Probabilidad y estadística

MAT 041, Primer semestre

Francisco Cuevas Pacheco

14 de noviembre de 2022

Contenidos

- ✓ Medidas de Forma
- ✓ Datos Tabulados
- ✓ Muestras Estratificadas
- ✓ Otros indicadores

Para definir ciertas medidas de forma, necesitamos la siguiente definición.

Definición (Momento Muestral)

El momento muestral central de orden r, para $r \in \mathbb{N}$, se define como

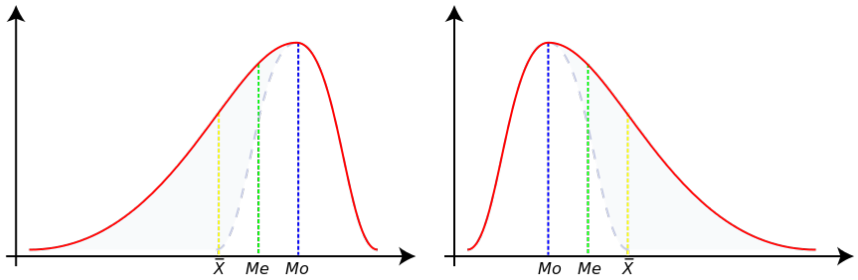
$$m_r = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^r.$$

Definición (Coeficiente de Simetría)

El **coeficiente de simetría de Fisher** *está dado por*

$$\gamma_1 = \frac{m_3}{S_n^3}.$$

- $\gamma_1 = 0 \Longrightarrow$ simetría con respecto a la media
- $\gamma_1 < 0 \Longrightarrow$ asimetría negativa
- $\gamma_1 > 0 \Longrightarrow$ asimetría positiva



Asimetría negativa | Asimetría positiva

Observación

Este coeficiente no tiene dimensión, es invariante bajo traslaciones del origen y transformaciones de escala.

Definición (Curtosis)

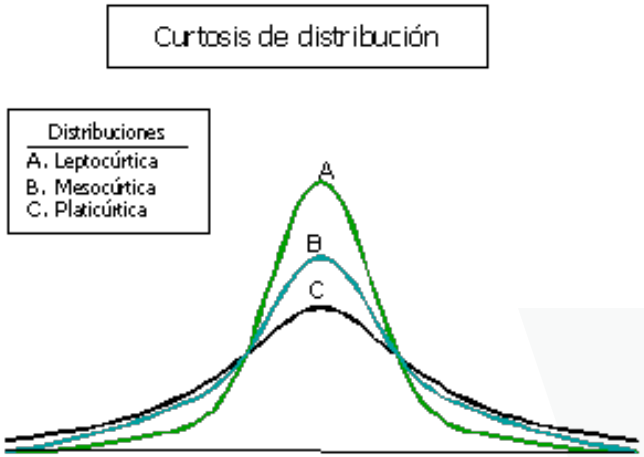
El **coeficiente de achatamiento (curtosis)** *se define como*

$$\gamma_2 = \frac{m_4}{S_n^4} - 3.$$

Observación

Este indicador ha sido diseñado utilizando como referencia la distribución normal (campana de Gauss).

- $\gamma_2 = 0 \Longrightarrow$ similar a una distribución normal (distribución mesocúrtica)
- $\gamma_2 > 0 \Longrightarrow$ más puntiaguda que una distribución normal (distribución leptocúrtica)
- $\gamma_2 < 0 \Longrightarrow$ menos puntiaguda que una distribución normal (distribución platicúrtica)



Tablas de Frecuencia

- Una tabla de frecuencias permite organizar la información contenida en la muestra de manera compacta.
- ¿Cómo se construye? Se divide el rango de la muestra en diferentes clases y se reportan las frecuencias clásicas y las frecuencias acumuladas.
- Este procedimiento es similar al utilizado en la construcción del histograma.

Una tabla de frecuencias tiene la forma

Clase	n_i	f_i	N_i	F_i	$\mathcal{M}_i$
I_1	n_1	f_1	N_1	F_1	$\mathcal{M}_1$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
I_k	n_k	f_k	N_k	F_k	$\mathcal{M}_k$

donde

- $N_j = \sum_{i=1}^{j} n_i$ son las frecuencias absolutas acumuladas;
- $F_j = \sum_{i=1}^{j} f_i$ son las frecuencias relativas acumuladas;
- $\mathcal{M}_i$ denota la marca de clase (un representante de la clase I_i).

Existen algunas reglas para estimar la cantidad de clases, por ejemplo:

- Regla de Sturges:
$$k = 1 + 3.3 \log_{10} n.$$
- Regla de la raíz:
$$k = \sqrt{n}.$$

Para construir las clases necesitamos saber la amplitud (longitud) de cada clase denotada por a. Una manera de obtener la amplitud consiste en dividir el rango de la muestra por la cantidad de clases.

Ejercicio

La tragedia que sufrió el transbordador espacial Challenger en 1986 condujo a varios estudios para investigar las razones de la falla. Las siguientes mediciones (ver el texto guía `Jay L. Devore`) corresponden a las temperaturas (°F) del sello anular en cada encendida de prueba o lanzamiento del motor del cohete:

84	49	61	40	83	67	45	66	70	69	80	58
68	60	67	72	73	70	57	63	70	78	52	67
53	67	75	61	70	81	76	79	75	76	58	31

Construír una tabla de frecuencias especificando el número de clases y la amplitud.

Solución: Tenemos $n = 36$ datos.

- Determinamos el **número de clases** utilizando la regla de la raíz:
$$k = \sqrt{n} = \sqrt{36} = 6.$$
- $X_{(1)} = 31$ y $X_{(36)} = 84 \Longrightarrow$ rango $= 53$.
- La **amplitud** está dada por
$$a = \frac{\text{rango}}{k} = \frac{53}{6} = 8.83.$$

Solución (continuación): Obtenemos la siguiente tabla:

Clase	n_i	f_i	N_i	F_i	$\mathcal{M}_i$
$[31, 39.83)$	1	0.027	1	0.027	35.41
$[39.83, 48.66)$	2	0.055	3	0.083	44.25
$[48.66, 57.5)$	4	0.111	7	0.194	53.08
$[57.5, 66.33)$	7	0.194	14	0.388	61.91
$[66.33, 75.16)$	14	0.388	28	0.777	70.75
$[75.16, 84]$	8	0.222	36	1	79.58

Observación

Con respecto a las investigaciones sobre el transbordador espacial, se concluyó que las temperaturas bajas están asociadas a fallas en el funcionamiento de los sellos anulares.

Indicadores para Datos Tabulados

- El **promedio**

$$\overline{X} = \sum_{i=1}^{k} f_i \mathcal{M}_i$$

- La **varianza**

$$S^2 = \sum_{i=1}^{k} f_i (\mathcal{M}_i - \overline{X})^2$$

- La **desviación estándar**

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

- La **mediana interpolada**

$$\text{Me} = L + \frac{a(\frac{n}{2} - N_{\text{Me}-1})}{n_{\text{Me}}},$$

donde

  - la **clase mediana** es la primera clase donde la frecuencia acumulada es al menos el $50\,\%$.
  - L es el límite inferior de la clase mediana.
  - $N_{\text{Me}-1}$ es la frecuencia absoluta acumulada hasta la clase anterior a la clase mediana.
  - n_{Me} es la frecuencia absoluta de la clase mediana.

- La **moda interpolada**

$$\text{Mo} = L + \frac{a\Delta_1}{\Delta_1 + \Delta_2},$$

donde

  - la **clase modal** es aquella clase que tiene la mayor frecuencia absoluta.
  - L es el límite inferior de la clase modal.
  - $\Delta_1 = n_{\text{Mo}} - n_{\text{Mo}-1}$, donde n_{Mo} es la frecuencia absoluta de la clase modal y $n_{\text{Mo}-1}$ es la frecuencia absoluta de la clase anterior a la clase modal.
  - $\Delta_2 = n_{\text{Mo}} - n_{\text{Mo}+1}$, donde $n_{\text{Mo}+1}$ es la frecuencia absoluta de la clase posterior a la clase modal.

Indicadores para Datos Tabulados

- El **coeficiente de variación**

$$\mathrm{CV} = \frac{S}{\overline{X}}$$

- El cálculo de las **medidas de forma** para datos agrupados es similar al caso de datos no agrupados teniendo en cuenta que en el cálculo de los momentos se usa la definición siguiente:

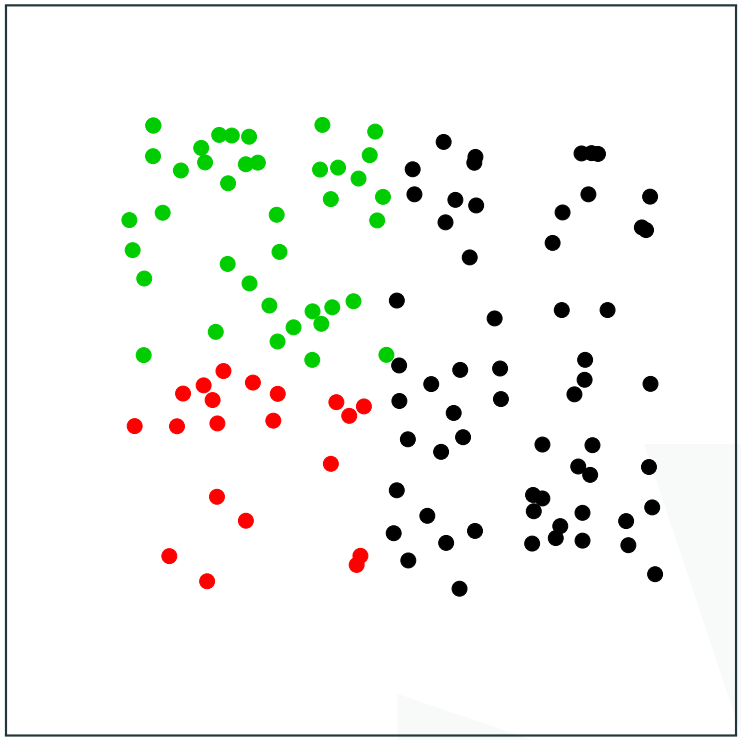
$$m_r = \sum_{i=1}^{k} f_i(\mathcal{M}_i - \overline{X})^r.$$

Ejercicio

Para la tabla de frecuencias obtenida en el ejercicio del transbordador espacial Challenger, verifique que:

1. *Promedio:* $\overline{X} = 66.4$
2. *Varianza:* $S^2 = 122.4$
3. *Mediana interpolada:* $Me = 68.8$
4. *Moda interpolada:* $Mo = 71.08$

Cuando tomamos una muestra desde una población estratificada, nos interesa saber qué relación existe entre las medidas de tendencia central y dispersión de **cada estrato** y las medidas de tendencia central y dispersión de la **muestra completa**.



Notación:

- Existen m estratos y se ha extraído una muestra total de tamaño n.
- En cada estrato se ha extraído una muestra de tamaño n_i. Luego,

$$\sum_{i=1}^{m} n_i = n.$$

- El peso (o ponderación) del estrato i-ésimo está dado por

$$w_i = \frac{n_i}{n}.$$

Sea $\overline{X}_i$ el promedio del estrato i-ésimo. El promedio total es

$$\overline{X}_{\text{total}} = \sum_{i=1}^{m} w_i \overline{X}_i.$$

Sea S_i^2 la varianza del estrato i-ésimo. Entonces, la varianza total es

$$S^2_{\text{total}} = \underbrace{\sum_{i=1}^{m} w_i S_i^2}_{\text{Varianza Intra}} + \underbrace{\sum_{i=1}^{m} w_i (\overline{X}_i - \overline{X}_{\text{total}})^2}_{\text{Varianza Inter}}.$$

Observación

La varianza intra mide la variabilidad que hay en el interior de cada estrato mientras que la varianza inter mide la variabilidad que existe entre los estratos.

Ejercicio

En una empresa conservera el transporte de la materia prima hacia las plantas de proceso se realiza mediante 5 camiones, cada uno de ellos con la capacidad de transportar hasta 45 toneladas. La empresa tiene nuevos directivos y se requiere de estudios estadísticos en el área de transporte. Se decide registrar la cantidad de materia prima (en toneladas) de siete viajes realizados por cada camión:

Camión 1	*39,61*	*41,27*	*44,46*	*41,11*	*43,66*	*43,90*	*38,41*
Camión 2	*39,42*	*44,06*	*43,04*	*39,20*	*43,13*	*39,80*	*40,91*
Camión 3	*38,19*	*40,56*	*43,66*	*38,33*	*44,30*	*44,31*	*43,89*
Camión 4	*38,11*	*44,90*	*40,16*	*43,66*	*40,12*	*43,57*	*42,54*
Camión 5	*43,14*	*39,60*	*40,16*	*39,70*	*40,85*	*42,61*	*43,85*

(a) ¿Cuál camión realiza un traslado menos homogéneo?

(b) ¿Qué porcentaje de la variabilidad total es explicada por la variabilidad de cada camión?

Solución :

¡Excel!

Solución:

(a) Tenemos los siguientes indicadores para cada camión:

$$\overline{X}_1 = 41.77; \quad S_1^2 = 5.33; \quad \text{CV}_1 = 0.055$$
$$\overline{X}_2 = 41.36; \quad S_2^2 = 4.05; \quad \text{CV}_2 = 0.048$$
$$\overline{X}_3 = 41.89; \quad S_3^2 = 7.82; \quad \text{CV}_3 = 0.066$$
$$\overline{X}_4 = 41.86; \quad S_4^2 = 5.97; \quad \text{CV}_4 = 0.058$$
$$\overline{X}_5 = 41.41; \quad S_5^2 = 3.07; \quad \text{CV}_5 = 0.042$$

El camión 3 es el que realiza un transporte menos homogéneo.

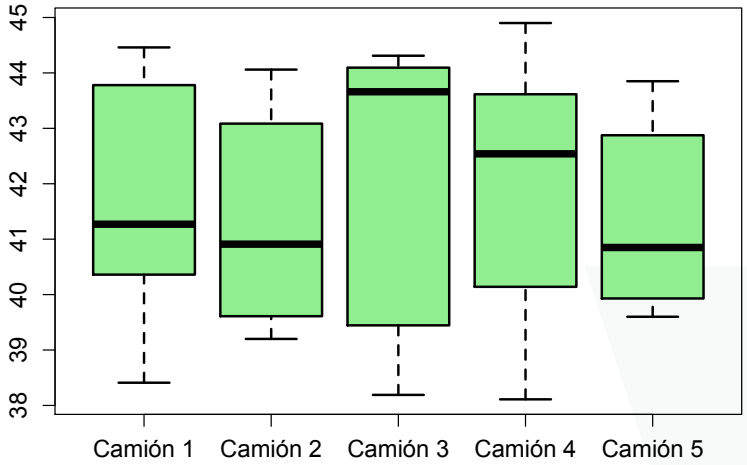
Solución (continuación):

(b) La varianza total es 4.55, mientras que la varianza intra es 4.50.

Haciendo el cuociente:

$$4.50/4.55 = 0.989.$$

Por lo tanto, el $98.9\,\%$ de la variabilidad se explica por la variabilidad de cada camión.



Otros indicadores

Como se ha dicho anteiormente existen factores que pueden afectaer el desempeño de ciertos indicadores de tendencia central o de variabilidad, estos factores pueden ser la presencia de datos extremos o que los datos presenten una alta asimetría.

A continuación listaremos algunos indicadores que pueden ser útiles en ciertos escenarios.

Cuantiles

Los cuantiles son medidas de posición que dividen a los datos en grupos bajo los cuales se encuentra una determinada proporción de éstos, por lo que se requiere que los datos estén al menos en escala ordinal.

- Datos no agrupados

Cuartil i	$Q_i = X_{\left(\frac{i(n+1)}{4}\right)}$	$i : 1, 2, 3, 4$
Quintil i	$K_i = X_{\left(\frac{i(n+1)}{5}\right)}$	$i : 1, 2, 3, 4, 5$
Decil i	$D_i = X_{\left(\frac{i(n+1)}{10}\right)}$	$i : 1, 2, \ldots, 10$
Percentil i	$P_i = X_{\left(\frac{i(n+1)}{100}\right)}$	$i : 1, 2, \ldots, 100$

Obs: El cuartil 2 (Q_2), el decil 5 (D_5) y el percentil 50 (P_{50}) son iguales a la mediana

Nota: Los cuantiles también son robustos ante observaciones extremas

- Datos agrupados

Para calcular cuantiles en datos agrupados se recurre a una interpolación lineal entre los extremos del intervalo que contiene al cuantil de interés.

De esta forma, una expresión para el cálculo de percentiles en datos agrupados:

$$P = LI + \left(\frac{\frac{n \times j}{100} - N_{i-1}}{n_i} \right) a$$

Índice de dispersión (medida de variabilidad)

Mientras más cercanas estén entre si las posiciones de la clase cuartil Q_1 y Q_3, más concentradas están las frecuencias alrededor de la clase mediana. Por tanto, es posible definir una medida de variabilidad en torno a la clase mediana considerando las posiciones de las clases Q_1 y Q_3.

Índice de dispersión (medida de variabilidad)

Mientras más cercanas estén entre si las posiciones de la clase cuartil Q_1 y Q_3, más concentradas están las frecuencias alrededor de la clase mediana. Por tanto, es posible definir una medida de variabilidad en torno a la clase mediana considerando las posiciones de las clases Q_1 y Q_3.

Índice de dispersión:

$$D = \frac{\text{Posición de } Q_3 - \text{Posición de } Q_1}{k-1},$$

donde k es la cantidad de clases

Índice de dispersión (medida de variabilidad)

Mientras más cercanas estén entre si las posiciones de la clase cuartil Q_1 y Q_3, más concentradas están las frecuencias alrededor de la clase mediana. Por tanto, es posible definir una medida de variabilidad en torno a la clase mediana considerando las posiciones de las clases Q_1 y Q_3.

Índice de dispersión:

$$D = \frac{\text{Posición de } Q_3 - \text{Posición de } Q_1}{k-1},$$

donde k es la cantidad de clases

Este índice tiene valores entre 0 y 1. Donde 0 indica baja dispersión y 1 indica alta dispersión.

Nota: Cuando hablamos de posiciones nos referimos al número de la clase a la cual pertenecen los cuartiles.

Esta medida den variabilidad puede ser utilizada para datos al menos ordinales

Coeficientes de asimetria

Coeficiente de asimetría de Bowly-Yule

Estos dos indicadores se basan en los cuartiles:

Coeficientes de asimetria

Coeficiente de asimetría de Bowly-Yule

Estos dos indicadores se basan en los cuartiles:

$$I_S = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1},$$

son adimencionales y no se ven afectados por datos extremos (robustos), por lo que estos enrealidad miden la simetría en el centro de los datos y no en el total de estos.

Coeficientes de asimetria

Coeficiente de asimetría de Bowly-Yule

Estos dos indicadores se basan en los cuartiles:

$$I_S = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1},$$

son adimencionales y no se ven afectados por datos extremos (robustos), por lo que estos enrealidad miden la simetría en el centro de los datos y no en el total de estos.

- Si $I_S < 0$ la distribución tiene una asimetría negativa.
- Si $I_S = 0$ la distribución es simétrica.
- Si $I_S > 0$ la distribución tiene una asimetría positiva.

Coeficiente de asimetría de Pearson

Cuando hay simetría ideal la mediana debiese coincidir con la media, esto lleva a definir el coeficiente de Pearson muestral y poblacional como:

Coeficiente de asimetría de Pearson

Cuando hay simetría ideal la mediana debiese coincidir con la media, esto lleva a definir el coeficiente de Pearson muestral y poblacional como:

$$A_S = \frac{3(\overline{x} - Me)}{S}, \qquad A_S = \frac{3(\mu - Me)}{\sigma}$$

este coeficiente está acotado teóricamente entre -3 y 3.

Coeficiente de asimetría de Pearson

Cuando hay simetría ideal la mediana debiese coincidir con la media, esto lleva a definir el coeficiente de Pearson muestral y poblacional como:

$$A_S = \frac{3(\overline{x} - Me)}{S}, \qquad A_S = \frac{3(\mu - Me)}{\sigma}$$

este coeficiente está acotado teóricamente entre -3 y 3.

- Si $A_S < 0$ la distribución tiene una asimetría negativa.
- Si $A_S = 0$ la distribución es simétrica.
- Si $A_S > 0$ la distribución tiene una asimetría positiva.

Coeficientes de curtosis

Coeficiente K_2

Se basa en cuantiles extremos

$$K_2 = \frac{D_9 - D_1}{1.9(Q_3 - Q_1)} - 1$$

Este no se ve afectado por observaciones extremas, el número 1,9 representa la distancia teórica entre los deciles 9 y 1.

Coeficientes de curtosis

Coeficiente K_2

Se basa en cuantiles extremos

$$K_2 = \frac{D_9 - D_1}{1.9(Q_3 - Q_1)} - 1$$

Este no se ve afectado por observaciones extremas, el número 1,9 representa la distancia teórica entre los deciles 9 y 1.

- Si $K_2 < 0 \Longrightarrow$ Platicúrtica
- Si $K_2 = 0 \Longrightarrow$ Mesocúrtica
- Si $K_2 > 0 \Longrightarrow$ Leptocúrtica