Estadística Descriptiva

Facultad de Ingeniería

Universidad Católica del Uruguay

Abril 2023

Índice

1 - Estadística Descriptiva

Estadística Descriptiva

- Para resumir información y poder sacar conclusiones sobre datos.

Inferencia Estadística

- Para estimar parámetros desconocidos, tanto del pasado como del futuro. Predicción.

1 - Estadística Descriptiva

Estadísticas de un jugador

Descriptiva: Resumir historial deportivo
Inferencial: Predecir desempeño futuro del jugador

Estadísticas de Jugador

Resumen | Defensivo | Ofensivo | Distribución | Detallado

General | Local | Visitante — Mínimo jgdos | Todos los jugadores

	Jugador	Jgdos	Mins	Goles	Asist	Amar	Roja	TpP	AP%	Aéreos	JdelP	Rating
1	Lionel Messi Barcelona, 33, MP(CD),DL	25(2)	2303	23	8	4	-	5.6	85.4	0.3	17	8.54
2	Robert Lewandowski Bayern, 32, DL	24(1)	2103	35	6	3	-	4.4	76.8	1.7	9	8.05
3	Harry Kane Tottenham, 27, MP(C),DL	28	2457	19	13	1	-	3.9	69.5	2.4	11	7.82
4	Gerard Moreno Villarreal, 28, MP(CD),DL	24	2068	19	5	3	-	3.3	68.7	2.1	11	7.78

1 - Estadística Descriptiva

¿Por qué estudiar estadística descriptiva?

- Permite una mejor comprensión de la información cuantitativa.
- Capacita a la persona para evaluar objetivamente y efectivamente la información que recibe (vía tablas, gráficos, porcentajes, tasas, etc.).
- Porque ejerce una profunda influencia en casi todos los campos de la actividad humana.

Estadística Descriptiva - Variables

- Tipos de Variables:

1. Cualitativas

2. Cuantitativas

Estadística Descriptiva - Variables

Cualitativas

Cada observación se puede clasificar en alguna categoría. (Estas deben ser exhaustivas)

Ejemplos: sexo, nivel educativo alcanzado, etc.

Variables Cualitativas:

1) Binarias (dummy, one-hot)
2) Nominales
3) Ordinales

Estadística Descriptiva - Variables

- 1) Variables Dummy (binarias)

  Se le asigna 0 o 1 a cada categoría.

  (Ej.: Fuma = 1, No_Fuma = 0)
- 2) Variables Nominales

  Más de dos categorías sin un orden establecido.

  (Ej.: País de origen, estado civil)
- 3) Variables Ordinales

  Existe un orden entre las categorías.

  (Ej.: Notas: regular, bueno, muy bueno; severidad de la patología: ausente, moderado, severo)

Estadística Descriptiva - Variables

2. Variables Cuantitativas

El resultado de la observación es un número.

Estas pueden ser:

1) Continuas
2) Discretas

Estadística Descriptiva – Variables Cuantitativas

1) Continuas:

Las mediciones pueden tomar teóricamente un conjunto infinito de valores posibles dentro de un rango. Ej.: altura, ingresos.

2) Discretas:

La variable sólo puede tomar un cierto conjunto de valores posibles. Ej.: números de integrantes del hogar.

Tipos de variables

- **En grupos, clasificar las siguientes variables:**

1. Marcas de autos que pasan por un peaje en una hora.
2. Número de autos que pasan por un peaje en una hora.
3. Tiempo de espera hasta que pasen 10 autos por un peaje.
4. Calidad de atención en una estación de servicio (MB, B, M, MM)
5. Observar si el auto es eléctrico o no.
6. Número de votos por un candidato.
7. Ganancias esperadas de un negocio.

Tipos de variables

wordwall.net/es/resource/54050025

Presentación de datos: Variables Cualitativas

- Tabla de Frecuencias

Un grupo de 30 estudiantes participaron en una prueba diagnóstica, clasificada en A, B y C.

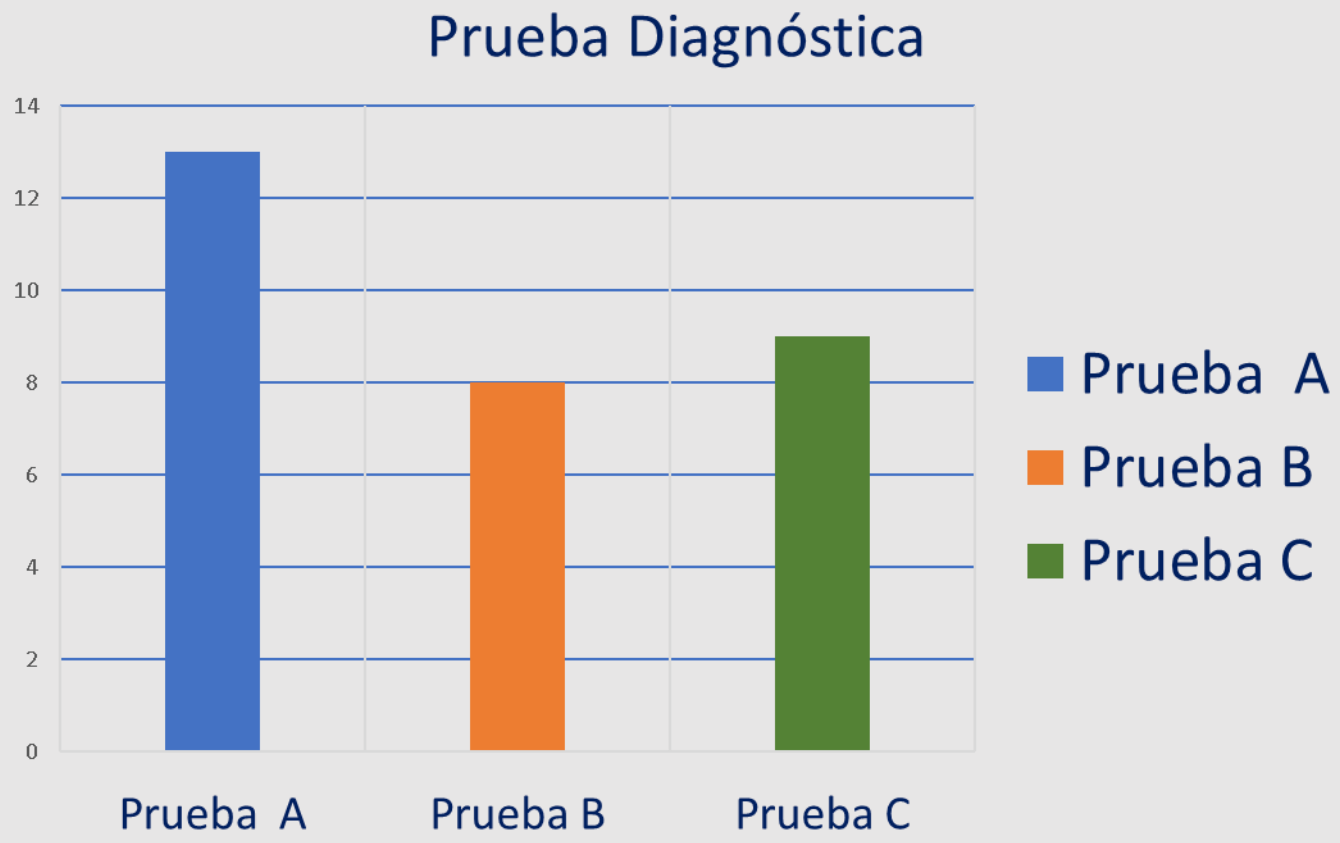
X = tipo de prueba n = 30

	ni	hi	F*i
Prueba A	13	0,43	0,43
Prueba B	8	0,27	0,7
Prueba C	9	0,3	1

- **ni** = Frecuencia absoluta
- **hi** = Frecuencia relativa
- **F*i** = Frecuencia acumulada

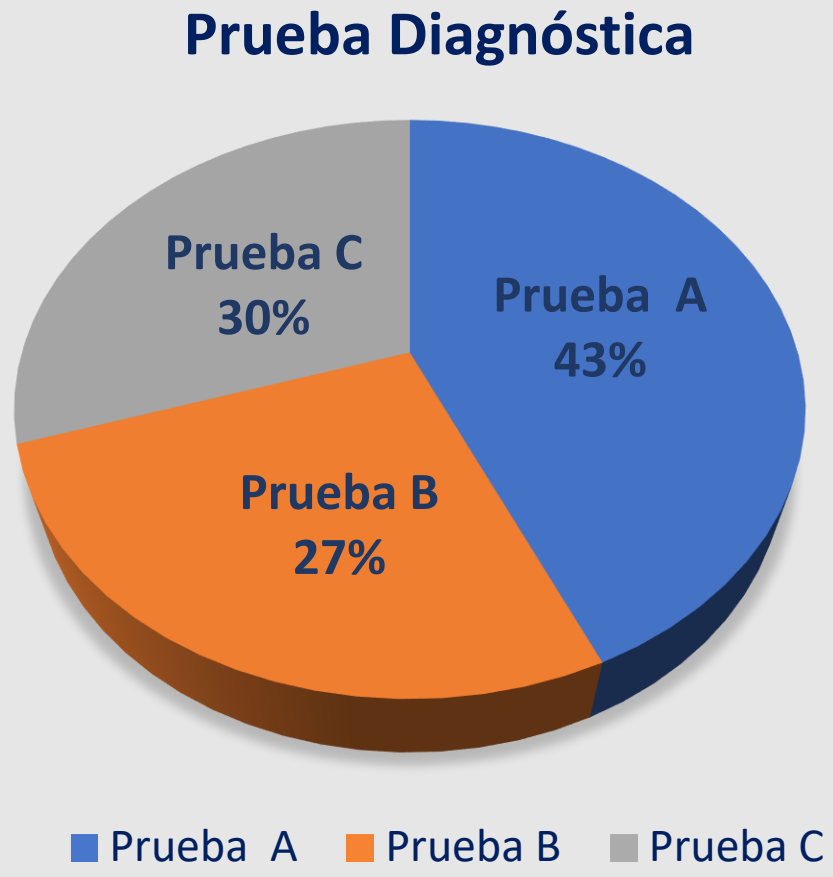
Presentación de datos: Variables Cualitativas

Gráfico de Barras



Presentación de datos: Variables Cualitativas

Gráfico Circular



Presentación de datos: Variables Cuantitativa

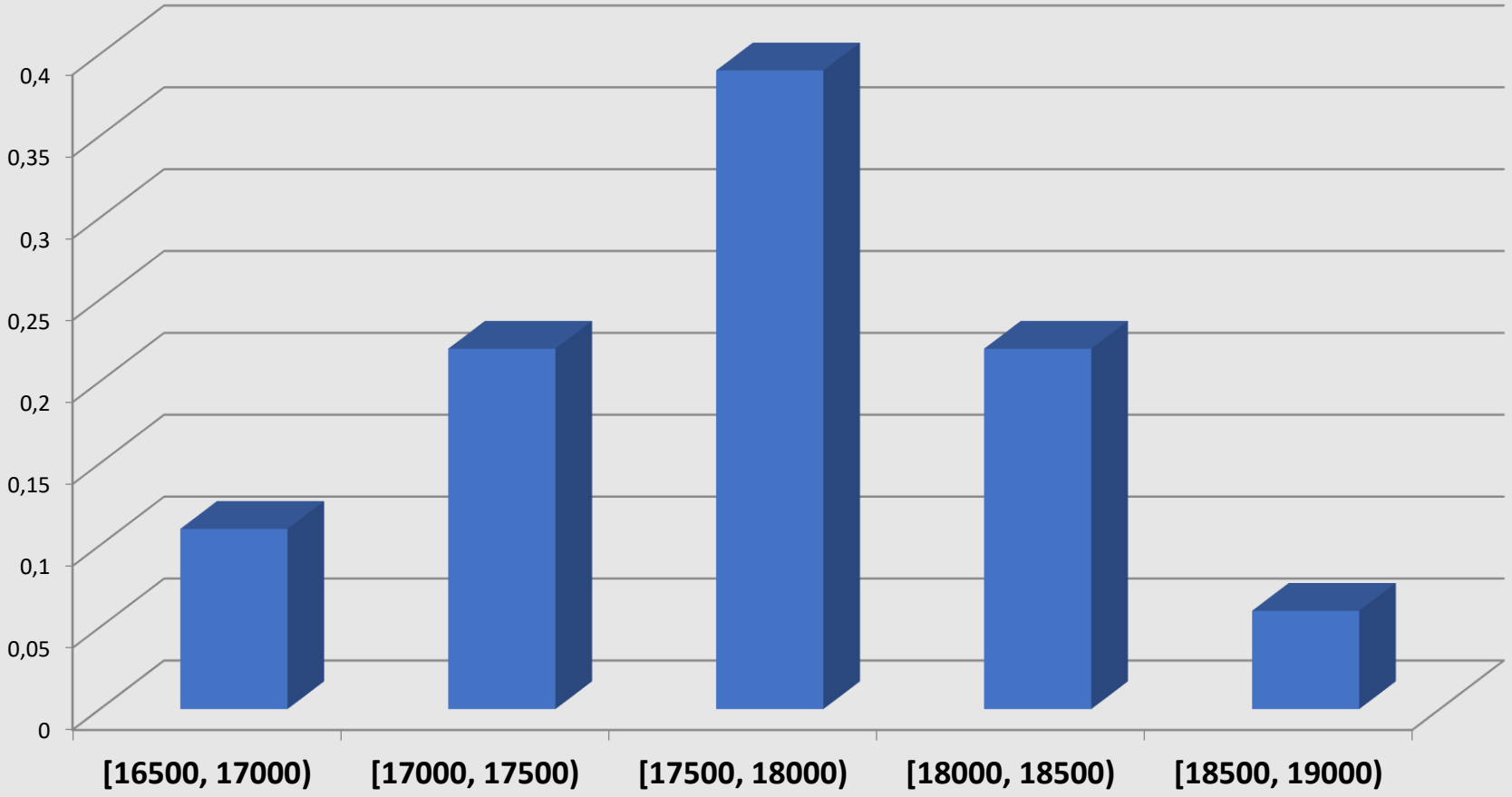
Información de los salarios (expresados en miles de pesos de 36 empleados)Tabla de Frecuencias

X = Salarios n = 36

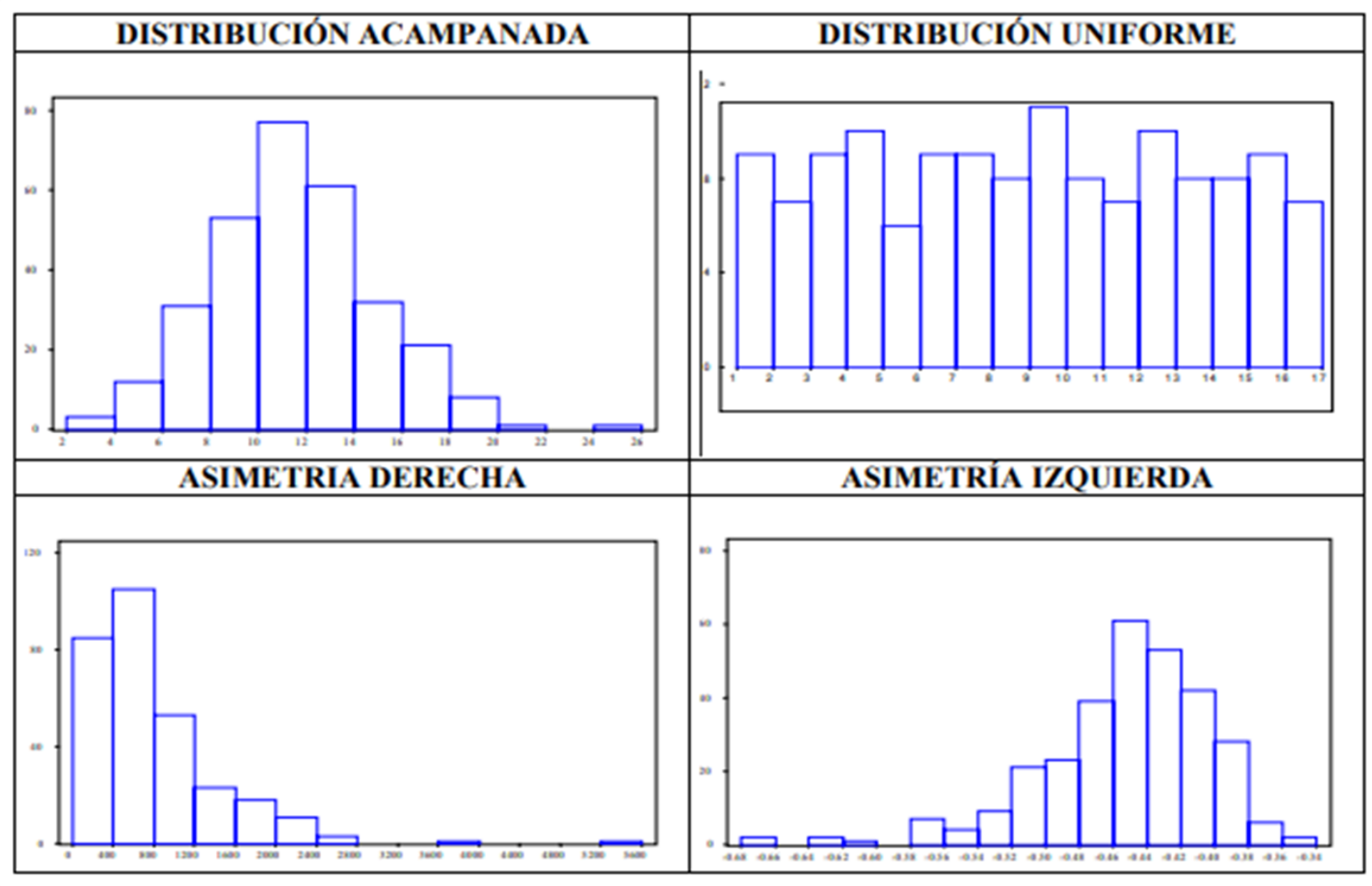
Intervalos	Marca de Clase	Frecuencia absoluta	Frecuencia relativa	Frecuencia relativa acumulada
[16500, 17000)	16750	4	0,11	0,11
[17000, 17500)	17250	8	0,22	0,33
[17500, 18000)	17750	14	0,39	0,72
[18000, 18500)	18250	8	0,22	0,94
[18500, 19000)	18750	2	0,06	1

Presentación de datos: Variables Cuantitativa

Gráfico de Barras



Presentación de datos: Variables Cuantitativa



Medidas.

Medidas de tendencia central:

- Media - Mediana - Moda

Medidas de dispersión:

- Varianza - Desvío Estándar – Coeficiente de Variación

Medidas de asociación:

- Covarianza - Correlación

Medidas de tendencia central

- MEDIA

$$\bar{X} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Ejemplo:

¿Cuál es la media del tiempo requerido para llegar a la UCU desde nuestros hogares? (En minutos)

Medidas de tendencia central

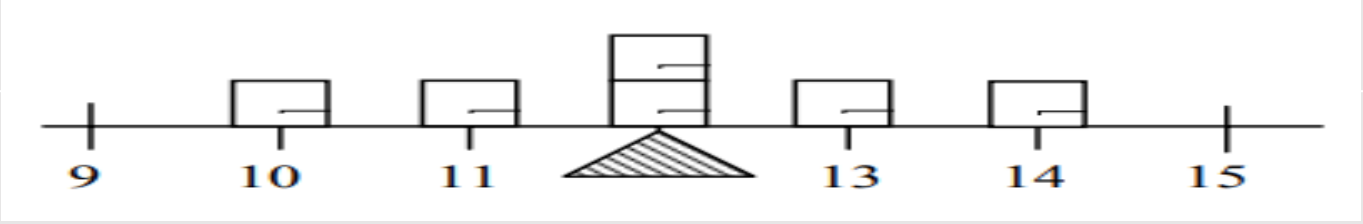
Características y propiedades de la media

- Se usa para datos numéricos.
- Representa el centro de gravedad o el punto de equilibrio de los datos.

Ejemplo.

$X_1 = 10 \qquad X_2 = 14 \qquad X_3 = 12 \qquad X_4 = 11 \qquad X_5 = 12 \qquad X_6 = 13$

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_6}{n} = \frac{10+14+12+11+12+13}{6} = \frac{72}{6} = 12$$

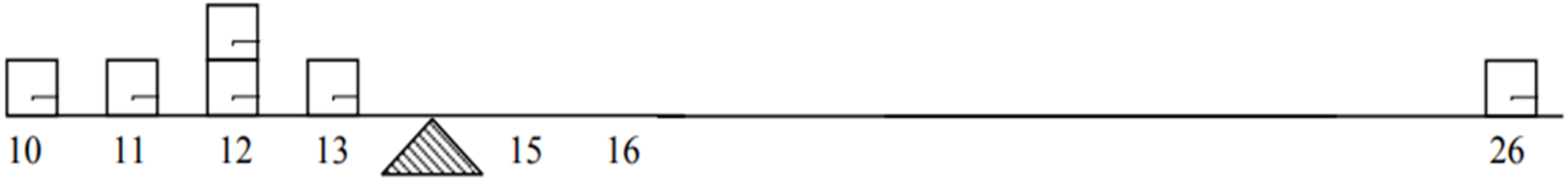


Medidas de tendencia central

Características y propiedades de la media:

- Es muy sensible a la presencia de datos atípicos

Modificamos 1 dato en el ejemplo anterior $X_2 = 14 \rightarrow X_2 = 26$ y $\overline{X} = 12 \rightarrow \overline{X} = 14$.



Medidas de tendencia central

- **Mediana**

Se obtiene ordenando primero las n observaciones de la más pequeña a la más grande (con cualquier valor repetido incluido de modo que cada observación muestral aparezca en la lista ordenada).

Medidas de tendencia central

- **Mediana**

➢ Si n = impar, la mediana es el dato que ocupa la posición central.

$$\tilde{X} = x_{\left(\frac{n+1}{2}\right)}$$

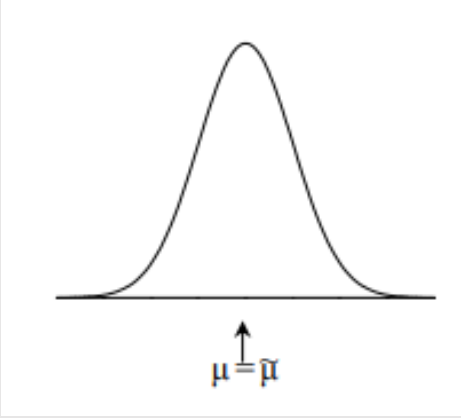
➢ Si n = par, la mediana es el promedio de los dos datos centrales.

$$\tilde{X} = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n+1}{2}\right)}}{2}$$

Medidas de tendencia central

Propiedades de la mediana

- Puede ser usada no sólo para datos numéricos sino además para datos ordinales, ya que para calcularla sólo es necesario establecer un orden en los datos.
- Si la distribución de los datos es aproximadamente simétrica la media y la mediana serán aproximadamente iguales.



Medidas de tendencia central

- **Moda**

El moda es el dato que ocurre con mayor frecuencia.

En el caso de variables agrupadas en intervalos se define el intervalo modal como aquel intervalo que presenta el mayor valor de frecuencia relativa .

Puede dar más de un valor.

Medidas de dispersión

RANGO

El rango de n observaciones $x_1, x_2, \ldots\ldots x_n$ es la diferencia entre la observación más grande y la más pequeña:

$$\boldsymbol{Rango = x_{máx} - x_{min}}$$

Medidas de dispersión

Características y propiedades del rango:

- Es muy simple de obtener.
- Es extremadamente sensible a la presencia de datos atípicos.
- Si hay datos atípicos, estos estarán en los extremos, que son los datos que se usan para calcular el rango.
- Ignora la mayoría de los datos.
- En general aumenta cuando aumenta el tamaño de la muestra (las observaciones atípicas tienen más chance de aparecer en una muestra con muchas observaciones).

Medidas de dispersión: varianza - desvío estándar

- **Varianza:** desvío cuadrático respecto a la media

$$S^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

- **Desvío Estándar:** desviación respecto a la media

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

Medidas de dispersión-Coeficiente de variación

Sí queremos medir la dispersión en porcentaje para comparar con otra variable que esté en distinta unidad de medida se usa el CV.

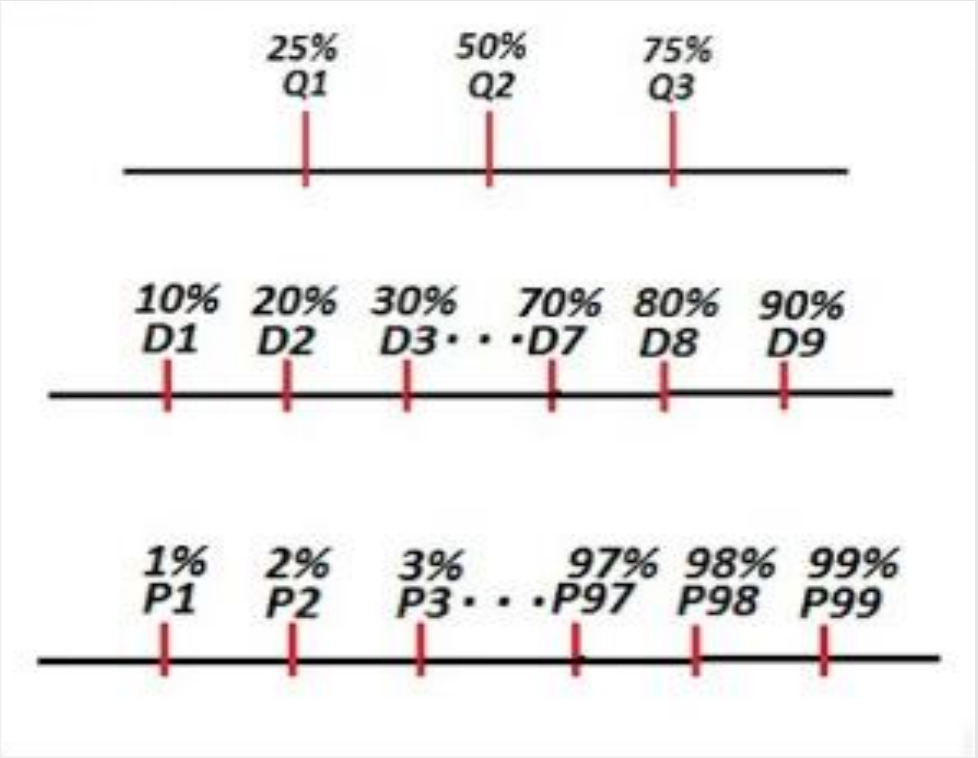
$$CV = \frac{S}{|\bar{x}|} \text{ si } \bar{x} \neq 0$$

Medidas de posición

Cuartil $Q_k = \frac{kN}{4}$

Decil $D_k = \frac{kN}{10}$

Percentil $Q_k = \frac{kN}{100}$



Medidas de dispersión: rango-intercuartílico

- Representa la distancia entre el primer y el tercer cuartil.

- Indica el rango donde se encuentra aproximadamente el 50% “central” de las observaciones.

$$RI = Q_3 - Q_1$$

Diagrama de caja (Box-Plot)

El gráfico de caja resultante se muestra en la figura siguiente.

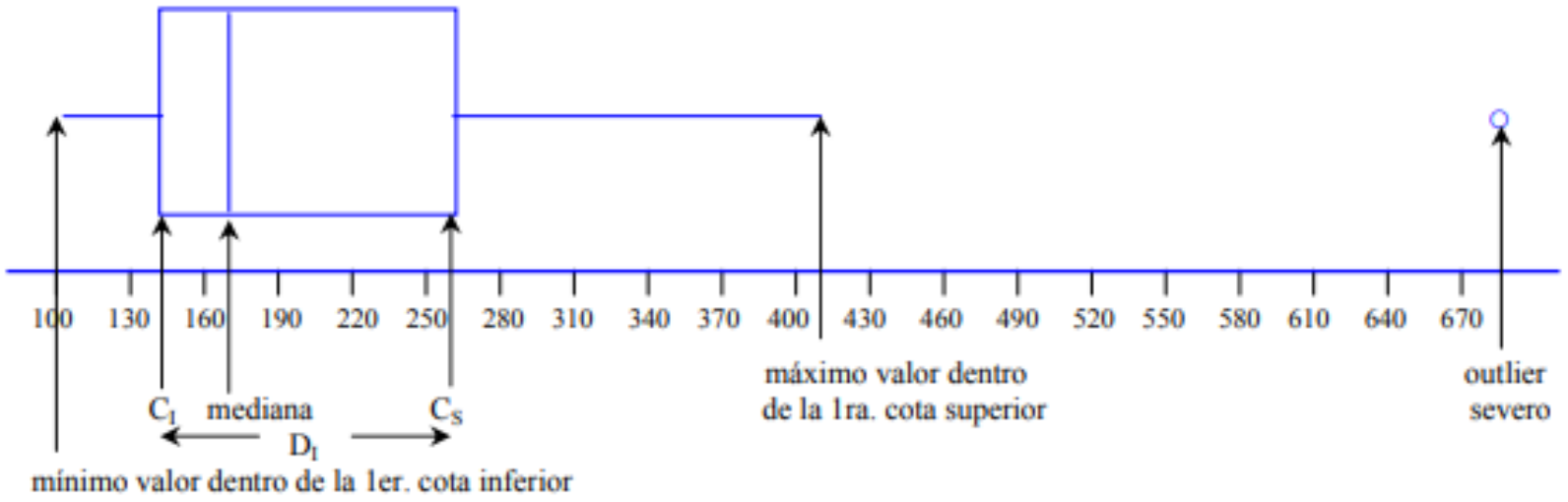


Diagrama de caja (Box-Plot)

Barreras:

- Interiores: $BII = Q_1 - 1{,}5RI \quad BIS = Q_3 + 1{,}5RI$
- Exteriores: $BEI = Q_1 - 3RI \quad BES = Q_3 + 3RI$

- Datos Atípicos:
- No extremos: Ubicados entre las barreras interiors y exteriores.
- Extremos: Ubicados más allá de las barreras exteriores.

- Nota: Si un dato coincide con el valor de una de las barreras exteriores se clasifican como atípico no extremo.

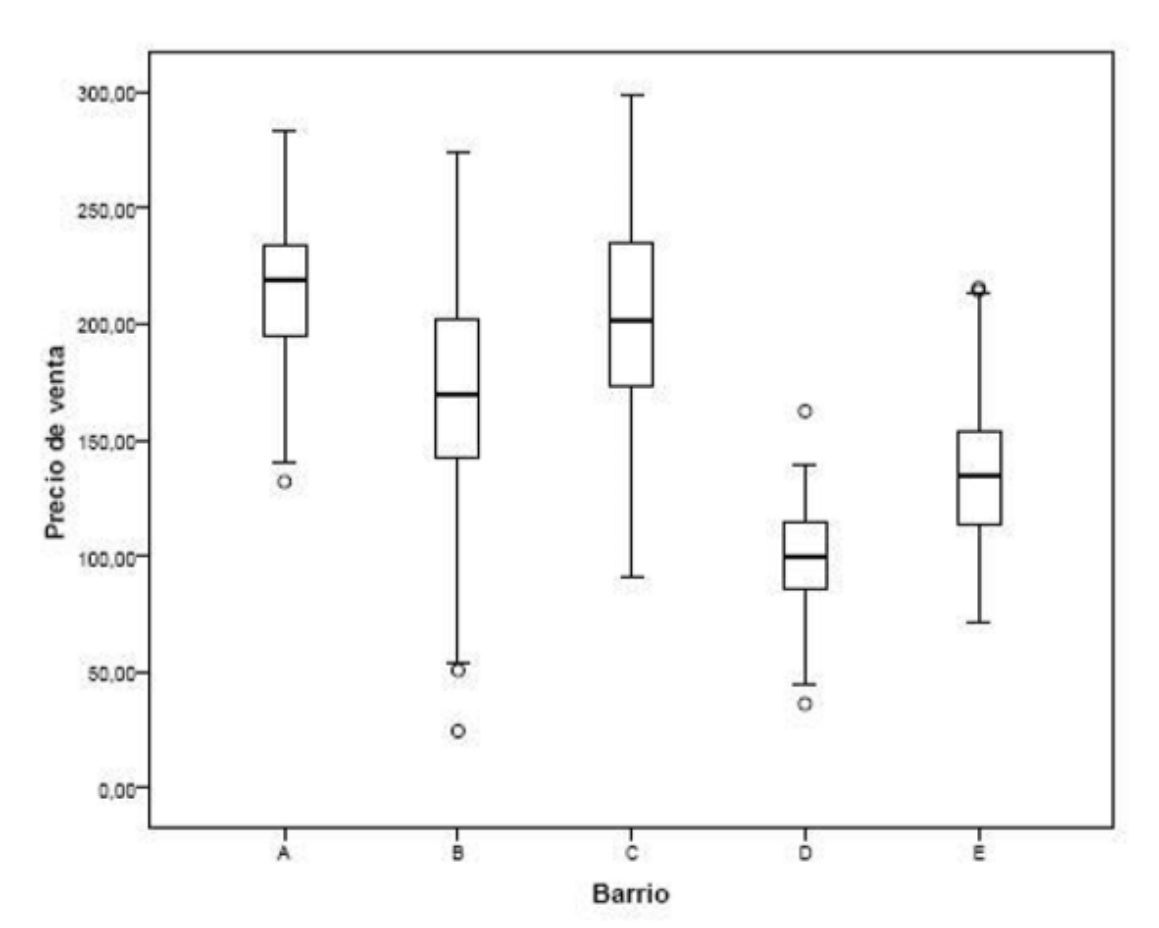
Diagrama de caja (Box-Plot)

¿Qué nos permite observar un box-plot?

- Muestra donde se posiciona el primer y tercer cuartil
- Muestra una medida de posición ⇒ MEDIANA
- Muestra una medida de dispersión ⇒ RECORRIDO INTERCUARTÍLICO
- Permite estudiar la simetría de la distribución
- Nos da un criterio de detección de datos.

Diagrama de caja (Box-Plot)

Comparación de datos a través del diagrama de caja



Medidas de asociación

Al estudiar la relación entre dos variables cuantitativas en general interesa:

- Investigar si existe asociación entre las dos variables.
- Cuantificar la fuerza de la asociación, a través de una medida de asociación denominada coeficiente de correlación.
- Estudiar la forma de la relación y en lo posible proponer un modelo matemático para la relación.
- Predecir una variable a partir de la otra usando el modelo propuesto (REGRESIÓN)

Medidas de asociación: covarianza y correlación

Covarianza

La relación entre dos variables también la podemos expresar de forma numérica. La covarianza es una medida de la asociación o relación lineal entre dos variables que resume la información existente en un diagrama de dispersión.

$$Cov(x,y) = \frac{\sum(x_i - \bar{X})(y_i - \bar{Y})}{n-1}$$

Medidas de asociación

- **Coeficiente de Correlación Lineal**

Una medida de la relación entre dos variables que no depende de las unidades de medida y a su vez, indique la fuerza de dicha relación es el coeficiente de correlación lineal de Pearson.

$$r_{XY} = \frac{Cov(x, y)}{s_x s_y}$$

Propiedades:

- $-1 \leq r \leq 1$
- $r > 0$ nos dice que existe una relación lineal positiva entre x e y
- $r < 0$ nos dice que existe una relación lineal negativa entre x e y
- $r = 0$ nos dice que no existe una relación lineal entre x e y,

Medidas de asociación - Coeficiente de Correlación Lineal