Tabla de contenido

1. Objetivo de la unidad 1

Apropiar conceptos básicos de la Estadística Descriptiva, para realizar análisis estadístico más sofisticados que facilite la toma de decisiones basadas en datos.

2. Tema 2: Estadísticas Descriptivas.

2.1 Distribuciones de Frecuencias

Una distribución de frecuencias es una tabla o gráfico que muestra con qué frecuencia ocurren valores diferentes en un conjunto de datos. Básicamente, resume la distribución de datos enumerando valores únicos y la cantidad de veces que aparece cada valor en el conjunto de datos.

Edad	Frecuencia
14	3
15	6
16	5
17	1
18	1

Distribuciones de Frecuencias con datos sin agrupar

Una distribución de frecuencias con datos sin agrupar es una forma de presentar la frecuencia con la que ocurre cada valor individual en un conjunto de datos. Es decir, se enumeran todos los valores únicos presentes en el conjunto de datos junto con el número de veces que cada uno ocurre.

Ejemplo:

Se desea analizar las edades de los miembros del club para comprender mejor la composición del grupo. Estos son los datos de las edades de los 30 miembros del club: 25, 28, 27, 30, 25, 30, 30, 32, 35, 30, 30, 30, 27, 29, 29, 32, 32 ,35, 36, 30, 35, 32, 27, 35, 30, 30, 25, 37, 30, 32. La distribución de Frecuencias de puede observar a continuación.

Edad	Frecuencia
25	3
27	3
28	1
29	2
30	10
32	5
35	4
36	1
37	1

Distribuciones de Frecuencias con datos agrupados

Una distribución de frecuencias con datos agrupados es una forma de organizar y resumir conjuntos extensos o continuos de datos. En lugar de listar cada valor único, los datos se agrupan en intervalos y se cuenta cuántas observaciones caen en cada intervalo.

Ejemplo:

Se está llevando un registro de las alturas de un grupo de estudiantes en una clase. Se tienen los siguientes datos de altura en centímetros: 152, 165, 158, 170, 155, 162, 168, 175, 160, 155, 165, 170, 168, 155, 162, 170, 165, 172, 158, 163, 160, 165, 168, 172, 167, 158, 166, 173, 159, 164.

Para determinar el número de intervalos, podemos usar la regla de Scott, que sugiere tomar aproximadamente la raíz cuadrada de n intervalos, donde n es el número total de datos. La distribución de Frecuencias de puede observar a continuación.

Intervalo de Altura	Frecuencia
150-153	1
154-157	3
158-161	5
162-165	7
166-169	8
170-173	4
174-177	2

2.2 Medidas de tendencia central

Las medidas de tendencia central son valores estadísticos que representan un punto central o típico en un conjunto de datos. Estas medidas son fundamentales en la estadística descriptiva debido a que proporcionan un resumen de los datos, ilustrando el valor alrededor del cual los datos tienden a agruparse. Las tres medidas de tendencia central más comunes son:

Media

La media, también conocida como promedio, es la suma de todos los valores en un conjunto de datos dividida por el número total de valores. Es una medida de tendencia central que se ve afectada por los valores extremos en los datos.

La fórmula matemática para calcular la media aritmética $\overline{x}$ de un conjunto de *n* datos $x_1, x_2, x_3, \ldots, x_n$ es:

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

Ejemplo:

Se desea calcular la media de las calificaciones obtenidas por un grupo de 10 estudiantes en un examen de matemáticas. Las calificaciones obtenidas por los 10 estudiantes son las siguientes: 85, 92, 78, 90, 88, 76, 85, 94, 89, 80.

Para calcular la media, sumamos los datos dados:

85 + 92 + 78 + 90 + 88 + 76 + 85 + 94 + 89 + 80 = 857.

Ahora, el resultado 857, debe ser dividido entre el número total de datos 10. Esto es,

857/10 = 85.7.

Por tanto, la media de las calificaciones del grupo de estudiantes en el examen de matemáticas es 85.7.

Mediana

La mediana es el valor central de un conjunto de datos ordenados. Es menos afectada por valores extremos que la media. En conjuntos con un número impar de datos, la mediana es el valor del centro; en conjuntos pares, es el promedio de los dos valores centrales.

Ejemplo:

Se desea calcular la mediana de los salarios mensuales, en dólares, de un grupo de 11 empleados en una empresa. Los salarios son: 2100, 2000, 2200, 1800, 2500, 1900, 2300, 2600, 1500, 1600, 2800.

Para calcular la mediana:

1. Ordenamos los datos de menor a mayor: 1500, 1600, 1800, 1900, 2000, 2100, 2200, 2300, 2500, 2600, 2800.

2. Debido a que el número de datos es impar, la mediana es el dato que se encuentra en el centro de la lista:

1500, 1600, 1800, 1900, 2000, **2100**, 2200, 2300, 2500, 2600, 2800.

Por lo tanto, la mediana de los salarios es 2100 dólares.

Moda

La moda es el valor que aparece con mayor frecuencia en un conjunto de datos. Puede haber una moda (unimodal), varias modas (multimodal) o ninguna moda. Es útil para identificar el valor más común en los datos.

Ejemplo:

Se realizó una encuesta a un grupo de 15 personas para conocer su color favorito. Se desea calcular la moda para determinar el color más popular entre los encuestados. Las respuestas de las 15 personas en la encuesta fueron las siguientes: Rojo, Azul, Verde, Azul, Rojo, Verde, Rojo, Amarillo, Azul, Verde, Rojo, Azul, Amarillo, Verde, Azul.

Para determinar la moda, se cuenta la frecuencia de cada respuesta y se identifica cuál tiene la mayor frecuencia. Las frecuencias son: Rojo: 4, **Azul: 5**, Verde: 3, Amarillo: 2.

El color que aparece con mayor frecuencia es el Azul, con 5 votos. Por lo tanto, la moda de las preferencias de color en la encuesta es el Azul.

2.3 Medidas de Dispersión

Una medida de dispersión, también conocida como medida de variación, es una estadística que indica el grado de dispersión o variabilidad dentro de un conjunto de datos. Estas medidas proporcionan información sobre la homogeneidad o heterogeneidad de los datos con respecto a la media. Las medidas de dispersión más comunes son:

Mínimo y Máximo

El Mínimo (Máximo) es el valor más pequeño (grande) en un conjunto de datos. Representa el límite inferior (superior) de los datos en términos de magnitud.

Ejemplo:

Supongamos que se tiene un conjunto de datos que representa las edades de un grupo de personas en una reunión:

18, 22, 30, 25, 27, 35, 40, 29, 21, 24 años.

Mínimo: El mínimo es el valor más pequeño en el conjunto de datos. En este caso, el mínimo es 18, ya que es la edad más baja entre las personas en la reunión.

Máximo: El máximo es el valor más grande en el conjunto de datos. En este caso, el máximo es 40, ya que es la edad más alta entre las personas en la reunión.

Cuantiles

Los cuantiles son valores que dividen un conjunto de datos ordenados en partes iguales. Específicamente, el p-ésimo cuantil, denotado como Q_p, es el valor por debajo del cual cae un porcentaje p de los datos. Los cuantiles son particularmente útiles para comprender la distribución y dispersión de un conjunto de datos porque proporcionan un punto de referencia sobre cómo se distribuyen los datos en diferentes segmentos.

Ejemplo:

Conjunto de datos: 10, 15, 20, 25, 30, 35, 40, 45, 50, 55.

- Mediana (segundo cuartil, Q_2):

1. Separa el conjunto de datos en dos partes iguales.
2. Q_2 es el valor que está en la posición central. En este caso, Q_2 =32.5.

- Primer cuartil (Q_1 o percentil 25):

1. El 25% de los datos son menores que Q_1.
2. Q_1 es el valor que está en la posición (1/4) × n. En este caso Q_1 =15.

- Tercer cuartil (Q_3 o percentil 75):

1. El 75% de los datos son menores que Q_3.
2. Q_3 es el valor que está en la posición (3/4)×n. Q_3=45.

- Percentil 10 (Q_{10}):

1. El 10% de los datos son menores que Q_{10}.
2. Q_{10} es el valor que está en la posición (10/100)×n. Q_{10}=15.

- Percentil 90 (Q_{90}):

1. El 90% de los datos son menores que Q_{90}.
2. Q_{90} es el valor que está en la posición (90/100)×n. Q_{90}=50.

Varianza

La varianza es una medida de dispersión que indica qué tan dispersos están los valores de un conjunto de datos en relación con su media. Es el promedio de las diferencias al cuadrado entre cada dato y la media del conjunto.

La fórmula de la varianza σ^2 para un conjunto de n datos $x_1, x_2, x_3, \ldots, x_n$ con media $\overline{x}$ es:

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

Donde

- x_i es cada dato individual.
- $\overline{x}$ es la media del conjunto de datos.
- n es el número total de datos.

Ejemplo:

Supongamos que tenemos las siguientes calificaciones de un examen de 5 estudiantes: **85, 90, 92, 88, 78** y deseamos calcular la varianza.

Paso 1: Calcular la media $\overline{x}$ de las calificaciones

$$\overline{x}=\frac{\sum_{i=1}^{5} x_i}{5}=\frac{85+90+92+88+78}{5}=\frac{433}{5}=86.6$$

Paso 2: Calcular la varianza con la fórmula

$$\sigma^2=\frac{\sum_{i=1}^{5}\left(x_i-\overline{x}\right)^2}{5}$$

Entonces,

$$\sigma^2=\frac{(85-86.6)^2+(90-86.6)^2+(92-86.6)^2+(88-86.6)^2+(78-86.6)^2}{5}$$

$$\sigma^2=\frac{2.56+11.56+31.36+1.96+72.96}{5}=\frac{120.4}{5}=24.08$$

Por lo tanto, la varianza es $\sigma^2=24.08$.

En este ejemplo, una varianza de 24.08 significa lo siguiente:

1. Los valores de las calificaciones están dispersos alrededor de la media de 86.6.

2. Un valor alto de varianza indica una mayor dispersión de los datos alrededor de la media.

3. En este caso, una varianza de 24.08 indica que las calificaciones de los estudiantes varían significativamente en comparación con la media de calificación, que es de 86.6.

Desviación estándar

La desviación estándar es una medida de dispersión o variabilidad que indica hasta qué punto los valores individuales en un conjunto de datos, en promedio, se desvían de la media de ese conjunto de datos. Esta es una medida importante de la variabilidad o heterogeneidad de un conjunto de datos.

Ejemplo:

En el ejemplo anterior asociado a la Varianza, se utilizó el conjunto de datos **85, 90, 92, 88, 78**. La Media fue calculada en 86.6 y la Varianza en 24.08.

Para obtener la Desviación Estándar, calculamos la raíz cuadrada de la varianza, la cual es aproximadamente 4.91. Lo que significa que, en promedio, las calificaciones de los estudiantes varían aproximadamente **±4.91** puntos alrededor de la media de **86.6.**

Coeficiente de variación

Es una medida de dispersión relativa que se utiliza para comparar la variabilidad de dos o más conjuntos de datos que tienen diferentes unidades o escalas. Se expresa como un porcentaje y se calcula como la desviación estándar dividida por la media, multiplicada por 100.

Matemáticamente, el coeficiente de variación *CV* se calcula con la siguiente formula:

$$CV = \left(\frac{\sigma}{\overline{x}}\right) \times 100$$

Donde:

- σ es la desviación estándar del conjunto de datos.
    - $\overline{x}$ es la media del conjunto de datos.

Ejemplo:

Se tienen datos de las alturas (en centímetros) y los pesos (en kilogramos) de dos grupos de personas:

Alturas: 160, 165, 170, 175, 180.

Pesos: 50, 55, 60, 65, 70.

La media de las alturas es 170 cm y de los pesos es 60 kg.

La desviación estándar para las alturas y los pesos se calcula en 7.07 centímetros y kilogramos, respectivamente.

El Coeficiente de Variación para cada grupo es:

$$CV_1 = \left(\frac{\sigma_1}{\overline{x}_1}\right) \times 100 = \left(\frac{7.07}{170}\right) \times 100 \approx 4.16\,\%$$

$$CV_2 = \left(\frac{\sigma_2}{\overline{x}_2}\right) \times 100 = \left(\frac{7.07}{60}\right) \times 100 \approx 11.78\,\%$$

El coeficiente de variación para las alturas es aproximadamente 4.16%.
El coeficiente de variación para los pesos es aproximadamente 11.78%.

El coeficiente de variación nos indica que las alturas tienen una dispersión relativa menor (4.16%) en comparación con las alturas medias del grupo (170 cm).

Los pesos tienen una dispersión relativa mayor (11.78%) en comparación con los pesos medias del grupo (60 kg).

Este ejemplo muestra cómo el coeficiente de variación puede ser útil para comparar la variabilidad relativa entre dos conjuntos de datos con diferentes unidades o escalas.

2.4 Autoevaluación

Pregunta 1:

Una Distribución de Frecuencia en Estadística es un conjunto de datos organizado que muestra la cantidad de veces que ocurre cada valor.

Opción 1: Verdadero

Opción 2: Falso

Pregunta 2:

¿Cuál de las siguientes opciones es una medida de tendencia central en estadística?

Opción 1: Rango

Opción 2: Mediana

Opción 3: Varianza

Opción 2: Desviación estándar

Pregunta 3:

¿Cuál de las siguientes opciones es una medida de dispersión en estadística?

Opción 1: Moda

Opción 2: Mediana

Opción 3: Desviación estándar

Opción 4: Media

3. Bibliografía

- ✓ Romero Ramos, E. (2016). Estadística para todos: análisis de datos: estadística descriptiva, teoría de la probabilidad e inferencia. Difusora Larousse - Ediciones Pirámide, (pp. 22-30). https://elibro-net.bibliotecavirtual.unad.edu.co/es/ereader/unad/49136?page=1
- ✓ Romero Ramos, E. (2016). Estadística para todos: análisis de datos: estadística descriptiva, teoría de la probabilidad e inferencia. Difusora Larousse - Ediciones Pirámide, (pp. 31-39).

  https://elibro-net.bibliotecavirtual.unad.edu.co/es/ereader/unad/49136?page=1
- ✓ Romero Ramos, E. (2016). Estadística para todos: análisis de datos: estadística descriptiva, teoría de la probabilidad e inferencia. Difusora Larousse - Ediciones Pirámide, (pp. 41-55). https://elibro-net.bibliotecavirtual.unad.edu.co/es/ereader/unad/49136?page=1
- ✓ Romero Ramos, E. (2016). Estadística para todos: análisis de datos: estadística descriptiva, teoría de la probabilidad e inferencia. Difusora Larousse - Ediciones Pirámide, (pp. 57-69). https://elibro-net.bibliotecavirtual.unad.edu.co/es/ereader/unad/49136?page=1
- ✓ Romero Ramos, E. (2016). Estadística para todos: análisis de datos: estadística descriptiva, teoría de la probabilidad e inferencia. Difusora Larousse - Ediciones Pirámide, (pp. 93-

100). https://elibro-net.bibliotecavirtual.unad.edu.co/es/ereader/unad/49136?page=1

- ✓ Cely, L. A. (2020). Tipos de muestreo y error muestral. [Objeto_virtual_de_Informacion_OVI]. Repositorio Institucional UNAD. https://repository.unad.edu.co/handle/10596/38474
- ✓ Ortegon, M. F. & Cabrera, F. (2018). Tablas de Frecuencia. [Objeto_virtual_de_Informacion_OVI]. Repositorio Institucional UNAD. https://repository.unad.edu.co/handle/10596/20458
- ✓ Ortegon, M. F. (2019). Medidas de Tendencia Central. [Objeto_virtual_de_aprendizaje_OVA]. Repositorio Institucional UNAD. https://repository.unad.edu.co/handle/10596/33842
- ✓ Camargo, I. (2022). *Medidas estadísticas univariantes: Muestreo*. [Objeto_virtual_de_Informacion_OVI]. Repositorio Institucional UNAD. https://repository.unad.edu.co/handle/10596/59205
- ✓ Camargo, I. (2022). Medidas univariantes para variables cuantitativas. [Objeto_virtual_de_aprendizaje_OVA]. Repositorio Institucional UNAD. https://repository.unad.edu.co/handle/10596/50133
- ✓ Camargo, I. (2023). Muestreo y caracterización de variables cualitativas. [Objeto_virtual_de_aprendizaje_OVA]. Repositorio Institucional UNAD. https://repository.unad.edu.co/handle/10596/59206