ l_1 &= \min\_val + 1 * \Delta I \\ l_2 &= \min\_val + 2 * \Delta I \\ &\vdots \\ l_k &= \min\_val + k * \Delta I \\ &\vdots \\ l_n &= \min\_val + n * \Delta I \end{aligned} \tag{1}$$

   Aquí, l_0 es el límite inferior del primer intervalo, l_1 es el límite superior del primer intervalo e inferior del segundo, y así sucesivamente.

4. **Crea los intervalos:** Utiliza los límites definidos anteriormente para crear los intervalos:

$$\begin{aligned} I_1 &= [l_0, l_1) \\ I_2 &= [l_1, l_2) \\ &\vdots \\ I_k &= [l_{k-1}, l_k) \\ &\vdots \\ I_n &= [l_{n-1}, l_n] \end{aligned} \tag{2}$$

   Cada intervalo I_k abarca un rango de valores desde un límite inferior hasta, pero sin incluir, el límite superior, excepto el último intervalo que incluye ambos límites.

5. **Cuenta cuántos datos caen dentro de cada intervalo:** Finalmente, revisa tus datos y cuenta cuántos valores caen dentro de cada intervalo. Registra estas frecuencias en una tabla.

6. **Construye la tabla de distribución de frecuencias:** donde las columnas son,
    - Intervalo: Lista los intervalos que has definido.
    - Frecuencia: Muestra cuántos datos caen en cada intervalo.

Esta técnica te permite organizar tus datos en grupos o intervalos, lo que facilita la comprensión de cómo se distribuyen los datos y cuántas veces aparecen en cada grupo. Esto es útil para analizar patrones y tendencias en tu conjunto de datos.

1.5.2 Distribución de Frecuencias Relativas

La distribución de frecuencias relativas es una variante de la distribución de frecuencias que te muestra cómo se distribuyen tus datos en intervalos, pero en lugar de contar el número de veces que aparece cada valor, calcula la proporción o el porcentaje de datos que caen en cada intervalo.

Para obtener la frecuencia relativa de un intervalo, simplemente divides la frecuencia de ese intervalo (es decir, cuántos datos están en ese grupo) entre el número total de datos en tu conjunto. La fórmula para calcular la frecuencia relativa es:

$$\text{Frecuencia Relativa} = \frac{\text{Frecuencia del Intervalo}}{\text{Total de Datos}}$$

La frecuencia relativa te indica el porcentaje de datos que se encuentra en cada intervalo y puedes interpretarla como la probabilidad de que un dato seleccionado al azar caiga en un intervalo específico. Esta medida es útil para comprender cómo se distribuyen tus datos de manera más proporcional en diferentes grupos.

1.6 Visualización de datos:

1.6.1 Histograma:

Un histograma Un histograma es una herramienta visual que permite representar la distribución de un conjunto de datos de manera gráfica. Se utiliza para mostrar cómo se agrupan o dispersan los valores de los datos en diferentes rangos o intervalos, llamados clases.

- **Eje horizontal (X):** En este eje se colocan los diferentes rangos o clases en los que se han dividido los datos.
- **Eje vertical (Y):** Aquí se representa la frecuencia, es decir, el número de veces que los datos caen dentro de cada clase.

Las barras en el histograma se dibujan sin espacios entre ellas, ya que los intervalos o clases son contiguos y representan un rango continuo de datos. La altura de cada barra refleja cuántas observaciones caen dentro del rango específico que representa la barra. Esto permite ver rápidamente la distribución de los datos, identificando patrones como concentraciones de valores, simetrías, asimetrías, o la presencia de valores atípicos.

1.6.2 Histograma de frecuencias relativas:

El histograma de frecuencias relativas es similar al histograma tradicional, pero en lugar de mostrar la frecuencia absoluta (el número total de veces que un valor o un rango de valores aparece en el conjunto de datos), muestra la frecuencia relativa o proporción.

Este tipo de histograma es útil cuando se quiere comparar la distribución de diferentes conjuntos de datos que no tienen el mismo tamaño, ya que permite ver la proporción de datos en cada intervalo en lugar de los valores absolutos.

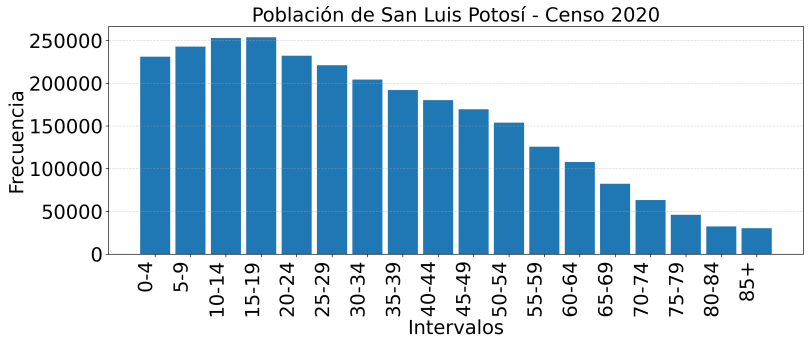


Figure 1: Ejemplo de histograma de la población de San Luis Potosí, censo 2020.

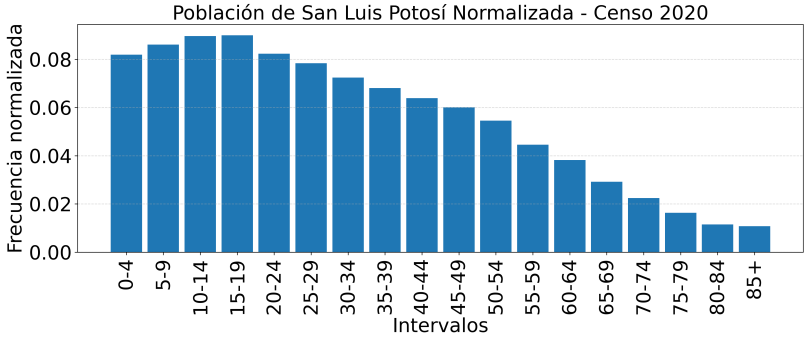


Figure 2: Ejemplo de histograma de frecuencias relativas de la población de San Luis Potosí, censo 2020.

1.6.3 Gráfica Circular:

Una gráfica circular, también conocida como gráfica de pastel, es una forma de visualizar datos cualitativos o categóricos. Cada categoría se representa como una porción o "rebanada" del círculo.

- **Categorías:** Cada rebanada del pastel representa una categoría diferente de los datos.
- **Tamaño de la rebanada:** El tamaño de cada rebanada es proporcional al porcentaje de datos que pertenece a esa categoría específica.

Para crear una gráfica circular:

1. Se comienza calculando las frecuencias relativas de cada categoría, lo que representa el porcentaje de datos en cada categoría.
2. Luego, para convertir estos porcentajes en ángulos que se utilizarán para dibujar las rebanadas del pastel, se multiplica cada frecuencia relativa por 360. Esto se debe a que un círculo tiene 360 grados, y multiplicar un porcentaje (frecuencia relativa) por 360 convierte la frecuencia relativa en la cantidad de grados correspondiente en el círculo.

Por ejemplo, si una categoría tiene una frecuencia relativa de 0.25 (porcentaje), la rebanada correspondiente tendrá un ángulo de $0.25 \times 360| = 90$ grados. Esto hace que la visualización sea intuitiva, permitiendo identificar rápidamente la proporción que cada categoría ocupa dentro del conjunto total de datos.

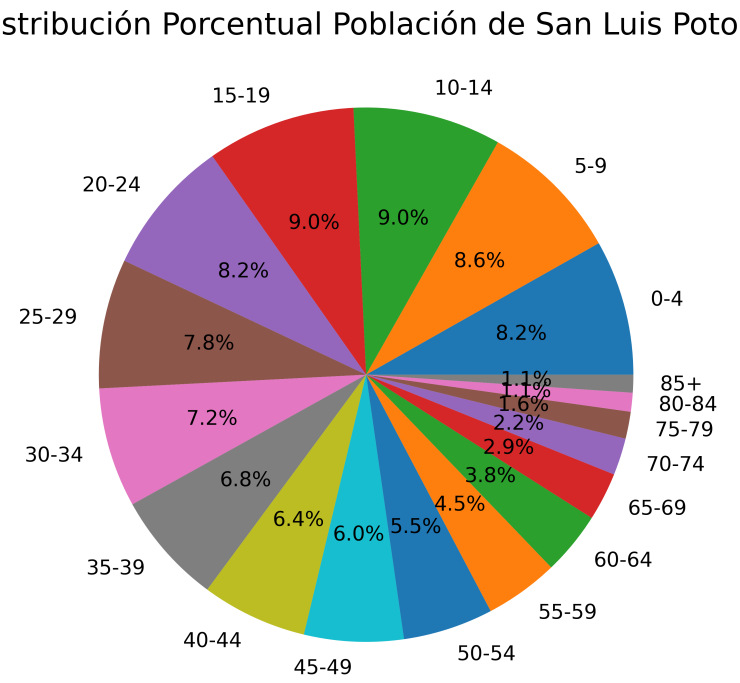


Figure 3: Ejemplo de gráfica circular de la población de San Luis Potosí, censo 2020.

1.7 Visualización de dos parámetros:

1.7.1 Diagramas de dispersión

Un diagrama de dispersión es una herramienta gráfica utilizada para analizar la relación entre dos variables diferentes, las cuales se representan mediante pares de valores (x, y). Este tipo de gráfico es especialmente útil para observar cómo una variable puede influir en otra, detectando patrones, tendencias, o correlaciones entre ellas.

1. **Eje Horizontal (X):** Representa una de las variables de tu conjunto de datos. Cada valor de esta variable se coloca en su posición correspondiente a lo largo de este eje.

2. **Eje Vertical (Y):** Representa la otra variable del conjunto de datos. Los valores de esta variable se colocan en su posición correspondiente a lo largo de este eje.

Cada par de valores (x, y) se representa como un punto en el gráfico. Por ejemplo, si en un experimento se mide la temperatura (variable X) y la presión (variable Y), un par de valores como $(30°C, 101.3kPa)$ se representaría como un punto en el gráfico.

1.7.2 Interpretación del Patrón de Puntos

- **Correlación Positiva:** Si los puntos tienden a agruparse en una línea que sube hacia la derecha, esto indica que a medida que la variable X aumenta, la variable Y también tiende a aumentar. Esto sugiere una correlación positiva.

- **Correlación Negativa:** Si los puntos se agrupan en una línea que desciende hacia la derecha, esto indica que a medida que la variable X aumenta, la variable Y tiende a disminuir. Esto sugiere una correlación negativa.

- **Sin Correlación:** Si los puntos están dispersos de manera aleatoria sin seguir un patrón claro, esto sugiere que no hay una relación lineal evidente entre las dos variables.

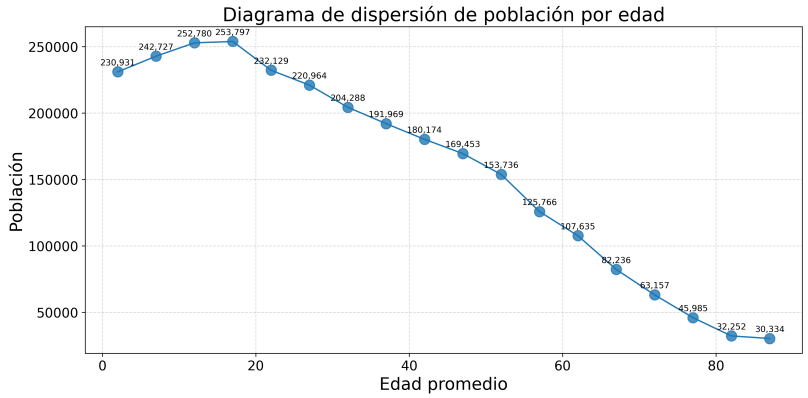


Figure 4: Ejemplo de diagrama de dispersión de la población de San Luis Potosí, censo 2020.

El diagrama de dispersión es una herramienta poderosa para identificar visualmente si existe algún tipo de relación entre las variables que estás analizando. Sin embargo, es importante tener en cuenta que la correlación observada no implica causalidad, es decir, que un cambio en una variable necesariamente cause un cambio en la otra.

1.8 EJERCICIOS

EJERCICIO 1: Para cada conjunto de datos, realice lo siguiente:

1. Identifique el tipo de dato.
2. Calcule las frecuencias, frecuencias relativas y construya un histograma para $n = 5$.
3. Elabore una gráfica circular.
4. ¿Qué conclusiones puede extraer de las frecuencias y las gráficas obtenidas?

- **Datos Reales 1**: Estatura de los estudiantes de la clase de Estadística Aplicada de la Facultad de Ciencias, UASLP.
- **Datos Reales 2**: Calificación final de los estudiantes en un curso determinado.
- **Datos Ficticios 3**: Encuesta sobre la satisfacción con un nuevo producto. Los resultados son los siguientes:
  Malo, Bueno, Bueno, Neutral, Neutral, Bueno, Malo, Bueno, Bueno, Bueno, Malo, Malo, Neutral, Bueno, Bueno, Bueno, Bueno, Bueno, Neutral

EJERCICIO 2: Lea el documento $- - - - - - - - - -$ y responda lo siguiente:

1. ¿Cuál es la población de estudio?
2. ¿Qué tipo de datos se obtuvieron (nominales, ordinales, etc.)?
3. ¿Qué técnicas se utilizaron para reportar las estadísticas?

1.9 SOLUCIONES

1.9.1 Ejercicio 1

SOLUCION Datos 1.

1. Se realizó un censo de los 15 estudiantes de la clase de Estadística Aplicada, midiendo la estatura en metros con una cinta métrica. Los resultados fueron los siguientes:

   Estaturas = [1.79, 1.69, 1.75, 1.70, 1.67, 1.50, 1.75, 1.80, 1.82, 1.70, 1.84, 1.73, 1.73, 1.60, 1.68]

2. **SOLUCIÓN identificación del tipo de dato:** Las estaturas son datos de tipo "razón", ya que son valores numéricos continuos, se pueden ordenar y comparar, y tienen un punto de origen significativo (el valor cero indica ausencia de estatura).

3. **SOLUCIÓN: Cálculo de frecuencias para** $n = 5$

   (a) **Paso 1: Ordenar datos de menor a mayor:** Primero, ordenamos los datos de estatura para facilitar la construcción de los intervalos:
   [1.50, 1.60, 1.67, 1.68, 1.69, 1.7, 1.7, 1.73, 1.73, 1.75, 1.75, 1.79, 1.8, 1.82, 1.84]

   (b) **Paso 2: Calcular la anchura del intervalo:** Para determinar la anchura del intervalo, utilizamos la fórmula:

   $$\Delta I = \frac{\max\_val - \min\_val}{n}$$

   Donde:
   $\max\_val = 1.84$ (el valor máximo de las estaturas)
   $\min\_val = 1.50$ (el valor mínimo de las estaturas)
   $n = 5$ (el número de intervalos deseado)
   Entonces, la anchura del intervalo es:

   $$\Delta I = \frac{1.84 - 1.50}{5} = \frac{0.34}{5} = 0.068$$

   (c) **Paso 3: Definir los límites de los intervalos**
   Los límites de los intervalos se determinan sumando múltiplos de ΔI al valor mínimo:

   $$\begin{aligned} l_0 &= \min\_val + 0 * \Delta I = 1.50 + 0 * 0.068 = 1.50 \\ l_1 &= \min\_val + 1 * \Delta I = 1.50 + 1 * 0.068 = 1.568 \\ l_2 &= \min\_val + 2 * \Delta I = 1.50 + 2 * 0.068 = 1.636 \\ l_3 &= \min\_val + 3 * \Delta I = 1.50 + 3 * 0.068 = 1.704 \\ l_4 &= \min\_val + 4 * \Delta I = 1.50 + 4 * 0.068 = 1.772 \\ l_5 &= \min\_val + 5 * \Delta I = 1.50 + 5 * 0.068 = 1.840 \end{aligned} \tag{3}$$

   (d) **Paso 4: Construir los intervalos**
   Con los límites definidos, los intervalos se construyen de la siguiente manera:

   $$\begin{aligned} I_1 &= [l_0, l_1) = [1.50, 1.568) \\ I_2 &= [l_1, l_2) = [1.568, 1.636) \\ I_3 &= [l_2, l_3) = [1.636, 1.704) \\ I_4 &= [l_3, l_4) = [1.704, 1.772) \\ I_5 &= [l_4, l_5] = [1.772, 1.840] \end{aligned} \tag{4}$$

   (e) **Paso 5: Obtener la frecuencia de datos que caen en cada intervalo:** Ahora, contamos cuántos datos caen dentro de cada intervalo:

Intervalo	Frecuencia
$[1.500, 1.568)$	1
$[1.568, 1.636)$	1
$[1.636, 1.704)$	5
$[1.704, 1.772)$	4
$[1.772, 1.840)$	4

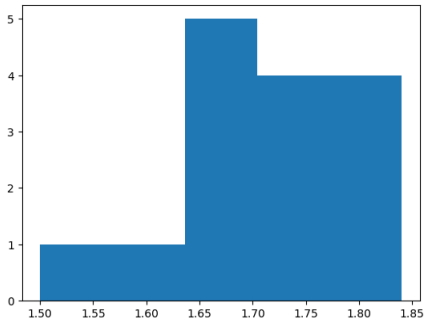
Por ejemplo, en el intervalo $[1.636, 1.704)$, buscamos todos los datos (ver **Paso 1**) que son mayores o iguales a 1.636 pero menores a 1.704. Los datos que cumplen esta condición son: $1.67, 1.68, 1.69, 1.70, 1.70$. Por lo tanto, la frecuencia es 5, ya que hay 5 datos en este intervalo.

4. **SOLUCION: FRECUENCIAS RELATIVAS**. Para calcular las frecuencias relativas, se debe dividir la frecuencia (tabla anterior) de cada intervalo por el número total de datos, que en este caso es 15.

Intervalo	Frecuencia Relativa
$[1.500, 1.568)$	$1/15 = 0.066$
$[1.568, 1.636)$	$1/15 = 0.066$
$[1.636, 1.704)$	$5/15 = 0.333$
$[1.704, 1.772)$	$4/15 = 0.266$
$[1.772, 1.840)$	$4/15 = 0.266$

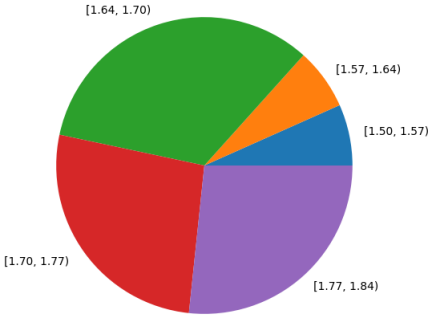
5. **SOLUCION: HISTOGRAMA**

   Gráfica de barras de la frecuencia de los datos. La altura de cada barra corresponde a la frecuencia del intervalo.



6. **SOLUCION: GRAFICA CIRCULAR.** Calcular los grados que le corresponden los cuales se obtienen multiplicando las frecuencias relativas x 360

Intervalo	Frecuencia Relativa	Grados Rebanada
$[1.5, 1.568)$	0.066	$0.066 \times 360 = 24$
$[1.568, 1.636)$	0.066	$0.066 \times 360 = 24$
$[1.636, 1.704)$	0.333	$0.333 \times 360 = 120$
$[1.704, 1.772)$	0.266	$0.266 \times 360 = 96$
$[1.772, 1.840)$	0.266	$0.266 \times 360 = 96$



7. **SOLUCION: ¿Qué puedes concluir de las frecuencias y las gráficas?** Al analizar las frecuencias y las gráficas obtenidas, podemos extraer las siguientes conclusiones sobre la distribución de estaturas entre los estudiantes:

   (a) Pocos estudiantes tienen una estatura menor a 1.636 metros:
   Interpretación: Según los datos y la gráfica de frecuencias, se observa que solo un pequeño número de estudiantes tienen una estatura por debajo de 1.636 metros. Esto se refleja en que los intervalos correspondientes a estas estaturas tienen barras más bajas en el histograma, lo que indica que hay menos datos en estos rangos de altura.

   (b) La mayoría de los estudiantes tienen una estatura mayor a 1.636 metros:
   Interpretación: La mayor parte de los estudiantes se encuentra en los intervalos de estatura superiores a 1.636 metros. Esto se evidencia en que las barras del histograma para los intervalos por encima de 1.636 metros son más altas, mostrando que un número significativo de estudiantes tiene estaturas en estos rangos.

   (c) Para estaturas mayores a 1.636 metros, la distribución sigue un patrón uniforme:
   Interpretación: Una distribución uniforme significa que los estudiantes están distribuidos de manera relativamente pareja entre los diferentes intervalos de estatura por encima de 1.636 metros. Esto se observa en la gráfica, donde las barras en estos intervalos tienen alturas similares, indicando que la frecuencia de estudiantes es aproximadamente la misma para cada rango de estaturas mayores a 1.636 metros.

1.9.2 Ejercicio 2

SOLUCION Datos 2.

1. Se solicitó al Dr. Paúl Hernández Herrera las calificaciones de algún curso que haya impartido. El profesor proporcionó las calificaciones finales de la materia de Probabilidad. Las calificaciones fueron las siguientes:

   C = [6, 6, 4, 7, 7, 5, 3, 9, 8, 10, 9, 9, 8, 8, 3, 9, 10, 10, 10, 10, 7, 5, 6, 10]

2. **SOLUCIÓN identificación del tipo de dato:** Las estaturas son datos de tipo "razón", ya que son valores numéricos continuos, se pueden ordenar y comparar, y tienen un punto de origen significativo (el valor cero indica ausencia de estatura).

3. **SOLUCION: FRECUENCIAS** $n = 5$

   (a) **Paso 1: Ordenar datos de menor a mayor:**
   [3, 3, 4, 5, 5, 6, 6, 6, 7, 7, 7, 8, 8, 8, 9, 9, 9, 9, 10, 10, 10, 10, 10, 10]

(b) **Paso 2: Calcular la anchura del intervalo:**

$$\Delta I = \frac{\max\_val - \min\_val}{n} = \frac{10-3}{5} = 1.4$$

(c) **Paso 3: Definir los límites de los intervalos**

$$\begin{aligned} l_0 &= \min\_val + 0 * \Delta I = 3 + 0 * 1.4 = 3.0 \\ l_1 &= \min\_val + 1 * \Delta I = 3 + 1 * 1.4 = 4.4 \\ l_2 &= \min\_val + 2 * \Delta I = 3 + 2 * 1.4 = 5.8 \\ l_3 &= \min\_val + 3 * \Delta I = 3 + 3 * 1.4 = 7.2 \\ l_4 &= \min\_val + 4 * \Delta I = 3 + 4 * 1.4 = 8.6 \\ l_5 &= \min\_val + 5 * \Delta I = 3 + 5 * 1.4 = 10 \end{aligned} \tag{5}$$

(d) **Paso 4: Construir los intervalos**

$$\begin{aligned} I_1 &= [l_0, l_1) = [3.0, 4.4) \\ I_2 &= [l_1, l_2) = [4.4, 5.8) \\ I_3 &= [l_2, l_3) = [5.8, 7.2) \\ I_4 &= [l_3, l_4) = [7.2, 8.6) \\ I_5 &= [l_4, l_5] = [8.6, 10] \end{aligned} \tag{6}$$

(e) **Paso 5: Obtener la frecuencia de datos que caen en cada intervalo:** Ahora, contamos cuántos datos caen dentro de cada intervalo:

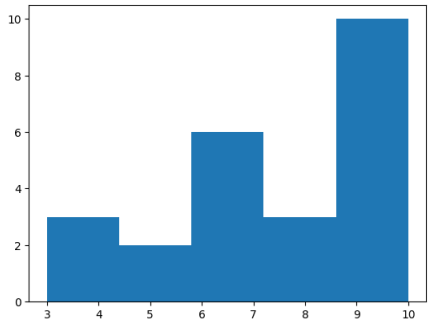
Intervalo	Frecuencia
$[3.0, 4.4)$	3
$[4.4, 5.8)$	2
$[5.8, 7.2)$	6
$[7.2, 8.6)$	3
$[8.6, 10)$	10

Por ejemplo, en el intervalo $[7.2, 8.6)$, buscamos todos los datos (ver **Paso 1**) que son mayores o iguales a 7.2 pero menores a 8.6. Los datos que cumplen esta condición son: $8, 8, 8$. Por lo tanto, la frecuencia es 3, ya que hay 3 datos en este intervalo.

4. **SOLUCION: FRECUENCIAS RELATIVAS**. Para calcular las frecuencias relativas, se debe dividir la frecuencia (tabla anterior) de cada intervalo por el número total de calificaciones, que en este caso son 24 calificaciones.

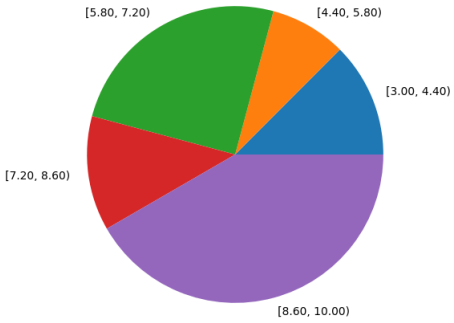
Intervalo	Frecuencia Relativa
$[3.0, 4.4)$	$3/24 = 0.125$
$[4.4, 5.8)$	$2/24 = 0.083$
$[5.8, 7.2)$	$6/24 = 0.25$
$[7.2, 8.6)$	$3/24 = 0.125$
$[8.6, 10)$	$10/24 = 0.416$

5. **SOLUCION: HISTOGRAMA** Gráfica de barras de la frecuencia de los datos.



6. **SOLUCION: GRAFICA CIRCULAR.** Calcular los grados que le corresponden los cuales se obtienen multiplicando las frecuencias relativas x 360

Intervalo	Frecuencia Relativa	Grados Rebanada
$[3, 4.4)$	0.125	$0.125 \times 360 = 45$
$[4.4, 5.8)$	0.083	$0.083 \times 360 = 30$
$[5.8, 7.2)$	0.25	$0.25 \times 360 = 90$
$[7.2, 8.6)$	0.125	$0.125 \times 360 = 45$
$[8.6, 10)$	0.416	$0.416 \times 360 = 150$



7. **SOLUCION: ¿Qué puedes concluir de las frecuencias y las gráficas?**

   (a) Pocos estudiantes tienen una calificación reprobatoria con calificación menor a 5.8.
   Solo 5 estudiantes obtuvo una calificación menor a 5.8. Esto sugiere que la mayoría de los estudiantes aprobaron el curso.

   (b) Aproximadamente el 42% de los estudiantes obtiene una calificación mayor a 8.6

   (c) Se forman dos grupos altos en el histograma:

   - **Primer grupo:** El primer grupo consiste en estudiantes con calificaciones en el rango de [8.6, 10], que incluye a 10 estudiantes. Este grupo representa a los estudiantes con las calificaciones más altas y muestra una fuerte concentración en las notas sobresalientes.
   - **Segundo grupo:** El segundo grupo se encuentra en el rango de [5.8, 7.2), con 6 estudiantes. Este grupo muestra una concentración en las calificaciones medias, lo que sugiere que hay un número significativo de estudiantes con un desempeño aceptable pero no excelente.

SOLUCION Datos 3.

1. **Identifique el tipo de dato:** Los datos de satisfacción que se presentan (”Malo”, ”Neutral”, ”Bueno”) son de tipo **Ordinal**. Esto significa que los valores pueden ordenarse en una secuencia lógica, donde ”Malo” es peor que ”Neutral”, y ”Bueno” es mejor que ambos. Sin embargo, la diferencia entre estos valores no tiene un significado numérico exacto; es decir, no podemos cuantificar la diferencia entre ”Malo” y ”Neutral” o entre ”Neutral” y ”Bueno”.

2. **SOLUCION: Frecuencias (para datos ordinales no se requiere n)**

   (a) **Paso 1: Ordenar datos:**
   ['Bueno', 'Bueno', 'Bueno', 'Bueno', 'Bueno', 'Bueno', 'Bueno', 'Bueno', 'Bueno', 'Bueno', 'Bueno", 'Neutral', 'Neutral', 'Neutral', 'Neutral', 'Malo', 'Malo', 'Malo', 'Malo]

   (b) **Paso 2: Calcular la anchura de intervalo:** No aplica para datos categóricos u ordinales.

   (c) **Paso 3: Definir los limites de intervalo** No aplica para datos categóricos u ordinales.

   (d) **Paso 4: Construir los intervalos** No aplica para datos categóricos u ordinales.

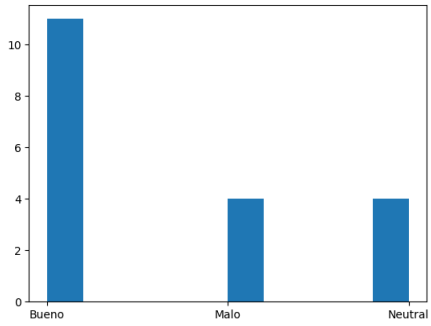
   (e) **Paso 5: Obtener la frecuencia de datos que caen en cada intervalo:**
   Para datos ordinales, el objetivo es identificar todas las categorías posibles y contar cuántas veces aparece cada una en el conjunto de datos.

Categoría	Frecuencia
Bueno	11
Neutral	4
Malo	4

3. **SOLUCION: FRECUENCIAS RELATIVAS**. Solamente dividir las frecuencias por el número total de datos = 19

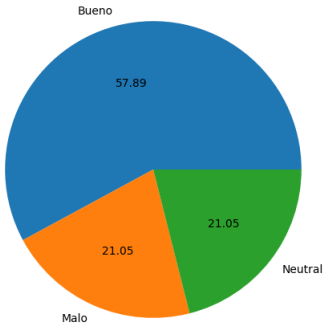
Categoría	Frecuencia Relativa
Bueno	$11/19 = 0.578$
Neutral	$4/19 = 0.21$
Malo	$4/19 = 0.21$

4. **SOLUCION: HISTOGRAMA** Gráfica de barras de la frecuencia de los datos.



5. **SOLUCION: GRAFICA CIRCULAR.** Calcular los grados que le corresponden los cuales se obtienen multiplicando las frecuencias relativas x 360

Categoría	Frecuencia Relativa	Grados Rebanada
Bueno	$11/19 = 0.578$	0.578*360= 208.42
Neutral	$4/19 = 0.21$	0.21*360 = 75.78
Malo	$4/19 = 0.21$	0.21* 360 = 75.78



6. ¿Qué puedes concluir de las frecuencias y las gráficas? Al analizar las frecuencias y las gráficas correspondientes a la evaluación de un nuevo producto, podemos llegar a las siguientes conclusiones:

   (a) **La mayoría de las personas (57%) aprueban el nuevo producto**
   Esta proporción indica que más de la mitad de los encuestados están satisfechos con el producto, lo que es un indicador favorable para su aceptación en el mercado.

   (b) **Aproximadamente el 21% de las personas desaprueban el producto**
   Este grupo representa una minoría significativa, lo que sugiere que aunque el producto tiene aceptación general, todavía hay un segmento considerable que no está satisfecho y que podría necesitar atención para mejorar su percepción del producto.

SOLUCION Ejercicio 2.

1. **SOLUCION: ¿Quién es la población de estudio?**La población de estudio está compuesta por estudiantes de la Universidad Veracruzana. En este caso, se trabajó con una muestra de 1087 estudiantes.

2. **¿Qué tipo de datos se obtuvieron (nominales, ordinales, . . .)?**

   SOLUCION:

   (a) **Nominales : Generos preferidos.** Clásica, Ranchera, Grupera, Pop, Rock en Español, etc.

   (b) **Nominales : Grupos de estudiantes:** Poperos, Rockeros, Bailadores, Tradicionalistas, Clasicos

   (c) **Nominales : Capital Cultural/Tipo de estudiante.** Herederos, Héroes, Pobres Exitosos, Riesgo, Alto Riesgo

   (d) **Nominal: Sexo.** Masculino o Femenino

   (e) **Ordinal: Trayectoria escolar.** Baja, Regular, Alta

3. **SOLUCION: ¿Qué técnicas se utilizaron para reportar las estadísticas?.**

   Los datos se presentaron en histogramas (por ejemplo, Gráfica 1), para observar la distribución de los datos. Además, se utilizaron tablas de frecuencia y frecuencia relativa (Cuadro 1, Cuadro 2, Cuadro 3 y Cuadro 4). Estas tablas muestran cuántas veces ocurre cada categoría (frecuencia) y la proporción que representa cada categoría en relación con el total (frecuencia relativa).

2 UNIDAD 2: MEDIDAS DESCRIPTIVAS

El objetivo de esta sección es adquirir las habilidades esenciales necesarias para analizar y comprender un conjunto de datos. Estas habilidades pueden incluir calcular promedios, identificar patrones, medir la dispersión de los datos y realizar otras tareas básicas para comprender y describir la información contenida en los datos. El objetivo final es equiparte con las herramientas necesarias para trabajar de manera efectiva con datos y extraer información valiosa de ellos.

OBJETIVOS: Al finalizar esta unidad, el estudiante será capaz de:

1. Diferenciar las principales medidas de tendencia central: media, mediana y moda.
2. Calcular e interpretar medidas de dispersión, como el rango, la varianza y la desviación estándar.
3. Calcular e interpretar cuartiles, percentiles y representar los datos mediante diagramas de caja.
4. Comprender el concepto de covarianza y el coeficiente de correlación de Pearson, así como aplicar la fórmula para su cálculo.
5. Integrar los conceptos y herramientas de medidas descriptivas para analizar conjuntos de datos y comunicar los resultados de manera clara y efectiva.

2.1 Medida de Tendencia Central:

Una medida de tendencia central es un valor que indica la posición central dentro de un conjunto de datos. Es un punto alrededor del cual se distribuyen los datos, y su función es resumir y representar un valor típico o promedio del conjunto de datos. Las medidas de tendencia central ayudan a entender cuál es el valor que mejor representa a todos los datos.

2.1.1 Media o Promedio

La media (o promedio) es una de las medidas de tendencia central más utilizadas. Se obtiene sumando todos los valores de un conjunto de datos y dividiendo esa suma entre el número total de valores en el conjunto.

Características de la media:

1. **Valor Típico:** La media representa un valor típico o promedio en el conjunto de datos. Por ejemplo, si tienes un conjunto de edades de personas, la media es la edad promedio de ese grupo.
2. **Centro de Gravedad:** La media puede considerarse como el "centro de gravedad" de los datos. Si imaginas los valores distribuidos a lo largo de una línea, la media sería el punto donde los datos estarían en equilibrio, es decir, donde las sumas de las diferencias respecto a la media se equilibran entre sí.
3. **Punto de Equilibrio:** Imagina una regla que representa tus datos numéricos. Si colocas la regla en el punto correspondiente a la media, la regla se mantendría en equilibrio, con los valores menores compensando a los valores mayores.
4. **Sensible a Valores Extremos:** La media es muy sensible a los valores extremos o atípicos (outliers). Por ejemplo, si una calificación es significativamente más alta o más baja que las demás en un conjunto de notas, la media se verá afectada, reflejando un valor mayor o menor de lo que sería sin ese valor extremo.

Fórmula de la Media:

Supongamos que tenemos un conjunto de datos $X = \{x_1, x_2, x_3, \ldots, x_n\}$. La media se calcula con la siguiente formula:

$$\text{media}(X) = \frac{\sum_{i=1}^{n} x_i}{n}, \tag{7}$$

la media se representa generalmente por $\bar{x}$

2.1.2 Mediana

La mediana es una medida de tendencia central que identifica el valor que se encuentra en el centro de un conjunto de datos cuando estos se ordenan de menor a mayor o de mayor a menor. Es un indicador importante porque divide el conjunto de datos en dos partes iguales: el 50% de los valores están por debajo de la mediana, y el otro 50% está por encima. A menudo se denota con la letra "$\tilde{x}$" con un tilde encima.

Cómo Calcular la Mediana:

El proceso para calcular la mediana varía ligeramente dependiendo de si el número de valores en el conjunto es impar o par

1. **Ordenar los Valores:** Coloca todos los valores del conjunto de datos en orden de magnitud, ya sea en orden ascendente (de menor a mayor) o descendente (de mayor a menor).
2. **Obtener la mediana dependiendo del número de datos**
   - **Número de Valores Impar:** Si tienes un número impar de valores, la mediana es el valor que está exactamente en el medio de la lista ordenada. No hay necesidad de calcular nada adicional; simplemente identifica el valor central.
   - **Número de Valores Par:** Si tienes un número par de valores, la mediana no será uno de los valores del conjunto, sino que se calculará como el promedio de los dos valores centrales. Para encontrar la mediana, suma esos dos valores y luego divide el resultado entre 2.

Características de mediana:

La mediana es el valor que se encuentra justo en el medio de un conjunto de datos ordenados, cuando estos se disponen en orden de magnitud creciente o decreciente. Esto significa que, al mirar el conjunto de datos, aproximadamente la mitad de los valores están por encima de la mediana y la otra mitad está por debajo.

La mediana tiene varias características importantes que la hacen útil para describir la tendencia central de un conjunto de datos:

- **Valor Central Típico**: La mediana es un valor que divide el conjunto de datos en dos partes iguales. En un conjunto de datos ordenados, el 50% de los valores estarán por encima de la mediana y el 50% por debajo. Por lo tanto, la mediana representa un "valor central" típico en el conjunto.
- **Robusta ante Valores Atípicos:** A diferencia de la media, la mediana no se ve afectada significativamente por valores extremos o atípicos (outliers). Por ejemplo, en un conjunto de datos de ingresos, un ingreso extremadamente alto no influirá en la mediana tanto como lo haría en la media, lo que la hace una medida más robusta en presencia de valores atípicos.
- **Adecuada para Datos Ordinales o Skewed**:La mediana es particularmente útil cuando se trabaja con datos ordinales (como clasificaciones) o con datos que tienen una distribución sesgada (donde la mayoría de los valores se agrupan en un extremo). En estos casos, la mediana proporciona una representación más precisa de la tendencia central que la media.
- **Ubicación Central**: Si tienes un conjunto de datos que representa edades, ingresos o cualquier otra variable, la mediana te dirá el valor que se encuentra en el punto medio de esa variable, lo que puede ayudarte a comprender cómo se distribuyen los datos en torno a esa ubicación central.

2.1.3 Moda

La moda es una medida de tendencia central que indica el valor que aparece con mayor frecuencia en un conjunto de datos. Es el valor que se repite más veces que cualquier otro, y por ello, destaca como el más común o popular dentro del conjunto.

Tipos de moda

- **Conjunto unimodal:** Si un solo valor tiene la mayor frecuencia en el conjunto de datos, se dice que el conjunto es unimodal. La moda en este caso es ese único valor que se repite más veces que los demás.
- **Conjunto bimodal:** Si hay dos valores que comparten la misma frecuencia máxima, el conjunto de datos es bimodal. Esto significa que hay dos modas, y ambos valores son considerados igualmente comunes.
- **Conjunto multimodal:** Si más de dos valores tienen la misma frecuencia más alta, todos esos valores se consideran modas, y el conjunto de datos se llama multimodal. En este caso, múltiples valores se repiten con la misma frecuencia máxima.
- **Sin moda:** Si cada valor en el conjunto es único y no se repite, se dice que el conjunto de datos no tiene moda. En este escenario, ningún valor destaca como más común que los demás.

Características de moda:

1. **Representación de la Tendencia Principal:** La moda es útil para identificar cuál es el valor más común o dominante en un conjunto de datos. Esto puede ser particularmente revelador en conjuntos de datos donde un valor específico se repite con mucha más frecuencia que otros, señalando una tendencia clara.
2. **Útil para Datos Categóricos y Cuantitativos:** a moda se puede aplicar tanto a datos cuantitativos (números, como edades o calificaciones) como a datos categóricos (categorías o grupos, como colores o tipos de productos). En datos categóricos, la moda muestra cuál es la categoría más popular, mientras que en datos cuantitativos indica el número que aparece más veces.

2.2 Medidas de Dispersión

Las medidas de dispersión son herramientas estadísticas esenciales que nos permiten entender cómo se distribuyen los datos en un conjunto. A diferencia de las medidas de tendencia central, que se enfocan en el valor típico o promedio, las medidas de dispersión se centran en la variabilidad de los datos, es decir, en qué tan dispersos o agrupados están los valores en relación con un punto central como la media o la mediana.

¿Porqué son importantes las medidas de dispersión?

Las medidas de dispersión son cruciales porque:

- Nos ofrecen una visión más completa de la distribución de los datos.
- Ayudan a evaluar la consistencia de los datos.
- Permiten identificar la presencia de valores atípicos (outliers) y su efecto en el conjunto de datos.
- Proporcionan contexto sobre la variabilidad, que es esencial para hacer comparaciones entre diferentes conjuntos de datos.

2.2.1 Rango

Es la diferencia entre el valor máximo y el valor mínimo en un conjunto de datos. Ofrece una idea general de cuán extensos son los datos.

$$\text{Rango} = \text{Valor máximo - Valor mínimo}$$

Interpretación del rango

1. **Extensión General:** El rango ofrece una idea general de cuán extendidos están los valores en el conjunto de datos. Un rango grande indica una mayor extensión o dispersión entre los valores, mientras que un rango pequeño sugiere que los valores están más agrupados.

2. **Limitación:** Aunque el rango es fácil de calcular e interpretar, no proporciona información sobre cómo están distribuidos los valores entre el mínimo y el máximo. Por ejemplo, dos conjuntos de datos diferentes pueden tener el mismo rango pero distribuciones internas muy distintas.

2.2.2 Desviación estándar

La desviación estándar es una medida de dispersión en estadísticas que proporciona información sobre cuán dispersos o agrupados están los valores en un conjunto de datos en relación con la media (promedio). La desviación estándar es una medida de dispersión precisa y ampliamente utilizada.
Cálculo de la desviación estándar Supongamos que tenemos un conjunto de datos $X = \{x_1, x_2, x_3, \ldots, x_n\}$. Dependiendo de como se obtuvieron los datos (censo o muestra), la desviación estándar se calcula de manera diferente.

- **Para una muestra:** La desviación estándar de una muestra se denota por s y se calcula con la siguiente fórmula:

$$\mathrm{s}(X) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}, \tag{8}$$

  Aquí, $\bar{x}$ es la media de la muestra, y n el número de observaciones.

- **Para una población:** La desviación estándar de una población se denota por σ y se calcula con la fórmula:

$$\sigma(X) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}. \tag{9}$$

  En este caso, μ es la media de la población, y n es el número total de observaciones en la población.

Interpretación de la Desviación Estándar

1. **Magnitud de la Variabilidad:** Una desviación estándar alta indica que los valores están ampliamente dispersos alrededor de la media. Esto sugiere una mayor variabilidad en los datos. Una desviación estándar baja sugiere que los valores están más agrupados cerca de la media, lo que indica menor variabilidad.

2. **Efecto de Valores Atípicos:** La desviación estándar es sensible a valores atípicos. Un valor extremadamente alto o bajo puede aumentar significativamente la desviación estándar, haciendo que parezca que hay más variabilidad en el conjunto de datos de lo que realmente hay.

3. **Unidad de Medida:** La desviación estándar se expresa en la misma unidad que los datos originales. Esto facilita la interpretación en un contexto práctico. Por ejemplo, si los datos son alturas en centímetros, la desviación estándar también estará en centímetros.

2.2.3 Varianza

La varianza es una medida de variación que se calcula como el cuadrado de la desviación estándar. A diferencia de la desviación estándar, que mide la dispersión en la misma unidad que los datos, la varianza se mide en unidades al cuadrado.
Cálculo de la varianza Supongamos que tenemos un conjunto de datos $X = \{x_1, x_2, x_3, \ldots, x_n\}$. Dependiendo de como se obtuvieron los datos (censo o muestra), la varianza se calcula de manera diferente.

- **Varianza Muestral**: La varianza muestral se refiere a la variación en un conjunto de datos específico, como una muestra de una población más grande. Se calcula como el cuadrado de la desviación estándar muestral (s^2), que es una medida de dispersión basada en la muestra.

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

- **Varianza Poblacional**: La varianza poblacional es la variación en una población completa. Se calcula como el cuadrado de la desviación estándar poblacional (σ), que es una medida de dispersión basada en toda la población. Se calcula por la formula:

$$\sigma^2(X) = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}. \tag{10}$$

Interpretación de la varianza:

1. **Unidad de Medida:** La varianza se expresa en unidades al cuadrado de los datos originales, lo que a menudo no tiene una interpretación directa en términos prácticos. Por esta razón, a menudo se prefiere la desviación estándar, que tiene la misma unidad que los datos originales.

2. **Relación con la Desviación Estándar:** La varianza es el cuadrado de la desviación estándar. Esto significa que mientras la desviación estándar nos da una medida directa de la dispersión, la varianza nos da una medida más amplia al considerar las diferencias al cuadrado.

2.2.4 Teorema de Chebyshev

El Teorema de Chebyshev es una regla matemática que describe cómo se distribuyen los datos en relación con la media en cualquier conjunto de datos, sin importar la forma de su distribución. Este teorema es particularmente útil porque se aplica a todos los tipos de distribuciones, ya sean simétricas, sesgadas o de cualquier otra forma.
Concepto principal: El teorema establece que la proporción de cualquier conjunto de datos que se encuentra dentro de K desviaciones estándar de la media es siempre al menos $1 - 1/K^2$, donde K es cualquier número positivo mayor que 1. Esto significa que, sin importar cómo estén distribuidos los datos:

1. Para $K = 2$, al menos
$$1 - \frac{1}{2^2} = 1 - \frac{1}{4} = 3/4 = 0.75$$
(o 75%) de todos los valores están dentro de dos desviaciones estándar de la media.

2. Para $K = 3$, al menos
$$1 - \frac{1}{3^2} = 1 - \frac{1}{9} = 8/9 = 0.888$$
(u 88.8%) de todos los valores están dentro de tres desviaciones estándar de la media.

2.3 Medidas de distribución de los datos adicionales

2.3.1 Cuartiles

Los cuartiles son una herramienta estadística que se utiliza para dividir un conjunto de datos en cuatro partes iguales. Al hacer esto, los cuartiles ayudan a describir cómo se distribuyen los valores dentro de un conjunto de datos y proporcionan una visión más detallada de la variabilidad y la dispersión de los datos. Los tres cuartiles principales son el primer cuartil (Q1), el segundo cuartil (Q2, que es también la mediana), y el tercer cuartil (Q3).

Conceptos principales de cuartiles

1. Q1 (Primer Cuartil): Este cuartil separa el 25% inferior de los datos del 75% superior cuando los datos están ordenados en orden ascendente. Esto significa que el 25% de los valores en el conjunto de datos son menores o iguales que Q1, mientras que el 75% restante de datos son mayores o iguales que Q1.

2. Q2 (Segundo Cuartil): Q2 es el valor que divide el conjunto de datos en dos mitades iguales, por lo que el 50% de los valores están por debajo de Q2 y el 50% están por encima. Este cuartil es equivalente a la mediana. Q2 representa el valor central de la distribución de los datos, proporcionando una medida de tendencia central.

3. Q3 (Tercer Cuartil): Q3 es el valor que separa el 75% inferior de los datos del 25% superior. Esto significa que al menos el 75% de los valores en el conjunto de datos son menores o iguales que Q3, mientras que al menos el 25% de los valores son mayores o iguales que Q3.

Importancia de cuartiles

- **Análisis de Distribución:** Los cuartiles permiten dividir el conjunto de datos en segmentos iguales, facilitando el análisis de cómo se distribuyen los valores y si existen concentraciones o dispersiones particulares en ciertas partes del conjunto de datos.

- **Detección de Valores Atípicos:** Al comparar los cuartiles, es posible identificar valores atípicos. Por ejemplo, valores que caen por debajo de Q1 o por encima de Q3 pueden ser considerados atípicos si se encuentran a una distancia significativa de estos cuartiles.

- **Diagrama de Caja (Box Plot):** Los cuartiles son fundamentales en la creación de diagramas de caja, que son representaciones gráficas que muestran la mediana, los cuartiles y los valores atípicos de manera visual. Un diagrama de caja facilita la comprensión de la distribución de los datos de un vistazo.

2.3.2 Diagrama de Caja

El diagrama de caja (también conocido como "box plot") es una herramienta gráfica que se utiliza para representar la distribución de un conjunto de datos. Este diagrama es especialmente útil para visualizar la dispersión, la simetría, y la presencia de valores atípicos en los datos. A continuación, se explica detalladamente cómo construir y entender un diagrama de caja.

Pasos para crear un diagrama de caja

1. **Organiza tus Datos:** Comienza con un conjunto de datos que deseas visualizar en el diagrama de caja. Asegúrate de que tus datos estén ordenados de manera ascendente o descendente. Cuenta el número total de datos n.

2. **Calcula los Cuartiles :** Los cuartiles son valores que dividen el conjunto de datos en 4 partes iguales. Los tres cuartiles principales son:

    - Q1 (Primer Cuartil): El valor que separa el 25% inferior de los datos del 75% superior. Se obtiene encontrando el valor de los datos correspondiente a la posición más cercana a $1 * n/4$

- Q2 (Segundo Cuartil o Mediana): El valor que divide el conjunto de datos en dos partes iguales, donde el 50% de los valores están por debajo y el otro 50% por encima. Se obtiene encontrando el valor de los datos correspondiente a la posición más cercana a $2 * n/4$
- Q3 (Tercer Cuartil): El valor que separa el 75% inferior de los datos del 25% superior. Se obtiene encontrando el valor de los datos correspondiente a la posición más cercana a $2 * n/4$
- **Ejemplo:** Supongamos que tenemos el conjunto de datos ordenados:
  2,3,4,5,6,7,15,15,17,20,21,30
  - Q1: El primer cuartil corresponde a la posición $1 * n/4 = 1 * 12/4 = 3$. En este caso, el dato en la tercera pocisión es 4. Por lo cual, Q1 = 4
  - Q2: El segundo cuartil corresponde a la posición $2 * n/4 = 2 * 12/4 = 6$. En este caso, el dato en la sexta pocisión es 7. Por lo cual, Q2 = 7
  - Q2: El tercer cuartil corresponde a la posición $3 * n/4 = 3 * 12/4 = 9$. En este caso, el dato en la novena pocisión es 17. Por lo cual, Q3 = 17

3. **Cuadro Central:** Dibuja un rectángulo que se extienda desde el valor del primer cuartil (Q1) hasta el tercer cuartil (Q3). Este rectángulo representa el 50% central de tus datos y se conoce como el "cajón" o la "caja". La parte inferior del rectángulo está en Q1, y la parte superior está en Q3.

4. **Línea en la Caja:** Dentro del rectángulo (caja), dibuja una línea horizontal en el valor del Q2 (mediana). Esta línea divide la caja en dos partes, cada una representando el 25% de los datos.

5. **Líneas Hacia Afuera:** Desde los extremos de la caja (Q1 y Q3), se dibujan líneas verticales llamadas "bigotes" que se extienden hasta los valores mínimo y máximo del conjunto de datos, que no se consideran valores atípicos.

El diagrama de caja es una herramienta visual poderosa para resumir y analizar la distribución de un conjunto de datos. Facilita la identificación de la dispersión, simetría, y valores atípicos, y es fundamental en la estadística descriptiva para comparar diferentes conjuntos de datos de manera efectiva.

2.3.3 Percentiles:

Los percentiles son medidas estadísticas que dividen un conjunto de datos en 100 partes iguales, permitiendo así una comprensión más detallada de la distribución de los valores. Al igual que los cuartiles, que dividen los datos en cuatro partes, los percentiles dividen los datos en 100 partes. Esto significa que hay 99 percentiles, denotados como $(P_1, P_2, \ldots, P_{99})$ donde cada percentil corresponde a un valor específico en el conjunto de datos.
Percentiles Son valores que indican la posición relativa de un dato dentro de un conjunto. Por ejemplo, el percentil P_1 es el valor por debajo del cual se encuentra el 1% de los datos, lo que significa que el 99% de los datos son mayores o iguales a P_1. De manera similar, el percentil P_{50} coincide con la mediana, lo que indica que el 50% de los datos están por debajo de este valor y el otro 50% están por encima. El percentil P_{99} indica el valor por debajo del cual se encuentra el 99% de los datos, y solo el 1% de los datos es mayor o igual a este valor.

Los percentiles proporcionan una manera detallada de analizar la posición de los datos dentro de un conjunto y son especialmente útiles en contextos donde se necesita una interpretación más granular de la distribución de datos, como en pruebas estandarizadas, análisis de rendimiento, y estudios de mercado.

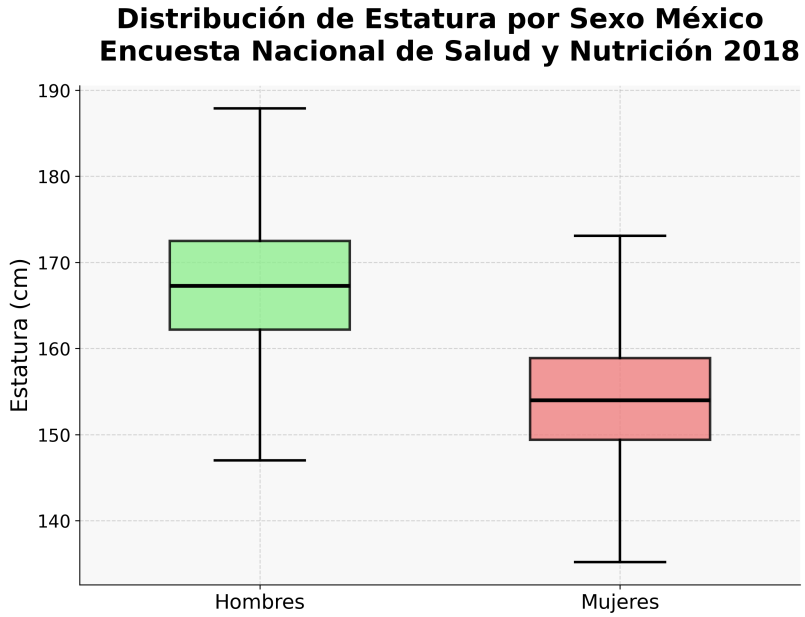


Figure 5: Ejemplo de Diagrama de Caja Estatura de Población Mexicana - ENSANUT 2018.

2.4 Medidas relaciones

2.4.1 Covarianza

La covarianza es una medida que indica cómo dos variables (por ejemplo, X y Y) cambian juntas. En otras palabras, nos dice si, cuando una variable aumenta, la otra también tiende a aumentar (covarianza positiva) o disminuir (covarianza negativa), o si no tienen una relación clara (covarianza cercana a cero). Puede tomar cualquier valor en los números reales. La formula para calcular la convariaza esta dada por:

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

donde $\bar{x}$ y $\bar{y}$ son el promedio de la variable X y Y, respectivamente.

1. Si la covarianza es negativa, significa que cuando X aumenta, Y tiende a disminuir y viceversa. Es una medida promedio, lo que significa que un solo par de valores donde X aumenta y Y disminuye significativamente puede tener un impacto grande en el resultado, incluso si la mayoría de los valores no siguen esta relación inversa, pero son pequeños.

X	Y
10	10
20	0
19	-1
20	0
18	-2

Table 1: Datos 1

En los datos de la tabla 1, la covarianza entre las variables X e Y es -19.20, lo que indica una relación inversa. Al observar los datos, se nota que solo un par de valores sigue esta relación inversa: cuando X aumenta de 10 a 20 (de la primera a la segunda línea), Y disminuye de 10 a 0. Sin embargo, para las otras líneas de datos, cuando X aumenta, Y también aumenta, y

cuando X disminuye, Y también disminuye, pero estos cambios son pequeños en comparación con la gran variación inversa observada en el primer par de valores. Por lo tanto, esta variación inversa dominante hace que la covarianza sea negativa y tenga un valor significativo.

2. La covarianza positiva indica que cuando el valor de X aumenta, el valor de Y tiende a aumentar también, y cuando X disminuye, Y tiende a disminuir. Es un indicador promedio, lo que implica que incluso si la mayoría de los valores no siguen esta relación específica pero son pequeños, un solo par de valores donde X y Y aumentan o disminuyen considerablemente puede tener un impacto significativo en el resultado global.

X	Y
0	0
10	1
9	0
11	11
15	15
10	13

Table 2: Datos 2

Para los datos en la tabla 2, la covarianza entre las variables X y Y es 23.86, indicando una relación directa. Esto significa que cuando X aumenta, Y también aumenta, y cuando X disminuye, Y también disminuye. Es importante notar que, de manera similar al ejemplo anterior, un solo valor positivo grande puede tener un impacto significativo en el resultado de la covarianza, incluso si hay muchos valores que siguen una relación inversa pero son pequeños.

3. Cuando la covarianza está cerca de cero, significa que no hay un patrón discernible en los datos que permita hacer conclusiones significativas sobre la relación entre las variables X y Y. En otras palabras, no hay una tendencia clara de cómo cambian juntas.

2.4.2 Coeficiente de correlación de Pearson

El coeficiente de correlación de Pearson es una medida estadística que se utiliza para evaluar la relación entre dos variables, X y Y. A diferencia de la covarianza, que puede tener cualquier valor real y es difícil de comparar, el coeficiente de correlación de Pearson normaliza los datos y proporciona un número entre -1 y 1, lo que hace que sea más fácil de interpretar. El coeficiente r de correlación de Pearson esta dado por:

$$r = \frac{cov(X,Y)}{s_x s_y}$$

donde $s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$ y $s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}}$ son las desviaciones estándar de la muestra.

El coeficiente de correlación de Pearson es un valor en el intervalo [-1,1]. Con las siguientes caracteristicas:

1. Si el coeficiente de correlación de Pearson es -1, significa que hay una correlación lineal inversa perfecta entre X e Y. Cuando una variable aumenta, la otra disminuye de manera proporcional, y viceversa. Esto se representa gráficamente como una línea recta descendente.

2. Si el coeficiente es 1, indica una correlación lineal directa perfecta. Cuando una variable aumenta, la otra también aumenta proporcionalmente, y viceversa. Esto se muestra gráficamente como una línea recta ascendente.

3. Un coeficiente de 0 significa que no hay correlación lineal entre las variables. En otras palabras, no hay un patrón claro en los datos cuando se grafican.

interpretar el valor del coeficiente de correlación de Pearson:

1. Si el valor absoluto del r está entre 0 y 0.09, indica que no hay una correlación lineal significativa entre las variables.
2. $|r|$ entre 0.10 y 0.29, la correlación es débil.
3. $|r|$ entre 0.30 y 0.50, la correlación es moderada.
4. $|r|$ entre 0.50 y 1.00, la correlación es fuerte.

En resumen, el coeficiente de correlación de Pearson nos dice si las variables están relacionadas de alguna manera y qué tan fuerte es esa relación. Es una herramienta valiosa para entender patrones en los datos y hacer predicciones basadas en esas relaciones.

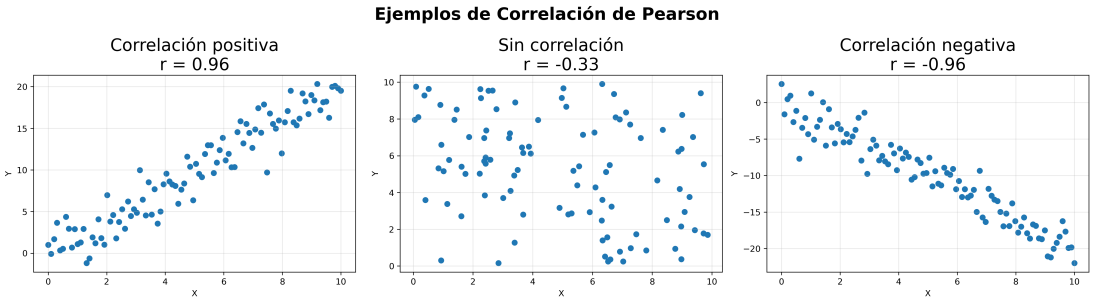


Figure 6: Ejemplo de coeficiente de correlación de Pearson.

2.5 EJERCICIOS

EJERCICIO 1: Medidas descriptivas

Para cada conjunto de datos, realice lo siguiente:

1. Calcule la media, mediana, moda y mitad de rango.
2. Interpreta los valores anteriores
3. Calcula la desviación estándar, varianza y cuartiles
4. Interpreta los valores anteriores

- **Datos Reales 1**: Estatura de los estudiantes de la clase de Estadística Aplicada de la Facultad de Ciencias, UASLP.
- **Datos Reales 2**: Calificación final de los estudiantes en un curso determinado.
- **Datos Ficticios 3**: Encuesta sobre la satisfacción con un nuevo producto. Los resultados son los siguientes:
  Malo, Bueno, Bueno, Neutral, Neutral, Bueno, Malo, Bueno, Bueno, Bueno, Malo, Malo, Neutral, Bueno, Bueno, Bueno, Bueno, Bueno, Neutral

EJERCICIO 2: Coeficiente de correlación

Instrucciones: Calcula el coeficiente de correlación de Pearson para los dos conjuntos de datos proporcionados. Estos coeficientes te permitirán entender la relación lineal entre las variables X y Y en cada conjunto de datos.

1. **Datos 1:**

X	Y
7.4	7.1
4.7	4.4
5.3	5.3
6.5	6.8
17.0	17.7
20.7	20.4

2. **Datos 2:**

X	Y
9.1	7.4
7.1	4.7
7.9	5.3
9.1	6.5
10.9	8.3
5.6	5.6

Analiza e interpreta los resultados

1. Explica qué significa el valor obtenido del coeficiente de correlación de Pearson para cada conjunto de datos. Por ejemplo, ¿indica una relación fuerte, débil o inexistente entre las variables X y Y?
2. Compara los dos coeficientes obtenidos: ¿Cuál de los dos conjuntos de datos muestra una mayor correlación entre X y Y?

2.6 SOLUCION 1: Medidas Descriptivas

2.6.1 Ejercicio 1

SOLUCION Datos 1.

1. Se realizó un censo de los 15 estudiantes de la clase de Estadística Aplicada, midiendo la estatura en metros con una cinta métrica. Los resultados fueron los siguientes:

   Estaturas = [1.79, 1.69, 1.75, 1.70, 1.67, 1.50, 1.75, 1.80, 1.82, 1.70, 1.84, 1.73, 1.73, 1.60, 1.68]

2. **SOLUCION: Media, mediana, moda, y mitad de rango.**

   Los datos ordenados de menor a mayor son:

   [1.50, 1.60, 1.67, 1.68, 1.69, 1.70, 1.70, 1.73, 1.73, 1.75, 1.75, 1.79, 1.8, 1.82, 1.84]

   (a) **Media o promedio:** Se calcula utilizando la fórmula:

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

   Sustituyendo los valores:

   $$\bar{x} = \frac{1.50 + 1.60 + 1.67 + 1.68 + 1.69 + 1.70 + 1.70 + 1.73 + 1.73 + 1.75 + 1.75 + 1.79 + 1.80 + 1.82 + 1.84}{15}$$
   $$= \frac{25.75}{15} = 1.71$$

   Por lo tanto, el promedio es $\bar{x} = 1.71$ metros.

   (b) **Mediana:** La mediana es el valor central que divide el conjunto de datos en dos partes iguales. Como hay 15 estaturas (un número impar), la mediana es el valor que ocupa la posición central. En un conjunto ordenado de 15 datos, la mediana está en la posición 8.

   En este caso, el valor en la posición 8 es 1.73 metros. Por lo tanto, la mediana es 1.73 metros.

   [1.50, 1.60, 1.67, 1.68, 1.69, 1.70, 1.70, **1.73**, 1.73, 1.75, 1.75, 1.79, 1.8, 1.82, 1.84]

   (c) **Moda:** La moda es el valor o los valores que más se repiten en un conjunto de datos. Para encontrar la moda, revisamos cuántas veces aparece cada valor en el conjunto de estaturas. En este caso, los valores 1.70, 1.73 y 1.75 se repiten cada uno dos veces, por lo que tenemos un conjunto multimodal con las modas 1.70, 1.73, 1.75.

   (d) **Mitad de rango:** La mitad de rango se calcula promediando el valor máximo y el valor mínimo del conjunto de datos. La fórmula es:

   $$\text{Mitad de rango} = \frac{\max val + \min val}{2}$$

   Sustituyendo los valores:

   $$= \frac{1.84 + 1.50}{2} = \frac{3.34}{2} = 1.67$$

   Por lo tanto, la mitad de rango es 1.67 metros.

3. **SOLUCION: Interpretación de las Medidas de Tendencia Central** Al observar las medidas calculadas:

- La media (1.717 metros) es una buena representación de la estatura promedio de los estudiantes.
- La mediana (1.73 metros) está muy cerca de la media, lo que sugiere que las estaturas están distribuidas de manera relativamente simétrica alrededor de la media.
- La moda indica que hay varios valores que se repiten, y todos están cerca de la media, reforzando la idea de que las estaturas son similares.
- La mitad de rango (1.67 metros) es un poco menor que la media, lo que sugiere que hay valores en el extremo inferior (por ejemplo, 1.50 metros) que están alejados del resto.

4. **SOLUCION: desviación estándar, varianza y cuartiles** Para analizar cómo se dispersan las estaturas alrededor de la media, calcularemos la desviación estándar, la varianza y los cuartiles.

   (a) **Desviación estandar:** La desviación estándar mide cuánta variación o dispersión existe en un conjunto de datos. Se calcula con la fórmula:

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

   En el ejercicio anterior se calculo el promedio el cual es igual a $\bar{x} = 1.71$, sustituyendo los valores:

   $$\mathrm{s}(X) = \sqrt{\frac{(1.50-1.71)^2 + (1.60-1.71)^2 + (1.67-1.71)^2 + \ldots + (1.80-1.71)^2 + (1.82-1.71)^2 + (1.84-1.71)^2}{15-1}}$$
   $$\mathrm{s}(X) = \sqrt{\frac{(-0.21)^2 + (-0.11)^2 + (-0.04)^2 + \ldots + (0.09)^2 + (0.11)^2 + (0.13)^2}{14}}$$
   $$= \sqrt{\frac{0.1068}{14}} = 0.08734$$

   (b) **Varianza:** La varianza es simplemente el cuadrado de la desviación estándar:

   $$v = s^2 = 0.08734^2 = 0.007628$$

   Por lo tanto, la varianza es $v = 0.007628$ metros cuadrados.

   (c) **Quartiles:** Representan el límite inferior de 25%, 50% y 75% de los datos menores al quartile. La posición esta dada por $n * 0.25, n * 0.50$ y $n * 0.75$ donde n representa el número total de datos. En nuestro ejercicio $n = 15$, por lo cual la posición de los cuartiles es 0.25*15, 0.5*15, 0.75*15 = 3.75, 7.5, 11.25, y redondiando se tiene 4,8,11. Que corresponde a los siguiente valores:

   $$Q_1 = 1.68$$

   $$Q_2 = 1.73$$

   $$Q_3 = 1.75$$

5. **SOLUCION: Interpreta los valores anteriores:** la desviación estándar baja (0.08734 metros) indica que las estaturas de los estudiantes están muy cerca de la media (1.717 metros), es decir, no hay mucha variación en las estaturas. En otras palabras, la mayoría de los estudiantes tienen estaturas muy cercanas al promedio. Además, Los cuartiles muestran que: El 50% de los estudiantes mide entre 1.68 y 1.75 metros. Esto refuerza la idea de que la mayoría de los estudiantes tienen estaturas cercanas al promedio, con una pequeña variación en el grupo.

2.6.2 Ejercicio 2

SOLUCION Datos 2.

1. Se solicitó al Dr. Paúl Hernández Herrera las calificaciones de algún curso que haya impartido. El profesor proporcionó las calificaciones finales de la materia de Probabilidad. Las calificaciones fueron las siguientes:

   C = [6, 6, 4, 7, 7, 5, 3, 9, 8, 10, 9, 9, 8, 8, 3, 9, 10, 10, 10, 10, 7, 5, 6, 10]

2. **SOLUCION: Media, mediana, moda, y mitad de rango.**

   Los datos ordenados de menor a mayor son:

   [3, 3, 4, 5, 5, 6, 6, 6, 7, 7, 7, 8, 8, 8, 9, 9, 9, 9, 10, 10, 10, 10, 10, 10]

   (a) **Media o promedio:** Se calcula utilizando la fórmula:

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

   Sustituyendo los valores:

   $$\bar{x} = \frac{3+3+4+5+5+6+6+6+7+7+7+8+8+8+9+9+9+9+10+10+10+10+10+10}{24}$$
   $$= \frac{179}{24} = 7.45$$

   Por lo tanto, el promedio es $\bar{x} = 7.45$ de calificación.

   (b) **Mediana:** El número de datos son 24, lo cual corresponde a un número par. Por lo debemos promediar los valores de posición 12 y 13, lo que corresponde a

   $$\text{mediana} = \frac{8+8}{2} = 8$$

   (c) **Moda:** En este caso, necesitamos identificar el dato que se repite con mayor frecuencia. La calificación de 10 es la que aparece más veces, con una frecuencia de 6. Por lo tanto, la moda es 10.

   (d) **Mitad de rango:** La mitad de rango se calcula promediando el valor máximo y el valor mínimo del conjunto de datos. La fórmula es:

   $$\text{Mitad de rango} = \frac{\max val + \min val}{2}$$

   Sustituyendo los valores:

   $$= \frac{10+3}{2} = \frac{13}{2} = 6.5$$

   Por lo tanto, la mitad de rango es una calificación de 6.5.

3. **SOLUCION: Interpretación de las Medidas de Tendencia Central** El análisis de las calificaciones de los estudiantes muestra lo siguiente:

   - **Promedio de calificación (7.45):** El promedio de todas las calificaciones es 7.45. Esto significa que, en general, se espera que los estudiantes hayan aprobado la materia, ya que el promedio está por encima del límite de aprobación.

- **Mediana (8):** La mediana es 8, lo que indica que el 50% de los estudiantes obtuvo una calificación de 8 o más. En otras palabras, la mitad de los estudiantes tiene calificaciones bastante buenas.
- **Moda (10):** La moda, que es la calificación que más se repite, es 10. Esto significa que un número significativo de estudiantes alcanzó la calificación máxima.

4. **SOLUCION: desviación estándar, varianza y cuartiles** Para analizar cómo se dispersan las estaturas alrededor de la media, calcularemos la desviación estándar, la varianza y los cuartiles.

   (a) **Desviación estandar:** La desviación estándar mide cuánta variación o dispersión existe en un conjunto de datos. Se calcula con la fórmula:

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

   En el ejercicio anterior se calculo el promedio de calificaciones el cual es igual a $\bar{x} = 7.45$, sustituyendo los valores:

   $$\mathrm{s}(X) = \sqrt{\frac{(3-7.45)^2 + (3-7.45)^2 + (4-7.45)^2 + (5-7.45)^2 + \ldots + (10-7.45)^2 + (10-7.45)^2 + (10-7.45)^2}{24-1}}$$
   $$\mathrm{s}(X) = \sqrt{\frac{(-4.45)^2 + (-4.45)^2 + (-3.45)^2 + (-2.45)^2 + \ldots + (2.55)^2 + (2.55)^2 + (2.55)^2}{14}}$$
   $$= \sqrt{\frac{119.96}{23}} = 2.2837$$

   (b) **Varianza:** La varianza es simplemente el cuadrado de la desviación estándar:

   $$v = s^2 = 2.2837^2 = 5.2152$$

   Por lo tanto, la varianza es $v = 5.2152$ y las unidades son puntos al cuadrado.

   (c) **Cuartiles:** Representan el límite inferior de 25%, 50% y 75% de los datos menores al quartile. La posición esta dada por $n * 0.25, n * 0.50$ y $n * 0.75$ donde n representa el número total de datos. En nuestro ejercicio $n = 24$, por lo cual la posición de los cuartiles es 0.25*24, 0.5*24, 0.75*24 = 6, 12, 18. Que corresponde a los siguiente valores:

   $$Q_1 = 6$$
   $$Q_2 = 8$$
   $$Q_3 = 9$$

   [3, 3, 4, 5, 5, Q_1 **= 6**, 6, 6, 7, 7, 7, Q_2**= 8**, 8, 8, 9, 9, 9, Q_3 **= 9**, 10, 10, 10, 10, 10, 10]

5. **SOLUCION: Interpreta los valores anteriores:**

   La desviación estándar es de 2.25, lo que indica que las calificaciones varían considerablemente con respecto al promedio de 7.45. Esto significa que las calificaciones no están muy concentradas alrededor del promedio; es decir, hay una gran dispersión, con algunas calificaciones mucho más altas y otras mucho más bajas que el promedio.

   Al analizar los cuartiles, obtenemos más información sobre la distribución de las calificaciones. El primer cuartil (Q_1) es 6, lo que indica que el 25% de los estudiantes obtuvo una calificación igual o menor a 6, sugiriendo que un cuarto de los estudiantes tuvo un desempeño bajo en la materia. Sin embargo, el hecho de que el 75% restante de los estudiantes obtuvo calificaciones superiores o iguales a 6 sugiere que la mayoría logró aprobar.

   El segundo cuartil (Q_2), que corresponde a la mediana, es 8. Esto significa que al menos la mitad de los estudiantes obtuvo una calificación igual o superior a 8, lo que refleja un rendimiento sólido y aprobatorio en la materia para una buena parte del grupo.

2.6.3 Ejercicio 3

SOLUCION Datos 3.

1. Los datos ficticios son:
   Malo, Bueno, Bueno, Neutral, Neutral, Bueno, Malo, Bueno, Bueno, Bueno, Malo, Malo, Neutral, Bueno, Bueno, Bueno, Bueno, Bueno, Neutral

2. **SOLUCION: Media, mediana, moda, y mitad de rango.** Los datos ordenados son:

   ['Bueno', 'Bueno', 'Bueno', 'Bueno', 'Bueno', 'Bueno', 'Bueno', 'Bueno', 'Bueno', 'Bueno', 'Bueno", 'Neutral', 'Neutral', 'Neutral', 'Neutral', 'Malo', 'Malo', 'Malo', 'Malo']

   (a) **Media o promedio:** No Aplica. El promedio es solamente para datos numéricos.

   (b) **Mediana:** Tenemos 19 datos, el dato a la mitad es el de la posición 10. Por lo cuál, la mediana es 'Bueno' ya que aparece en la posición número 10.
   ['Bueno', 'Bueno', 'Bueno', 'Bueno', 'Bueno', 'Bueno', 'Bueno', 'Bueno', 'Bueno',**'Bueno'**, 'Bueno", 'Neutral', 'Neutral', 'Neutral', 'Neutral', 'Malo', 'Malo', 'Malo', 'Malo']

   (c) **Moda:** en este caso, tenemos que encontrar el dato que más se repite. La moda es 'Bueno' ya que es la que más se repite con una frecuencia de 11 veces.

   (d) **Mitad de rango:** No Aplica. El promedio es solamente para datos numéricos.

3. **SOLUCIÓN: Interpretación de las medidas de tendencia central:**

   La moda es la categoría que aparece con mayor frecuencia en los datos. En este contexto, "Bueno" es la moda, lo que significa que la mayoría de los encuestados o evaluadores han calificado el producto como "Bueno". Esto sugiere que, en general, el producto tiene una percepción positiva entre los usuarios o consumidores.

   La mediana, en este caso, también es "Bueno", lo que indica que si ordenamos las respuestas de menor a mayor calidad (suponiendo que existe un orden implícito entre las categorías), la calificación "Bueno" es la que ocupa la posición central. Esto refuerza la idea de que "Bueno" es representativo del sentir general de los evaluadores.

4. **SOLUCION: Desviación estándar, varianza y cuartiles**

   (a) **Desviación estándar:** No aplica.

   (b) **Varianza:** No Aplica.

   (c) **Quartiles:** No Aplica.

   (d) **Interpreta los valores anteriores:** No Aplica.

2.7 SOLUCION 2: Medidas relacionales

2.7.1 Ejercicio 1

SOLUCION Datos 1.

Datos: Los datos para el primer ejercicio están dados por:

X	Y
7.4	7.1
4.7	4.4
5.3	5.3
6.5	6.8
17.0	17.7
20.7	20.4

1. **Coeficiente de Correlación de Pearson:** Para calcular el coeficiente de correlación de Pearson, se deben seguir tres pasos clave: primero, calcular la media de cada una de las variables X e Y; segundo, calcular la covarianza entre las variables; y por último, determinar la desviación estándar de cada variable.

   (a) **Paso 1: Calcular el promedio de cada variable.** El primer paso consiste en encontrar el promedio (media) de las variables X e Y.

   $$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{7.4 + 4.7 + 5.3 + 6.5 + 17.0 + 20.7}{6} = \frac{61.6}{6} \approx 10.27$$

   $$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{7.1 + 4.4 + 5.3 + 6.8 + 17.7 + 20.4}{6} = \frac{61.7}{6} \approx 10.28$$

   (b) **Paso 2: Calcular la covarianza.** La convarianza de los datos esta dada por la formula:

   $$cov(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

   Con los promedios calculados, procedemos a calcular la diferencia de las variables con respecto al promedio ($x_i - \bar{x}$ y $y_i - \bar{y}$). Para ello, creamos una tabla que también nos servirá para calcular la desviación estándar.

X	Y	X - $\bar{x}$	Y - $\bar{y}$
7.4	7.1	7.4-10.27 = -2.87	7.1-10.28 = -3.18
4.7	4.4	4.7 - 10.27 = -5.57	4.4-10.28 = -5.88
5.3	5.3	5.3 - 10.27 = -4.97	5.3 -10.28 = -4.98
6.5	6.8	6.5 - 10.27 = -3.77	6.8 - 10.28 = -3.48
17.0	17.7	17-10.27 = 6.73	17.7 -10.28 = 7.42
20.7	20.4	20.7-10.27 =10.43	20.4-10.28 = 10.12

Table 3

Utilizando Tabla 3 podemos calcular la covarianza:

$$
\begin{aligned}
cov(X,Y) &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1} \\
&= \frac{(-2.87*-3.18)+(-5.57*-5.88)+(-4.97*-4.98)+(-3.77*-3.48)+(6.73*7.42)+(10.43*10.12)}{6-1} \\
&= \frac{9.12+32.75+24.75+13.12+49.94+105.55}{5} = \frac{235.23}{5} \\
&\approx 47.05
\end{aligned}
$$

(c) **Paso 3: Calcular la desviación estándar de las variables X e Y.** Utilizando la tabla 3, calculamos las desviaciones estándar de X e Y, que son necesarias para el próximo paso.

$$
\begin{aligned}
s_x &= \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}} \\
&= \sqrt{\frac{(-2.87)^2+(-5.57)^2+(-4.97)^2+(-3.77)^2+(6.73)^2+(10.43)^2}{6-1}} \\
&= \sqrt{\frac{8.24+31.02+24.70+14.21+45.29+108.78}{5}} = \sqrt{\frac{232.24}{5}} \\
&= \sqrt{46.44} \approx 6.81
\end{aligned}
$$

$$
\begin{aligned}
s_y &= \sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}} \\
&= \sqrt{\frac{(-3.18)^2+(-5.88)^2+(-4.98)^2+(-3.48)^2+(7.42)^2+(10.12)^2}{6-1}} \\
&= \sqrt{\frac{10.11+34.55+24.80+12.11+55.06+102.41}{5}} = \sqrt{\frac{239.04}{5}} \\
&= \sqrt{47.81} \approx 6.91
\end{aligned}
$$

(d) **Paso 4: Calcular el coeficiente de correlación.** Con los valores obtenidos, aplicamos la fórmula del coeficiente de correlación de Pearson.

$$r = \frac{cov(X,Y)}{s_x s_y} = \frac{47.05}{6.81*6.91} = \frac{47.05}{47.06} = 0.999$$

Ten en cuenta que los resultados pueden variar ligeramente debido a las pequeñas diferencias en las décimas, dependiendo de cuántos decimales se utilizaron para calcular los valores. El valor más cercano al resultado real es (utilizando precisión doble):

$$r = 0.9983044613502591$$

2.7.2 Ejercicio 1

SOLUCION Datos 2.

Datos: Los datos para el primer ejercicio están dados por:

X	Y
9.1	7.4
7.1	4.7
7.9	5.3
9.1	6.5
10.9	8.3
5.6	5.6

1. **Paso 1: Calcular el promedio de cada variable.** El primer paso consiste en encontrar el promedio (media) de las variables X e Y.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{9.1 + 7.1 + 7.9 + 9.1 + 10.9 + 5.6}{6} = \frac{49.7}{6} \approx 8.28$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{7.4 + 4.7 + 5.3 + 6.5 + 8.3 + 5.6}{6} = \frac{37.8}{6} \approx 6.30$$

2. **Paso 2: Calcular la covarianza.** La convarianza de los datos esta dada por la formula:

$$cov(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

Con los promedios calculados, procedemos a calcular la diferencia de las variables con respecto al promedio ($x_i - \bar{x}$ y $y_i - \bar{y}$). Para ello, creamos una tabla que también nos servirá para calcular la desviación estándar.

X	Y	X - $\bar{x}$	Y - $\bar{y}$
9.1	7.4	9.1-8.28 = 0.82	7.4 - 6.3 = 1.1
7.1	4.7	7.1-8.28 = -1.18	4.7-6.3 = -1.6
7.9	5.3	7.9 - 8.28 = -0.38	5.3-6.3 = -1.0
9.1	6.5	9.1-8.28 =0.82	6.5-6.3 = 0.2
10.9	8.3	10.9-8.28 = 2.62	8.3-6.3 = 2.0
5.6	5.6	5.6-8.28 = -2.68	5.6-6.3 = -0.7

Table 4

Utilizando Tabla 4 podemos calcular la covarianza:

$$\begin{aligned} cov(X, Y) &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n - 1} \\ &= \frac{(0.82 * 1.1) + (-1.18 * -1.6) + (-0.38 * -1.0) + (0.82 * 0.2) + (2.62 * 2.0) + (-2.68 * -0.7)}{6 - 1} \\ &= \frac{0.902 + 1.888 + 0.38 + 0.164 + 5.24 + 1.876}{5} = \frac{10.45}{5} \\ &= 2.09 \end{aligned} \tag{11}$$

3. **Paso 3: Calcular la desviación estándar de las variables X e Y.** Utilizando la Tabla 4, calculamos las desviaciones estándar de X e Y, que son necesarias para el próximo paso.

$$\begin{aligned} s_x &= \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \\ &= \sqrt{\frac{(0.82)^2 + (-1.18)^2 + (-0.38)^2 + (0.82)^2 + (2.62)^2 + (-2.68)^2}{6-1}} \\ &= \sqrt{\frac{0.6724 + 1.3924 + 0.1444 + 0.6724 + 6.8644 + 7.1824}{5}} = \sqrt{\frac{16.9284}{5}} \\ &= \sqrt{3.38568} \approx 1.84 \end{aligned}$$

$$\begin{aligned} s_y &= \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}} \\ &= \sqrt{\frac{(1.1)^2 + (-1.6)^2 + (-1.0)^2 + (0.2)^2 + (2.0)^2 + (-0.7)^2}{6-1}} \\ &= \sqrt{\frac{1.21 + 2.56 + 1 + 0.04 + 4 + 0.49+}{5}} = \sqrt{\frac{9.3}{5}} \\ &= \sqrt{1.86} \approx 1.3638 \end{aligned}$$

Paso 4: Utilizar la formula para calcular el coeficiente de correlación

$$r = \frac{cov(X,Y)}{s_x s_y} = \frac{2.09}{1.84 * 1.3638} = \frac{2.09}{2.509392} \approx 0.83287$$

Ten en cuenta que los resultados pueden variar ligeramente debido a las pequeñas diferencias en las décimas, dependiendo de cuántos decimales se utilizaron para calcular los valores. El valor más cercano al resultado real es: $r = 0.8328517739961736$

2.7.3 Ejercicio 1

SOLUCION 3.

¿Qué conclusiones obtienes de los valores del coeficiente de correlación de Pearson?

- En el primer conjunto de datos, los valores del coeficiente de correlación están muy cerca de 1. Esto indica una fuerte correlación lineal, es decir, los datos se pueden representar fácilmente utilizando la ecuación de una línea.
- En el segundo conjunto de datos, aunque también existe una alta correlación, no se puede describir adecuadamente con una línea recta. Aunque el valor del coeficiente de correlación pueda parecer menor, como 0.83, sigue indicando una correlación muy fuerte.

3 ESTIMADORES E INTERVALOS DE CONFIANZA

En este capítulo, exploraremos cómo la estadística inferencial nos permite realizar estimaciones sobre una población completa utilizando únicamente una muestra representativa. En lugar de analizar a cada individuo de una población grande —lo que a menudo resulta impráctico o incluso imposible— trabajamos con un subconjunto más pequeño y manejable. A partir de este subconjunto, extrapolamos conclusiones que aplicamos a toda la población.

Este enfoque es fundamental en diversas áreas, como la investigación médica, las encuestas políticas y el análisis financiero, donde las decisiones importantes se toman basándose en los resultados obtenidos de muestras.

OBJETIVOS

Al finalizar esta unidad, el estudiante será capaz de:

1. Comprender el concepto de estimador e intervalo de confianza, así como distinguir sus diferencias.
2. Calcular estimadores para proporción, media y desviación estándar a partir de muestras.
3. Calcular intervalos de confianza para proporción, media y desviación estándar con un nivel de confianza determinado.
4. Resolver ejercicios prácticos que involucren el cálculo e interpretación de estimadores e intervalos de confianza.

3.1 ¿Qué es un estimador e intervalo de confianza?

Un estimador y un intervalo de confianza son conceptos fundamentales en estadística inferencial, pero tienen significados y roles diferentes en el proceso de hacer inferencias sobre una población basada en una muestra.

3.1.1 Estimador:

Un estimador es una función o procedimiento utilizado para aproximar un parámetro poblacional (como la media, varianza o proporción) basado en una muestra. En otras palabras, es un valor calculado a partir de los datos de la muestra que sirve como una estimación del valor desconocido del parámetro en la población completa.

Ejemplo:

- Si tenemos una muestra de datos, la media muestral (denotada como $\bar{x}$) es un estimador de la media poblacional (μ).
- Si estamos estimando la proporción de éxito en una población, la proporción muestral ($\hat{p}$) es un estimador de la proporción poblacional (p).

Un estimador puntual es simplemente un valor único que se obtiene al aplicar la fórmula o el método adecuado. Por ejemplo, la media muestral es un estimador puntual de la media poblacional.

3.1.2 Intervalo de Confianza:

Un intervalo de confianza es un rango de valores dentro del cual se cree que se encuentra el verdadero parámetro de la población con una cierta probabilidad o nivel de confianza. El intervalo no da un único valor, sino un rango de valores, lo que refleja la incertidumbre inherente a la estimación basada en una muestra.

Un intervalo de confianza se expresa generalmente como: $[\hat{\theta} - E, \hat{\theta} + E]$, Donde:

$\hat{\theta}$ es el estimador puntual (por ejemplo, la media muestral), E es el margen de error, que depende de la variabilidad de los datos y el tamaño de la muestra. Ejemplo:

- Si calculamos un intervalo de confianza del 95% para la media poblacional, eso significa que estamos 95% seguros de que el valor verdadero de la media poblacional está dentro del rango especificado por el intervalo.

3.1.3 Diferencias clave:

Estimador: Es un valor puntual que estima el parámetro de la población. Ejemplo: la media muestral.

Intervalo de confianza: Es un rango de valores que estima el parámetro de la población, considerando la variabilidad de la muestra. Ejemplo: el intervalo de confianza del 95% para la media poblacional.

En resumen, el estimador proporciona un valor puntual basado en los datos de la muestra, mientras que el intervalo de confianza ofrece un rango de posibles valores para el parámetro poblacional con un nivel de confianza determinado.

3.2 Estimación de Proporciones (p) o Porcentaje de una población dada una muestra

Aquí aprenderemos a calcular el porcentaje o proporción de una población que cumple una característica específica, basándonos en los datos de una muestra. En lugar de estudiar toda la población, examinamos una muestra representativa y, a partir de esos datos, inferimos la proporción en toda la población que cumple con la característica específica que nos interesa. Por ejemplo, si quieres estimar el porcentaje de personas en una ciudad que prefieren transporte público, no es necesario encuestar a cada habitante; basta con una muestra aleatoria bien elegida.

Para utilizar las técnicas descritas en esta sección, se deben de cumplir las siguientes condiciones necesarias para una estimación válida:

1. **Muestra Aleatoria Simple**:La selección de la muestra debe ser aleatoria, asegurando que cada miembro de la población tenga la misma probabilidad de ser elegido. Esto evita sesgos y hace que la muestra sea representativa del conjunto completo.

2. **Distribución Binomial**: Los eventos deben ser categóricos con solo dos resultados posibles: "éxito" (la característica está presente) o "fracaso" (la característica no está presente). Además, las probabilidades de éxito (p) y fracaso ($q = 1 - p$) deben permanecer constantes para cada observación.

3. **Condiciones para la Aproximación Gaussiana**: Para aplicar técnicas basadas en la distribución normal, es necesario que se cumplan: (1) el producto (multiplicación) del número de muestras (n) y la probabilidad de éxito (p) debe ser mayor o igual a 5 ($np \geq 5$), así como el producto del número de muestras (n) y la probabilidad de fracaso (q) también debe ser mayor o igual a 5 ($nq \geq 5$). Estas condiciones son necesarias para aproximar la distribución binomial a una distribución normal, lo que simplifica los cálculos y análisis.

Notación para proporciones

- $p : s$ Proporción o **porcentaje de la población** que cumple la característica de interés.
- $\hat{p} = \frac{x}{n}$: Proporción observada en la muestra, donde x es el número de "éxitos" (o casos que cumplen la característica de interés) en una muestra de tamaño n. En otras palabras, $\hat{p}$ representa el porcentaje estimado basado en los datos de la muestra.
- $\hat{q} = 1 - \hat{p}$: Proporción de "fracasos" en la muestra, es decir, la parte de la muestra que no cumple la característica de interés.

3.2.1 Estimador puntual:

es un valor específico que se calcula a partir de una muestra para aproximar un parámetro desconocido de la población. Este valor único proporciona una estimación directa del parámetro, como la media, proporción o varianza poblacional. Por ejemplo, Si deseamos estimar la proporción de personas que prefieren un producto, el porcentaje $\hat{p}$ obtenido de una muestra de consumidores es un estimador puntual de la proporción verdadera p de toda la población.

La proporción muestral $\hat{p}$ es el mejor estimado puntual en la proporción de la población. Esto significa que, cuando se dispone de una muestra aleatoria de una población, el valor calculado como:

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

donde: x es el número de éxitos (casos que cumplen con la características de interés) y n el tamaño total de la muestra. $\hat{p}$ es la mejor aproximación directa del valor real de la proporción en la población completa.

3.3 Intervalo de confianza descripción detallada

Un intervalo de confianza (IC) es un rango de valores calculado a partir de los datos muestrales, que se utiliza para hacer una estimación informada del valor verdadero de un parámetro de la población.

En otras palabras, un IC proporciona un rango dentro del cual es probable que se encuentre el valor real del parámetro que estamos tratando de estimar (por ejemplo, la proporción o la media poblacional), con un cierto nivel de certeza.

El nivel de confianza es la probabilidad, denotada por $1 - \alpha$ (donde α es el nivel de significancia), y se expresa generalmente como un porcentaje. Representa la frecuencia con la que los intervalos de confianza generados a partir de múltiples muestras contendrían el verdadero valor del parámetro poblacional si se repitiera el proceso de muestreo muchas veces.

Interpretación práctica: Un nivel de confianza del 95% significa que hay un 95% de seguridad de que el intervalo de confianza incluye el valor verdadero del parámetro poblacional. Un nivel de confianza del 99% ofrece una mayor certeza, pero los intervalos calculados suelen ser más amplios.

El nivel de confianza es crucial para medir la fiabilidad de las inferencias estadísticas. Cuanto mayor sea el nivel de confianza, más seguros podemos estar de que el intervalo incluye el parámetro verdadero, pero también más amplio será el rango del intervalo de confianza. Por el contrario, un nivel de confianza más bajo produce intervalos más estrechos, pero con una menor certeza de que contienen el valor verdadero.

El valor crítico: es un número clave en estadística que establece un límite entre resultados que son considerados habituales y aquellos que son inusuales o extremos. Este valor permite determinar qué tan alejados están los datos observados del valor esperado bajo una hipótesis nula. En el caso de una *distribución normal estándar*, el valor crítico $z_{\alpha/2}$ es una puntuación z que delimita un área de probabilidad igual a $\alpha/2$ en la cola derecha de la distribución. Esto implica que solo el $\alpha/2\%$ de los valores estarán en esa región extrema, mientras que el $1 - \alpha\%$ restante se encuentra en la región central. Por ejemplo, Si utilizamos un nivel de confianza del 95% ($\alpha = 0.05$), los valores críticos $z_{\alpha/2}$ corresponderán a las posiciones que separan el 2.5% de los datos en cada cola de la distribución, dejando un área central del 95% donde se espera que caigan la mayoría de los resultados.

Para la distribución normal estándar $N(0, 1)$, el valor critico $z_{\alpha/2}$ satisface:

$$\int_{-\infty}^{z_{\alpha/2}} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx = 1 - \alpha/2$$

Los valores críticos se utilizan para calcular intervalos de confianza y para realizar pruebas de hipótesis, ayudando a evaluar si un resultado es suficientemente extremo como para rechazar una hipótesis nula.

El margen de error (E) El margen de error es una medida clave en estadística inferencial que indica la posible variabilidad de los resultados obtenidos de una muestra aleatoria con respecto al verdadero valor poblacional. Representa la diferencia máxima probable entre la proporción observada en una muestra ($\hat{p}$) y la proporción real de la población (p). El margen de error E Proporciona

una estimación de la precisión de los resultados basados en muestras y permite establecer un rango alrededor de la proporción muestral dentro del cual se espera que se encuentre la verdadera proporción poblacional.

Para calcular el margen de error, multiplicamos el valor crítico $z_{\alpha/2}$ (correspondiente al nivel de confianza deseado) por la desviación estándar de las proporciones muestrales, también conocida como error estándar.

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

Interpretación Práctica: Si calculamos un margen de error E para una muestra con $\hat{p} = 0.6$, un tamaño de muestra $n = 100$ y un nivel de confianza 95% ($z_{0.025} \approx 1.96$), el margen de error se calcularía como:

$$E = 1.96\sqrt{\frac{0.6 * 0.4}{100}} = 1.96 * 0.049 = 0.096$$

Tamaño de muestra: Muestras más grandes n reducen el margen de error (E)y mejoran la precisión de la estimación estadística. Esto se debe a que, al aumentar el tamaño de la muestra, la variabilidad de las proporciones muestrales disminuye, lo que permite obtener estimaciones más cercanas al verdadero parámetro poblacional. En términos prácticos, si deseamos reducir el margen de error, es fundamental aumentar el tamaño de la muestra. Aunque aumentar el tamaño de la muestra mejora la precisión, también implica costos y tiempos mayores, por lo que es crucial equilibrar la precisión deseada con los recursos disponibles para el estudio.

3.4 Intervalo de confianza para la proporción poblacional p

El intervalo de confianza para una proporción poblacional se puede expresar de las siguientes formas equivalentes:

$$\hat{p} - E < p < \hat{p} + E$$

o

$$(\hat{p} - E, \hat{p} + E)$$

o también más compacto como

$$\hat{p} \pm E$$

Donde:

- $\hat{p}$ es la proporción muestral, es decir, el estimador puntual para la proporción poblacional p.
- E es el margen de error, que determina la distancia entre la proporción muestral $\hat{p}$ y los límites del intervalo.

3.4.1 Procedimiento para construir el intervalo de confianza de proporciones

Para calcular el intervalo de confianza para una proporción poblacional, sigue estos pasos:

1. **Calcular proporción muestral $\hat{p}$:**

   Determina la proporción de éxitos en la muestra, que se calcula dividiendo el número de éxitos observados (x) entre el tamaño de la muestra (n):

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

2. **Verificar que los condiciones para construir el intervalo de confianza se cumplen:**

   Asegúrate de que las condiciones necesarias para usar un intervalo de confianza de proporciones estén cumplidas, como que la muestra sea aleatoria y que $n * \hat{p} \geq 5$ y $n * (1 - \hat{p}) \geq 5$ para que se pueda aplicar una aproximación normal.

3. **Determinar el Nivel de Confianza y Encontrar el Valor Crítico**:

   Decide el nivel de confianza que deseas para tu intervalo de confianza (por ejemplo, 95% o 99%) y encuentra el valor crítico $z_{\alpha/2}$ que corresponde a la puntuación z en la distribución normal estándar. Este valor se puede consultar en tablas estadísticas o calcular mediante software especializado.

4. **Calcular el margen de error E:**

   Utiliza el valor crítico y la desviación estándar de la proporción muestral para calcular el margen de error:

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

5. **Construir el Intervalo de Confianza:** Finalmente, establece los límites superior e inferior del intervalo de confianza sumando y restando el margen de error a la proporción muestral:

$$(\hat{p} - E, \hat{p} + E)$$

3.4.2 Determinación del tamaño de muestra n

Si no se conoce la proporción poblacional, el tamaño de muestra n se puede estimar utilizando el siguiente fórmula, que parte de un valor conservador de 0.25 para maximizar el tamaño de la muestra:

$$n = \frac{0.25(z_{\alpha/2})^2}{E^2}$$

Este cálculo proporciona el tamaño de muestra necesario para estimar la proporción poblacional con el margen de error especificado y el nivel de confianza deseado.

3.5 Estimación de la media o promedio (μ) de una población cuando σ es desconocida

En esta sección, aprenderemos cómo estimar el promedio de una población. En lugar de estudiar toda la población, lo cual podría ser prácticamente imposible debido a su tamaño, tomamos una muestra representativa. Utilizamos los datos obtenidos de esta muestra para hacer nuestras estimaciones. De esta forma, inferimos el promedio de toda la población basándonos en los datos de una muestra cuidadosamente seleccionada.

Para utilizar las técnicas descritas en esta sección, se deben de cumplir las siguientes condiciones necesarias para una estimación válida:

1. **Muestra Aleatoria Simple**:

   La selección de la muestra debe ser aleatoria, asegurando que cada miembro de la población tenga la misma probabilidad de ser elegido. Esto evita sesgos y hace que la muestra sea representativa del conjunto completo.

2. **La muestra debe provenir de una distribución que sea aproximadamente normal o tener un tamaño de muestra $n \geq 30$**:

   Si el tamaño de la muestra es suficientemente grande (mayor que 30), podemos aplicar el Teorema Central del Límite, lo que nos permite hacer inferencias sobre el promedio de la población aunque la distribución original de la población no sea normal.

La media muestral $\bar{x}$ es el mejor estimado puntual de la media poblacional μ. ya que proporciona una estimación precisa del valor promedio de toda la población basada en los datos obtenidos de una muestra representativa.

3.6 Intervalo de confianza para la media poblacional μ

Cuando estimamos la media de una población (μ) a partir de una muestra, es importante tener en cuenta que el valor que obtenemos de la muestra ($\bar{x}$) es solo una aproximación de la verdadera media de la población. Dado que las muestras son solo subconjuntos de la población, siempre habrá cierta incertidumbre sobre el valor exacto de μ. Un intervalo de confianza nos permite expresar esta incertidumbre de manera cuantitativa, proporcionando un rango de valores dentro del cual es probable que se encuentre la media poblacional real.

3.6.1 Valor crítico $t_{\alpha/2}$

Cuando deseamos estimar la media de una población pero **no conocemos la desviación estándar** σ de esa población, recurrimos a la distribución t de Student. Esto se debe a que, cuando la desviación estándar de la población es desconocida, la estimación de la media se ve influenciada por la variabilidad en la muestra, lo que hace que la distribución t sea adecuada para reflejar esa incertidumbre adicional. La distribución t de Student se utiliza para calcular valores críticos (denotados como $t_{\alpha/2}$) que nos permiten construir intervalos de confianza para la media poblacional

Los valores críticos $t_{\alpha/2}$ dependen del nivel de confianza que elijamos y del tamaño de la muestra n. A medida que aumentamos el tamaño de la muestra, la distribución t se aproxima a la normal, pero para muestras pequeñas, la distribución t tiene colas más gruesas, lo que refleja una mayor incertidumbre sobre el valor estimado de la media.

3.6.2 Margen de error E para la estimación de μ

El margen de error E para estimar la media poblacional μ se calcula con la siguiente fórmula:

$$E = t_{\alpha/2}\frac{s}{\sqrt{n}}$$

donde s es la desviación estándar de la muestra y $t_{\alpha/2}$ es el valor crítico obtenido de la distribución t de Student, el cuál se determina utilizando $n-1$ grados de libertad, donde n es el tamaño de la muestra.

3.6.3 Construcción de intervalo de confianza

El intervalo de confianza para la media poblacional μ se calcula de la siguiente manera:

$$\bar{x} - E < \mu < \bar{x} + E$$

donde $\bar{x}$ es la media muestral y E es el margen de error. El intervalo de confianza también se puede expresar de diferentes maneras, pero todas representan la misma idea. Además de la forma estándar:

$$(\bar{x} - E, \bar{x} + E)$$

$$\bar{x} \pm E$$

3.6.4 Interpretación del intervalo de confianza

El intervalo de confianza nos da un rango de valores dentro del cual, con un nivel de confianza determinado (por ejemplo, 95%), se espera que se encuentre el valor real de la media poblacional μ. Si se repitiera el proceso de muestreo muchas veces, aproximadamente el 95% de los intervalos de confianza construidos de esta manera contendrían la verdadera media poblacional.

Es importante recordar que el nivel de confianza indica qué tan seguro estamos de que el intervalo calculado incluye el valor verdadero de la media poblacional. Un nivel de confianza del 95% significa que, si repitieras el procedimiento de muestreo 100 veces, aproximadamente 95 de esos intervalos de confianza incluirían el valor verdadero de μ.

3.7 Estimación de Varianza o Desviación Estándar de una Población

En esta sección, aprenderemos cómo estimar la varianza (o desviación estándar) de una población. Dado que estudiar toda la población puede ser prácticamente imposible, recurrimos a una muestra representativa de la población y usamos los datos obtenidos de esa muestra para realizar nuestras estimaciones. Es decir, inferimos el valor de la varianza de toda la población basándonos en los datos de una muestra cuidadosamente seleccionada.

Para aplicar las técnicas de estimación que describimos en esta sección, es importante que se cumplan los siguientes requisitos:

1. **Muestra Aleatoria Simple**:

   La selección de la muestra debe ser aleatoria, asegurando que cada miembro de la población tenga la misma probabilidad de ser elegido. Esto evita sesgos y hace que la muestra sea representativa del conjunto completo.

2. **La población debe tener una distribución normal**:

   Si la población es normal, se puede usar la distribución Chi-cuadrada para estimar la varianza y la desviación estándar de manera precisa. Además, en una distribución normal, la media y la varianza están bien definidas y no hay distribuciones sesgadas o asimétricas que puedan alterar los resultados.

La varianza muestral s^2 es el mejor estimado puntual de la varianza poblacional σ^2.

La desviación estándar muestral s suele utilizarse como un estimado puntual de σ (aunque es un estimado sesgado). . Es decir, la desviación estándar muestral tiene una ligera tendencia a subestimar la desviación estándar real de la población, especialmente cuando el tamaño de la muestra es pequeño.

3.8 Intervalo de confianza Varianza (σ^2) y Desviación Estándar (σ)

Uso de la distribución Chi-cuadrada Cuando queremos estimar el intervalo de confianza para la varianza de una población, utilizamos la distribución Chi-cuadrada (X^2). La distribución Chi-cuadrada es adecuada porque está basada en la suma de los cuadrados de variables aleatorias independientes y distribuidas de manera normal. Es la distribución que se emplea para estimar la varianza y para construir intervalos de confianza sobre la varianza poblacional.

3.8.1 Valores críticos:

Cuando calculamos un intervalo de confianza para la varianza poblacional, necesitamos encontrar los valores críticos en la distribución Chi-cuadrada. Estos valores se obtienen al dividir el área total $(1 - \alpha)$ de la distribución Chi-cuadrada de manera equitativa entre las dos colas de la distribución (izquierda y derecha).

- El **valor crítico de la cola izquierda (X_I^2)** marca el punto donde comienza el área asignada a la cola izquierda, que tiene un área de $(\alpha/2)$
- El **valor crítico de la cola derecha (X_D^2)** marca el punto donde comienza el área asignada a la cola derecha, con un área de $\alpha/2$

Estos valores críticos son de suma importancia ya que se utilizan para construir el intervalo de confianza para la varianza y la desviación estándar de la población.

3.8.2 Intervalo de confianza para la varianza poblacional σ^2

El intervalo de confianza para la varianza poblacional σ^2 se calcula utilizando los valores críticos X_I^2 y X_D^2 de la distribución Chi-cuadrada, de acuerdo con la siguiente formula:

$$\frac{(n-1)s^2}{X_D^2} < \sigma^2 < \frac{(n-1)s^2}{X_I^2}$$

Este intervalo de confianza proporciona un rango dentro del cual se espera que se encuentre la verdadera varianza poblacional σ^2.

El intervalo de confianza puede expresarse en los siguientes formatos equivalentes:

$$(\frac{(n-1)s^2}{X_D^2}, \frac{(n-1)s^2}{X_I^2})$$

3.8.3 Intervalo de confianza para la desviación estándar poblacional σ

Para obtener un intervalo de confianza para la desviación estándar poblacional σ, simplemente calculamos la raíz cuadrada de los límites inferior y superior del intervalo de confianza para la varianza. Es decir:

$$\sqrt{\frac{(n-1)s^2}{X_D^2}} < \sigma < \sqrt{\frac{(n-1)s^2}{X_I^2}}$$

De esta manera, podemos estimar el rango en el que se encuentra la verdadera desviación estándar de la población con un nivel de confianza determinado.

3.9 Muestreo repetido por bootstrap

El método bootstrap es una técnica estadística utilizada para estimar la distribución de una estadística de interés (como la media, la mediana, la varianza, etc.) a partir de los datos de una muestra. Este método es especialmente útil cuando los métodos estadísticos tradicionales son difíciles de aplicar o cuando no podemos hacer suposiciones sobre la distribución de los datos.

A continuación, se explica de manera detallada cómo funciona el método bootstrap:

1. **Recopilación de Datos**:

   Comenzamos con un conjunto de datos que representa una muestra de la población de interés.

2. **Resampling:**

   En lugar de hacer suposiciones sobre la población subyacente, generamos múltiples muestras de bootstrap a partir de la muestra original mediante el proceso de remuestreo con reemplazo. Esto significa que tomamos aleatoriamente observaciones de la muestra original y las colocamos de vuelta en la muestra antes de seleccionar la siguiente observación. Como resultado, algunas observaciones aparecerán varias veces en una muestra de bootstrap, mientras que otras pueden no aparecer en absoluto.

3. **Cálculo de la Estadística de Interés**: Para cada una de estas muestras de bootstrap, calculamos la estadística que queremos estimar (como la media, la mediana, la varianza, etc.). Esto nos da una colección de valores de la estadística calculados a partir de las muestras de bootstrap.

4. **Estimación de la Distribución y el Error Estándar:**

   Usamos la colección de valores obtenidos para estimar la distribución de la estadística de interés y calcular el error estándar. Además, los percentiles de esta distribución pueden utilizarse para construir intervalos de confianza.

3.9.1 Ventajas y limitaciones del bootstrap

El método bootstrap es potente porque no requiere hacer suposiciones fuertes sobre la forma de la población subyacente y puede proporcionar estimaciones precisas incluso con muestras pequeñas. Sin embargo, es importante tener en cuenta que el método bootstrap no puede corregir las limitaciones de los datos originales. Si los datos de la muestra son sesgados o incompletos, las estimaciones obtenidas mediante bootstrap también reflejarán esas limitaciones.

3.10 EJERCICIOS:

1. Un investigador desea estimar la proporción de personas mayores de 40 años en una población. Para ello, selecciona una muestra aleatoria de 100 personas y encuentra que 30 de ellas tienen más de 40 años. Con base en esta información, calcule el intervalo de confianza para la proporción real de personas mayores de 40 años en la población, utilizando un nivel de confianza del 96%.

2. Un estudio busca estimar la estatura promedio de una población. Para ello, se seleccionó aleatoriamente una muestra de 500 personas y se midió la estatura de cada individuo. Los resultados muestran que la estatura promedio en la muestra es de 1.72 metros, con una desviación estándar de 0.2 metros. Utilizando esta información:

   - calcule el intervalo de confianza para la estatura promedio de la población con un nivel de confianza del 99%.
   - ¿Cuál es el intervalo de confianza si solamente se hubieran seleccionado 20 personas?

3. Un grupo de investigadores está interesado en estimar la variabilidad en las estaturas de una población. Para ello, seleccionaron aleatoriamente una muestra de 80 personas y calcularon que la desviación estándar de las estaturas en la muestra es de 0.2 metros.

   Pregunta: Utilizando un nivel de confianza del 90%, determine el intervalo de confianza para la desviación estándar poblacional. Además, interprete cómo este intervalo nos ayuda a entender la variabilidad real de las estaturas en toda la población.

3.11 SOLUCIONES

3.11.1 Ejercicio 1.

Problema: Un investigador desea estimar la proporción de personas mayores de 40 años en una población. Para ello, selecciona una muestra aleatoria de 100 personas y encuentra que 30 de ellas tienen más de 40 años. Con base en esta información, calcule el intervalo de confianza para la proporción real de personas mayores de 40 años en la población, utilizando un nivel de confianza del 96%.

SOLUCION EJERCICIO 1:

Paso 1: Calcular la población muestral . Para obtener el intervalo de confianza necesitamos obtener la proporción muestral $\hat{p}$, se calcula dividiendo el número de éxitos (# mayores de 40 años) entre el tamaño total de la muestra (n):

$$\hat{p} = \frac{\#exitos}{n} = \frac{30}{100} = 0.30,$$

Esto significa que, en la muestra, el 30% de las personas tienen más de 40 años.

Paso 2: Verificar las condiciones para construir el intervalo de confianza Para calcular un intervalo de confianza para una proporción, necesitamos asegurarnos de que:

$$n * \hat{p} > 5 \text{ y } n * (1 - \hat{p}) > 5$$

Para los datos dados, se tiene que:

$$n\hat{p} = 100 * 0.3 = 30 > 5 \text{Si se cumple}$$

$$n(1 - \hat{p}) = 100 * 0.7 = 70 > 5 \text{Si se cumple}$$

Ambas condiciones se cumplen, por lo que podemos proceder.

Paso 3: Determinar el valor crítico $z_{\alpha/2}$

1. **Nivel de confianza y área de las colas:**

   El nivel de confianza es del 96%, lo que significa que el área central bajo la curva normal estándar es 0.96. Esto deja una probabilidad de $\alpha = 1 - 0.96 = 0.04$ distribuida en las dos colas de la distribución.

2. **Área de una sola cola:**

   Debido a que solo necesitamos el área de una cola, tenemos que dividir la probabilidad de las colas en dos, ya que la distribución normal estándar es simétrica. Esto nos da:

   $$\frac{\alpha}{2} = \frac{0.04}{2} = 0.02$$

3. **Buscar valor crítico** $z_{\alpha/2}$

   Usamos una tabla de valores z de la distribución normal para encontrar el puntaje z que deja un área acumulada de 0.02 en la cola izquierda. El valor crítico correspondiente es:

   $$z_{\alpha/2} \approx -2.05$$

   Nota: El valor no es exacto, ya que el área acumulada asociada con $z_{\alpha/2} \approx -2.05$ es aproximadamente 0.021, que es cercano al 0.2 que buscamos.

Paso 4: Calcular el margen de error E

El margen de error E nos indica cuánto podría variar nuestra estimación de la proporción muestral $\hat{p}$ con respecto a la proporción poblacional p. Para calcularlo, utilizamos la fórmula específica para las proporciones:

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

Donde:

- $z_{\alpha/2} = 2.05$ (valor critico obtenido en el paso anterior).
- $\hat{p} = 0.30$ (la proporción muestral, ya que de 30 de las 100 personas son mayores de 40 años).
- $n = 100$(tamaño de muestra).

Ahora sustituimos estos valores en la fórmula:

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 2.05\sqrt{\frac{(0.3)(0.7)}{100}} = 2.05\sqrt{0.0021} \approx 0.09394$$

El margen de error es $E \approx 0.09394$, este margen de error nos indica cuánto puede variar nuestra estimación de la proporción de la población y es fundamental para construir el intervalo de confianza

Paso 5: Construir el intervalo de confianza Ahora que tenemos el margen de error $E = 0.09394$, podemos calcular el intervalo de confianza para la proporción poblacional p. El intervalo de confianza se obtiene sumando y restando el margen de error E a la proporción muestral $\hat{p}$, de la siguiente manera:

$$\hat{p} - E < p < \hat{p} + E$$

Sustituimos los valores conocidos:

$$0.3 - 0.09394 < p < 0.3 + 0.09394$$

Realizamos las operaciones:

$$0.20606 < p < 0.39394 \quad \text{con un nivel de confianza de } 96\%$$

Interpretación: Con un nivel de confianza del 96%, podemos afirmar que la proporción de personas mayores de 40 años en la población se encuentra entre el 20.6% y el 39.4%. Esto significa que, si repetimos este proceso muchas veces, en el 96% de los casos, el intervalo calculado contendría la proporción verdadera de la población.

3.11.2 Ejercicio 2.

Problema: Un estudio busca estimar la estatura promedio de una población. Para ello, se seleccionó aleatoriamente una muestra de 500 personas y se midió la estatura de cada individuo. Los resultados muestran que la estatura promedio en la muestra es de 1.72 metros, con una desviación estándar de 0.2 metros. Utilizando esta información:

- A) calcule el intervalo de confianza para la estatura promedio de la población con un nivel de confianza del 99%.
- B) ¿Cuál es el intervalo de confianza si solamente se hubieran seleccionado 20 personas?

SOLUCION EJERCICIO 2: problema A.

Paso 1: Identificar los datos que se nos proporcionan. En este caso, se nos proporciona la siguiente información:

- $n = 500$: El tamaño de la muestra
- $\bar{x} = 1.72$: . Este valor se obtuvo al calcular el promedio de las estaturas de las 500 personas seleccionadas aleatoriamente y obtener su promedio muestral.
- $s = 0.2$: la desviación estandar de la muestra.
- El nivel de confianza es del 99%, por lo tanto $1-\alpha = 0.99$, lo que implica que $\alpha = 1-0.99 = 0.01$

Con estos datos, podemos proceder a calcular el intervalo de confianza para la media poblacional μ
Paso 2: Verificación de los criterios. En este paso, verificamos que se cumplan los criterios necesarios para construir el intervalo de confianza. En nuestro caso, el criterio principal es:

- $n > 30$: El tamaño de la muestra debe ser suficientemente grande para aplicar la distribución normal. En este caso, se nos proporciona $n = 500$, lo que satisface ampliamente este criterio, ya que $500 > 30$

Por lo tanto, podemos proceder a calcular el intervalo de confianza para la media poblacional.
Paso 3: Determinar el valor crítico $t_{\alpha/2}$
Para este paso, necesitamos calcular el valor crítico de la distribución t-Student. Dado que el nivel de confianza es del 99%, sabemos que $\alpha = 0.01$, por lo que dividimos α entre 2 para determinar el área que se encuentra en una de las colas de la distribución normal estándar.

- **Calculo de $\alpha/2$:**
$$\frac{\alpha}{2} = \frac{0.01}{2} = 0.005$$
Esto nos da el área de la cola izquierda de la distribución.

- **Usamos la tabla de valores críticos t-Student**: Buscamos en la tabla de puntuaciones $t-$Student el valor que corresponde a un área de 0.005 en la cola izquierda de la distribución. Como el número de grados de libertad es $n - 1 = 500 - 1 = 499$, pero la tabla no tiene un valor exacto para 499 grados de libertad, tomamos el valor más cercano, que corresponde a 500 grados de libertad.

De acuerdo con la tabla, el valor crítico para $t_{\alpha/2}$ con 500 grados de libertad y un área de 0.005 en la cola izquierda es:
$$t_{\alpha/2} = 2.586$$
Este valor es el que usaremos en los siguientes cálculos para determinar el margen de error y el intervalo de confianza.
Paso 4: Calculo de margen de error E medias Para calcular el margen de error, utilizamos la fórmula específica para las medias, que es:
$$E = t_{\alpha/2}\frac{s}{\sqrt{n}}$$
Donde:

1. $t_{\alpha/2} = 2.586$ es el valor crítico obtenido en el paso anterior.
2. $s = 0.2$ es la desviación estándar de la muestra.
3. $n = 500$ es el tamaño de la muestra

Sustituyendo estos valores nos da:
$$E = t_{\alpha/2}\frac{s}{\sqrt{n}} = 2.586\frac{0.2}{\sqrt{500}} \approx 0.5172 * 0.04472 \approx 0.02313$$
Este margen de error de aproximadamente 0.02313 nos indica cuánto puede variar nuestra estimación de la media poblacional en función de los datos de la muestra. El margen de error es esencial para construir el intervalo de confianza, ya que nos ayuda a definir el rango dentro del cual podemos estar razonablemente seguros de que se encuentra la media poblacional.
Paso 5: Construcción del intervalo de confianza
Ahora que hemos calculado el margen de error (E), podemos construir el intervalo de confianza para la media poblacional (μ) utilizando la fórmula:
$$(\bar{x} - E, \bar{x} + E)$$

Sustituyendo los valores:

$$(\bar{x} - E, \bar{x} + E)$$

$$1.72 - 0.02313 < \mu < 1.72 + 0.02313$$

$$1.6969 < \mu < 1.7431 \quad \text{con un nivel de confianza de } 99\%$$

Interpretación: El intervalo de confianza obtenido,

$$1.6969 < \mu < 1.7431,$$

nos indica que, si repitiésemos el experimento en muchas ocasiones, en el 99% de los casos la media poblacional de las estaturas de las personas de esta población se encontraría entre 1.6969 y 1.7431 metros.

Esto significa que, aunque no conocemos la media exacta de toda la población, basándonos en nuestra muestra de 500 personas, podemos afirmar con un alto nivel de confianza que la media de estaturas de la población general está dentro de este rango.

SOLUCION EJERCICIO 2: problema B.

Para el problema B, se requiere que asumamos que el tamaño de la muestra es $n = 100$ en lugar de $n = 500$. Es decir, asumimos que la media y la desviación estándar muestral se obtuvieron a partir de una muestra de solo 100 personas seleccionadas de manera aleatoria.

Paso 3: Determinar el valor crítico $t_{\alpha/2}$

Para este paso, necesitamos calcular el valor crítico de la distribución t-Student. Dado que el nivel de confianza es del 99%, sabemos que $\alpha = 0.01$, por lo que dividimos α entre 2 para determinar el área que se encuentra en una de las colas de la distribución normal estándar.

- **Calculo de $\alpha/2$:**

  $$\frac{\alpha}{2} = \frac{0.01}{2} = 0.005$$

  Esto nos da el área de la cola izquierda de la distribución.

- **Usamos la tabla de valores críticos t-Student**: Buscamos en la tabla de puntuaciones $t-$Student el valor que corresponde a un área de 0.005 en la cola izquierda de la distribución. Como el número de grados de libertad es $n - 1 = 100 - 1 = 99$, pero la tabla no tiene un valor exacto para 99 grados de libertad, tomamos el valor más cercano, que corresponde a 100 grados de libertad.

De acuerdo con la tabla, el valor crítico para $t_{\alpha/2}$ con 100 grados de libertad y un área de 0.005 en la cola izquierda es:

$$t_{\alpha/2} = 2.626$$

Este valor es el que usaremos en los siguientes cálculos para determinar el margen de error y el intervalo de confianza.

Paso 4: Calculo de margen de error E medias Para calcular el margen de error, utilizamos la fórmula específica para las medias, que es:

$$E = t_{\alpha/2} \frac{s}{\sqrt{n}}$$

Donde:

1. $t_{\alpha/2} = 2.626$ es el valor crítico obtenido en el paso anterior.
2. $s = 0.2$ es la desviación estándar de la muestra.
3. $n = 100$ es el tamaño de la muestra

Sustituyendo estos valores nos da:

$$E = t_{\alpha/2}\frac{s}{\sqrt{n}} = 2.626\frac{0.2}{\sqrt{100}} \approx 2.626 * 0.02 \approx 0.04524$$

Este margen de error de aproximadamente 0.04524 nos indica cuánto puede variar nuestra estimación de la media poblacional en función de los datos de la muestra. El margen de error es esencial para construir el intervalo de confianza, ya que nos ayuda a definir el rango dentro del cual podemos estar razonablemente seguros de que se encuentra la media poblacional.

Paso 5: Construcción del intervalo de confianza

Ahora que hemos calculado el margen de error (E), podemos construir el intervalo de confianza para la media poblacional (μ) utilizando la fórmula:

$$(\bar{x} - E, \bar{x} + E)$$

Sustituyendo los valores:

$$(\bar{x} - E, \bar{x} + E)$$
$$1.72 - 0.04524 < \mu < 1.72 + 0.04524$$
$$1.67476 < \mu < 1.76524 \quad \text{con un nivel de confianza de } 99\%$$

Conparación intervalo de confianza: Problema A ($n = 500$) vs Problema B ($n = 100$).

- Para $n = 100$: El intervalo de confianza es:

  $$1.67476 < \mu < 1.76524$$

  con un nivel de confianza del 99%. Este intervalo es relativamente amplio, lo que refleja una mayor incertidumbre en la estimación de la media poblacional. A pesar de que el intervalo incluye el valor de la media estimada ($\bar{x} = 1.72$), tiene un rango más extenso, lo que es típico cuando el tamaño de la muestra es pequeño.

- Para $n = 500$ El intervalo de confianza es:

  $$1.6969 < \mu < 1.7431$$

  con un nivel de confianza del 99%. Este intervalo es más estrecho, lo que indica que, con una muestra más grande, la estimación de la media poblacional es más precisa. El margen de error es menor, lo que refleja una mayor certeza sobre el valor de la media de la población.

Conclusiones:

- **Efecto del tamaño de la muestra:** La comparación entre los dos intervalos de confianza muestra claramente que, a medida que aumenta el tamaño de la muestra, el intervalo de confianza se hace más estrecho. Esto es un resultado esperado porque con una muestra más grande, tenemos más información y, por lo tanto, una estimación más precisa de la media poblacional.
- **Mayor incertidumbre con muestras pequeñas:** El intervalo más amplio para $N = 100$ refleja la mayor variabilidad inherente a las estimaciones basadas en una muestra más pequeña. Con una muestra más pequeña, hay más incertidumbre sobre la estimación de la media poblacional, y es por esto que el intervalo de confianza es más amplio.
- **Mayor precisión con muestras grandes**: Con $N = 500$, el intervalo de confianza es más estrecho y preciso. Esto significa que la estimación de la media poblacional es más confiable y el margen de error es más pequeño. Esto también sugiere que, con muestras grandes, tenemos mayor certeza de que la media de la población está dentro de ese rango.
- **Conclusión general**: Estos resultados subrayan la importancia de contar con un tamaño de muestra adecuado en estudios estadísticos. Cuanto mayor sea el tamaño de la muestra, mayor será la precisión de la estimación y más pequeño será el margen de error.

3.11.3 Ejercicio 3.

Problema: Un grupo de investigadores está interesado en estimar la variabilidad en las estaturas de una población. Para ello, seleccionaron aleatoriamente una muestra de 80 personas y calcularon que la desviación estándar de las estaturas en la muestra es de 0.2 metros.

Pregunta: Utilizando un nivel de confianza del 90%, determine el intervalo de confianza para la desviación estándar poblacional. Además, interprete cómo este intervalo nos ayuda a entender la variabilidad real de las estaturas en toda la población.

SOLUCION Ejercicio 3.

Paso 1 Identificar los datos que se nos proporcionan: En este caso, se nos proporciona la siguiente información:

- $n = 80$: el tamaño de la muestra.
- $s = 0.2$: la desviación estándar de la muestra.
- El nivel de confianza es 90%, lo que implica que $1 - \alpha = 0.90$, de donde obtenemos $\alpha = 0.10$

Paso 2: Verificación de los criterios para construir el intervalo de confianza

Para construir un intervalo de confianza para la desviación estándar, necesitamos suponer que los datos siguen una distribución normal. Aunque no se menciona explícitamente que los datos sigan una distribución normal, asumimos que esta condición se cumple para poder continuar con el cálculo. Si no tuviéramos esta información, deberíamos utilizar un enfoque diferente, como el método de **bootstrap**, que no requiere suposiciones sobre la distribución de los datos.

Paso 3: Encontrar los valores críticos de la distribución chi-cuadrada

En este paso, necesitamos encontrar los valores críticos de la distribución chi-cuadrada X_I^2 y X_D^2 para un nivel de confianza del 90% (lo que implica un valor de $\alpha = 0.10$). Para calcular estos valores críticos, primero dividimos α por 2, ya que la distribución chi-cuadrada se divide en dos colas (izquierda y derecha) de igual área. Esto nos da:

$$\frac{\alpha}{2} = \frac{0.10}{2} = 0.05$$

Ahora, necesitamos buscar los valores críticos correspondientes a un área de 0.05 en las colas izquierda y derecha de la distribución chi-cuadrada. Dado que tenemos una muestra de tamaño $n = 80$, los grados de libertad serán:

$$n = 80 - 1 = 79$$

Usamos la tabla de la distribución chi-cuadrada para encontrar los valores correspondientes a un área de 0.05 en las colas de la distribución chi-cuadrada para 79 grados de libertad.

- El valor crítico para la cola izquierda X_I^2 con un área de 0.05 es: $X_I^2 = 60.391$
- El valor crítico para la cola derecha X_D^2 con un área de 0.05 es: $X_D^2 = 101.879$

Nota: Dado que la tabla no tiene valores exactos para $n = 79$, utilizamos el valor correspondiente para $n = 80$, ya que es muy cercano.

Paso 4: Construir el intervalo de confianza para la varianza poblacional

Usamos la fórmula para calcular el intervalo de confianza para la varianza poblacional (σ^2) utilizando la distribución chi-cuadrada:

$$\frac{(n-1)s^2}{X_D^2} < \sigma^2 < \frac{(n-1)s^2}{X_I^2}$$

Sustituyendo los valores que ya tenemos:

$$\frac{(80-1)(0.2)^2}{101.879} < \sigma^2 < \frac{(80-1)(0.2)^2}{60.39}$$

Lo que se simplifica a:

$$\frac{(79)(0.04)}{101.879} < \sigma^2 < \frac{(79)(0.04)}{60.39}$$

Calculamos los valores:

$$0.031017 < \sigma^2 < 0.052326 \quad \text{con un nivel de confianza de } 90\%$$

Interpretación:
El intervalo de confianza obtenido es:

$$0.031017 < \sigma^2 < 0.052326$$

Esto significa que, con un nivel de confianza del 90%, podemos afirmar que la varianza de las estaturas de la población de donde se extrajo la muestra está en algún lugar entre 0.031017 y 0.052326.

Es decir, si repitieramos este proceso muchas veces, 90 de cada 100 intervalos de confianza calculados con muestras similares de tamaño 80 incluirían el valor verdadero de la varianza poblacional. Este intervalo refleja la precisión con la que estimamos la varianza de la población a partir de los datos de la muestra.

Paso 5 (Opcional):

Si deseas calcular el intervalo de confianza para la desviación estándar en lugar de la varianza, simplemente debes tomar la raíz cuadrada de ambos límites del intervalo de la varianza. Esto te proporcionará el intervalo de confianza para la desviación estándar:

$$\sqrt{0.031017} < \sqrt{\sigma^2} < \sqrt{0.052326} \quad \text{con un nivel de confianza de } 90\%$$

Esto nos da:

$$0.1761 < \sigma < 0.2287 \quad \text{con un nivel de confianza de } 90\%$$

Este intervalo nos indica que, con un 90% de confianza, la desviación estándar de la población se encuentra entre 0.1761 y 0.2287.

4 INFERENCIA ESTADISTICA: UNA MUESTRA

OBJETIVOS

Al finalizar esta unidad, el estudiante será capaz de:

1. Comprender el concepto de prueba de hipótesis e identificar correctamente la hipótesis nula y la hipótesis alternativa.
2. Calcular el estadístico de prueba para una muestra en el caso de proporciones, medias y desviación estándar.
3. Identificar y analizar los elementos clave de una prueba de hipótesis: región crítica, nivel de significancia, valor crítico y valor-p.
4. Aplicar los conceptos de inferencia estadística para resolver problemas prácticos con una sola muestra, interpretando adecuadamente los resultados.

4.1 Pruebas de Hipótesis

4.2 Definiciones básicas:

4.2.1 Hipótesis:

En estadística, una hipótesis es simplemente una declaración que hacemos acerca de un parámetro o característica de una población. Es como hacer una suposición sobre algo que no conocemos con certeza. Por ejemplo, podríamos hipotetizar que la media de estaturas de los estudiantes de la UASLP es 1.65 metros.

4.2.2 Prueba de hipótesis:

Ahora, una prueba de hipótesis es un procedimiento que usamos para evaluar si nuestra hipótesis sobre la población es válida o no. Imagina que tienes una idea acerca de una población, como que su media es 50. La prueba de hipótesis te ayuda a determinar si los datos que has recolectado apoyan esa afirmación o si tal vez estás equivocado. Básicamente, se trata de un proceso sistemático para comprobar si una suposición es razonable según los datos disponibles.

4.3 Hipótesis nula y Hipótesis alternativa:

4.3.1 Hipótesis Nula

La hipótesis nula (denotada por H_0) es una afirmación inicial sobre una población que asumimos como verdadera hasta que se demuestre lo contrario con evidencia estadística. Es como un punto de partida. Por ejemplo, podemos suponer que no hay diferencia entre dos grupos, o que la media de una población es igual a cierto valor. Para hacerlo más claro:

- $H_0 : p = 0.3$ significa que se está asumiendo que la proporción de una población es 0.3
- $H_0 : \mu = 1.75$ implica que se está asumiendo que la media de una población es 1.75.
- $H_0 : \sigma = 1.2$ indica que la desviación estándar de una población es 1.2.

Estas afirmaciones representan la hipótesis nula en diferentes contextos (proporciones, medias y desviaciones estándar). En otras palabras, la hipótesis nula es la afirmación que se pone a prueba y, si no tenemos suficiente evidencia para rechazarla, se acepta como cierta.

4.3.2 Hipótesis alternativa:

La hipótesis alternativa (denotada por H_1 o H_a) es la afirmación opuesta a la hipótesis nula. Representa lo que creemos que es cierto en lugar de la hipótesis nula. Si se obtiene suficiente evidencia para rechazar la hipótesis nula, entonces se acepta la hipótesis alternativa. Por ejemplo, la hipótesis alternativa podría afirmar que:

- $H_a : p > 0.1$ significa que creemos que la proporción de una población es mayor a 0.1.
- $H_a : \mu \neq 1.50$ sugiere que creemos que la media de una población es diferente a 1.50.
- $H_a : \sigma =< 3.5$ implica que creemos que la desviación estándar de una población es menor a 3.5.

En la hipótesis alternativa, solo se pueden utilizar los símbolos mayor que ($>$), menor que ($<$), y diferente de ($\neq$). Estos símbolos indican que estamos buscando una diferencia significativa en alguna dirección o en cualquier dirección, dependiendo del caso.

4.3.3 Pasos para identificar H_0 y H_a:

1. **Paso 1: Identificar la Aseveración o Hipótesis Especifica** H. El primer paso es identificar la afirmación específica que deseas probar. Esta afirmación debe ser expresada de manera simbólica. Por ejemplo, si la afirmación es: *"La proporción de estudiantes que egresan de la facultad de ciencias UASLP es mayor al 70%",* la expresión simbólica sería $H : p > 0.7$. En este caso, se utiliza la letra p porque la aseveración se refiere a una proporción dentro de la población. Es importante tener en cuenta que la expresión simbólica solo puede utilizar los símbolos $>, <, =, \leq, \geq, \neq$. Esto es fundamental para que la hipótesis sea clara y matemáticamente válida.

2. **Paso 2: Negar la Aseveración original:** El siguiente paso es negar la afirmación obtenida en el paso anterior. Si la hipótesis original es $H : p > 0.7$, la negación sería $H : p \leq 0.7$. La negación tiene como objetivo representar lo contrario de la hipótesis planteada.

3. **Definir H_a y H_0:** A partir de las dos expresiones simbólicas obtenidas en los pasos 1 y 2, se definen las hipótesis nula y alternativa. La hipótesis alternativa, denotada como H_a, es la que contiene alguno de los símbolos $>, <, \neq$, ya que estos símbolos indican una diferencia o desviación significativa respecto a la hipótesis nula. Por otro lado, la hipótesis nula H_0 corresponde a la expresión simbólica con el signo de igualdad. Siguiendo el ejemplo de la afirmación sobre la proporción de estudiantes, tenemos dos expresiones simbólicas:

   (a) Paso 1: $H : p > 0.7$
   (b) Paso 2: $H : p \leq 0.7$

   De estas dos expresiones, la hipótesis alternativa será la que contiene alguno de los símbolos $>, <, \neq$. En este caso, la expresión del Paso 1 tiene el símbolo $>$,por lo que corresponde a la hipótesis alternativa. Así, definimos:

   $$H_a : p > 0.7,$$

   Y la hipótesis nula es la expresión con el signo de igualdad:

   $$H_0 : p = 0.7,$$

Estos pasos te ayudarán a establecer claramente la hipótesis nula y alternativa para tu prueba estadística, proporcionando una base sólida para el análisis y la interpretación de los resultados.

Nota Importante: Las Aseveraciones o Hipótesis se Aplican a la POBLACIÓN. Utiliza los Símbolos Correspondientes: Es fundamental recordar que las hipótesis o aseveraciones que formulamos están dirigidas a la población en estudio, no a una muestra. Por lo tanto, es crucial utilizar los símbolos apropiados que representan las características de la población. Los símbolos pertinentes son:

1. Proporciones de la población: p
2. Media de la población: μ
3. Desviación estándar de la población: σ

Por ejemplo, si estamos realizando una prueba sobre la proporción de personas con una característica específica, usaríamos p como el parámetro de la población. Si estamos tratando de probar una hipótesis sobre la media de una característica, utilizaríamos μ, y para la desviación estándar de la población, emplearíamos σ.

Esto asegura que estamos trabajando con las características de la población, no con valores muestrales.

4.4 Estadístico de Prueba

El estadístico de prueba es un valor calculado a partir de los datos muestrales que se utiliza para tomar decisiones sobre el rechazo o aceptación de la hipótesis nula en un análisis estadístico. Se obtiene transformando el estadístico muestral (como la proporción, la media o la desviación estándar muestral) en una puntuación, como z, t o x^2, bajo la premisa de que la hipótesis nula es verdadera.

En otras palabras, el estadístico de prueba nos ayuda a evaluar si la evidencia recolectada en la muestra es lo suficientemente fuerte como para **CONTRADECIR** la hipótesis nula. Al convertir los datos en una puntuación asumiendo que la hipótesis nula es cierta, podemos determinar si existe evidencia suficiente para rechazarla.

La elección del estadístico de prueba depende del problema específico a resolver. Por ejemplo, se utilizan proporciones para comparaciones de frecuencias, medias para comparaciones de promedios, y desviación estándar para estudios de variabilidad. Cada estadístico tiene un proceso único para determinar la relevancia estadística y tomar decisiones basadas en los datos muestrales.

Los estadísticos de prueba se obtienen de la siguiente manera:

1. **Estadístico de prueba para proporciones:**

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

   Donde $\hat{p}$ es la proporción muestral, p es la proporción poblacional bajo la hipotesis nula, $q = 1 - p$, y n es el tamaño de la muestra.

2. **Estadístico de prueba para medias:**

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

   Donde $\bar{x}$ es la media muestral, μ es la media poblacional bajo la hipótesis nula, s es la desviación estándar muestral, y n es el tamaño de la muestra.

3. **Estadístico de prueba para desviación estándar:**

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

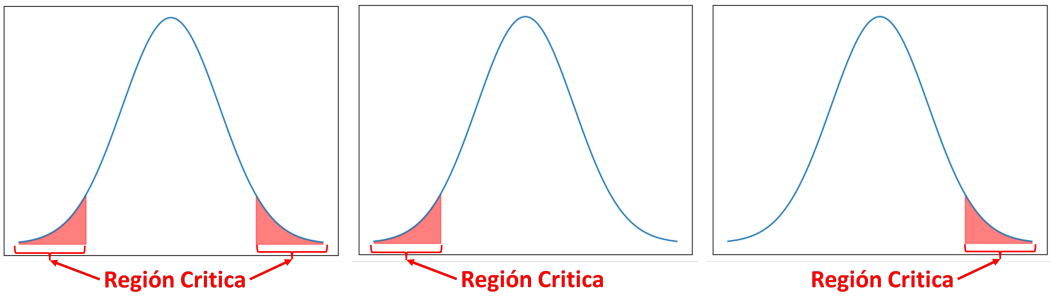
   Donde s^2 es la varianza muestral, σ^2 es la varianza poblacional bajo la hipótesis nula, y n es el tamaño de la muestra.

Estos estadísticos nos permiten determinar si la diferencia observada entre los datos muestrales y los valores esperados (bajo la hipótesis nula) es suficientemente significativa como para rechazar la hipótesis nula.

4.5 Región crítica, nivel de significancia, valor crítico y valor p:

4.5.1 Región crítica:

La región crítica (o región de rechazo) es el conjunto de valores posibles para el estadístico de prueba que nos llevarían a rechazar la hipótesis nula. En otras palabras, son los valores que sugieren evidencia suficiente en contra de la hipótesis nula. Si el valor del estadístico de prueba cae dentro de la región crítica, podemos concluir que hay evidencia suficiente para rechazar la hipótesis nula. La región crítica se determina a partir del nivel de significancia (α) y de la dirección de la hipótesis alternativa. Dependiendo del tipo de prueba (bilateral, unilateral izquierda o unilateral derecha), la región crítica será diferente. Aquí te presentamos ejemplos visuales de cómo se ve la región crítica en diferentes tipos de pruebas:



(a) H_a : símbolo $\neq$ (b) H_a : símbolo $<$ (c) H_a : símbolo $>$

4.5.2 Nivel de significancia:

El nivel de significancia (α) es la probabilidad de que el estadístico de prueba caiga en la región crítica. La región crítica está representada por el área sombreada en rojo en las representaciones gráficas anteriores. En términos simples, α es la posibilidad de que el resultado observado sea tan raro que conduzca al rechazo de la hipótesis nula.

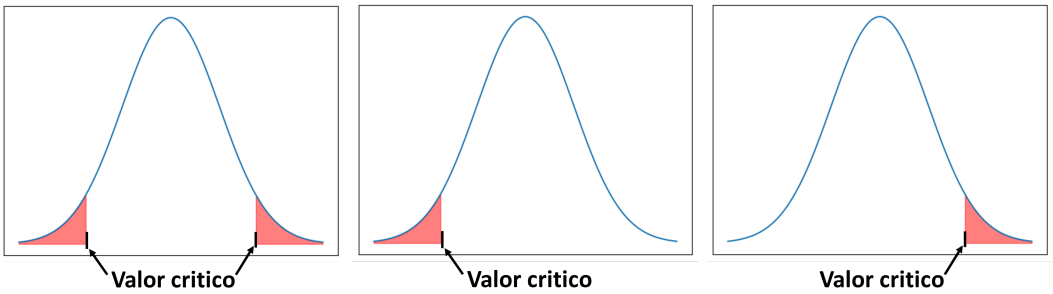
El nivel de significancia α es un valor que determina cuánto riesgo estamos dispuestos a asumir al rechazar la hipótesis nula, considerando que esta sea en realidad cierta. Los valores más comunes para α son 0.05, 0.01 y 0.10, siendo 0.05 el más utilizado. Esto implica que si $\alpha = 0.05$, estamos dispuestos a aceptar un 5% de probabilidad de cometer un error tipo I.

Por ejemplo, si realizamos una prueba de hipótesis con $\alpha = 0.05$, y obtenemos un valor de p menor que 0.05, rechazaremos la hipótesis nula, ya que la probabilidad de observar los resultados obtenidos es suficientemente baja con respecto a la hipótesis nula.

4.5.3 Valor Crítico:

Un valor crítico es cualquier valor que marca la separación entre la región crítica (donde se rechaza la hipótesis nula) y los valores del estadístico de prueba que no llevan al rechazo de la hipótesis nula. Este valor es fundamental porque nos ayuda a determinar si los datos observados caen dentro o fuera de la región crítica. Si el valor del estadístico de prueba excede el valor crítico, se rechaza la hipótesis nula; si no lo excede, no se rechaza.

- Para las **proporciones** (p), se utiliza la distribución normal (Gaussiana), y los valores críticos se encuentran en la tabla A2.
- Para las **medias** (μ), se recurre a la distribución t de Student, y los valores críticos se obtienen de la tabla A3.
- En el caso de la **desviación estándar** (σ), se utiliza la distribución chi-cuadrado (χ^2), y los valores críticos se encuentran en la tabla A4. Es importante destacar que para obtener el valor



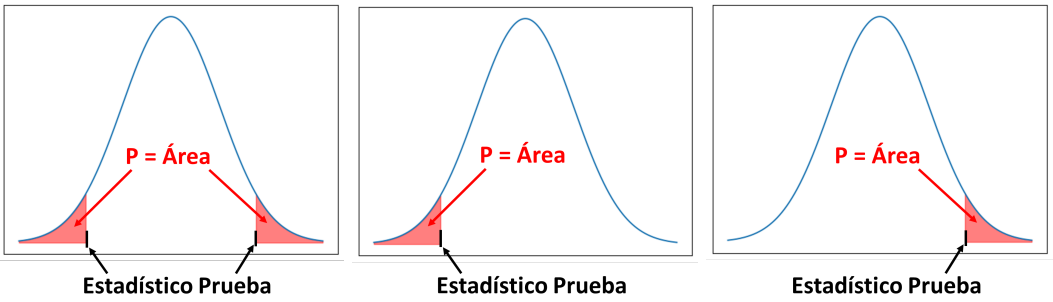
(a) H_a : símbolo $\neq$ (b) H_a : símbolo $<$ (c) H_a : símbolo $>$

crítico en la cola derecha, buscamos un área de valor α, mientras que para una prueba de cola izquierda, buscamos un área de $1 - \alpha$.

Esta distinción es clave para interpretar correctamente la tabla A4 y determinar los valores críticos correspondientes en pruebas de hipótesis sobre la desviación estándar.

4.6 Valor P en pruebas de Hipótesis

El valor P es la probabilidad de obtener un valor del estadístico de prueba que sea al menos tan extremo como el valor observado en los datos muestrales. *En términos sencillos, el valor P nos indica qué tan inusual o improbable es el* ***estadístico de prueba*** *observado.* Si el valor P es muy pequeño, generalmente menor o igual a 0.05, se considera evidencia suficiente para rechazar la hipótesis nula. Para calcular el valor P, primero se obtiene el estadístico de prueba y luego se determina el área bajo la curva de la distribución que corresponde a ese estadístico.



(a) H_a : símbolo $\neq$ (b) H_a : símbolo $<$ (c) H_a : símbolo $>$

4.7 Tipos de colas en Pruebas de Hipótesis: Dos colas, Izquierda y Derecha

- **Prueba de Dos Colas:** En esta prueba, la región crítica se encuentra en ambas extremidades (colas) bajo la curva de la distribución.
- **Prueba de Cola Izquierda:** En este caso, la región crítica se sitúa en la cola izquierda (extremo izquierdo) bajo la curva.
- **Prueba de Cola Derecha:** En una prueba de cola derecha, la región crítica se encuentra en la cola derecha (extremo derecho) bajo la curva.

Comprender los diferentes tipos de colas es fundamental para definir las áreas de interés bajo la curva de distribución y para determinar en qué extremo buscamos evidencia significativa en una prueba de hipótesis específica. Tabla 5 resumen los tipos de pruebas que se pueden realizar.

H_a Símbolo	Tipo de Prueba
$>$	Prueba de Cola Derecha
$<$	Prueba de Cola Izquierda
$\neq$	Prueba de Dos Colas

Table 5: Resumen del tipo de prueba que se pueden aplicar. La hipótesis nula solo puede incluir los símbolos ($>, <, \neq$).

4.7.1 Tipos de distribuciones

Las pruebas de hipótesis e inferencias estadísticas se basan en la identificación de distribuciones específicas dependiendo del tipo de datos analizados y la estadística que se está estudiando.

- Proporciones (p): Cuando se analizan proporciones o porcentajes de una población, como la proporción de personas que cumplen un criterio o la fracción de objetos que presentan un defecto. La distribución asociada es una distribución normal (Gaussiana) cuando el tamaño muestral es suficientemente grande ($np > 5$ y $n(1-p) > 5$, donde p es la proporción poblacional). En estos casos, el estadístico estándar Z es el adecuado.

- Medias (μ): cuando se desea realizar inferencias sobre la media poblacional μ, como determinar si la media de una característica específica en una muestra difiere de un valor teórico. La distribución asociada para tamaños muestrales pequeños ($n < 30$) y cuando la desviación estándar poblacional σ es desconocida, se utiliza la distribución t-Student. Esta distribución tiene en cuenta la variabilidad adicional que proviene de estimar σ con la desviación estándar muestral s. A medida que el tamaño muestral aumenta ($n \geq 30$), la distribución t converge hacia la normal.

- Desviación estándar (σ): cuando el objetivo es realizar inferencias sobre la variabilidad de una población, representada por la desviación estándar σ o la varianza σ^2. La varianza de una muestra sigue una distribución X^2 (chi-cuadrado), que es asimétrica y depende del tamaño de la muestra. La asimetría disminuye a medida que aumentan los grados de libertad ($gl = n-1$).

Tabla 6 resume el tipo de inferencia y la distribución asociada.

Tipo de Inferencia/Hipotesis	Tipo de Distribución	Tabla a Utilizar
Proporciones (p)	Distribución Gaussiana/Normal	Tabla A2
Medias (μ)	Distribución t-students	Tabla A3
Desviación Estándar	Distribución χ^2	Tabla A4

Table 6: Tipo de problema y distribución a utilizar.

4.8 Decisiones sobre la hipótesis nula

El propósito principal de esta sección es evaluar la hipótesis nula (H_0) en función de los resultados obtenidos y con base en la teoría desarrollada en este capítulo. Este proceso permite llegar a conclusiones fundamentadas sobre la hipótesis planteada inicialmente.

Al analizar la hipótesis nula, existen dos posibles decisiones:

- **Rechazar la hipótesis nula:** Esto implica que hay suficiente evidencia estadística en los datos para apoyar la hipótesis alternativa (H_a)

- **No Rechazar la hipótesis nula:** Esto indica que no hay suficiente evidencia estadística para rechazar (H_0), pero no necesariamente confirma que H_0 sea verdadera.

4.8.1 Criterios de Decisión en Pruebas de Hipótesis

En las pruebas de hipótesis, la decisión de rechazar o no la hipótesis nula se puede tomar utilizando diferentes métodos y enfoques. Los métodos tradicionales, el método del valor P y la utilización de intervalos de confianza son opciones comunes. A continuación, se describen estos métodos de manera más clara:

1. **Método Tradicional:**

   Criterio de Decisión: Rechaza la hipótesis nula (H_0) si el estadístico de prueba cae dentro de la región crítica. No la rechaza si el estadístico de prueba no está en la región crítica.

2. **Método del Valor P:**

   Criterio de Decisión: Rechaza H_0 si el valor de P es menor que el nivel de significancia (α, comúnmente 0.05). No la rechaza si el valor de P es mayor o igual a α.

3. **Intervalo de Confianza:**

   Criterio de Decisión: Si el valor bajo la hipótesis nula está fuera del intervalo de confianza, entonces hay evidencia para rechazar la hipótesis nula. En caso contrario, no hay suficiente evidencia para rechazarla.

4.8.2 Significado de Rechazar H_0 y no rechazar H_0

En el contexto de una prueba de hipótesis, H_0 (hipótesis nula) representa la afirmación inicial que se somete a prueba. Por su parte, H_a (hipótesis alternativa) es la afirmación contraria, que se apoya si los datos ofrecen suficiente evidencia en contra de H_0

RECHAZAR H_0:
Rechazar H_0 significa que los datos proporcionan suficiente evidencia estadística para concluir que H_a es más plausible. En otras palabras:

- Existe una diferencia significativa o un efecto estadístico detectable en los datos.
- Implica que el resultado observado es lo suficientemente extremo como para que sea improbable bajo la suposición de que H_0 es verdadera.

Conclusión Típica: "Hay evidencia estadísticamente significativa para apoyar la hipótesis alternativa (H_a)."

NO RECHAZAR H_0:
No rechazar H_0 significa que los datos no ofrecen suficiente evidencia estadística en contra de H_0. H_0 Sin embargo, esto no implica que H_0 sea verdadera, sino que no se puede descartar con los datos disponibles. En otras palabras:

- No hay suficiente evidencia para concluir que H_a es cierta.
- Puede deberse a que el efecto o la diferencia es muy pequeña, el tamaño de la muestra es insuficiente o el nivel de significancia (α) es demasiado estricto.

Conclusión Típica: "No hay evidencia suficiente para rechazar la hipótesis nula (H_0)."

4.9 EJERCICIOS

EJERCICIO 1: Hipótesis nula y Hipótesis Alternativa

Determine la hipótesis nula (H_0) y la hipótesis alternativa (H_a) para los siguientes problemas cotidianos:

1. Madison Advertising Company afirma que el 70% de los hombres no ven golf en televisión.
2. En la Facultad de Ciencias (UASLP), se afirma que el 30% de los estudiantes son foráneos.
3. Una empresa asegura que la desviación estándar σ de tamaño de tornillos es mayor a 1.5.

EJERCICIO 2: Cálculo de valores críticos

En cada problema, calcule los valores críticos necesarios para tomar decisiones sobre las hipótesis planteadas.

1. Se desea evaluar si la proporción de estudiantes foraneos p es diferente de 0.30, con un nivel de significancia $\alpha = 0.02$.
2. Deseamos verificar si el promedio de edad μ de estudiantes de la FC-UASLP es menor a 25. El nivel de significancia es $\alpha = 0.025$, y se seleccionan 21 estudiantes de manera aleatoria como muestra.
3. Una empresa asegura que la desviación estándar σ de tamaño de tornillos es mayor a 1.5. Usando un nivel de significancia $\alpha = 0.05$ y con una muestra de 91 tornillos:

EJERCICIO 3: Cálculo de estadístico de prueba.

Utilice los datos proporcionados para calcular el estadístico de prueba en cada caso.

1. La aseveración (hipótesis) es que la proporción de adultos que compra a través de internet es menor a 0.50 (50%), y el estadístico de muestra (datos obtenidos de una muestra aleatoria) que incluyen $n = 1025$ sujetos, se obtiene que 29% dice que utiliza internet para realizar compras.
2. La aseveración (hipótesis) es que el promedio de calificación de estudiantes de la FC-UASLP es de 8.0, y el estadístico de muestra obtenidos de una muestra de $n = 50$ se obtiene que el promedio es 8.5 y la desviación estándar muestral es 1.5.

EJERCICIO 4: Identificación de tipo de prueba

En los siguientes ejercicios identifique si la región critica corresponde a cola izquierda, cola derecha o dos colas.

1. Un investigador está analizando la proporción de defectos en una línea de producción. La proporción histórica de defectos es del 10%. Sin embargo, tras implementar mejoras en el proceso, desea determinar si la proporción actual de defectos es significativamente diferente del 10%.
2. Una empresa de transporte afirma que el tiempo promedio para entregar paquetes es de 25 minutos. Un grupo de clientes sospecha que los tiempos promedio han aumentado debido a cambios recientes en la logística. Desean evaluar si el tiempo promedio actual es mayor a 25 minutos.
3. Un laboratorio farmacéutico ha desarrollado un nuevo método para medir la variabilidad (σ) en la concentración de un medicamento en lotes de producción. El método estándar tiene una variabilidad de 1 unidad, pero desean probar si la variabilidad es menor con el nuevo método.

EJERCICIO 5: Prueba de Hipótesis

En los siguientes ejercicios identifique si se rechaza la hipótesis nula o no se rechaza utilizando el método tradicional y el del valor p

1. En una encuesta a 300 conductores seleccionados de manera aleatoria, el 55% admitió haberse pasado la luz roja. Se hace la afirmación:
   "La mayoría (más del 50%) de los conductores se pasan la luz roja."
   Determinar si existe evidencia estadística para respaldar esta afirmación usando un nivel de significancia de $\alpha = 0.05$.

2. En una prueba realizada con una muestra de 40 pelotas nuevas, rebotaron en promedio 92.67 pulgadas, con una desviación estándar de 1.79 pulgadas. Determine si existe evidencia suficiente para sustentar la afirmación de que las pelotas tienen rebotes con una media distinta a 92.48 pulgadas, usando un nivel de significancia de $\alpha = 0.05$.

4.10 SOLUCION 1: Hipótesis nula y Hipótesis Alternativa

Determine la hipótesis nula (H_0) y la hipótesis alternativa (H_a) para los siguientes problemas cotidianos:

4.10.1 Ejercicio 1

Problema: Madison Advertising Company afirma que el 70% de los hombres no ven golf en televisión.

SOLUCION:

- **Paso 1: Identificar la Aseveración o Hipótesis Especifica** H.
  - En este problema, La afirmación inicial sugiere que "El porcentaje no es mayor al 70%". Esto es lo mismo que afirmar que "El porcentaje es menor o igual a 70%"
  - Representamos esta afirmación utilizando el símbolo p para el porcentaje de hombres que no ven golf en televisión:
$$H : p \leq 0.70$$
- **Paso 2: Negar la Aseveración original:**

  La negación del símbolo $\leq$ es $>$, por lo cual, la hipótesis negada es
$$H : p > 0.70$$
- **Paso 3: Definir H_a (hipótesis alternativa) y H_0 (hipótesis nula):**
  - La hipótesis alternativa (H_a) debe contener alguno de los símbolos $>, <$, o $\neq$. Por lo tanto, negar la aseveración original (Paso 2) corresponde a la hipótesis alternativa
$$H_a : p > 0.70$$
  - La hipótesis nula (H_0) corresponde a la igualdad:
$$H_0 : p = 0.70$$

Conclusión:

Para este problema, las hipótesis formuladas son:

- Hipótesis nula $H_0 : p = 0.7$
- Hipótesis alternativa $H_a : p > 0.7$

Estas hipótesis pueden ahora ser utilizadas para analizar los datos y determinar si hay evidencia suficiente para rechazar la hipótesis nula.

4.10.2 Ejercicio 2

Problema: En la Facultad de Ciencias (UASLP), se afirma que el 30% de los estudiantes son foráneos.

SOLUCION:

- **Paso 1: Identificar la Aseveración o Hipótesis Especifica** H.
  - En este problema, La afirmación inicial indica que "el porcentaje de estudiantes foráneos es igual al 30
  - Utilizamos el símbolo p para representar el porcentaje de estudiantes foráneos en la población:
$$H : p = 0.30$$

- **Paso 2: Negar la Aseveración original:**

  La negación de "igual a 30%" (=) es "diferente de 30%" ($\neq$). Por lo tanto, la hipótesis negada es:

  $$H : p \neq 0.30$$

- **Paso 3: Definir H_a (hipótesis alternativa) y H_0 (hipótesis nula):**

  - La hipótesis alternativa (H_a) debe contener alguno de los símbolos $>, <$, o $\neq$. Por lo tanto, la hipótesis correspondiente a Paso 2, es la hipótesis alternativa

    $$H_a : p \neq 0.30$$

  - La hipótesis nula (H_0) corresponde a la igualdad:

    $$H_0 : p = 0.30$$

Conclusión:

Para este problema, las hipótesis formuladas son:

- Hipótesis nula $H_0 : p = 0.30$
- Hipótesis alternativa $H_a : p \neq 0.30$

4.10.3 Ejercicio 3

Problema: Una empresa asegura que la desviación estándar σ de tamaño de tornillos es mayor a 1.5.

SOLUCION:

- **Paso 1: Identificar la Aseveración o Hipótesis Especifica H.**

  - En este problema, La afirmación inicial indica que "la desviación estándar de tamaño de tornillos es mayor a 1.5".
  - Utilizamos el símbolo σ para representar la desviación estándar de la población. La expresión simbólica es:

    $$H : \sigma > 1.5$$

- **Paso 2: Negar la Aseveración original:**

  La negación de "mayor que 1.5" (¿) es "menor o igual que 1.5" ($\neq$). Por lo tanto, la hipótesis negada es:

  $$H : \sigma \leq 1.5$$

- **Paso 3: Definir H_a (hipótesis alternativa) y H_0 (hipótesis nula):**

  - La hipótesis alternativa (H_a) debe contener alguno de los símbolos $>, <$, o $\neq$. Por lo tanto, la hipótesis correspondiente a Paso 1, es la hipótesis alternativa

    $$H_a : \sigma > 1.5$$

  - La hipótesis nula (H_0) corresponde a la igualdad:

    $$H_0 : \sigma = 1.5$$

Conclusión:

Para este problema, las hipótesis formuladas son:

- Hipótesis nula $H_0 : \sigma = 1.5$
- Hipótesis alternativa $H_a : \sigma > 1.5$

4.11 SOLUCION 2: Cálculo de Valores Críticos

En cada problema, calcule los valores críticos necesarios para tomar decisiones sobre las hipótesis planteadas.

4.11.1 Ejercicio 1

Problema: Se desea evaluar si la proporción de estudiantes foraneos p es diferente de 0.30, con un nivel de significancia $\alpha = 0.02$.

SOLUCION:

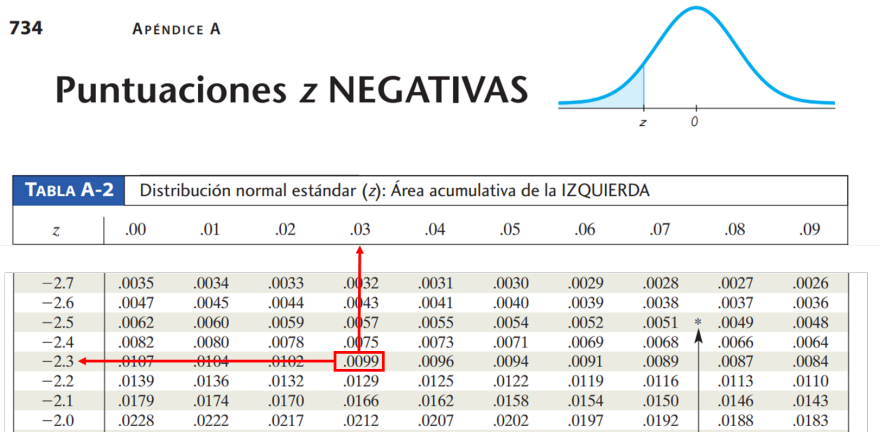
- **Paso 1: Definir hipótesis nula** Hipotesis: la proporción de estudiantes foraneos p es diferente de 0.30. En este caso, se tiene que $H_a : p \neq 0.30$ y $H_0 : p = 0.30$. *Nota: El estudiante debe confirmar este planteamiento siguiendo la metodología explicada en ejercicios anteriores.*

- **Paso 2: Determinar los valores críticos**

  1. **Distribución utilizada:** Para encontrar el valor crítico en pruebas de hipótesis para proporciones (p), utilizamos la distribución normal estándar.
  2. **Nivel de significancia:** El nivel se significancia es $\alpha = 0.02$. Dado que la hipótesis alternativa H_a tiene el símbolo $\neq$, indica que es una prueba de dos colas, el área en cada cola es:
  $$\alpha/2 = 0.02/2 = 0.01.$$
  3. **Buscar el valor crítico en la tabla z:** En la Figura 13 se presenta parte de la tabla A2 (área para distribución normal estándar). Buscamos un área igual a 0.01. Aunque no encontramos exactamente 0.01, hallamos el valor más cercano, 0.0099.

     Para hallar el valor crítico, buscamos en la tabla hacia la izquierda (2.3) y hacia arriba (0.03). Al sumar estos valores (2.3+0.03 = 2.33), **obtenemos que los valores críticos son** $z = \pm 2.33$

     Nota: El signo positivo-negativo se aplica porque es una prueba de dos colas.



734 APÉNDICE A

Puntuaciones *z* NEGATIVAS

TABLA A-2 Distribución normal estándar (*z*): Área acumulativa de la IZQUIERDA

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183

Figure 7: Tabla A2: para determinar valores críticos de proporciones.

4.11.2 Ejercicio 2

Problema: Deseamos verificar si el promedio de edad μ de estudiantes de la FC-UASLP es menor a 25. El nivel de significancia es $\alpha = 0.025$, y se seleccionan 21 estudiantes de manera aleatoria como muestra.

SOLUCION:

- **Paso 1: Definir hipótesis nula** Hipotesis: la promedio de edad μ de estudiantes de la FC-UASLP es menor a 25. En este caso, se tiene que $H_a : \mu < 25$ y $H_0 : \mu = 25$. *Nota: El estudiante debe confirmar este planteamiento siguiendo la metodología explicada en ejercicios anteriores.*

- **Paso 2: Determinar los valores críticos**

  1. **Distribución utilizada:** Dado que la muestra tiene un tamaño pequeño ($n = 21$) y no se conoce la desviación estándar de la población, utilizamos la distribución t-Student.
  2. **Nivel de significancia:** Dado que la hipótesis alternativa H_a tiene el símbolo $<$, se trata de una prueba de una cola, por lo que el área de interés esta en la cola izquiera de la distribución con un área igual a:

  $$\alpha = 0.0251.$$

  3. **Buscar el valor crítico en la tabla t-Student:** En la Figura 8 se presenta parte de la tabla A3 correspondiente a valores críticos para la distribución t-Student. Buscamos un área igual a 0.025 con $n - 1 = 20$ grados de libertad. Encontramos el valor crítico buscando en la fila de 20 grados de libertad y hacia arriba el área correspondiente, 0.025, para una cola. **El valor en el recuadro nos da el valor crítico, siendo**

  $$t = -2.086$$

  Nota: El signo negativo se aplica porque es una prueba de una cola hacia la izquierda.

4.11.3 Ejercicio 3

Problema: Una empresa asegura que la desviación estándar σ de tamaño de tornillos es mayor a 1.5. Usando un nivel de significancia $\alpha = 0.05$ y con una muestra de 91 tornillos:

SOLUCION:

- **Paso 1: Definir hipótesis nula** Hipotesis: la desviación estándar σ de tamaño de tornillos es mayor a 1.5. En este caso, se tiene que $H_a : \sigma > 1.5$ y $H_0 : \sigma = 1.5$. *Nota: El estudiante debe confirmar este planteamiento siguiendo la metodología explicada en ejercicios anteriores.*

- **Paso 2: Determinar los valores críticos**

  1. **Distribución utilizada:** Para encontrar el valor crítico en pruebas de hipótesis para desviación estándar (σ), utilizamos la distribución χ^2.
  2. **Nivel de significancia:** La hipotesis alternativa H_a tiene el símbolo $>$, indicando que es una prueba de cola derecha (una cola), el valor de significancia es:

  $$\alpha = 0.05$$

  .

  3. **Buscar el valor crítico en la tabla t-Student:** En la Figura 9 se presenta parte de la tabla A4 correspondiente a valores críticos para la distribución χ^2. Buscamos un área igual a 0.05 con 90 grados de libertad. Encontramos el valor crítico buscando en la fila de 90 grados de libertad y hacia arriba el área correspondiente a $\alpha = 0.05$. El valor en el recuadro nos da el valor crítico, siendo $\chi^2 = 113.145$

  Nota: El signo positivo se aplica porque es una prueba de una cola hacia la derecha.

TABLA A-3 Distribución *t*: Valores críticos *t*

	Área en una cola				
	0.005	0.01	0.025	0.05	0.10
Grados de libertad	Área en dos colas 0.01	0.02	0.05	0.10	0.20
1	**63.657**	31.821	**12.706**	6.314	3.078
2	**9.925**	6.965	**4.303**	2.920	1.886
3	**5.841**	4.541	**3.182**	2.353	1.638
4	**4.604**	3.747	**2.776**	2.132	1.533
5	**4.032**	3.365	**2.571**	2.015	1.476
6	**3.707**	3.143	**2.447**	1.943	1.440
7	**3.499**	2.998	**2.365**	1.895	1.415
8	**3.355**	2.896	**2.306**	1.860	1.397
9	**3.250**	2.821	**2.262**	1.833	1.383
10	**3.169**	2.764	**2.228**	1.812	1.372
11	**3.106**	2.718	**2.201**	1.796	1.363
12	**3.055**	2.681	**2.179**	1.782	1.356
13	**3.012**	2.650	**2.160**	1.771	1.350
14	**2.977**	2.624	**2.145**	1.761	1.345
15	**2.947**	2.602	**2.131**	1.753	1.341
16	**2.921**	2.583	**2.120**	1.746	1.337
17	**2.898**	2.567	**2.110**	1.740	1.333
18	**2.878**	2.552	**2.101**	1.734	1.330
19	**2.861**	2.539	**2.093**	1.729	1.328
20	**2.845**	2.528	**2.086**	1.725	1.325
21	**2.831**	2.518	**2.080**	1.721	1.323
22	**2.819**	2.508	**2.074**	1.717	1.321
23	**2.807**	2.500	**2.069**	1.714	1.319
24	**2.797**	2.492	**2.064**	1.711	1.318

Figure 8: Tabla A3: para determinar valores críticos de medias.

TABLA A-4 Distribución chi cuadrada (χ^2)

	Área a la derecha del valor crítico									
Grados de libertad	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Figure 9: Tabla A4: para determinar valores críticos de desviación estándar.

4.12 SOLUCION 3: Cálculo de Estadístico de Prueba.

Utilice los datos proporcionados para calcular el estadístico de prueba en cada caso.

4.12.1 Ejercicio 1

Problema: La aseveración (hipótesis) es que la proporción de adultos que compra a través de internet es menor a 0.50 (50%), y el estadístico de muestra (datos obtenidos de una muestra aleatoria) que incluyen $n = 1025$ sujetos, se obtiene que 29% dice que utiliza internet para realizar compras.

SOLUCION:

- **Paso 1: Identificar Hipótesis Nula (H_0).** En esta caso. la hipótesis especifica es $H : p < 0.50$, y su negación es $H : p \geq 0.50$. Esto lleva a $H_a : p < 0.50$ y $H_0 : p = 0.50$.
- **Paso 2: Determinar la proporción muestral.** La muestra consta de $n = 1025$ sujetos, y la proporción muestral ya esta dada por $\hat{p} = 0.29$
- **Paso 3: calcular estadístico de prueba.** debido a que el problema corresponde a proporciones, entonces debemos de utilizar el estadístico de prueba z (correspondiente a distribución normal estándar), y sustituir los valores adecuados

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

Sustituimos los valores:

$$\begin{aligned} z &= \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \\ &= \frac{0.29 - 0.50}{\sqrt{\frac{0.50*(1-0.50)}{1025}}} \\ &= \frac{-0.21}{\sqrt{\frac{0.25}{1025}}} \\ &= \frac{-0.21}{\sqrt{0.0002439}} \\ &= \frac{-0.21}{0.0156172} \\ &= -13.44 \end{aligned}$$

El resultado $z = -13.44$ es el valor del estadístico de prueba.

Conclusión (ejercicio adicional):

El valor del estadístico de prueba obtenido es $z = -13.44$. Este valor debe compararse con el valor crítico correspondiente al nivel de significancia (α) en una prueba de una cola (lado izquierdo).

Dado que:

1. Asumiendo un nivel de significancia de $\alpha = 0.05$, el valor crítico para una cola es aproximadamente $z_c = -1.645$ (según la tabla z).
2. El estadístico de prueba $z = -13.44$ es mucho menor que el valor critico $z_c = -1.645$

Por lo tanto, se rechaza la hipótesis nula (H_0) en favor de la hipótesis alternativa (H_a). Esto significa que existe evidencia estadística suficiente, con un nivel de significancia del 5%, para afirmar que la proporción de adultos que realiza compras por internet es menor al 50%.

4.12.2 Ejercicio 2

Problema: La aseveración (hipótesis) es que el promedio de calificación de estudiantes de la FC-UASLP es de 8.0, y el estadístico de muestra obtenidos de una muestra de n=50 se obtiene que el promedio es 8.5 y la desviación estándar muestral es 1.5

SOLUCION:

- **Paso 1: Identificar las hipótesis.**

  El primer paso es establecer las hipótesis que se someterán a prueba. La hipótesis original es "El promedio de calificación es de 8.0", en expresión simbólica corresponde a: $H : \mu = 8.0$, y su negación es: $H : \mu neq 8.0$. De las expresiones anteriores se concluye que: La hipotesis nula es: $H_0 : \mu = 8.0$, mientras que la hipótesis alternativa es: $H_a : \mu \neq 8.0$

- **Paso 2: Identificar los valores conocidos y determinar la media muestral** De los datos proporcionados:

  - Tamaño de muestra: $n = 50$
  - Media muestral: $\bar{x} = 8.5$
  - Desviación estándar muestral: $s = 1.5$
  - Media poblacional hipotética: $\mu = 8.0$

  La muestra consta de $n = 50$ sujetos, y la media muestral ya esta dada por $\mu = 8.5$, y las desviación estándar por $s = 1.5$

- **Paso 3: Calcular el estadístico de prueba**

  El estadístico de prueba para una media muestral, cuando se desconoce la desviación estándar poblacional, se calcula utilizando la distribución t-Student

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

  Sustituyendo los valores conocidos:

  $$\begin{aligned} t &= \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \\ &= \frac{8.5 - 8.0}{\frac{1.5}{\sqrt{50}}} \\ &= \frac{0.5}{\frac{1.5}{7.071}} \\ &= \frac{0.5}{\frac{1.5}{7.071}} \\ &= \frac{0.5}{0.212134} \\ &= 2.357 \end{aligned}$$

  El valor del estadístico de prueba es:

  $$t = 2.357$$

- **Interpretación del resultado:**

  El valor calculado de t debe compararse con los valores críticos de la distribución t-Student para $n - 1 = 49$ grados de libertad y el nivel de significancia especificado (por ejemplo, se puede suponer $\alpha = 0.05$ para una prueba bilateral). Con base en esto, se determinará si se rechaza o no H_0.

4.13 SOLUCION 4: Identificación de tipo de prueba.

En los siguientes ejercicios identifique si la región critica corresponde a cola izquierda, cola derecha o dos colas.

4.13.1 Ejercicio 1

Problema: Un investigador está analizando la proporción de defectos en una línea de producción. La proporción histórica de defectos es del 10%. Sin embargo, tras implementar mejoras en el proceso, desea determinar si la proporción actual de defectos es significativamente diferente del 10%.

SOLUCION:

- **Paso 1: Identificar Hipótesis Alternativa:**
  La hipótesis es: "La proporción actual de defectos es diferente del 10%". Esto se traduce simbolicamente a: $H_a : p \neq 0.10$
- **Paso 2: Identificar tipo de prueba**
  Dado que la hipótesis alternativa tiene el símbolo $\neq$, esto indica una PRUEBA DE DOS COLAS (ver Fig. 10). En una prueba de dos colas, la región crítica está distribuida en ambas colas de la distribución.

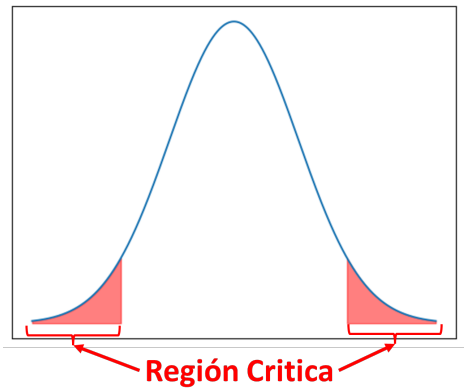


Figure 10: PRUEBA DE DOS COLAS. Región critica ambos lados de la distribución.

4.13.2 Ejercicio 2

Problema: Una empresa de transporte afirma que el tiempo promedio para entregar paquetes es de 25 minutos. Un grupo de clientes sospecha que los tiempos promedio han aumentado debido a cambios recientes en la logística. Desean evaluar si el tiempo promedio actual es mayor a 25 minutos.

SOLUCION:

- **Paso 1: Identificar Hipótesis Alternativa:**
  La hipótesis es: "El tiempo promedio actual es mayor a 25 minutos". Esto se traduce simbolicamente a: $H_a : \mu > 25$
- **Paso 2: Identificar tipo de prueba**
  Dado que la hipótesis alternativa tiene el símbolo $>$, esto indica una PRUEBA DE COLA DERECHA (ver Fig. 11). En una prueba de cola derecha, la región crítica está distribuida a la derecha de la distribución.

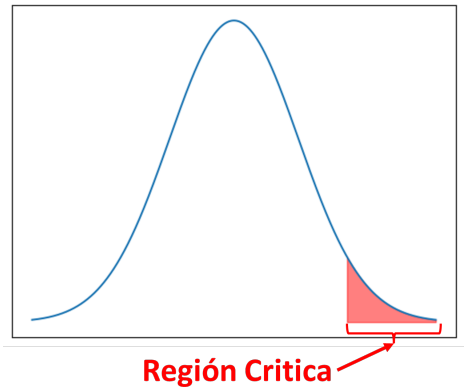


Figure 11: PRUEBA DE COLA DERECHA. Región critica a la derecha de la distribución.

4.13.3 Ejercicio 3

Problema: Un laboratorio farmacéutico ha desarrollado un nuevo método para medir la variabilidad (σ) en la concentración de un medicamento en lotes de producción. El método estándar tiene una variabilidad de 1 unidad, pero desean probar si la variabilidad es menor con el nuevo método.

SOLUCION:

- **Paso 1: Identificar Hipótesis Alternativa:**
  La hipótesis es: "Variabilidad es menor con el nuevo método". Esto se traduce simbolicamente a: $H_a : \sigma < 1$
- **Paso 2: Identificar tipo de prueba**
  Dado que la hipótesis alternativa tiene el símbolo $<$, esto indica una PRUEBA DE COLA IZQUIERDA (ver Fig. 12). En una prueba de cola izquierda, la región crítica está distribuida a la izquierda de la distribución.

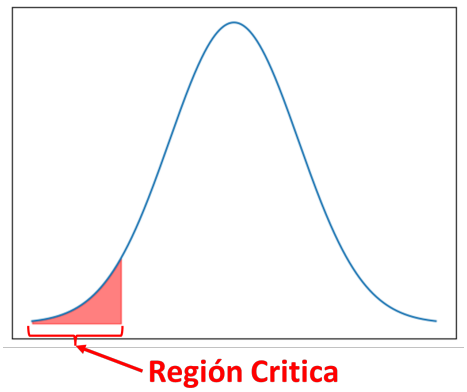


Figure 12: PRUEBA DE COLA IZQUIERDA. Región critica a la izquierda de la distribución.

4.14 SOLUCION 5: Prueba de Hipótesis.

En los siguientes ejercicios identifique si se rechaza la hipótesis nula o no se rechaza utilizando el método tradicional y el del valor p.

4.14.1 Ejercicio 1

Problema: En una encuesta a 300 conductores seleccionados de manera aleatoria, el 55% admitió haberse pasado la luz roja. Se hace la afirmación:
"La mayoría (más del 50%) de los conductores se pasan la luz roja."
Determinar si existe evidencia estadística para respaldar esta afirmación usando un nivel de significancia de $\alpha = 0.05$.

SOLUCION:

- **Paso 1: Identificar hipótesis alternativa:** En este problema, la hipótesis es que "La mayoría de los conductores se pasan la luz roja". Utilizamos el símbolo p para representar porcentajes de población, por lo tanto, la hipótesis específica es $H : p > 0.50$ y su negación es $H : p \leq 0.50$. Esto lleva a $H_a : p > 0.50$ y $H_0 : p = 0.50$.

- **Paso 2: Determinar datos muestrales:** La muestra consta de $n = 300$ sujetos, la proporción muestral ya esta dada por $\hat{p} = 0.55$, y el nivel de significancia es $\alpha = 0.05$.

- **Paso 3: Identificar tipo de prueba:** dado que $H_a : p > 0.50$, se trata de una PRUEBA DE COLA DERECHA (ver Tabla 5).

- **Paso 4: Calcular estadístico de prueba:** debido a que el problema corresponde a proporciones, entonces debemos de utilizar el estadístico de prueba z:

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

Sustituyendo los valores:

$$\begin{aligned} z &= \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \\ &= \frac{0.55 - 0.50}{\sqrt{\frac{0.50*(1-0.50)}{300}}} \\ &= \frac{0.05}{\sqrt{\frac{0.25}{300}}} \\ &= \frac{0.05}{\sqrt{0.0008333}} \\ &= \frac{0.05}{0.028809} \\ &= 1.7355 \end{aligned}$$

Por lo tanto, el valor del estadístico de prueba es:

$$z = 1.7355$$

- **Paso 5: Identificar si se rechaza la hipótesis nula**

1. **Método tradicional (Región crítica):** Para encontrar la región crítica en pruebas de hipótesis para proporciones (p), utilizamos la distribución gaussiana. En la Figura 13 se presenta parte de la tabla A2. Buscamos un área igual a 0.05. Aunque no encontramos exactamente 0.05, hallamos el valor más cercano, 0.0505.

TABLA A-2 Distribución normal estándar (z): Área acumulativa de la IZQUIERDA

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823

Figure 13: Tabla A2: para determinar valores críticos de proporciones.

Para hallar el valor crítico, buscamos en la tabla hacia la izquierda (1.6) y hacia arriba (0.04). Al sumar estos valores (1.6+0.04 = 1.64), obtenemos que el valor crítico es 1.64
Finalmente, identificamos si el estadístico de prueba $z = 1.7355$ cae en la región critica. Una representación visual en Figura 14 muestra que el estadístico de prueba cae en la región critica. Por lo tanto, se rechaza la hipótesis nula y los datos muestrales respaldan la afirmación de la hipótesis alternativa. Por lo tanto, podemos afirmar con alta confianza que en $H_a : p > 0.5$ es verdadera. Entonces, los datos respaldan que la mayoría de los conductores se pasan la luz roja.

Figure 14: Identificando si el estadístico de prueba cae en región critica.

2. **Método valor p:** Para aplicar el método del valor p debemos calcular el área bajo la curva normal estándar que corresponde al estadístico de prueba z. Dado que la hipótesis alternativa es $H_a : p > 0.5$, el área que nos interesa calcular es a la DERECHA (prueba de cola derecha) del estadístico de prueba $z = 1.7355$. Primero, buscamos en la tabla de distribución normal estándar (Tabla A2) el área acumulada correspondiente al valor $z = 1.7355$. Como este valor no está en la tabla, utilizamos el valor más cercano, que es $z = 1.74$. Para este valor, el área acumulada a la izquierda es $p = 0.0409$ (ver Figura 15). Ahora, comparamos el valor p con el nivel de significancia:

$$p = 0.0409 < \alpha = 0.05$$

Dado que p es menor que el nivel de significancia, rechazamos la hipótesis nula H_0. Esto significa que existe suficiente evidencia estadística para aceptar la hipótesis alternativa H_a **la mayoría de los conductores (más del 50%) se pasan la luz roja.** En conclusión, el método del valor p lleva al mismo resultado que el método tradicional basado en el estadístico de prueba.

TABLA A-2 Distribución normal estándar (z): Área acumulativa de la IZQUIERDA

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823

Figure 15: Calculando el área correspondiente al estadístico de prueba..

4.14.2 Ejercicio 2

Problema: En una prueba realizada con una muestra de 40 pelotas nuevas, rebotaron en promedio 92.67 pulgadas, con una desviación estándar de 1.79 pulgadas. Determine si existe evidencia suficiente para sustentar la afirmación de que las pelotas tienen rebotes con una media distinta a 92.48 pulgadas, usando un nivel de significancia de $\alpha = 0.05$.

SOLUCION:

- **Paso 1: Identificar hipótesis alternativa:** En este problema, la hipótesis es que "Las pelotas tienen una media distinta a 92.48. Utilizamos el símbolo μ para representar medias de población, por lo tanto, la hipótesis específica es $H : \mu \neq 92.48$ y su negación es $H : \mu = 92.48$. Esto lleva a $H_a : \mu \neq 92.48$ y $H_0 : \mu = 92.48$.
- **Paso 2: Determinar datos muestrales:**
  - La muestra consta de $n = 40$ pelotas
  - La media muestral ya esta dada por $\bar{x} = 92.67$
  - La desviación muestral es $s = 1.79$
  - El nivel de significancia es $\alpha = 0.05$.
  - Hipótesis Nula $H_0 : \mu = 92.48$
- **Paso 3: Identificar tipo de prueba:** dado que $H_a : \mu \neq 0.50$, se tiene que es una PRUEBA DE DOS COLAS (ver Tabla 5) y el valor $\alpha = 0.05$ se modifica por $\alpha/2 = 0.05/2 = 0.025$.
- **Paso 4: Calcular estadístico de prueba:** debido a que el problema corresponde a medias, entonces debemos de utilizar el estadístico de prueba t para distribución t-students:

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

Sustituyendo los valores:

$$\begin{aligned} t &= \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \\ &= \frac{92.67 - 92.48}{\frac{1.79}{\sqrt{40}}} \\ &= \frac{0.19}{\frac{1.79}{6.3245}} \\ &= \frac{0.19}{0.283026} \\ &= 0.6713 \end{aligned}$$

Por lo tanto, el valor del estadístico de prueba es:

$$t = 0.6713$$

- **Paso 5: Identificar si se rechaza la hipótesis nula**
  - **MÉTODO TRADICIONAL (Región crítica):** Para encontrar la región crítica en pruebas de hipótesis para medias (μ), utilizamos la distribución t-student (ver Tabla 6). El método tradicional utiliza la distribución t-Student para determinar si el estadístico de prueba cae en la región crítica. Este enfoque es apropiado para pruebas de hipótesis sobre medias (μ), especialmente cuando el tamaño muestral es pequeño ($n < 30$) o la desviación estándar poblacional (σ) es desconocida.

TABLA A-3 Distribución t: Valores críticos t

Grados de libertad	Área en una cola 0.005	0.01	0.025	0.05	0.10
	Área en dos colas 0.01	0.02	0.05	0.10	0.20
34	**2.728**	2.441	**2.032**	1.691	1.307
36	**2.719**	2.434	**2.028**	1.688	1.306
38	**2.712**	2.429	**2.024**	1.686	1.304
40	**2.704**	2.423	**2.021**	1.684	1.303
45	**2.690**	2.412	**2.014**	1.679	1.301
50	**2.678**	2.403	**2.009**	1.676	1.299
55	**2.668**	2.396	**2.004**	1.673	1.297

Figure 16: Tabla A3: para determinar valores críticos de medias.

Dado que el nivel de significancia es $\alpha = 0.05$, y la hipótesis alternativa ($H_a : \mu \neq 92.48$) indica una prueba de dos colas (ver Tabla 5), dividimos α entre 2:

$$\alpha/2 = \frac{0.05}{2} = 0.025$$

Los grados de libertad se calculan como $gl = n - 1 = 40 - 1 = 39$.

Con $gl = 39$ y un área en cada cola de $\alpha/2 = 0.025$, buscamos el valor crítico en la tabla de distribución t-Student (Tabla A3). Aunque $gl = 39$ no aparece directamente en la tabla, tomamos el valor más cercano, que corresponde a $gl = 40$. El valor crítico obtenido es 16):

$$t_{critico} = 2.021$$

La región crítica esta formada por los valores de t menores que -2.021 o mayores que 2.021. Esto significa que si el estadístico de prueba cae en fuera del intervalo [-2.021, 2.021], rechazaremos la hipótesis nula H_0

El estadístico de prueba calculado previamente:

$$t = 0.6713$$

Al comparar este valor con los límites de la región crítica [-2.021, 2.021]:

$$0.6713 \in [-2.021, 2.021]$$

Por lo tanto, el estadístico de prueba no cae en la región crítica. Por lo tanto, no se puede rechazar la hipótesis nula.

Visualmente, esto se puede confirmar observando la representación en la Figura 17. Dado que el estadístico de prueba no se encuentra en la región crítica, no rechazamos la hipótesis nula H_0. **Esto indica que no hay suficiente evidencia estadística para concluir que la media de los rebotes sea diferente de 92.48 pulgadas.**

– **MÉTODO VALOR p:**

Calcular el área (valor p): El objetivo es calcular el valor p asociado al estadístico de prueba $t = 0.6713$ en una prueba de DOS COLAS ($H_a : \mu \neq 92.48$) con 39 grados de libertad. Este valor representa la probabilidad de observar un valor tan extremo como $t = 0.6713$, bajo la suposición de que la hipótesis nula H_0 es verdadera.

Nos fijamos en la fila correspondiente a $gl = 39$ en la tabla t-Student (Tabla A3). Buscamos un valor cercano a $t = 0.6713$. Sin embargo, este valor no aparece directamente en la tabla, ya que los valores mínimos registrados son mayores (ver Figura 18), como

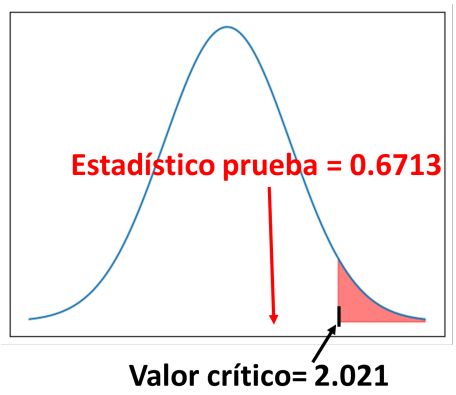


Figure 17: Identificando si el estadístico de prueba cae en región critica.

$t = 1.303$ (correspondiente a un área de dos colas $p = 0.20$). Dado que $t = 0.6713$ es menor que $t = 1.303$, concluimos que el área de dos colas asociada al estadístico $t = 0.6713$ es mayor que 0.20. Por lo tanto, el valor p es:

$$p > 0.20$$

Comparar con el nivel de significancia α: El valor p obtenido ($p > 0.20$) es claramente mayor que el nivel de significancia $\alpha = 0.20$. Esto implica que no existe suficiente evidencia para rechazar la hipótesis nula H_0.
Conclusión: Dado que $p > \alpha$, no se puede rechazar la hipótesis nula. Este resultado indica que no hay evidencia estadística significativa para concluir que la media de los rebotes de las pelotas sea diferente de 92.48 pulgadas.
En resumen, el método del valor p respalda la misma decisión que el método tradicional: no rechazamos H_0

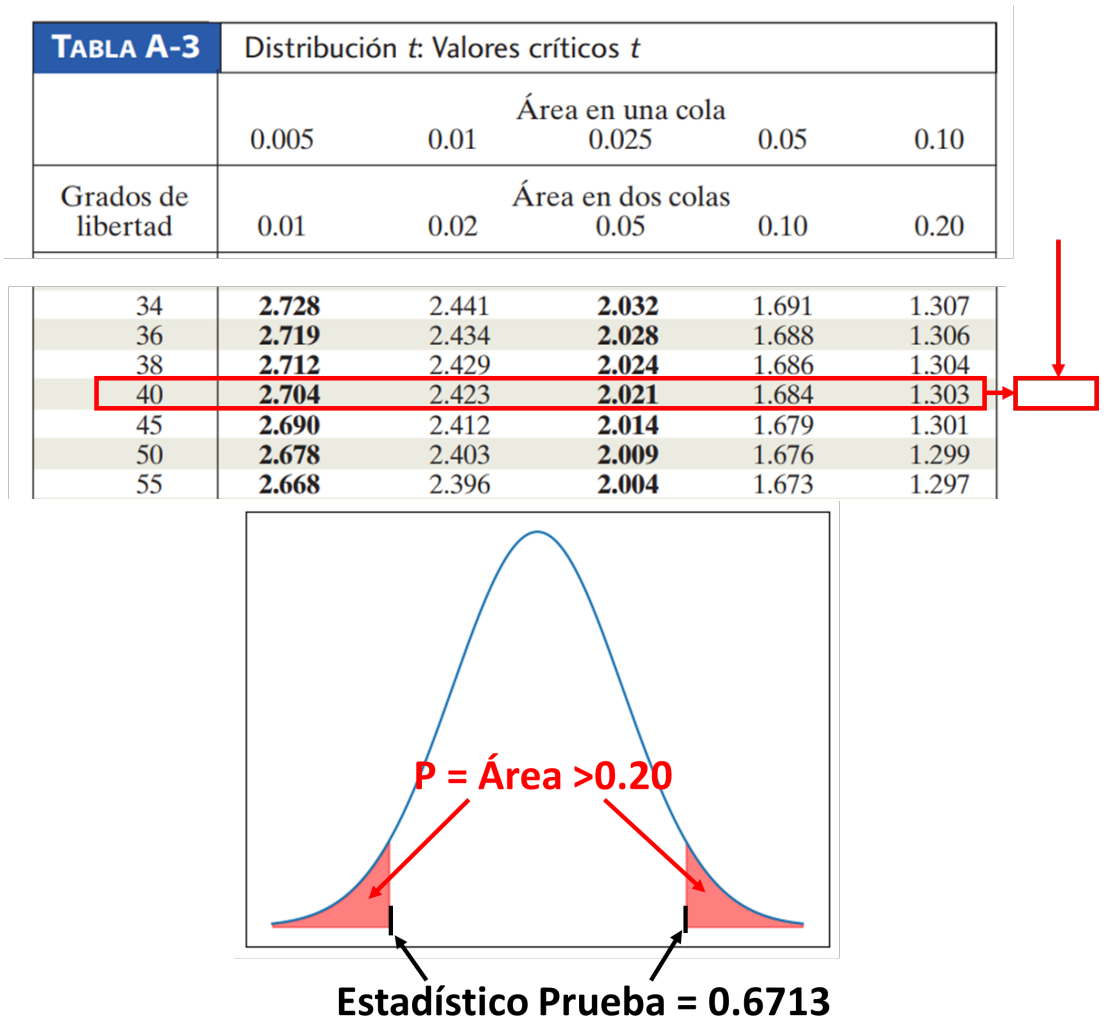


Figure 18: Calculando el área correspondiente al estadístico de prueba.

5 PRUEBA DE HIPOTESIS: DOS MUESTRAS

OBJETIVOS

Al finalizar esta unidad, el estudiante será capaz de:

1. Realizar pruebas de hipótesis para comparar proporciones entre dos muestras independientes.
2. Realizar pruebas de hipótesis para comparar medias entre dos muestras, tanto independientes como relacionadas.
3. Realizar pruebas de hipótesis para comparar desviaciones estándar entre dos muestras.
4. Aplicar adecuadamente las pruebas de hipótesis con dos muestras para resolver problemas prácticos, analizando e interpretando los resultados con rigor estadístico.

H_a Símbolo	Tipo de Prueba
$>$	Prueba de Cola Derecha
$<$	Prueba de Cola Izquierda
$\neq$	Prueba de Dos Colas

Tipo de Inferencia/Hipotesis	Tipo de Distribución	Tabla a Utilizar
Proporciones (p)	Distribución Gaussiana/Normal	Tabla A2
Medias (μ)	Distribución t-students	Tabla A3
Desviación Estándar	Distribución χ^2	Tabla A4

5.1 Tipos de pruebas

5.2 Tipos de distribuciones

5.3 Introducción

En diversas situaciones significativas y concretas, resulta fundamental emplear datos de muestras para contrastar dos poblaciones distintas. De hecho, podríamos afirmar con énfasis que esta sección es una de las más cruciales, ya que aborda métodos específicos para abordar comparaciones entre dos proporciones extraídas de muestras. Por ejemplo:

1. **Estudio de preferencias políticas en dos regiones:** En una investigación electoral, se pueden tomar muestras de dos regiones geográficas distintas para comparar las proporciones de votantes que apoyan a diferentes candidatos o partidos políticos. Esto puede proporcionar información valiosa sobre las preferencias políticas en cada área.

2. **Evaluación del desempeño de dos equipos de ventas:** Se podría analizar la media de las ventas mensuales generadas por cada equipo de ventas. Esta métrica proporcionaría una visión general del rendimiento promedio en términos de generación de ingresos. Una media más alta indicaría un rendimiento generalmente mejor en comparación con un equipo cuya media de ventas sea más baja.

5.4 Prueba para dos muestras PROPORCIONES (p)

. La prueba de hipótesis de dos muestras para proporciones es una técnica estadística utilizada para comparar dos proporciones muestrales con el objetivo de determinar si hay evidencia estadística suficiente para concluir que las proporciones en la población subyacente son diferentes. Este tipo de prueba es comúnmente utilizado en investigaciones y estudios que involucran dos grupos o condiciones distintas.

5.4.1 Supuestos

1. Se tienen proporciones de dos muestras aleatorias simples que son independientes.

2. Es necesario que se cumpla la condición $np \geq 5$ y $nq \geq 5$ para cada muestra. En otras palabras, se requiere que haya al menos cinco éxitos y cinco fracasos en cada muestra.

Notación

Para la población 1:

- p_1: proporción de la POBLACION
- n_1: tamaño de muestra
- x_1: número de éxitos en la muestra

- $\hat{p_1} = \frac{x_1}{n_1}$: proporción MUESTRAL
- $\hat{q_1} = 1 - \hat{p_1}$: Complemento de la proporción muestral

La misma notación se aplica para la muestra 2, utilizando las variables $p_2, n_2, x_2, \hat{p_2}$ y $\hat{q_2}$

5.4.2 Estadístico de Prueba

Se prueba la aseveración $H_0 : p_1 = p_2$.

Nota: Las posibles hipótesis alternativas H_a que se pueden corroborar en caso de rechazar la hipótesis nula son

$$H_a : p_1 \neq p_2, \qquad H_a : p_1 < p_2, \qquad H_a : p_1 > p_2$$

5.4.3 Estimado apareado

El estimado apareado de p_1 y p_2 se denota por $\bar{p}$ y se calcula como:

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

su complemento es $\bar{q} = 1 - \bar{p}$. Este estimado apareado representa la proporción de éxitos combinados de las dos muestras.

5.4.4 Estadístico de prueba:

Con la hipótesis nula $H_0 : p_1 = p_2$, el estadístico de prueba z se define como:

$$z = \frac{(\hat{p_1} - \hat{p_2}) - (p1 - p_2)}{\sqrt{\frac{\bar{p}\bar{q}}{n_1} + \frac{\bar{p}\bar{q}}{n_2}}}$$

Dado que asumimos en la hipótesis nula que $p_1 = p_2$, la ecuación anterior se simplifica al estadístico de prueba:

$$z = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\frac{\bar{p}\bar{q}}{n_1} + \frac{\bar{p}\bar{q}}{n_2}}}$$

5.4.5 Tipos de Prueba de hipótesis dos muestras proporciones (p):

- **Valor P:** Para obtener el Valor P, utilice la Tabla A2. Esto implica tomar el valor calculado del estadístico de prueba z y encontrar el área correspondiente a z en la Tabla A2. Rechace H_0 si el valor P es menor que el nivel se significación ($P < \alpha$)
- **Método tradicional:** determine los valores críticos basándose en el nivel de significancia α. Rechace H_0 si el estadístico de prueba z cae en la región crítica.

5.5 Prueba de hipótesis dos MEDIAS (μ_1 y μ_2): Muestras Independientes

La prueba de hipótesis de dos muestras para medias es una técnica estadística utilizada para comparar las medias de dos grupos o poblaciones diferentes y determinar si hay evidencia suficiente para concluir que estas medias son estadísticamente diferentes.

5.5.1 Muestras independientes:

Dos muestras son consideradas independientes cuando los valores seleccionados de una población no tienen ninguna relación, emparejamiento o asociación con los valores seleccionados de la otra población. En otras palabras, no hay ninguna conexión entre los elementos de una muestra y los de la otra. Ejemplo: Imagina que estás comparando la altura de estudiantes de dos universidades diferentes. Seleccionas aleatoriamente a estudiantes de una universidad y mides sus alturas, y luego

haces lo mismo con estudiantes de la otra universidad. Las muestras son independientes porque no hay relación o emparejamiento entre los estudiantes seleccionados de una universidad y los de la otra.

5.5.2 Muestras apareadas:

En contraste, si hay una relación tal que cada valor en una muestra está emparejado con un valor correspondiente en la otra muestra, entonces las muestras se consideran dependientes. Comúnmente, se denominan muestras dependientes o datos emparejados, y a menudo se utilizan términos como "datos emparejados" para describir con mayor precisión la naturaleza de estos datos. Ejemplo: Supongamos que estás evaluando el efecto de un nuevo medicamento. Mides la presión arterial de un grupo de pacientes antes de administrar el medicamento y luego mides la presión arterial nuevamente después de que hayan tomado el medicamento. En este caso, las muestras son emparejadas, ya que cada medición después del tratamiento está vinculada a una medición antes del tratamiento para el mismo paciente.

Supuestos

1. Se tienen proporciones de dos muestras aleatorias simples que son independientes.
2. Es necesario que se cumpla alguna de las siguientes condiciones:
   - Las muestras son grandes ($n_1 > 30$ y $n_2 > 30$)
   - Ambas muestras provienen de poblaciones con distribución normal

5.5.3 Estadístico de prueba de hipótesis para dos medias: Muestras Independientes

Se prueba la aseveración $H_0 : \mu_1 = \mu_2$.

Nota: Las posibles hipótesis alternativas H_a que se pueden corroborar en caso de que se rechazar la hipótesis nula son

$$H_a : \mu_1 \neq \mu_2, \qquad H_a : \mu_1 < \mu_2, \qquad H_a : \mu_1 > \mu_2$$

5.5.4 Estadístico de prueba:

Con la hipótesis nula $H_0 : \mu_1 = \mu_2$, el estadístico de prueba t se define como:

$$t = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Dado que asumimos en la hipótesis nula que $\mu_1 = \mu_2$, la ecuación anterior se simplifica al estadístico de prueba:

$$t = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Grados de libertad: Cuando calcule valores críticos o valores P, utilice lo siguiente para determinar el número de grados de libertad, denotados por gl = el más pequeño de $n_1 - 1$ y $n_2 - 1$.

5.5.5 Tipos de Pruebas de hipótesis dos muestras medias (μ): Muestras independientes

- **Valor P:** Para obtener el Valor P, utilice la Tabla A3. Esto implica tomar el valor calculado del estadístico de prueba t y encontrar el área correspondiente a t en la Tabla A3. Rechace H_0 si el valor P es menor que el nivel de significancia ($P < \alpha$)
- **Método tradicional:** determine los valores críticos basándose en el nivel de significancia α. Rechace H_0 si el estadístico de prueba t cae en la región crítica.

5.6 Prueba de hipótesis dos MEDIAS (μ_1 y μ_2): Muestras Apareadas

Supuestos

1. Se tienen proporciones de dos muestras aleatorias simples
2. Los datos muestrales consisten en datos apareados
3. Es necesario que se cumpla alguna de las siguientes condiciones:
   - Las muestras son grandes ($n > 30$)
   - Los pares de valores tienen diferencias que se toman de una distribución aproximadamente normal

Notación

- d : Es la diferencia individual entre dos valores en un solo par de datos apareados. Por ejemplo, si estás comparando la altura de una persona antes y después de un tratamiento, d sería la diferencia entre estas dos alturas para ese individuo específico.
- μ_d : Es el valor medio de todas las diferencias d para la POBLACIÓN COMPLETA de datos apareados. Representa la media de las diferencias entre todos los pares de datos apareados en la POBLACIÓN.
- $\bar{d}$: Es el valor medio de las diferencias d para la MUESTRA de datos apareados.
- s_d : Es la desviación estándar de las diferencias d para la muestra de datos apareados. Indica la medida de dispersión o variabilidad de las diferencias en la muestra.
- n : Representa el número total de pares de datos apareados en la muestra.

5.6.1 Estadístico de prueba de hipótesis para dos medias: datos apareados

Se prueba la aseveración $H_0 : \mu_d = 0$. Es decir, que no hay cambio en los datos apareados.

Nota: Las posibles hipótesis alternativas H_a que se pueden corroborar en caso de que se rechazar la hipótesis nula son

$$H_a : \mu_d \neq 0, \qquad H_a : \mu_d < 0, \qquad H_a : \mu_d > 0$$

5.6.2 Estadístico de prueba:

Con la hipótesis nula $H_0 : \mu_d = 0$, el estadístico de prueba t se define como:

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}}$$

donde se utilizan $n - 1$ grados de libertad.

Dado que asumimos en la hipótesis nula que $\mu_d = 0$, la ecuación anterior se simplifica al estadístico de prueba:

$$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}$$

5.6.3 Tipos de Pruebas de hipótesis dos muestras medias (μ Tabla A3): Muestras Apareadas

- **Valor P:** Para obtener el Valor P, utilice la Tabla A3. Esto implica tomar el valor calculado del estadístico de prueba t y encontrar el área correspondiente a t en la Tabla A3. Rechace H_0 si el valor P es menor que el nivel de significancia ($P < \alpha$)
- **Método tradicional:** determine los valores críticos basándose en el nivel de significancia α. Rechace H_0 si el estadístico de prueba t cae en la región crítica.

5.7 Prueba de Hipótesis dos muestra DESVIACIÓN ESTÁNDAR (σ_1 y σ_2)

Supuestos

1. Las dos poblaciones son independientes
2. Se tienen proporciones de dos muestras aleatorias simples
3. Las dos poblaciones están DISTRIBUIDAS NORMALMENTE.

Notación

- s_1^2 : Representa la varianza MUESTRAL más grande entre dos muestras.
- n_1 : Es el tamaño de la muestra correspondiente a la varianza más grande (s_1).
- σ_1^2 : Es la varianza de la POBLACIÓN de la cual se extrae la muestra que tiene la varianza más grande (s_1)

Se utilizan símbolos similares, como s_2^2, n_2 y σ_2^2 , para referirse a la otra muestra y población respectivamente.

5.7.1 Estadístico de prueba de hipotesis con dos varianzas

$$F = \frac{s_1^2}{s_2^2} \text{ donde } s_1 > s_2$$

5.7.2 Prueba de hipótesis

Utilizar tabla A5 para encontrar valores críticos de F

1. Utilice alguno de los niveles de significancia posibles ($\alpha = 0.025$, $\alpha = 0.05$)
2. Grados de libertad en el numerador $n_1 - 1$
3. Grados de libertad en el denominador $n_2 - 1$

5.8 EJERCICIOS

EJERCICIO 1: Comparación de satisfacción en el lugar de trabajo

Una empresa está interesada en saber si la proporción de satisfacción laboral entre sus trabajadores es igual a la de los jefes. Para ello, se realizó una encuesta a dos grupos

- **Trabajadores:** Se encuestaron de manera aleatoria 436 empleados, de los cuales 192 dijeron estar satisfechos con su trabajo.
- **Supervisores:** Se encuestaron de manera aleatoria 121 supervisores, de los cuales 40 dijeron estar satisfechos con su trabajo.

La empresa desea determinar, utilizando un nivel de significancia de $\alpha = 0.05$, si existe evidencia suficiente para concluir que las proporciones de satisfacción laboral son diferentes entre trabajadores y jefes.

EJERCICIO 2: Evaluación del efecto del consumo de marihuana en estudiantes universitarios

Se han realizado numerosos estudios para investigar cómo el consumo de marihuana afecta las capacidades cognitivas. En uno de estos estudios, se analizó la capacidad de recuperación de memoria en dos grupos de estudiantes universitarios:

- Consumidores ocasionales (uso esporádico de marihuana).
- Consumidores frecuentes (uso habitual y sostenido de marihuana).

A ambos grupos se les aplicó una prueba de recuperación de memoria, obteniendo los siguientes resultados (datos tomados de "The Residual Cognitive Effects of Heavy Marijuana Use in College Students", de Pope y Yurgelun-Todd, Journal of the American Medical Association, vol. 275, núm. 7):

- Artículos ordenados correctamente por consumidores ocasionales de marihuana: $n = 64$, $\bar{x} = 53.3$ y $s = 3.6$
- Artículos ordenados correctamente por consumidores frecuentes de marihuana: $n = 65$, $\bar{x} = 51.3$ y $s = 4.5$

Se desea probar la aseveración de que los consumidores frecuentes tienen una capacidad de recuperación de memoria significativamente (con un nivel de significancia de $\alpha = 0.01$) más baja que los consumidores ocasionales.

5.9 SOLUCIONES

5.9.1 Ejercicio 1

Problema: Una empresa está interesada en saber si la proporción de satisfacción laboral entre sus trabajadores es igual a la de los jefes. Para ello, se realizó una encuesta a dos grupos

- **Trabajadores:** Se encuestaron de manera aleatoria 436 empleados, de los cuales 192 dijeron estar satisfechos con su trabajo.
- **Supervisores:** Se encuestaron de manera aleatoria 121 supervisores, de los cuales 40 dijeron estar satisfechos con su trabajo.

La empresa desea determinar, utilizando un nivel de significancia de $\alpha = 0.05$, si existe evidencia suficiente para concluir que las proporciones de satisfacción laboral son diferentes entre trabajadores y jefes.

SOLUCION:

1. **Paso 1: Identificar datos:** La información que se proporciona es la siguiente:
   - **Trabajadores**
     - Número de trabajadores: $n_1 = 436$
     - Número de satisfechos: $x_1 = 192$
   - **Supervisores**
     - Número de supervisores: $n_2 = 121$
     - Número de satisfechos: $x_2 = 40$
   - **Nivel de significancia:** $\alpha = 0.05$
2. **Paso 2: Formulación de la hipótesis**
   - Hipótesis nula (H_0): $p_1 = p_2$, las proporciones de satisfacción laboral son iguales
   - Hipótesis alternativa (H_a): $p_1 \neq p_2$, las proporciones de satisfacción laboral son diferentes.
3. **Paso 3: Cálculo del Estimador Agrupado:**
   El estimador agrupado $\bar{p}$ combina las proporciones muestrales de ambos grupos:
   $$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$$
   Sustituimos los valores:
   $$\begin{aligned} \bar{p} &= \frac{192 + 40}{436 + 121} \\ &= \frac{232}{557} \\ &= 0.4165 \end{aligned}$$
   Por lo tanto, $\bar{p} = 0.4165$ y su complemento $\bar{q} = 1 - \bar{p} = 0.5835$
4. **Paso 4: Cálculo del estadístico de prueba** z Primero, calculamos las proporciones muestrales de cada grupo:
   $$\hat{p_1} = \frac{x_1}{n_1} = \frac{192}{436} = 0.4403$$
   $$\hat{p_2} = \frac{x_2}{n_2} = \frac{40}{121} = 0.3305$$

Luego, el estadístico de prueba z se calcula como:

$$\begin{aligned} z &= \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\frac{\bar{p}\bar{q}}{n_1} + \frac{\bar{p}\bar{q}}{n_2}}} \\ &= \frac{0.4403 - 0.3305}{\sqrt{\frac{0.4165*0.5835}{436} + \frac{0.4165*0.5835}{121}}} \\ &= \frac{0.1098}{\sqrt{0.0005574 + 0.002008}} \\ &= \frac{0.1098}{\sqrt{0.0025654}} \\ &= \frac{0.1098}{0.050649} \\ &= 2.1678 \end{aligned}$$

Por lo tanto, el estadístico de prueba es:

$$z = 2.1678$$

5. **Paso 5: Decisión sobre hipótesis:**

   - **MÉTODO TRADICIONAL:** El método tradicional consiste en determinar la región crítica correspondiente al nivel de significacia α y verificar si el estadístico de prueba $z = \pm 2.1678$ cae dentro de esta región para tomar una decisión. Para ello, primero es necesario identificar el valor crítico correspondiente.

     Determinación del valor critico

     Como la hipótesis alternativa es de dos cola ($H_a : p_1 \neq p_2$), el nivel de significancia se divide entre dos:

     $$\alpha/2 = 0.05/2 = 0.025$$

     Para encontrar el valor crítico en pruebas de hipótesis para proporciones, utilizamos la distribución normal estándar para un área de $\alpha = 0.025$. En la Figura 19 se presenta parte de la tabla A2. Buscamos un área igual a 0.025.

     Para hallar el valor crítico, buscamos en la tabla hacia la izquierda (1.9) y hacia arriba (0.06). Al sumar estos valores (1.9+0.06 = 1.96), obtenemos que el valor crítico es ± 1.96

TABLA A-2 Distribución normal estándar (z): Área acumulativa de la IZQUIERDA

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367

Figure 19: Tabla A2: para determinar valores críticos de proporciones.

Decisión sobre la hipótesis nula:

Verificamos si el estadístico de prueba $z = -2.1678$ cae en la región critica. Debido a que $z = -2.1675$ esta a la izquierda del valor critico $z_c = -1.96$, entonces se rechaza la hipótesis nula. Esto es verificado con la representación visual en la Figura 20 que muestra claramente que el estadístico de prueba está dentro de la región crítica. Por lo tanto, se rechaza la hipótesis nula, y los datos muestrales respaldan la hipótesis alternativa. En consecuencia, podemos afirmar con alta confianza que $H_a : p_1 \neq p_2$ es verdadera.

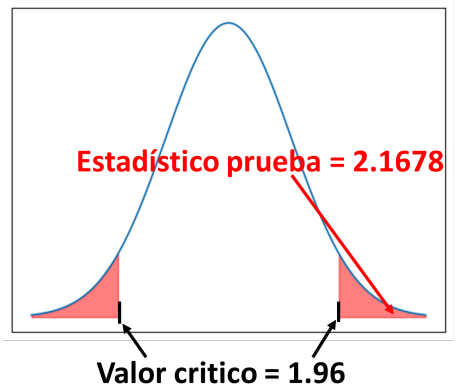


Figure 20: Identificando si el estadístico de prueba cae en región critica.

- **MÉTODO DEL VALOR** p

  La hipótesis alternativa es $H_a : p_1 \neq p_2$, lo que indica una prueba de dos colas. Para realizar la prueba utilizando el valor p, es necesario calcular el área bajo la curva en ambos extremos (izquierdo y derecho) de la distribución normal. Debido a la simetría de esta distribución, basta con calcular el área bajo la curva del lado izquierdo y multiplicar ese valor por dos.

  Para determinar el área correspondiente al estadístico de prueba $z = -2.1678$, primero se redondea este valor a dos decimales, obteniendo $z = -2.17$. A continuación, se descompone en dos partes:

  – Las décimas (-2.1) se localizan en las filas de la tabla.
  – Las centésimas (0.07) se identifican en las columnas.

  El cruce entre la fila -2.1 y la columna 0.07 en la tabla (ver Figura 21) de la distribución normal estándar permite encontrar el valor $p = 0.0150$, que representa el área bajo la curva para $z = -2.17$. Debido a que es una prueba de dos colas este valor se tiene que multiplicar por dos:

$$p = 2 * 0.0150 = 0.03$$

TABLA A-2 (*continuación*) Área acumulativa de la IZQUIERDA

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233

Figure 21: Tabla A2: para determinar valores P.

Decisión sobre la hipótesis nula:

Dado que el valor $p = 0.03$ es menor que el valor de significancia $\alpha = 0.05$, se rechaza la hipótesis nula $H_0 : p_1 = p_2$. Esto implica que hay suficiente evidencia estadística para concluir que las proporciones de satisfacción laboral entre trabajadores y jefes son significativamente diferentes.

5.9.2 Ejercicio 2

PROBLEMA: Se han realizado numerosos estudios para investigar cómo el consumo de marihuana afecta las capacidades cognitivas. En uno de estos estudios, se analizó la capacidad de recuperación de memoria en dos grupos de estudiantes universitarios:

- Consumidores ocasionales (uso esporádico de marihuana).
- Consumidores frecuentes (uso habitual y sostenido de marihuana).

A ambos grupos se les aplicó una prueba de recuperación de memoria, obteniendo los siguientes resultados (datos tomados de "The Residual Cognitive Effects of Heavy Marijuana Use in College Students", de Pope y Yurgelun-Todd, Journal of the American Medical Association, vol. 275, núm. 7):

- Artículos ordenados correctamente por consumidores ocasionales de marihuana: $n = 64$, $\bar{x} = 53.3$ y $s = 3.6$
- Artículos ordenados correctamente por consumidores frecuentes de marihuana: $n = 65$, $\bar{x} = 51.3$ y $s = 4.5$

Se desea probar la aseveración de que los consumidores frecuentes tienen una capacidad de recuperación de memoria significativamente (con un nivel de significancia de $\alpha = 0.01$) más baja que los consumidores ocasionales.

SOLUCION:

- **Paso 1: Identificar datos:**
  Se analizan dos grupos de consumidores de marihuana, con los siguientes datos muestrales resumidos en la Tabla 7

	Consumidores Ocasionales	Consumidores Frecuentes
Tamaño Muestral	$n_2 = 64$	$n_1 = 65$
Media Muestral	$\bar{x}_2 = 53.3$	$\bar{x}_1 = 51.3$
Desviación Estándar Muestral	$s_2 = 3.6$	$s_1 = 4.5$

Table 7

- **Paso 2: Formulación de la hipótesis:**
  El objetivo de este análisis es determinar si la media de rendimiento o desempeño en una tarea específica es significativamente más baja en consumidores frecuentes de marihuana en comparación con consumidores ocasionales. Para ello, se plantea la siguiente hipótesis:
  - Hipótesis nula (H_0): $\mu_1 < \mu_2$, es decir, las medias poblacionales son iguales.
  - Hipótesis alternativa ($H_a : \mu_1 < \mu_2$), lo que indica que la media de los consumidores frecuentes (μ_1) es menor que la de los consumidores ocasionales (μ_2)

  Este es un caso de prueba de una cola debido a la dirección específica de la hipótesis alternativa.

- **Paso 3: Calculo del estadístico de prueba:** Para comparar las medias muestrales, calculamos el estadístico (t) utilizando la fórmula:

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

Sustituyendo los valores:

$$\begin{aligned} t &= \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \\ &= \frac{51.3 - 53.3}{\sqrt{\frac{4.5^2}{65} + \frac{3.6^2}{64}}} \\ &= \frac{-2}{\sqrt{0.311538 + 0.2025}} \\ &= \frac{-2}{\sqrt{0.311538 + 0.2025}} \\ &= \frac{-2}{0.716964} \\ &\approx -2.79 \end{aligned}$$

Por lo tanto, el estadístico de prueba es:

$$t = -2.79$$

- **Paso 4: Decisión sobre hipótesis. Método Valor p:**
  El estadístico de prueba es $t = -2.79$. Para determinar si se rechaza la hipótesis nula, se utiliza la distribución t-Student con grados de libertad $gl = n_1 - 1 = 65 - 1 = 64$ para obtener el valor p.

  Debido a que no se dispone exactamente de 64 grados de libertad en la tabla t, se aproxima a 65 grados de libertad. Para una prueba de una cola con $t = -2.79$, buscamos el valor más cercano en la tabla:

  - Con $gl = 65$, el valor crítico es $t = -2.654$, corresponde a un área de $p = 0.005$.

  Dado que $t = -2.79$ es menor que $t_c = -2.654$, se concluye que $p < 0.005$. Comparando esto con el nivel de significancia $\alpha = 0.01$, se observa que $p < \alpha$, lo cual lleva a rechazar la hipótesis nula.

- **Conclusión:**
  Con base en los resultados, se concluye que los consumidores frecuentes de marihuana tienen un rendimiento promedio significativamente menor que los consumidores ocasionales. Este hallazgo respalda la idea de que el consumo frecuente de marihuana podría estar asociado con efectos negativos en el rendimiento. Por lo tanto, es motivo de preocupación, especialmente en el contexto de estudiantes universitarios, ya que podría afectar su desempeño académico y en otras actividades.

6 ANALISIS DE VARIANZA - ANOVA

OBJETIVOS

Al finalizar esta unidad, el estudiante será capaz de:

1. Comprender el propósito del análisis de varianza (ANOVA) y su utilidad para comparar tres o más medias de grupos.
2. Identificar los supuestos básicos del ANOVA y verificar su cumplimiento en un conjunto de datos.
3. Calcular e interpretar el estadístico F en un análisis de varianza de una vía (ANOVA de un factor).
4. Formular e interpretar correctamente las hipótesis nula y alternativa en el contexto de ANOVA.
5. Analizar los resultados del ANOVA e interpretar su significado en el contexto del problema, incluyendo la comparación de grupos si es necesario.

6.1 ANOVA

En temas previos, aprendiste a realizar pruebas de hipótesis para una y dos poblaciones en casos relacionados con proporciones (p), medias (μ) y desviación estándar (σ). Ahora exploraremos una metodología diseñada para probar hipótesis cuando se comparan tres o más poblaciones, específicamente en el caso de las MEDIAS (μ).

El Análisis de Varianza (ANOVA) es una técnica utilizada para determinar si las medias de tres o más poblaciones son iguales, analizando la variabilidad entre las muestras.

Algunos ejemplos comunes donde se requiere comparar tres o más medias poblacionales son los siguientes:

1. **Comparación de métodos de enseñanza:** Supongamos que queremos evaluar si el rendimiento en exámenes varía significativamente entre tres métodos de enseñanza: Método A, Método B y Método C. Contamos con muestras de estudiantes que han seguido cada método y deseamos determinar si existen diferencias significativas en sus resultados.
2. **Evaluación de dietas en adolescentes:** Imaginemos que investigamos la influencia de tres dietas diferentes (Dieta A, Dieta B y Dieta C) en el crecimiento en estatura de adolescentes. Tomamos muestras de tres grupos, cada uno siguiendo una dieta distinta, y registramos sus alturas después de seis meses para comparar los resultados.

El ANOVA compara dos estimaciones diferentes de la variabilidad observada entre los grupos o poblaciones:

- La variabilidad **entre grupos**, que mide las diferencias en las medias de los distintos grupos.
- La variabilidad **dentro de los grupos**, que mide las diferencias individuales dentro de cada grupo.

El objetivo es determinar si las diferencias observadas entre las medias de los grupos son lo suficientemente grandes como para no atribuirse al azar, lo que indicaría la existencia de diferencias reales entre las poblaciones.

En una prueba de hipótesis ANOVA, se plantea como punto de partida la hipótesis nula H_0, que asume que todas las poblaciones tienen medias iguales. Matemáticamente, esto se expresa como:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$

donde $\mu_1, \mu_2, \ldots, \mu_k$ representan las medias de las k poblaciones o grupos que se están analizando.

En resumen, el ANOVA es una herramienta clave para evaluar si las diferencias en las medias observadas entre múltiples grupos son estadísticamente significativas, ofreciendo un enfoque riguroso y sistemático para estas comparaciones.

6.2 Estadístico de Prueba

El análisis de varianza (ANOVA) se basa en un concepto clave: asumimos que las k poblaciones tienen la misma varianza (s^2). Para estimar este valor común de s^2utilizamos dos métodos diferentes. El estadístico de prueba F es crucial en este análisis; se define como la proporción entre estos dos estimados.

Un valor significativamente grande de F, ubicado en la cola derecha de la distribución F, sugiere evidencia en contra de la hipótesis nula H_0, que afirma que las medias poblacionales son iguales.

Ahora, los dos métodos para estimar s^2:

1. La varianza entre muestras (también llamada variación debida al tratamiento) es una estimación de $\bar{s}^2$, basada en la variabilidad entre las medias de las diferentes muestras. Representa cuánto difieren las medias de las muestras unas de otras.

2. La varianza dentro de las muestras (también llamada variación debida al error) es otra estimación de s^2, pero basada en la variabilidad dentro de cada muestra individual. Representa las diferencias entre los valores observados y las medias de las muestras.

Estadístico de Prueba del ANOVA

El estadístico de prueba se calcula como:

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

Un valor de F alto indica que la variabilidad entre las medias de los grupos es significativamente mayor que la variabilidad dentro de los grupos, lo que podría evidenciar diferencias reales entre las medias poblacionales.

Grados de libertad

Para el análisis de F, los grados de libertad son los siguientes:

- k: número de grupos o muestras.
- n: tamaño total de la muestra.
- Grados de libertad entre grupos: $k - 1$
- Grados de libertad dentro de los grupos: $n - k$

6.3 Muestras con tamaños muestrales iguales (n)

Cuando se trabaja con muestras que tienen el mismo tamaño (n), el cálculo de la varianza entre muestras y la varianza dentro de las muestras se realiza siguiendo estos pasos:

1. **La varianza entre muestras:**

   - **Calcula la media de cada grupo:** Para cada grupo o tratamiento, calcula la media de las observaciones dentro del grupo:

     $$\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k$$

     donde $\bar{X}_i$ representa la media de muestra i.
   - **Calcula la varianza entre muestras** $s^2_{\bar{X}}$: es igual a la *varianza de los datos anteriores.*

2. **Varianza dentro de las muestras:**
   La varianza dentro de las muestras (también llamada variación debida al error) es otra estimación de s^2, esta vez fundamentada en las varianzas de las muestras individuales.

- **Calcula la varianza de cada grupo:** Para cada grupo o tratamiento, calcula la varianza de las observaciones dentro de ese grupo:

$$s_1^2, s_2^2, \ldots, s_k^k$$

donde s_i^2 representa la varianza de muestra i.

- **Calcula la varianza dentro de las muestras s_P^2:** Esta es la media de las varianzas calculadas para los grupos.

6.3.1 Estadístico de Prueba del ANOVA

El estadístico de prueba se calcula como:

$$F = \frac{\text{Varianza entre muestras}}{\text{Varianza dentro de muestras}} = \frac{n * s_{\bar{X}}^2}{s_P^2}$$

Este estadístico compara la variabilidad entre las medias de los grupos con la variabilidad dentro de los grupos para determinar si las diferencias en las medias son estadísticamente significativas.

6.3.2 Valor crítico:

Con los grados de libertad y el nivel de significancia α, es posible obtener el valor crítico utilizando la **Tabla A5**. Es importante destacar que los valores críticos solo están disponibles para los niveles de significancia $\alpha = 0.05$ y $\alpha = 0.025$

6.4 Muestras con tamaños muestrales desiguales

Cuando trabajas con muestras que tienen tamaños diferentes $(n_1, n_2, \ldots, n_k)$, la determinación de la varianza entre muestras y la varianza dentro de muestras implica los siguientes pasos:

1. **Varianza entre muestras ($S_{\bar{X}}^2$):** La varianza entre las muestras, o variabilidad debida a los tratamientos, se calcula de manera ponderada, ya que los tamaños de las muestras pueden ser diferentes. La fórmula para calcularla es:

   - **Calcula la media de cada grupo:** Para cada grupo o tratamiento, calcula la media de las observaciones en ese grupo.

$$\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k$$

   donde $\bar{X}_i$ representa la media de muestra i.

   - **Varianza entre muestras (PONDERADA)**
   La varianza entre muestras ponderada se calcula como

$$s_{\bar{X}}^2 = \frac{\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X}_{mean})^2}{k-1}$$

   donde

$$\bar{X}_{mean} = \frac{\bar{X}_1 + \bar{X}_2 + \ldots + \bar{X}_k}{k}$$

   Este valor $\bar{X}_{mean}$ es el promedio de las medias de todos los grupos. La diferencia $(\bar{X}_i - \bar{X}_{mean})$ refleja cuánto se desvía cada media grupal de la media general.

2. **La varianza dentro de las muestras (S_P^2):** también llamada variación debida al error es otra estimación de s^2, esta vez fundamentada en las varianzas de las muestras individuales.

- **Calcula la varianza de cada grupo:** Para cada grupo o tratamiento, calcula la varianza de las observaciones en ese grupo.

$$s_1^2, s_2^2, \ldots, s_k^k$$

donde s_i^2 es la varianza de muestra i

- **Varianza dentro de muestras (PONDERADA)**

$$s_P^2 = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{\sum_{i=1}^{k}(n_i - 1)}$$

La fórmula refleja que las varianzas de los grupos más grandes tienen más peso en el cálculo de la varianza global, lo cual es importante cuando los tamaños de muestra son desiguales.

6.4.1 Estadístico de Prueba del ANOVA

Sustituyendo las fórmulas de las varianzas, la fórmula completa del estadístico F para tamaños de muestras desiguales es:

$$F = \frac{\text{Varianza entre muestras}}{\text{Varianza dentro de muestras}} = \frac{s_{\bar{X}}^2}{s_P^2} = \frac{\frac{\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X}_{mean})^2}{k-1}}{\frac{\sum_{i=1}^{k}(n_i-1)s_i^2}{\sum_{i=1}^{k}(n_i-1)}}$$

donde

$$\bar{X}_{mean} = \frac{\bar{X}_1 + \bar{X}_2 + \ldots + \bar{X}_k}{k}$$

Este valor de F se utiliza para determinar si las diferencias entre las medias de los grupos son estadísticamente significativas. Si el valor de F es mayor que el valor crítico de la distribución F correspondiente a un nivel de significancia preestablecido, se puede rechazar la hipótesis nula.

Grados de libertad

Los grados de libertad (GL) en un análisis de varianza se distribuyen entre el numerador y el denominador de la fórmula del estadístico

- Grados de libertad para el numerador: Se corresponden con el número de grupos menos uno, es decir, $k - 1$. Estos grados de libertad reflejan la cantidad de variabilidad explicada por las diferencias entre las medias de los grupos.
- Grados de libertad para el denominador: Se calculan sumando $(n_i - 1)$ para cada grupo i, lo cual refleja la variabilidad interna dentro de cada grupo. La fórmula es:

$$\sum_{i=1}^{k}(n_i - 1)$$

6.4.2 Valor crítico:

Con los grados de libertad y el nivel de significancia α, es posible obtener el valor crítico utilizando la **Tabla A5**. Es importante destacar que los valores críticos solo están disponibles para niveles de significancia de $\alpha = 0.05$ y $\alpha = 0.025$

6.4.3 Interpretación del Estadístico F

Una vez calculado el valor F, se compara con el valor crítico F_α obtenido de la tabla de distribución F para el nivel de significancia α elegido (por ejemplo, 0.05) y los grados de libertad correspondientes. Si F es mayor que F_α, se rechaza la hipótesis nula y se concluye que hay diferencias significativas entre las medias de los grupos. Si F es menor que F_α, no se rechaza la hipótesis nula, lo que sugiere que las medias de los grupos no son significativamente diferentes.

6.5 EJERCICIOS

EJERCICIO 1: Energía solar en diferentes climas Un alumno del autor vive en una casa con un sistema eléctrico solar. A la misma hora, cada día, reunió las lecturas de voltaje de un medidor que conectó al sistema, cuyos resultados se listan en la tabla adjunta. Utilice un nivel de significancia de 0.05 para probar la aseveración de que la lectura media de voltaje es la misma en los tres tipos distintos de días. ¿Hay evidencia suficiente para sustentar una aseveración de medias poblacionales diferentes? Esperaríamos que un sistema solar proporcione más energía eléctrica los días soleados que los días nublados o lluviosos. ¿Concluiríamos que los días soleados dan como resultado mayores cantidades de energía eléctrica?

Días Soleados	Días Nublados	Días lluviosos
13.5	12.7	12.1
13.0	12.5	12.2
13.2	12.6	12.3
13.9		

6.6 SOLUCION

6.6.1 Ejercicio 1

EJERCICIO 1: Energía solar en diferentes climas Un alumno del autor vive en una casa con un sistema eléctrico solar. A la misma hora, cada día, reunió las lecturas de voltaje de un medidor que conectó al sistema, cuyos resultados se listan en la tabla adjunta. Utilice un nivel de significancia de 0.05 para probar la aseveración de que la lectura media de voltaje es la misma en los tres tipos distintos de días. ¿Hay evidencia suficiente para sustentar una aseveración de medias poblacionales diferentes? Esperaríamos que un sistema solar proporcione más energía eléctrica los días soleados que los días nublados o lluviosos. ¿Concluiríamos que los días soleados dan como resultado mayores cantidades de energía eléctrica?

Días Soleados	Días Nublados	Días lluviosos
13.5	12.7	12.1
13.0	12.5	12.2
13.2	12.6	12.3
13.9		

SOLUCION: Para probar la aseveración de que la lectura media de voltaje es la misma en los tres tipos distintos de días, puedes utilizar un análisis de varianza (ANOVA) ya que tienes más de dos grupos para comparar. El planteamiento de hipótesis es el siguiente:

1. **Paso 1: Hipótesis nula (H_0)**: La hipótesis nula plantea que no hay diferencias en las medias de voltaje entre los días soleados, nublados y lluviosos. Esto implica que cualquier variación observada en los datos se debe al azar y no a diferencias significativas en las condiciones meteorológicas.

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

2. **Paso 2: Hipótesis alternativa (H_1)**: La hipótesis alternativa propone que al menos una de las medias de voltaje es significativamente diferente de las demás. Esto significa que las condiciones climáticas influyen en el voltaje de los paneles solares.

$$H_a : \text{Al menos una de las medias es diferente.}$$

3. **Paso 3: Calculo de promedios y varianzas:** Se tienen datos de 3 muestras, cada una correspondiente a un grupo (soleado, nublado, lluvioso). Es necesario recopilar los siguientes valores para cada grupo: el numero de sujetos (n), medias ($\bar{X}$) y varianzas s^2

   (a) **Días soleados**
   Se tienen los siguientes datos: 13.5, 13.0, 13.2, 13.9

   Cálculo de la media:

$$n_1 = 4$$
$$\bar{X}_1 = \frac{13.5 + 13.0 + 13.2 + 13.9}{4} = \frac{53.6}{4} = 13.4$$

   Cálculo de varianza:

$$\begin{aligned} s_1^2 &= \frac{(13.5 - 13.4)^2 + (13.0 - 13.4)^2 + (13.2 - 13.4)^2 + (13.9 - 13.4)^2}{4 - 1} \\ &= \frac{0.01 + 0.16 + 0.04 + 0.25}{3} \\ &= \frac{0.46}{3} \\ &= 0.1533 \end{aligned} \tag{12}$$

Por lo tanto:

$$n_1 = 4, \bar{X}_1 = 13.4, s_1^2 = 0.1533$$

(b) Días Nublados Se tienen los siguientes datos: 12.7, 12.5, 12.6

Cálculo de la media:

$$n_2 = 3$$
$$\bar{X}_2 = \frac{12.7 + 12.5 + 12.6}{3} = \frac{37.8}{3} = 12.6$$

Cálculo de varianza:

$$\begin{aligned} s_2^2 &= \frac{(12.7 - 12.6)^2 + (12.5 - 12.6)^2 + (12.6 - 12.6)^2}{3 - 1} \\ &= \frac{0.01 + 0.01 + 0.0}{2} \\ &= \frac{0.02}{2} \\ &= 0.01 \end{aligned}$$

Por lo tanto:

$$n_2 = 3, \bar{X}_2 = 12.6, s_2^2 = 0.01$$

(c) Días Lluviosos Se tienen los siguientes datos: 12.1, 12.2, 12.3

Cálculo de la media:

$$n_3 = 3$$
$$\bar{X}_3 = \frac{12.1 + 12.2 + 12.3}{3} = \frac{36.6}{3} = 12.2$$

Cálculo de varianza:

$$\begin{aligned} s_3^2 &= \frac{(12.1 - 12.2)^2 + (12.2 - 12.2)^2 + (12.3 - 12.2)^2}{3 - 1} \\ &= \frac{0.01 + 0.0 + 0.01}{2} \\ &= \frac{0.02}{2} \\ &= 0.01 \end{aligned}$$

Por lo tanto:

$$n_3 = 3, \bar{X}_3 = 12.2, s_3^2 = 0.01$$

En resumen se tiene:

Días Soleados	Días Nublados	Días lluviosos
$n_1 = 4$	$n_2 = 3$	$n_3 = 3$
$\bar{X}_1 = 13.4$	$\bar{X}_2 = 12.6$	$\bar{X}_3 = 12.2$
$s_1^2 = 0.1533$	$s_2^2 = 0.01$	$s_3^2 = 0.01$

4. **Paso 4: Calcular $\bar{X}_{mean}$:** Cuando las muestras tienen tamaños diferentes, calcular la media global ($\bar{X}_{mean}$) implica tomar el promedio de las medias de las muestras individuales. Esto se hace de la siguiente manera:

$$\bar{X}_{mean} = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3}{3}$$

Sustituyendo los valores obtenidos:

$$\bar{X}_{mean} = \frac{13.4 + 12.6 + 12.2}{3} = \frac{38.2}{3} = 12.73$$

Por lo tanto, el valor de la media global es:

$$\bar{X}_{mean} = 12.73$$

5. **Paso 5: Estadístico de prueba.** Con los datos obtenidos en los pasos anteriores, podemos calcular el estadístico de prueba F, que está dado por:

$$F = \frac{\frac{\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X}_{mean})^2}{k-1}}{\frac{\sum_{i=1}^{k}(n_i-1)s_i^2}{\sum_{i=1}^{k}(n_i-1)}}$$

Sustituyendo los valores específicos para $k = 3$ (número de grupos):

$$F = \frac{\frac{n_1(\bar{X}_1 - \bar{X}_{mean})^2 + n_2(\bar{X}_2 - \bar{X}_{mean})^2 + n_3(\bar{X}_3 - \bar{X}_{mean})^2}{3-1}}{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + (n_3-1)s_3^2}{(n_1-1)+(n_2-1)+(n_3-1)}}$$

Sustituyendo los valores de cada parámetro:

$$\begin{aligned}
F &= \frac{\frac{4(13.4-12.73)^2 + 3(12.6-12.73)^2 + 3(12.2-12.73)^2}{3-1}}{\frac{(4-1)0.1533 + (3-1)0.01 + (3-1)0.01}{(4-1)+(3-1)+(3-1)}} \\
&= \frac{\frac{4(0.4489) + 3(0.0169) + 3(0.2809)}{2}}{\frac{(3)(0.1533) + (2)(0.01) + (2)(0.01)}{3+2+2}} \\
&= \frac{\frac{1.7956 + 0.0507 + 0.8427}{2}}{\frac{0.4599 + 0.02 + 0.02}{7}} \\
&= \frac{\frac{2.689}{2}}{\frac{0.4999}{7}} \\
&= \frac{1.3445}{0.0714} \\
&= 18.83
\end{aligned}$$

Por lo cual, el estadístico de prueba esta dado por $F = 18.82$

Nota: Este valor indica qué tan grande es la variabilidad entre las medias de los grupos en comparación con la variabilidad dentro de los grupos. Un valor F alto puede sugerir diferencias significativas entre las medias.

6. **Paso 6: Obtener el valor crítico para** $\alpha = 0.05$. Para determinar el valor crítico del estadístico F, debemos calcular primero los grados de libertad correspondientes al numerador y al denominador:

- **Grados de libertad del numerador $(k-1)$:**

$$k - 1 = 3 - 1 = 2$$

- **Grados de libertad del denomidador $\sum_{i=1}^{k}(n_i - 1)$:**

$$\sum_{i=1}^{k}(n_i - 1) = (4-1) + (3-1) + (3-1) = 7$$

Con los grados de libertad calculados ($gl_{numerador} = 2$ y $gl_{denominador} = 7$), utilizamos la Tabla A5 de valores críticos del estadístico F para un nivel de significancia de $\alpha = 0.05$

(a) De acuerdo con la tabla (ver Fig. 22), el valor crítico es:

$$F_{critico} = 4.7374$$

Este valor indica el punto de corte para rechazar la hipótesis nula. Si el valor del estadístico F calculado es mayor que $F_{critico}$, rechazaremos la hipótesis nula en favor de la alternativa.

TABLA A-5 Distribución F ($\alpha = 0.025$ en la cola derecha)

	Grados de libertad del numerador (gl_1)								
	1	2	3	4	5	6	7	8	9
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385
3	10.128	9.5521	9.2766	9.1172	9.0135	8.9406	8.8867	8.8452	8.8123
4	7.7086	6.9443	6.5914	6.3882	6.2561	6.1631	6.0942	6.0410	6.9988
5	6.6079	5.7861	5.4095	5.1922	5.0503	4.9503	4.8759	4.8183	4.7725
6	5.9874	5.1433	4.7571	4.5337	4.3874	4.2839	4.2067	4.1468	4.0990
7	5.5914	4.7374	4.3468	4.1203	3.9715	3.8660	3.7870	3.7257	3.6767
8	5.3177	4.4590	4.0662	3.8379	3.6875	3.5806	3.5005	3.4381	3.3881
9	5.1174	4.2565	3.8625	3.6331	3.4817	3.3738	3.2927	3.2296	3.1789
10	4.9646	4.1028	3.7083	3.4780	3.3258	3.2172	3.1355	3.0717	3.0204
11	4.8443	3.9823	3.5874	3.3567	3.2039	3.0946	3.0123	2.9480	2.8962
12	4.7472	3.8853	3.4903	3.2592	3.1059	2.9961	2.9134	2.8486	2.7964
13	4.6672	3.8056	3.4105	3.1791	3.0254	2.9153	2.8321	2.7669	2.7144
14	4.6001	3.7389	3.3439	3.1122	2.9582	2.8477	2.7642	2.6987	2.6458

minador (gl_2)

Figure 22: Tabla A2: para determinar valores críticos de proporciones.

7. **Conclusiones:**

- **Decisión sobre la hipótesis nula:**
  El valor calculado del estadístico de prueba $F = 18.83$ es mayor que el valor crítico $F_{critico} = 4.7374$. Esto implica que el estadístico de prueba se encuentra dentro de la región crítica, lo que nos lleva a **rechazar la hipótesis nula (H_0)**.
- **Interpretación general:**
  Rechazar la hipótesis nula significa que tenemos evidencia suficiente para afirmar que al menos una de las medias poblacionales de las tres categorías (días soleados, nublados y lluviosos) es significativamente diferente de las demás.
- **Limitaciones de la prueba ANOVA:**
  Aunque la prueba ANOVA confirma que existen diferencias significativas entre las medias, no identifica específicamente cuáles medias son diferentes entre sí ni en qué dirección están las diferencias. Por ejemplo, no podemos concluir directamente si los días soleados generan sistemáticamente mayores cantidades de energía eléctrica en comparación con los días nublados o lluviosos.

7 REGRESION LINEAL SIMPLE

OBJETIVOS

Al finalizar esta unidad, el estudiante será capaz de:

1. Comprender el concepto de regresión lineal simple y su utilidad para modelar la relación entre dos variables cuantitativas.
2. Estimar los parámetros del modelo de regresión lineal simple a partir de datos muestrales.
3. Utilizar el modelo de regresión para hacer predicciones e interpretar los resultados en el contexto del problema.

7.1 Calculo de coeficientes de regresión

La regresión lineal simple es un método estadístico ampliamente utilizado para modelar la relación entre dos variables: una variable dependiente (y), que se desea predecir o explicar, y una variable independiente (x), que se utiliza como base para realizar dicha predicción. Este modelo asume que la relación entre x e y es lineal.

- **Variable dependiente (y):** Es la variable que intentamos explicar o predecir mediante el modelo.
- **Variable independiente (x):** Es la variable que utilizamos para hacer la predicción de y.

Dado un conjunto de datos muestrales (x_i, y_i) para $i = 1, 2, \ldots, n$, el objetivo es encontrar la ecuación de la línea:

$$\hat{y} = b_0 + b_1 x,$$

donde:

- b_0 es el intercepto o término independiente, que representa el valor de y cuando $x = 0$.
- b_1 es la pendiente de la línea, que indica el cambio en y por cada unidad de cambio en x.

El criterio para ajustar la línea es minimizar el error cuadrático total entre los valores observados (y_i) y los valores predichos $(\hat{y}_i)$:

$$E = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2.$$

Para encontrar los valores óptimos de b_0 y b_1, derivamos E con respecto a ambos coeficientes y resolvemos el sistema de ecuaciones resultante. Esto lleva a las siguientes fórmulas:

$$b_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2},$$

$$b_0 = \frac{\sum_{i=1}^{n} y_i}{n} - b_1 \frac{\sum_{i=1}^{n} x_i}{n}$$

Estas fórmulas permiten calcular los coeficientes de regresión a partir de los datos muestrales.

7.2 Uso de la regresión lineal

El modelo de regresión lineal puede emplearse para:

- **Predicción:** Estimar nuevos valores de y a partir de valores de x, siempre y cuando la relación entre ambas variables sea lineal.
- **Interpretación:** Evaluar cómo la variable independiente afecta a la variable dependiente, analizando la pendiente (b_1).

Sin embargo, es importante considerar que el modelo es confiable solo si el coeficiente de correlación (r) es cercano a ± 1, lo que indica una fuerte relación lineal. Si $|r|$ es pequeño, la regresión lineal no será adecuada para describir los datos.

7.3 Consideraciones para el uso de regresión lineal

Al aplicar la regresión lineal simple, se deben tener en cuenta las siguientes consideraciones:

1. **Coeficiente de correlación:** Si el coeficiente de correlación es bajo, el modelo lineal no será adecuado.
2. **Intervalo de predicción:** Las predicciones son válidas solo dentro del rango de valores de x observado en los datos muestrales. Extrapolar más allá de este intervalo puede llevar a resultados inexactos.
3. **Población analizada:** La ecuación de regresión lineal es válida únicamente para la población de la cual se obtuvieron los datos. Si se cambia la población, los coeficientes calculados podrían no ser aplicables.

7.4 Impacto de datos atípicos

Los datos atípicos (outliers) pueden influir significativamente en los coeficientes de regresión, distorsionando el modelo. Por ello:

- Es importante identificar y analizar los puntos distantes antes de ajustar el modelo.
- Si se justifica, se pueden eliminar o tratar estos puntos mediante técnicas de preprocesamiento, asegurándose de que su exclusión esté debidamente argumentada.

El análisis adecuado de los datos y la validación del modelo son pasos esenciales para garantizar que las predicciones obtenidas sean confiables.

7.5 Ejemplo numérico

Para ilustrar el cálculo de los coeficientes de regresión, consideremos el siguiente conjunto de datos muestrales:

x	y
1	2
2	3
3	5
4	7
5	11

Cálculo del coeficiente b_1 (pendiente): Primero calculamos las sumas necesarias:

$$\sum x = 1 + 2 + 3 + 4 + 5 = 15, \quad \sum y = 2 + 3 + 5 + 7 + 11 = 28,$$

$$\sum x^2 = 1^2 + 2^2 + 3^2 + 4^2 + 5^2 = 55, \quad \sum xy = (1)(2) + (2)(3) + (3)(5) + (4)(7) + (5)(11) = 109.$$

Sustituyendo en la fórmula de b_1:

$$b_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2},$$

$$b_1 = \frac{5(109) - (15)(28)}{5(55) - (15)^2} = \frac{545 - 420}{275 - 225} = \frac{125}{50} = 2.5.$$

Cálculo del coeficiente b_0 (intercepto):

$$b_0 = \frac{\sum y - b_1 \sum x}{n},$$

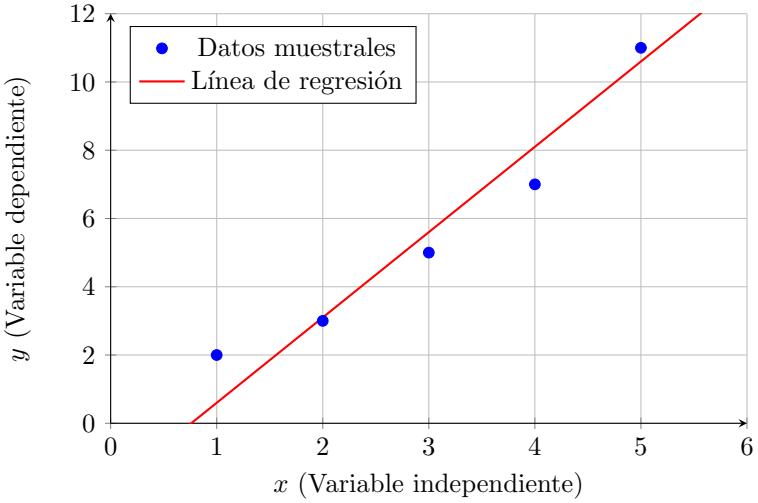
$$b_0 = \frac{28 - (2.5)(15)}{5} = \frac{28 - 37.5}{5} = \frac{-9.5}{5} = -1.9.$$

Por lo tanto, la ecuación de la recta de regresión es:

$$\hat{y} = -1.9 + 2.5x.$$

7.6 Ejemplo gráfico

En la siguiente figura, se muestra la relación entre los datos muestrales y la línea de regresión ajustada:



La línea de regresión se ajusta de tal forma que minimiza el error cuadrático entre los puntos y la línea. Observa cómo la línea sigue la tendencia general de los datos.

7.7 Uso práctico

Supongamos que deseamos predecir el valor de y cuando $x = 6$. Usando la ecuación ajustada:

$$\hat{y} = -1.9 + 2.5(6) = -1.9 + 15 = 13.1.$$

Por lo tanto, el valor estimado de y es 13.1.

7.8 Impacto de datos atípicos: Ejemplo

Si añadimos un dato atípico $(x, y) = (6, 25)$ al conjunto original, el modelo se verá afectado significativamente. Esto puede observarse en el gráfico, donde el dato atípico se aleja de la tendencia general y provoca que la línea de regresión se incline para tratar de minimizar el error total.

El análisis de datos atípicos es crucial para decidir si es válido incluirlos o eliminarlos del modelo.

Bibliografía

[1] Triola F. Mario (2018). *Estadística*. 12a. Edición, Pearson Education, México.

[2] Jay L. Devore (2015). *Probabilidad y Estadística para Ingenierías y Ciencias*. 9a. Edición, Cengage Learning.

[3] Ronald E. Walpole, Raymond H. Myers, Sharon L. Myers & Keying Ye (2012). *Probabilidad y Estadística para Ingenierías y Ciencias*. 9a. Edición, Pearson Education, México.

[4] Thomas Haslwanter (2016). *An Introduction to Statistics with Python*. Springer.

[5] David Spiegelhalter (2021). *The Art of Statistics: How to Learn from Data*. Basic Books.

[6] Austin Warpenter (2025). *Esenciales de Probabilidad y Estadística: Lecciones Introductorias, 30 Ejemplos y Práctica de Examen*. Independently published.

[7] Manuel Gómez Navarrete (2025). *Estadística Y Probabilidad Con Python Para Stem*. Independently published.

[8] Joel Grus (2019). *Data Science from Scratch First Principles with Python*. O'Reilly

[9] Miguel Ghebré Ramírez Elías (2025). *Manual de Prácticas Estadística Aplicada*. Facultad de Ciencias, UASLP, Material Didactico.
`http://www.fc.uaslp.mx/academicos/material-didactico/ManualEA_2025_Final.pdf`