Análisis Teórico de Conceptos Estadísticos (Págs. 1-15) Basado en "Probabilidad y Estadística - 2025-A"

1. Estadística (Páginas 4-5)

Definición: La estadística es la ciencia que proporciona métodos para la recopilación, organización, presentación, análisis e interpretación de datos. Su propósito es describir un fenómeno, inferir comportamientos, predecir resultados y facilitar la toma de decisiones informada.

Antecedentes Históricos: Surgió del latín *status* como una herramienta de los gobernantes para conocer el "estado" de sus dominios (población, riquezas). Hoy es un pilar del método científico para manejar la variabilidad de los fenómenos observables.

Ramas Principales:

1. **Estadística Descriptiva:**
    - Se enfoca en resumir y describir las características fundamentales de un conjunto de datos.
    - Utiliza tablas, gráficos e indicadores.
    - Su método es la **agregación** y el **resumen** de lo observado.
2. **Estadística Inferencial:**
    - Utiliza los resultados de una muestra para hacer predicciones, generalizaciones y obtener conclusiones sobre la población completa.
    - Su método es la **deducción** basada en la probabilidad.

Interpretación de un Resultado:

- **Descriptiva** (ej. un gráfico): *"Así es como se ven y se comportan los datos que hemos recogido".*
- **Inferencial** (ej. una predicción): *"Basándonos en los datos recogidos, es probable que el fenómeno completo se comporte de esta manera".*

---

2. Población, Muestra e Individuo (Página 5)

Estos conceptos definen el alcance de un estudio.

- **Población:** Es el conjunto o colección total de objetos (personas, cosas, mediciones) al que se refiere el estudio. (Ej. Todos los alumnos inscritos en

la universidad).
- **Individuo:** Es cada uno de los elementos que componen la población. (Ej. Cada estudiante).
- **Muestra:** Es un subconjunto de la población que se selecciona para el estudio. (Ej. 3000 estudiantes).

Relación Matemática: $Muestra \subset Población$

Justificación: Estudiar a la población completa es a menudo imposible o demasiado costoso. La validez de la inferencia depende críticamente de cómo se seleccione la muestra.

Interpretación de un Resultado: Un resultado obtenido de la muestra (ej. "el promedio de la muestra fue 8.5") no es la verdad absoluta, sino una **estimación** del resultado de la población (ej. "estimamos que el promedio de toda la universidad es cercano a 8.5").

---

3. Variables Estadísticas (Páginas 6-7)

Una variable es la característica o propiedad específica que se estudia en cada individuo y que puede tomar diferentes valores. (Ej. Si el individuo es un "estudiante", las variables pueden ser "Estatura", "Facultad" o "Promedio").

Clasificación: Es crucial porque define qué análisis matemáticos se pueden aplicar.

1. **Cualitativas (o Categóricas):** No se describen numéricamente (atributos).
    - **Nominales:** Solo nombran, no tienen un orden inherente. (Ej. Sexo, Estado Civil).
    - **Ordinales:** Sugieren una ordenación o jerarquía. (Ej. Nivel de estudios, Grado de satisfacción).
2. **Cuantitativas (o Numéricas):** Se describen con números.
    - **Discretas:** Toman valores enteros; se "cuentan". (Ej. Número de hermanos).
    - **Continuas:** Toman cualquier valor intermedio; se "miden". (Ej. Peso, Altura).

Interpretación de un Resultado: Un resultado de este concepto es un **"dato"**. Es el valor específico que toma la variable para un individuo (ej. la variable "Peso" tomó el valor "70.5 kg"). El análisis estadístico se realiza sobre el conjunto de todos esos datos.

---

4. Tabla de Frecuencias (Páginas 7-9)

Es una herramienta que resume y organiza un conjunto de datos mostrando cuántas veces aparece cada valor distinto de la variable. Permite ver la **distribución** de los datos.

Componentes Principales:

- **Frecuencia Absoluta (f_i):** Conteo directo de ocurrencias.
- **Frecuencia Relativa (f_r):** Proporción ($f_r = f_i/N$). Permite comparar conjuntos de diferente tamaño.
- **Frecuencia Absoluta Acumulada (F_i):** Suma de las f_i hasta ese valor ($F_i = f_1 + ... + f_i$).
- **Frecuencia Relativa Acumulada (F_r):** Suma de las f_r hasta ese valor.

Propiedades: $\sum f_i = N$ (total de datos) y $\sum f_r = 1$ (100%).

Interpretación de un Resultado:

- $f_i = 7$ (para $x_i = 2$ hijos): *"Exactamente 7 familias tienen 2 hijos".*
- $f_r = 0.35$**:** *"El 35% de las familias de la muestra tienen 2 hijos".*
- $F_i = 13$ (para $x_i = 2$ hijos): *"13 familias tienen 2 hijos o menos".*
- $F_r = 0.65$**:** *"El 65% de las familias de la muestra tienen 2 hijos o menos" (este valor es el percentil 65).*

¿Cuándo es "alto" o "bajo"?

- Una f_i o f_r **alta** indica un valor común o frecuente (es o está cerca de la **moda**).
- Una f_i o f_r **baja** indica un valor raro o poco frecuente.
- La frecuencia **acumulada** (F_i o F_r) representa el porcentaje de datos que está por debajo de ese valor.

---

5. Datos Agrupados (Intervalos de Clase) (Págs. 8-10)

Técnica usada para variables continuas o discretas con muchos valores. Los datos se agrupan en rangos (intervalos) para resumirlos. Se pierde precisión pero se gana capacidad de resumen.

Conceptos Clave:

- **Amplitud (A):** Define el ancho de cada intervalo ($A = Rango/k$, donde k es el número de intervalos).
- **Marca de Clase (m_i):** Punto medio del intervalo ($m_i = (L_i + L_{i+1})/2$). Es el valor representativo de todo el intervalo para cálculos (como la media).

Interpretación de un Resultado:

- $f_i = 9$ para el intervalo [7.47 - 8.17): *"9 mediciones cayeron dentro de este rango".*
- $m_i = 7.82$**:** *"Para cualquier cálculo futuro, trataremos esas 9 mediciones como si todas hubieran sido 7.82".*

¿Cuándo es "alto" o "bajo"?

- Una f_i **alta** en un intervalo indica una alta **concentración** de datos en ese rango (un pico en el histograma).
- Una f_i **baja** indica una baja concentración de datos en ese rango.

---

6. Representaciones Gráficas (Páginas 11-15)

Es la visualización de la tabla de frecuencias. Permite entender la forma, localización y dispersión de los datos de manera inmediata.

Tipos de Gráficos:

- **Diagrama de Barras** (Var. Discretas/Cualitativas):
  - Rectángulos de altura proporcional a la frecuencia (f_i).
  - Las barras están **SEPARADAS** para indicar categorías distintas.
- **Histograma** (Var. Continuas/Agrupadas):
  - Las barras están **JUNTAS**, representando que la variable es continua.
  - Es clave para ver la "forma" de la distribución.
- **Ojiva** (Frecuencias Acumuladas):
  - Línea siempre creciente que va de 0 a N (o 0 a 1).
  - Útil para encontrar visualmente la **mediana** ($F_r = 0.5$) y otros percentiles.
- **Diagrama de Sectores (Pastel)** (Var. Cualitativas):
  - Círculo (100%) dividido en sectores proporcionales a f_r.
  - Fórmula: $Ángulo = f_r \times 360°$.

Interpretación de un Resultado:

- **Histograma (Pág. 13):** *"La mayoría de las mediciones se concentran en valores bajos y hay muy pocas con valores altos (distribución sesgada a la derecha)".*
- **Pastel (Pág. 15):** *"Casi la mitad de los estudiantes (46%) están Suspensos".*
- **Ojiva (Pág. 13):** *"El 75% ($F_r = 0.75$) de las mediciones son menores a 8.87".*

¿Cuándo es "alto" o "bajo"?

- En **Barras/Histograma:** Una barra "alta" indica alta frecuencia (**moda**).
- En **Pastel:** Un sector "grande" (ángulo alto) indica una alta proporción de esa categoría.
- En **Ojiva:** La **pendiente** de la línea es lo relevante. Una pendiente "alta" (casi vertical) significa que muchos datos están concentrados en un rango pequeño. Una pendiente "baja" (casi horizontal) significa que los datos están muy dispersos.

---

7. Medidas de Tendencia Central (Página 15)

Son procedimientos (media, mediana, moda) cuyo objetivo es hallar un solo valor que actúe como el "representante" o "centro" de todo el conjunto de datos. Es la forma más extrema de resumir la información.

Elección del Método: Depende del tipo de variable y de la forma de la distribución de los datos.

Interpretación de un Resultado: Un resultado (ej. "la media es 2.25") es el valor que se considera el **centro** de la distribución o el valor más **característico** que representa a todos los datos de la muestra.

¿Cuándo es "alto" o "bajo"? Un valor de tendencia central no significa nada por sí solo. Su valor ("alto" o "bajo") solo adquiere significado cuando se **compara** con algo:

- Con un **estándar de referencia** (ej. una media de 80 es "alta" si la nota máxima es 100, pero "baja" si el máximo es 1000).
- Con **otro grupo** (ej. la media del Grupo A fue 80, "más alta" que la del Grupo B, que fue 70).
- Con su propia **dispersión** (concepto que se ve más adelante).

Análisis Teórico de Conceptos Estadísticos (Parte 2: Págs. 15-35) Basado en "Probabilidad y Estadística - 2025-A"

---

8. Media Aritmética (Págs. 15-16)

Comúnmente conocida como "promedio", es la medida de tendencia central más utilizada. Representa el "centro de gravedad" de los datos; es el valor que tomaría cada individuo si el total se repartiera en partes iguales.

Fórmulas (derivadas del concepto de "reparto equitativo"):

- **Para datos sin agrupar** (n observaciones): Se suma el valor de todas las observaciones (x_i) y se divide por el número total de ellas (n).

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

- **Para datos en tabla de frecuencias** (m categorías): Es un promedio ponderado. Se multiplica cada valor (x_i) por las veces que aparece (f_i), se suma todo, y se divide por el total de datos (n).

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{m} x_i \cdot f_i$$

- **Para datos agrupados en intervalos** (k clases): Igual que el anterior, pero se usa la marca de clase (m_i) como el valor representativo de todo el intervalo.

$$\overline{x} = \sum_{i=1}^{k} m_i \cdot f_{r_i} \quad \text{o} \quad \overline{x} = \frac{1}{n} \sum_{i=1}^{k} m_i \cdot f_i$$

Interpretación de un Resultado: Un resultado como "la media de hijos por familia es 2.25" nos dice que, si bien nadie puede tener 2.25 hijos, este valor representa el punto de equilibrio del conjunto de datos. Si tomáramos 40 familias, esperaríamos que tuvieran un total de $40 \times 2.25 = 90$ hijos.

Importante: La media es muy sensible a valores atípicos (extremadamente altos o bajos), lo que puede hacerla poco representativa en distribuciones sesgadas.

¿Cuándo es "alto" o "bajo"? Como toda medida de tendencia central, su valor ("alto" o "bajo") solo tiene sentido al compararlo con un estándar (ej. una media de 90 es "alta" en un examen sobre 100) o con otro grupo (ej. la media del Grupo A es "más alta" que la del Grupo B).

---

9. Mediana (Me) (Págs. 16-17)

Es el valor que ocupa la posición central de un conjunto de datos, una vez que estos han sido ordenados. Divide la distribución en dos partes iguales.

Procedimiento de Cálculo (Localización):

1. Ordenar los n datos.
2. Localizar la posición central:
   - Si n es **impar**, la mediana es el valor en la posición $\frac{n+1}{2}$.
   - Si n es **par**, la mediana es el promedio de los dos valores centrales (posiciones $\frac{n}{2}$ y $\frac{n}{2}+1$).

Para datos agrupados, la mediana se estima por interpolación dentro del "intervalo mediano" (el primero cuya $F_i \geq n/2$). La fórmula busca qué valor dentro de ese intervalo corresponde exactamente al 50%:

$$Me = L_{i-1} + \left(\frac{\frac{n}{2} - F_{i-1}}{f_i} \right) A$$

Interpretación de un Resultado: Un resultado como "la mediana del número de hijos es 2" nos dice que el 50% de las familias tiene 2 hijos o menos, y el otro 50% tiene 2 hijos o más.

A diferencia de la media, la mediana **no es sensible a valores atípicos**. Es una medida "robusta" de tendencia central, preferida en distribuciones asimétricas (ej. salarios).

---

10. Moda (Mo) (Página 17)

Es la medida de tendencia central más simple: representa el valor que más se repite (el que tiene la mayor frecuencia absoluta, f_i).

- Es la única medida de tendencia central que se puede usar para variables cualitativas nominales (ej. la moda de "color de ojos" es "café").
- Una distribución puede no tener moda (amodal), tener una (unimodal), dos (bimodal) o más.

Interpretación de un Resultado: Un resultado como "la moda del número de hijos es 2" nos dice que el valor más común o frecuente en la muestra es "2 hijos".

En un histograma, la moda corresponde al punto medio del intervalo con la barra más alta (el "pico" de la distribución).

---

11. Cuantiles (Percentiles y Cuartiles) (Págs. 17-18)

Los cuantiles son medidas de **posición no central**. Son valores que dividen la distribución ordenada de datos en partes iguales.

- **Percentiles (P_k):** Dividen la distribución en 100 partes iguales. P_k es el valor que deja al $k\%$ de los datos por debajo de él.
- **Cuartiles (Q_k):** Dividen la distribución en 4 partes iguales (cada una con el 25% de los datos).
  - Q_1 (Cuartil 1) = P_{25}. El 25% de los datos es menor que Q_1.
  - Q_2 (Cuartil 2) = P_{50}. Coincide exactamente con la **Mediana**.
  - Q_3 (Cuartil 3) = P_{75}. El 75% de los datos es menor que Q_3.

Las fórmulas para encontrarlos son similares a la mediana, pero en lugar de buscar la posición $n/2$, se busca la posición $nk/100$ (para el percentil k) o $nk/4$ (para el cuartil k).

Interpretación de un Resultado:

- $Q_1 = 1$ nos dice: "El 25% de las familias tiene 1 hijo o menos".
- $Q_3 = 3$ nos dice: "El 75% de las familias tiene 3 hijos o menos".
- **Combinados ($Q_1 = 1$ y $Q_3 = 3$)**, nos dicen que el 50% central de las familias (la "caja") tiene entre 1 y 3 hijos.
- Un $P_{90} = 50000$ en salarios nos dice: "El 90% de los empleados gana 50000 o menos" (o, equivalentemente, "solo el 10% gana más de 50000").

---

12. Diagrama de Cajas (Boxplot) (Página 18)

Es una representación gráfica que resume visualmente la distribución de los datos usando **cinco números**: el valor mínimo, Q_1, la mediana (Q_2), Q_3, y el valor máximo.

Permite identificar rápidamente la simetría y la dispersión. La "caja" central representa al 50% de los datos (Q_1 a Q_3). Los "bigotes" se extienden a los valores mínimos y máximos que no se consideran atípicos.

- **Rango Intercuartil (RIQ):** $RIQ = Q_3 - Q_1$. Es la "longitud" de la caja.

- **Límites para datos atípicos:** Se definen $L_i = Q_1 - 1.5 \cdot RIQ$ y $L_s = Q_3 + 1.5 \cdot RIQ$. Cualquier dato fuera de estos límites se considera atípico y se grafica como un punto individual.

Interpretación de un Resultado:

- Una caja **corta** (RIQ bajo) indica que el 50% central de los datos está muy **concentrado**.
- Una caja **larga** (RIQ alto) indica que el 50% central está muy **disperso**.
- Una mediana (Q_2) centrada en la caja y bigotes de igual longitud sugieren una distribución **simétrica**.
- Una mediana no centrada o bigotes desiguales sugieren **asimetría**. (Ej. un bigote derecho largo indica asimetría positiva o sesgo a la derecha).
- Puntos fuera de los bigotes indican **valores atípicos**, que merecen investigación.

---

13. Medidas de Dispersión (Págs. 25-26)

Miden el grado de variabilidad o "esparcimiento" de los datos. Nos dicen si los datos están muy juntos (**homogéneos**) o muy separados (**heterogéneos**) alrededor del centro.

Un promedio (*media*) solo tiene sentido si la dispersión es baja. (Ej. "El promedio del examen fue 50" es poco informativo si las notas fueron 0 y 100, pero muy informativo si fueron 49 y 51).

---

14. Varianza (s^2) y Desviación Estándar (s) (Pág. 25)

Son las medidas de dispersión más importantes. Miden la distancia "promedio" de cada dato con respecto a la media ($\overline{x}$).

- **Varianza (s^2):** Es el promedio de las distancias al cuadrado de cada dato a la media.

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

(Se usa $n-1$ en lugar de n por razones de inferencia estadística, se llama "grados de libertad"). La fórmula es así porque $(x_i - \overline{x})$ es la "desviación" de un dato. Se eleva al cuadrado para que las desviaciones positivas y negativas no se cancelen entre sí.

- **Desviación Estándar (s):** Es simplemente la raíz cuadrada de la varianza.

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

  Se crea la desviación estándar porque las unidades de la varianza están al cuadrado (ej. $hijos^2$ o $dólares^2$), lo cual es difícil de interpretar. Al sacar la raíz cuadrada, s vuelve a tener las mismas unidades que los datos originales (ej. hijos o dólares).

Interpretación de un Resultado:

- $s = 0$ nos dice: No hay dispersión. Todos los datos son idénticos.
- $s = 1.16$ **hijos** (Ej. 9) nos dice: "En promedio, el número de hijos de cualquier familia se desvía de la media (2.25) en aproximadamente 1.16 hijos".

¿Cuándo es "alto" o "bajo"?

- Una desviación estándar **"baja"** (cercana a 0) indica que los datos están muy agrupados alrededor de la media. El promedio es un representante muy fiable.
- Una desviación estándar **"alta"** indica que los datos están muy dispersos. El promedio es un representante menos fiable de la muestra. (El "qué tan alto es alto" es relativo y se responde mejor con el Coeficiente de Variación).

---

15. Rango y Rango Intercuartil (RIQ) (Págs. 25-26)

Son medidas de dispersión más simples:

- **Rango:** Es la diferencia entre el valor máximo y el mínimo ($Rango = x_{max} - x_{min}$).
  - Es muy fácil de calcular, pero muy sensible a valores atípicos.
- **Rango Intercuartil (RIQ):** Es la diferencia entre el cuartil 3 y el cuartil 1 ($RIQ = Q_3 - Q_1$).
  - Representa la dispersión del **50% central** de los datos.
  - Es una medida de dispersión **robusta**, ya que no le afectan los valores atípicos.

Interpretación de un Resultado:

- $Rango = 4$ **hijos** (Ej. 9) nos dice: "La diferencia entre la familia con más hijos (5) y la que tiene menos (1) es de 4 hijos".
- $RIQ = 2$ **hijos** (Ej. 7, $Q_3 = 3, Q_1 = 1$) nos dice: "El 50% central de las familias tiene una dispersión de 2 hijos (están entre 1 y 3)".

---

16. Coeficiente de Variación (CV) (Página 26)

Es una medida de **dispersión relativa**. No tiene unidades (es adimensional). Su propósito es comparar la dispersión de dos conjuntos de datos diferentes, incluso si tienen medias o unidades distintas (ej. comparar la variabilidad del peso en kg con la de la altura en cm).

Se calcula dividiendo la desviación estándar por la media.

$$CV = \frac{s}{|\overline{x}|}$$

(Usualmente se multiplica por 100 para expresarlo como porcentaje).

Interpretación de un Resultado: Un $CV = 51.5\%$ ($s = 1.16, \overline{x} = 2.25$) nos dice que la desviación estándar es el 51.5% de la media.

¿Cuándo es "alto" o "bajo"? Aquí sí hay una regla general:

- Un CV **"bajo"** (ej. $CV < 20\%$ o 30%) indica un conjunto de datos **homogéneo**. La media es muy representativa.
- Un CV **"alto"** (ej. $CV > 30\%$) indica un conjunto de datos **heterogéneo**. La media es menos representativa.

Si el CV del Grupo A es 15% y el del Grupo B es 40%, podemos concluir que los datos del Grupo B son relativamente más variables que los del Grupo A.

---

17. Medidas de Forma (Págs. 28-29)

Describen la "forma" del histograma de la distribución. Las dos medidas principales son la **asimetría (sesgo)** y la **curtosis (apuntamiento)**.

---

18. Coeficiente de Asimetría (A_s) (Págs. 28-29)

Mide el grado de simetría (o falta de ella) de la distribución alrededor de su media.

$$A_s = \frac{\frac{1}{n}\Sigma(x_i - \overline{x})^3}{s^3}$$

La fórmula eleva las desviaciones al cubo. Esto es clave, porque mantiene el signo:

- Datos muy a la derecha ($\overline{x}$ positivo grande) $\rightarrow (+)^3 \rightarrow A_s$ **positivo**.
- Datos muy a la izquierda ($\overline{x}$ negativo grande) $\rightarrow (-)^3 \rightarrow A_s$ **negativo**.

Interpretación de un Resultado:

- $A_s \approx 0$**:** Simétrica. La media, mediana y moda son similares. Las colas a ambos lados son iguales.
- $A_s > 0$**:** Asimétrica positiva (sesgo a la derecha). Hay valores atípicos altos que "estiran" la cola derecha. En este caso: $Moda < Mediana < Media$. (El salario es un ejemplo clásico).
- $A_s < 0$**:** Asimétrica negativa (sesgo a la izquierda). Hay valores atípicos bajos que "estiran" la cola izquierda. En este caso: $Media < Mediana < Moda$. (Notas en un examen muy fácil).

¿Cuándo es "alto" o "bajo"?

- **"Bajo"** (cercano a 0) $\rightarrow$ simétrica.
- **"Alto"** (positivo o negativo) $\rightarrow$ muy sesgada.

---

19. Coeficiente de Apuntamiento o Curtosis (A_c, A_p) (Pág. 29)

Mide el grado de concentración de los datos alrededor de la media (qué tan "puntiaguda" es la distribución) y qué tan "pesadas" son las colas. Se compara contra la distribución Normal (la campana de Gauss), que tiene $A_c = 3$.

$$A_c = \frac{\frac{1}{n}\sum(x_i - \overline{x})^4}{s^4}$$

La fórmula eleva a la cuarta potencia. Esto hace que las desviaciones (tanto positivas como negativas) tengan un peso enorme, enfatizando los valores en las colas.

Para facilitar la interpretación, se usa el **exceso de curtosis**: $A_p = A_c - 3$.

Interpretación de un Resultado:

- $A_p \approx 0$ $(A_c \approx 3)$**: Mesocúrtica.** Es tan puntiaguda como la distribución Normal. (Ej. Sueldos, Pág. 35).
- $A_p > 0$ $(A_c > 3)$**: Leptocúrtica.** Es más puntiaguda que la Normal. Hay una gran concentración de datos en la media y colas "pesadas" (más valores atípicos de lo esperado).

- $A_p < 0$ $(A_c < 3)$**: Platicúrtica.** Es más plana que la Normal. Los datos están más repartidos, hay poca concentración en la media y colas "ligeras". (Ej. Núm. de hijos, Pág. 31).

¿Cuándo es "alto" o "bajo"?

- **"Alto" (Leptocúrtica)** $\rightarrow$ muchos datos en el centro y en las colas (muchos atípicos).
- **"Bajo" (Platicúrtica)** $\rightarrow$ pocos datos en el centro y pocos en las colas (pocos atípicos).