Probabilidad y estadística

MAT 041, Primer semestre

Francisco Cuevas Pacheco

14 de noviembre de 2022

Contenidos

- ✓ Introducción
- ✓ Medidas de Tendencia Central
- ✓ Medidas de Dispersión
- ✓ Gráficos
- ✓ Medidas de Forma
- ✓ Datos Tabulados
- ✓ Muestras Estratificadas
- ✓ Otros indicadores

¿Qué es la Estadística Descriptiva?

La estadística descriptiva es una técnica matemática dedicada a obtener, organizar y describir un conjunto de datos.

Los primeros conceptos que debemos identificar son población y muestra.

Definición (Población)

Una población es un conjunto que contiene la totalidad de individuos a ser estudiados por sus características. Una característica o variable estadística de interés se denota generalmente como X.

Definición (Muestra)

Una muestra es un subconjunto de una población seleccionada de acuerdo a algún método de muestreo. Generalmente, n denotará el tamaño de la muestra y $X_1, X_2, \ldots, X_n$ las observaciones pertenecientes a la muestra.

Población/Muestra

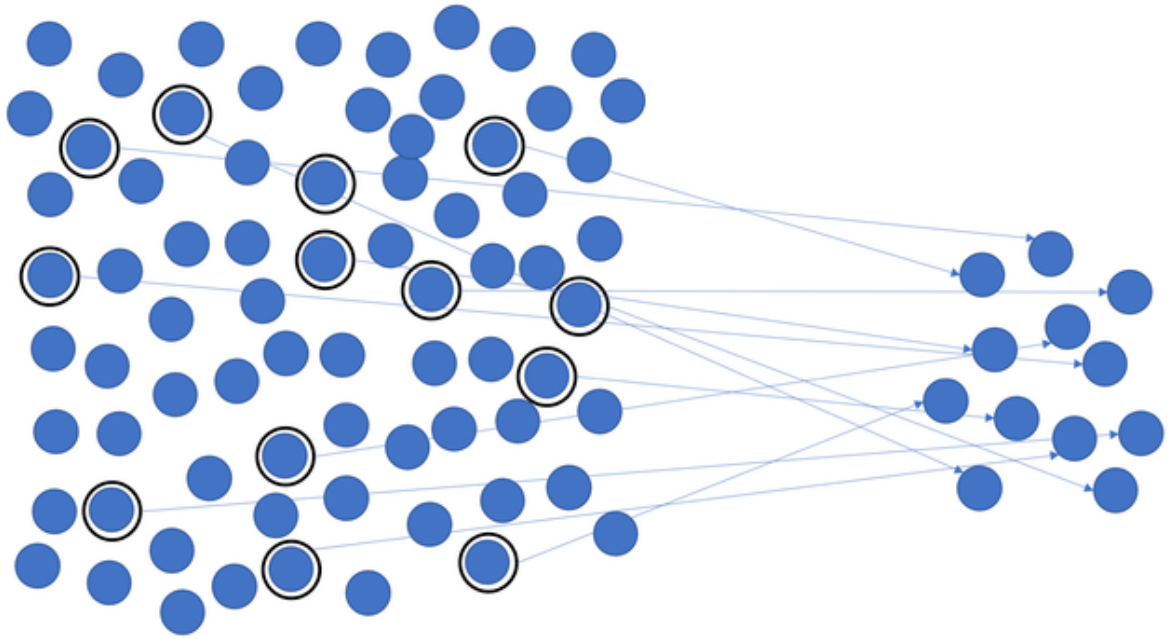


Figura 1: Población (Izquierda) y muestra (Derecha)

¿Porqué muestrear?

¿Porqué muestrear?

1. Porque es muy caro.

¿Porqué muestrear?

1. Porque es muy caro.
2. Porque no se puede recolectar toda la información.

¿Porqué muestrear?

1. Porque es muy caro.
2. Porque no se puede recolectar toda la información.
3. Porque son mas fáciles de recolectar.

¿Porqué muestrear?

1. Porque es muy caro.
2. Porque no se puede recolectar toda la información.
3. Porque son mas fáciles de recolectar.

⋆ Porque no se puede modelar toda la información

Ejemplo de una muestra

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	#	Name	Type 1	Type 2	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary	
2	1	Bulbasaur	Grass	Poison	318	45	49	49	65	65	45	1	False	
3	2	Ivysaur	Grass	Poison	405	60	62	63	80	80	60	1	False	
4	3	Venusaur	Grass	Poison	525	80	82	83	100	100	80	1	False	
5	3	VenusaurMe	Grass	Poison	625	80	100	123	122	120	80	1	False	
6	4	Charmander	Fire		309	39	52	43	60	50	65	1	False	
7	5	Charmeleon	Fire		405	58	64	58	80	65	80	1	False	
8	6	Charizard	Fire	Flying	534	78	84	78	109	85	100	1	False	
9	6	CharizardMe	Fire	Dragon	634	78	130	111	130	85	100	1	False	
10	6	CharizardMe	Fire	Flying	634	78	104	78	159	115	100	1	False	
11	7	Squirtle	Water		314	44	48	65	50	64	43	1	False	
12	8	Wartortle	Water		405	59	63	80	65	80	58	1	False	
13	9	Blastoise	Water		530	79	83	100	85	105	78	1	False	
14	9	BlastoiseMe	Water		630	79	103	120	135	115	78	1	False	
15	10	Caterpie	Bug		195	45	30	35	20	20	45	1	False	
16	11	Metapod	Bug		205	50	20	55	25	25	30	1	False	
17	12	Butterfree	Bug	Flying	395	60	45	50	90	80	70	1	False	

Figura 2: Ejemplo de dato

Tipos de dato para X

Los datos se suelen clasificar en:

1. Cualitativos
    1) Nominal.
    2) Ordinal.
2. Cuantitativos
    1) Discretos.
    2) Contínuos.

Tipos de dato para X

1. Dato Nominal

Las observaciones son categorías sin orden.

Ejemplo

- ✓ *Color de pelo: rubio, castaño, rojo, negro, etc*
- ✓ *Tipo de dieta: vegetariano, no vegetariano*

2. Dato Ordinal

Las observaciones se pueden ordenar aunque no son necesariamente números

Ejemplo

- ✓ *Satisfacción con algún servicio: (1) insatisfecho (2) neutral (3) satisfecho*
- ✓ *Prioridades luego de un desastre: (1) suministro de agua (2) comida, etc.*

3. Dato Discreto

Las observaciones pertenecen a algún subconjunto de $\mathbb{N}$.

Ejemplo

- ✓ *Edad.*
- ✓ *Número de hijos.*
- ✓ *Número de vehículos.*

4. Dato Contínuo

Las observaciones pertenecen a algún subconjunto de $\mathbb{R}$.

Ejemplo

- ✓ *Nivel de colesterol*
- ✓ *Tiempo de duración de un equipo electrónico*

Discreto vs. Continuo

Ejemplo

- *El número de clientes que llegan diariamente a un restaurante o el número anual de nacimientos en Chile son variables discretas.*
- *La temperatura del aire o el valor del dólar son variables continuas.*

Resúmenes estadísticos

Debido a que tenemos una muestra, necesitamos generar resúmenes que nos permitan comprender la realidad de estos datos.

Para esto, la estadística nos proporciona varias herramientas:

1. Indicadores
2. Gráficos
3. Tablas

Las medidas de tendencia central representan un centro en torno al cual se encuentra ubicado el conjunto de datos.

- El **promedio** (o media aritmética):

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

- El **promedio ponderado**:

$$\overline{X}_w = \sum_{i=1}^{n} w_i X_i,$$

donde las constantes $w_i \geq 0$ son ponderaciones o pesos tales que $\sum_{i=1}^{n} w_i = 1$.

- La **moda** es aquel valor que más se repite en la muestra (observe que no necesariamente es única).

- La **mediana** (Me) es un valor que divide la muestra en dos partes iguales. Usamos la notación $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$, donde $X_{(1)} = \min\{X_i\}$ y $X_{(n)} = \max\{X_i\}$, para representar las observaciones ordenadas de menor a mayor. Entonces,

$$\text{Me} = \begin{cases} X_{\left(\frac{n+1}{2}\right)} & \text{cuando } n \text{ es impar} \\ \\ \frac{1}{2}\left(X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}\right) & \text{cuando } n \text{ es par} \end{cases}$$

Ejercicio

Considere la muestra $\{9.2,\ 4,\ 7,\ 4,\ 3.5,\ 200\}$. *¿Como impacta la observación* 200 *en las diferentes medidas de tendencia central?*

Solución:

- Sin la observación 200: $\overline{X} = 5.54$; Me $= 4$; moda $= 4$.
- Con la observación 200: $\overline{X} = 37.95$; Me $= 5.5$; moda $= 4$.

Conclusión: se dice que la mediana es más *robusta* que el promedio, ya que se ve menos afectada por la presencia de datos atípicos.

Las medidas de dispersión representan la variabilidad de las observaciones respecto de algún punto de referencia.

- La **varianza muestral** está dada por

$$S^2_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

  La **desviación estándar muestral** es

$$S_{n-1} = \sqrt{S^2_{n-1}}.$$

- La **varianza** se define como

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

  La **desviación estándar** se define como

$$S_n = \sqrt{S_n^2}.$$

Observación

La única diferencia entre la **varianza** y la **varianza muestral** es el factor que multiplica la suma. Más adelante veremos que la varianza muestral es un *estimador insesgado* de la *varianza poblacional*.

Observaciones Adicionales

1. A diferencia de la varianza, la desviación estándar tiene las mismas unidades que los datos.

2. La varianza mide las desviaciones cuadráticas respecto al promedio. Se pueden definir medidas de dispersión adicionales reemplazando el promedio por otra medida de tendencia central (por ejemplo, la mediana).

- El **coeficiente de variación (CV)** está dado por

$$\text{CV} = \frac{S_{n-1}}{\overline{X}}.$$

Observación

El CV no tiene dimensiones y es útil para comparar dos o más muestras. Mientras más pequeño es el CV, más homogénea es la muestra.

- El **rango** es la diferencia entre el elemento más grande y más pequeño de la muestra:
$$\text{rango} = X_{(n)} - X_{(1)}.$$

- El **rango modificado** es la diferencia entre el elemento más grande y más pequeño de la muestra:
$$\text{RM}_j = X_{(n/2-j)} - X_{(n/2+j)}.$$

- El **rango intercuartil (RIQ)** se define como
$$\text{RIQ} = Q_3 - Q_1,$$
donde Q_1 es el valor medio entre el valor mínimo de la muestra y la mediana (cuartil 1) y Q_3 es el valor medio entre la mediana y el valor máximo de la muestra (cuartil 3). En otras palabras, el RIQ corresponde al rango luego de remover el $25\,\%$ más grande y el $25\,\%$ más pequeño de la muestra.

Fórmula alternativa para la varianza

La siguiente ecuación permite calcular la varianza (muestral) y la desviación estándar (muestral) de una manera más sencilla:

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\overline{X})^2 &= \sum_{i=1}^{n}\left(X_i^2-2X_i\overline{X}+\overline{X}^2\right)\\ &= \sum_{i=1}^{n}X_i^2-2\overline{X}\sum_{i=1}^{n}X_i+n\overline{X}^2\\ &= \sum_{i=1}^{n}X_i^2-n\overline{X}^2.\end{aligned}$$

Por ejemplo, dividiendo por n en la ecuación anterior, concluímos

$$\boxed{S_n^2=\frac{1}{n}\sum_{i=1}^{n}X_i^2-\overline{X}^2}$$

Ejercicio

Para una muestra compuesta de 21 observaciones se calculó el promedio ($\overline{X} = 1$) y la varianza ($S_n^2 = 50$). Posteriormente, descubrieron un error en los datos originales. Una observación con valor -5, en realidad tenía un valor igual a 16. Calcular los valores correctos para $\overline{X}$ y S_n^2.

Solución: Como $n = 21$ y el promedio incorrecto es igual 1, la suma incorrecta de los datos es 21. Debemos remover el -5 erróneo y agregar el 16, es decir, la suma corregida será

$$\sum_{i=1}^{21} X_i = 21 - (-5) + 16 = 42.$$

Luego, el promedio corregido es igual $\overline{X} = 42/21 = 2$.

Solución (continuación): Desde la fórmula alternativa para la varianza, se tiene que

$$\sum_{i=1}^{n} X_i^2 = nS_n^2 + n\overline{X}^2.$$

Luego, la suma incorrecta de los cuadrados es

$$21 \times 50 + 21 \times (1)^2 = 1071.$$

Entonces, la suma de los cuadrados corregida es

$$\sum_{i=1}^{21} X_i^2 = 1071 - (-5)^2 + 16^2 = 1302.$$

Por lo tanto, la varianza corregida es $S_n^2 = 1302/21 - 2^2 = 58$.

Propiedades

Sea $X_1, \ldots, X_n$ una muestra de la variable X, desde donde se obtuvo el promedio $\overline{X}$ y la varianza S_X^2. Si transformamos los valores X_i's de la siguiente forma:

$$Y_i = aX_i + b,$$

donde a y b son constantes. Entonces,

$$\begin{aligned} \overline{Y} &= a\overline{X} + b \\ S_Y^2 &= a^2 S_X^2. \end{aligned}$$

Ejercicio

Los empleados de una compañía británica reciben sus salarios mensuales en libras esterlinas (£). Una división de la compañía será re-ubicada en Francia, donde sus salarios serán pagados en euros (€). Una libra esterlina es igual a 1.27 euros. Mientras los empleados están en Francia, cada uno recibirá también un bono mensual de €325. La siguiente tabla contiene información sobre los salarios originales en Gran Bretaña.

Mínimo	***£800***
Primer cuartil	***£1250***
Mediana	***£1470***
Tercer cuartil	***£2250***
Máximo	***£4500***
Promedio	***£2025***
Desviación Estándar	***£475***

(a) Calcule el rango y el rango intercuartil.

(b) Calcule el promedio y la desviación estándar de los salarios mensuales en euros luego de re-ubicarse en Francia.

Solución:

(a) rango $= 4500 - 800 = 3700$; RIQ $= 2250 - 1250 = 1000$.

(b) Denotemos por X los sueldos en Gran Bretaña y denotemos por Y los sueldos en Francia. Luego,

$$\begin{aligned} \overline{Y} &= 1.27 \times \overline{X} + 325 \\ &= 1.27 \times 2025 + 325 \\ &= 2896.75, \end{aligned}$$

mientras que $S_Y^2 = 1.27^2 \times S_X^2$. Luego,

$$\begin{aligned} S_Y &= 1.27 \times S_X \\ &= 1.27 \times 475 \\ &= 603.25. \end{aligned}$$

Ahora veremos algunos gráficos de interés.

Definición (Frecuencia)

El rango de los datos se divide en k intervalos: $I_1, I_2, \ldots, I_k$.

- *La cantidad de observaciones en I_i se denomina la frecuencia absoluta de la clase I_i y se denota por n_i.*
- *La frecuencia relativa de la clase I_i se define como $f_i = n_i/n$.*

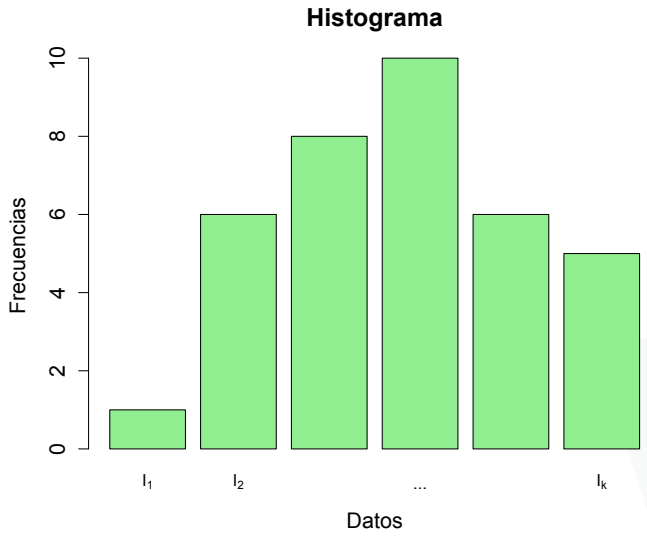
Observación

Note que

$$\sum_{i=1}^{k} n_i = n \quad \text{y} \quad \sum_{i=1}^{k} f_i = 1.$$

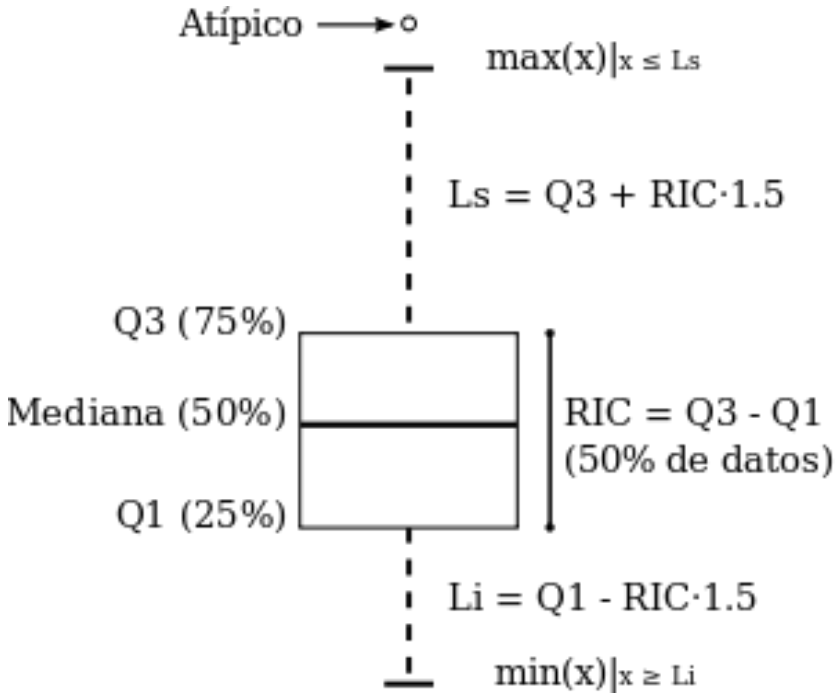
Definición (Histograma)

Un histograma es un gráfico de n_i versus I_i (ó f_i versus I_i) que permite resumir la cantidad de observaciones por unidad de longitud.



Boxplot (Diagrama de Caja)

Permite visualizar la simetría, las observaciones atípicas (outliers), la dispersión respecto a la mediana y el rango. El bigote superior es $\mín\{L_s, X_{(n)}\}$. El bigote inferior es $\máx\{L_i, X_{(1)}\}$.



Ejercicio

Se han obtenido las siguientes calificaciones:

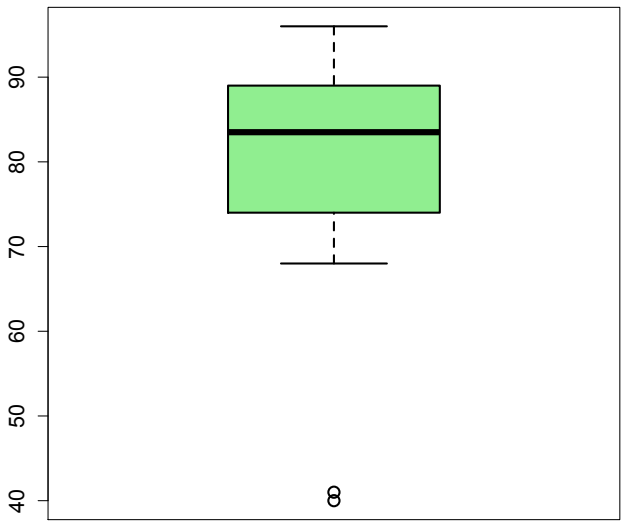
40	*41*	*68*	*69*	*72*	*76*	*79*	*79*	*80*	*82*
85	*86*	*87*	*88*	*88*	*90*	*91*	*92*	*93*	*96*

Construír el boxplot e identificar los datos atípicos.

Solución: $Q_1 = (X_{(5)} + X_{(6)})/2 = 74$; Me $= (X_{(10)} + X_{(11)})/2 = 83.5$; $Q_3 = (X_{(15)} + X_{(16}$

$\Rightarrow$ RIC $= 15$; $L_i = 51.5$; $L_s = 111.5$.

Para realizar el gráfico debemos utilizar los valores $\text{mín}\{X_{(20)}, L_s\} = 96$ y $\text{máx}\{X_{(1)}, L_i\} = 51.5$.



Observación

En el rango comprendido entre 51.5 y 68 no hay datos. El software R sube el bigote inferior hasta 68. Este paso adicional permite que el gráfico sea más informativo. Esto muestra que existen diferentes variantes para construír este tipo de gráficos. Notar que existen 2 valores atípicos en la muestra.