Estadística descriptiva

Jessica Nathaly Pulzara Mora
jessica.pulzara@udea.edu.co

Departamento de ingeniería de sistemas

UNIVERSIDAD
DE ANTIOQUIA

Medidas de Tendencia Central
Datos agrupados

Intervalo de Clase	m_i	f_i	F_i	f_{Ri}	F_{Ri}	$f_{Ri}\%$	$F_{Ri}\%$

Media

Se calcula con esta fórmula:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{k} m_i f_i,$$

donde,

- k: número de clases (intervalos).
- m_i: marca de clase del intervalo i
- f_i: frecuencia absoluta del intervalo i

Retomemos el ejemplo del préstamo de bicicletas:

Intervalo de Clase	m_i	f_i	F_i	f_{Ri}	F_{Ri}	$f_{Ri}\%$	$F_{Ri}\%$
$[12, 22)$	17	6	6	0,11	0,11	11%	11%
$[22, 32)$	27	22	28	0,38	0,49	38%	49%
$[32, 42)$	37	6	34	0,11	0,60	11%	60%
$[42, 52)$	47	16	50	0,28	0,88	28%	88%
$[52, 62)$	57	1	51	0,02	0,90	2%	90%
$[62, 72)$	67	4	55	0,07	0,97	7%	97%
$[72, 82)$	77	2	57	0,03	1	3%	100%

Moda

Para distribuciones agrupadas, se define como la marca de clase de la clase con mayor frecuencia absoluta. En el ejemplo anterior se tiene:

Intervalo de Clase	m_i	f_i	F_i	f_{Ri}	F_{Ri}	$f_{Ri}\%$	$F_{Ri}\%$
$[12, 22)$	17	6	6	0,11	0,11	11%	11%
$[22, 32)$	27	22	28	0,38	0,49	38%	49%
$[32, 42)$	37	6	34	0,11	0,60	11%	60%
$[42, 52)$	47	16	50	0,28	0,88	28%	88%
$[52, 62)$	57	1	51	0,02	0,90	2%	90%
$[62, 72)$	67	4	55	0,07	0,97	7%	97%
$[72, 82)$	77	2	57	0,03	1	3%	100%

Medidas de Posición
Datos agrupados

Percentil

Para calcular el percentil i, se requiere la columna de frecuencias relativas y se utiliza la siguiente fórmula:

$$P_i = L_{inf} + \frac{\left(\frac{i \cdot n}{100} - a\right) \cdot A}{f_i}$$

donde,

- i: Número del percentil deseado.
- A: Amplitud de los intervalos.
- f_i: frecuencia absoluta del intervalo que contiene al percentil.
- a: frecuencia acumulada del intervalo anterior al del percentil.

Para identificar la clase (o intervalo) del percentil se identifica la primera clase que tiene una frecuencia relativa acumulada igual o superior a $\frac{i}{100}$.

Ejemplo

Usando los datos correspondientes al número de bicicletas prestadas durante varias semanas por una estación del sistema de bicicletas públicas calcule el P_{50}.

Intervalo de Clase	m_i	f_i	F_i	f_{Ri}	F_{Ri}	$f_{Ri}\%$	$F_{Ri}\%$
$[12, 22)$	17	6	6	0,11	0,11	11%	11%
$[22, 32)$	27	22	28	0,38	0,49	38%	49%
$[32, 42)$	37	6	34	0,11	0,60	11%	60%
$[42, 52)$	47	16	50	0,28	0,88	28%	88%
$[52, 62)$	57	1	51	0,02	0,90	2%	90%
$[62, 72)$	67	4	55	0,07	0,97	7%	97%
$[72, 82)$	77	2	57	0,03	1	3%	100%

$$P_i = L_{inf} + \frac{\left(\frac{i \cdot n}{100} - a\right) \cdot A}{f_i}$$

Medidas de Dispersión (o variabilidad)
Datos agrupados

Varianza

Se calcula con esta fórmula:

$$S^2 = \frac{1}{n} \sum_{i=1}^{k} f_i (m_i - \bar{X})^2,$$

donde,

- k: número de clases (intervalos).
- m_i: marca de clase del intervalo i.
- f_i: frecuencia absoluta del intervalo i.

Para calcular, puede crear columnas adicionales en una tabla de frecuencia:

Intervalo de Clase	m_i	f_i	F_i	f_{Ri}	F_{Ri}	Diferencia	Producto
...	...	...	...	...	...	...	...

Entonces,

- Diferencia corresponde a $(m_i - \bar{X})^2$
- Producto corresponde a $f_i(m_i - \bar{X})^2$
- Al final, basta con sumar las filas de la columna "Producto", y dividir el resultado por n.

Ejemplo

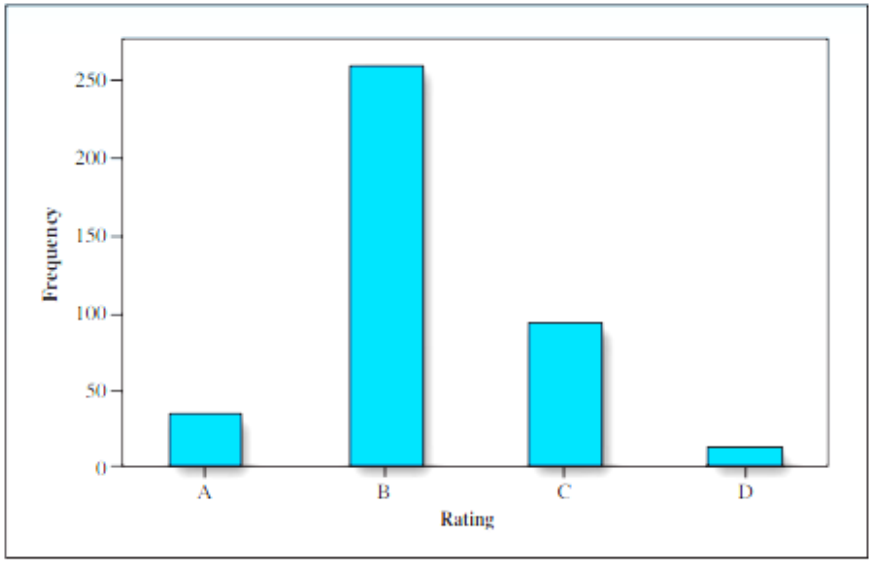
Intervalo de Clase	m_i	f_i	F_i	f_{Ri}	F_{Ri}	Diferencia	Producto
$[12, 22)$	17	6	6	0,11	0,11		
$[22, 32)$	27	22	28	0,38	0,49		
$[32, 42)$	37	6	34	0,11	0,60		
$[42, 52)$	47	16	50	0,28	0,88		
$[52, 62)$	57	1	51	0,02	0,90		
$[62, 72)$	67	4	55	0,07	0,97		
$[72, 82)$	77	2	57	0,03	1		
						SUMATORIA	

$$S^2 = \frac{\text{SUMATORIA}}{n} =$$

Técnicas gráficas para explorar datos

Variable Cualitativa

Diagrama de barras

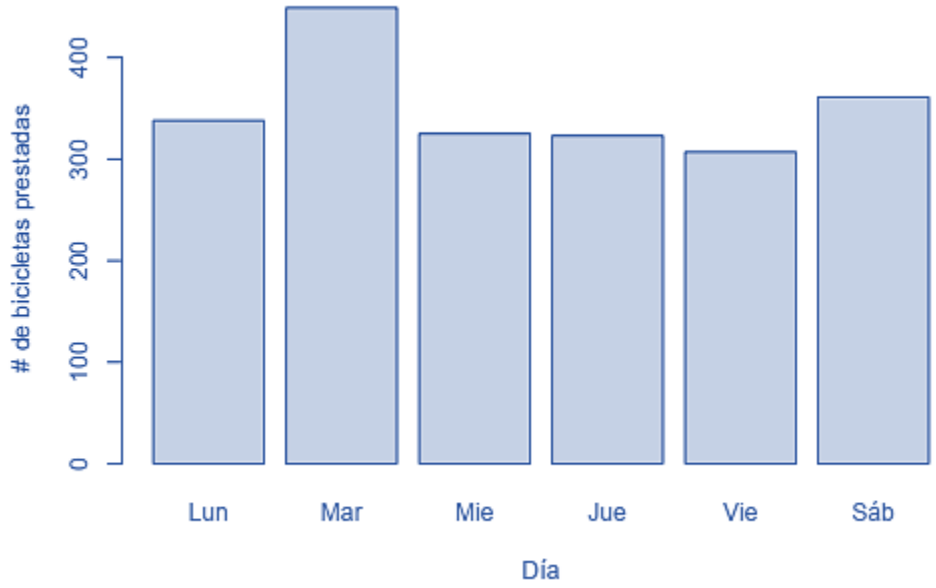


- Eje x: Variable categórica.
- Eje y: Frecuencia (absoluta o relativa), conteo.
- Los ejes son inercambiables.

Ejemplo

Por ejemplo, consideremos la cantidad de bicicletas que se prestaron cada día:

Lun	Mar	Mie	Jue	Vie	Sáb
338	449	325	323	307	361



Barras de dos variables

- Permite graficar una variable numérica con dos categóricas.
- Alternativamente, permite graficar tablas de contingencia de dos variables.
- Eje x: Variable categórica.
- Eje y: Variable numérica (o Frecuencia).
- Colores: otra variable categórica.

Número de Bicicletas prestadas por día:

	Sem.1	Sem.2	Sem.3	Sem.4	Sem.5	Sem.6	Sem.7	Sem.8
Lun	68	63	42	27		30	36	28
Mar	65	43	25	74	38	51	36	42
Mie	12	32	49	38	21	42	27	31
Jue	22	43	27	49	16	28	23	19
Vie	79	27	22	23		24	25	44
Sáb	31	28	25	45		12	57	51

Número de Bicicletas prestadas por día:

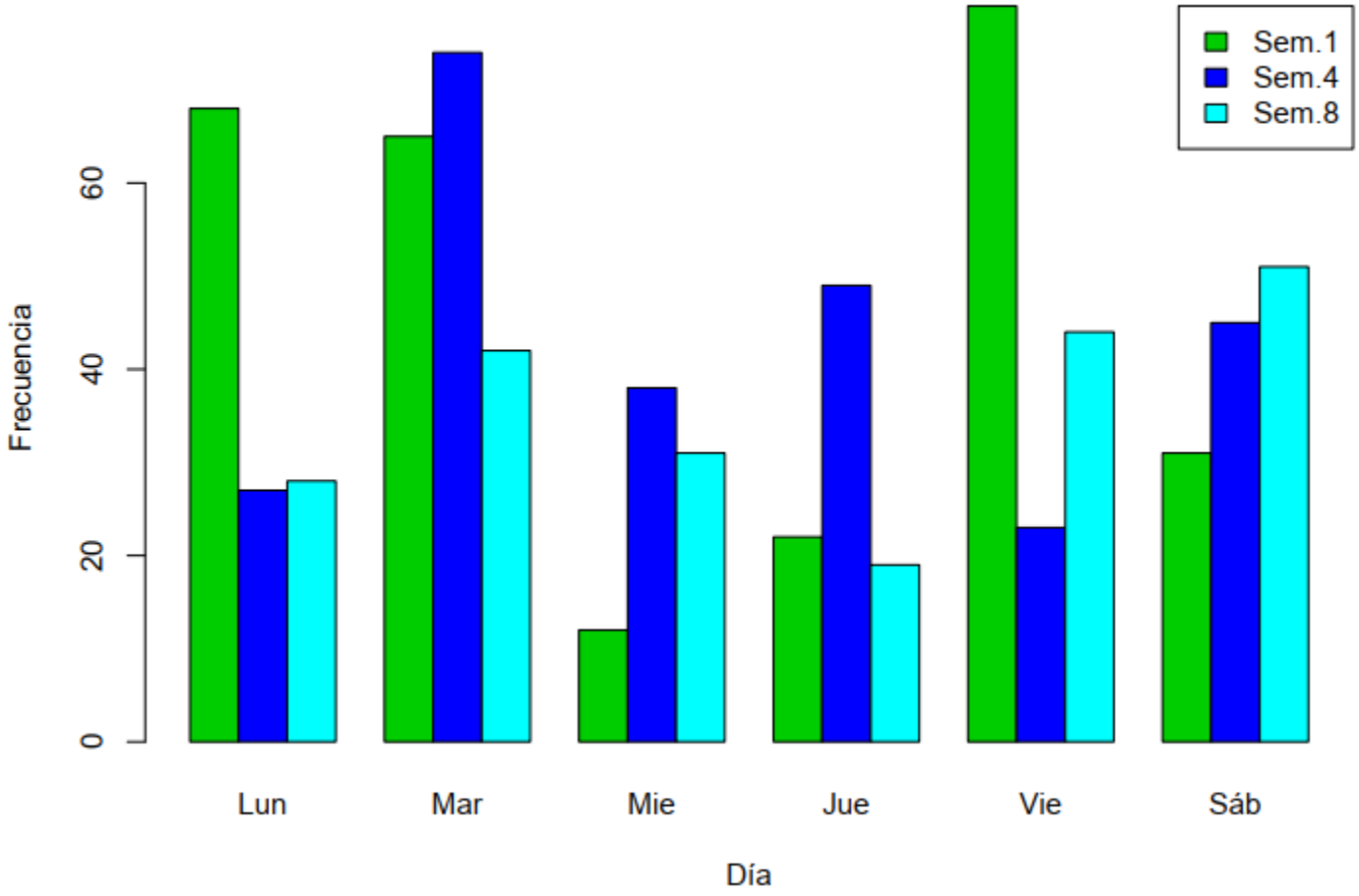


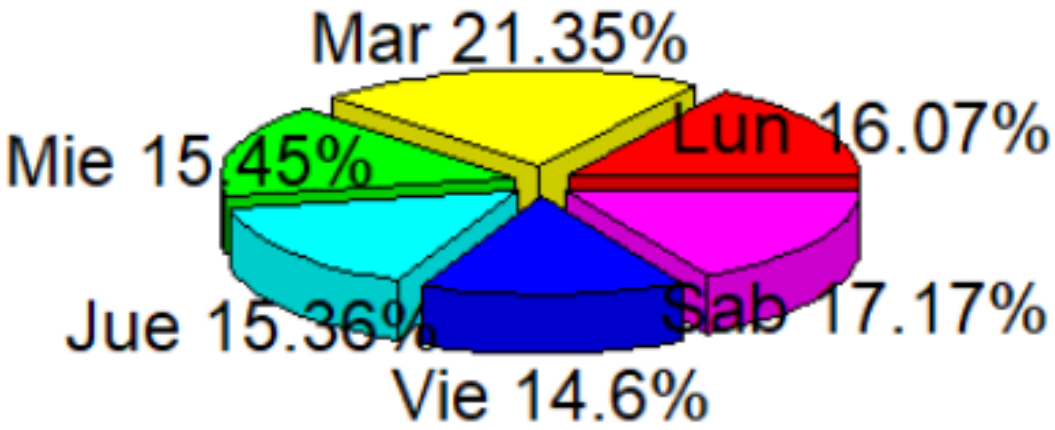
Diagrama circular

- Es una representación circular.
- También se conoce como *Torta*.
- Las tajadas representan la frecuencia (absoluta o relativa).
- Los colores representan la modalidad (valores) de la variable categórica.

Ejemplo

Por ejemplo, consideremos la cantidad de bicicletas que se prestaron cada día:

Lun	Mar	Mie	Jue	Vie	Sáb
338	449	325	323	307	361



Variable Cuantitativa

Diagrama de puntos

Un diagrama de puntos muestra datos y sus frecuencias a lo largo de una recta numérica.

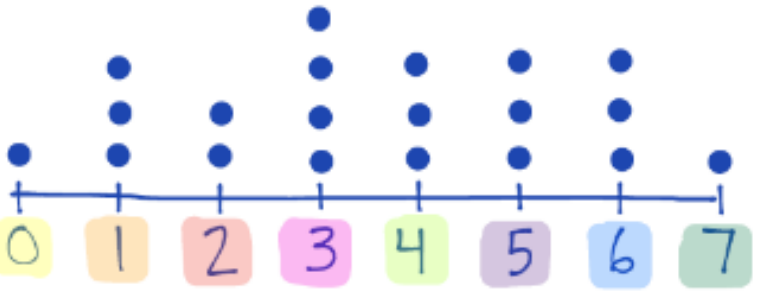
- Recomendados para muestras pequeñas, la gráfica permite ver rápidamente la tendencia y variabilidad de los datos.
- Cada punto representa un dato. Su valor se lee en el eje x.

Ejemplo

conjunto de datos:

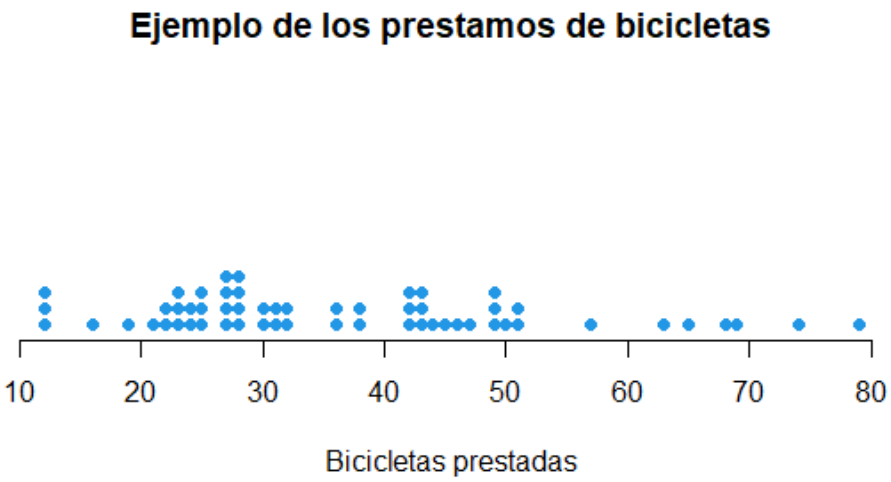
6, 3, 6, 3, 5,
7, 4, 6, 5, 3
4, 4, 5, 1, 0,
3, 2, 2, 1, 1

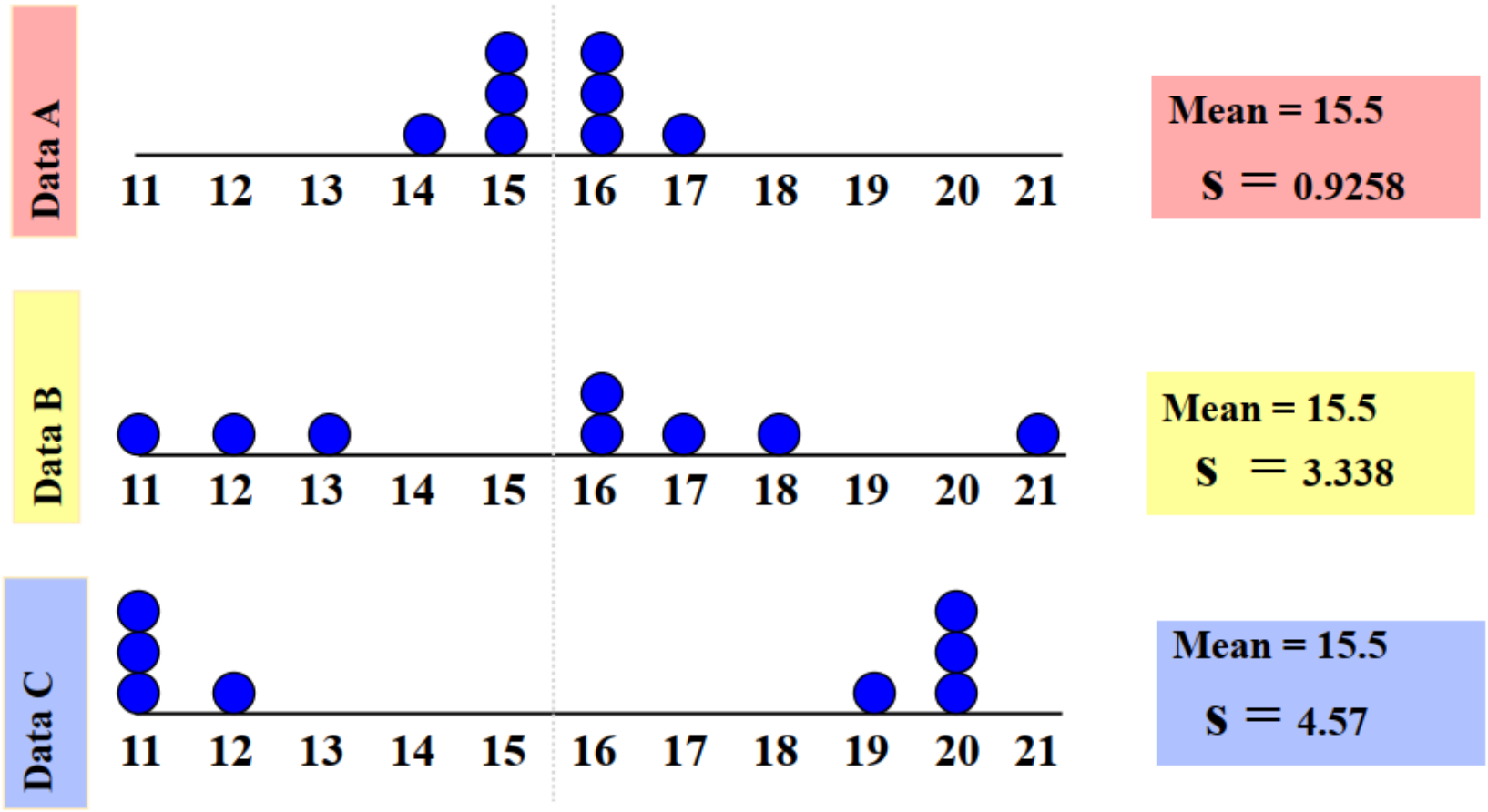
número de goles	frecuencia
0	1
1	3
2	2
3	4
4	3
5	3
6	3
7	1



Ejemplo

Lun.	Mar.	Mier.	Jue.	Vier.	Sáb.
68	65	12	22	79	31
63	43	32	43	27	28
42	25	49	27	22	25
27	74	38	49	23	45
30	51	42	28	24	12
36	36	27	23	25	57
28	42	31	19	44	51
32	28	50	46	30	43
12	38	21	16	24	69
	47	23	49		

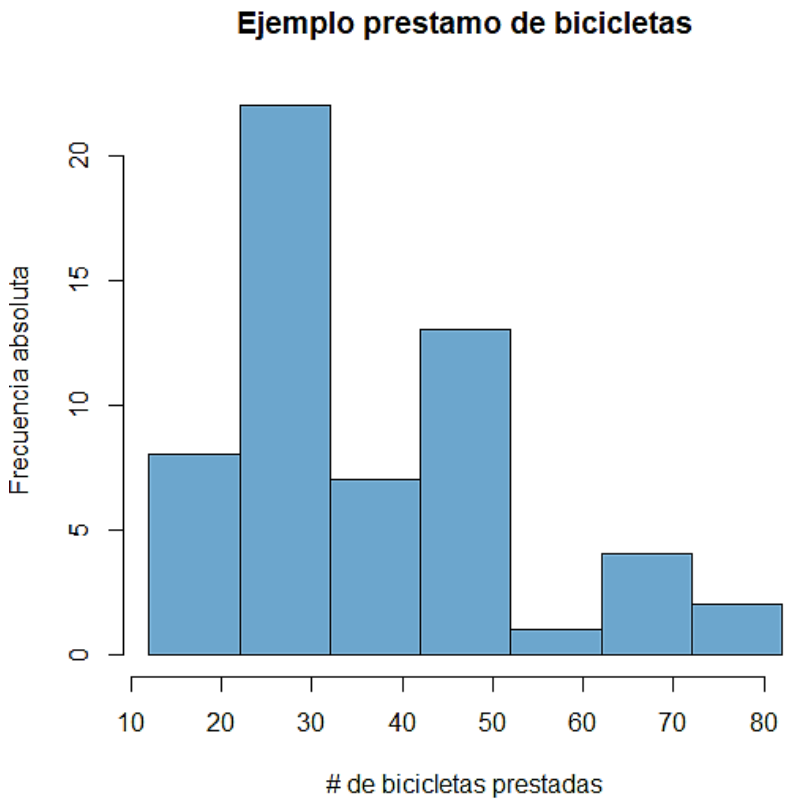



Data A
11 12 13 14 15 16 17 18 19 20 21
Mean = 15.5
S = 0.9258
Data B
11 12 13 14 15 16 17 18 19 20 21
Mean = 15.5
S = 3.338
Data C
11 12 13 14 15 16 17 18 19 20 21
Mean = 15.5
S = 4.57

Histograma

- Representación gráfica de las tablas de frecuencia.
- Se recomienda para muestras grandes.
- Es un gráfico formado por barras.
- Eje x: Variable numérica.
- Eje y: Frecuencia (absoluta o relativa)

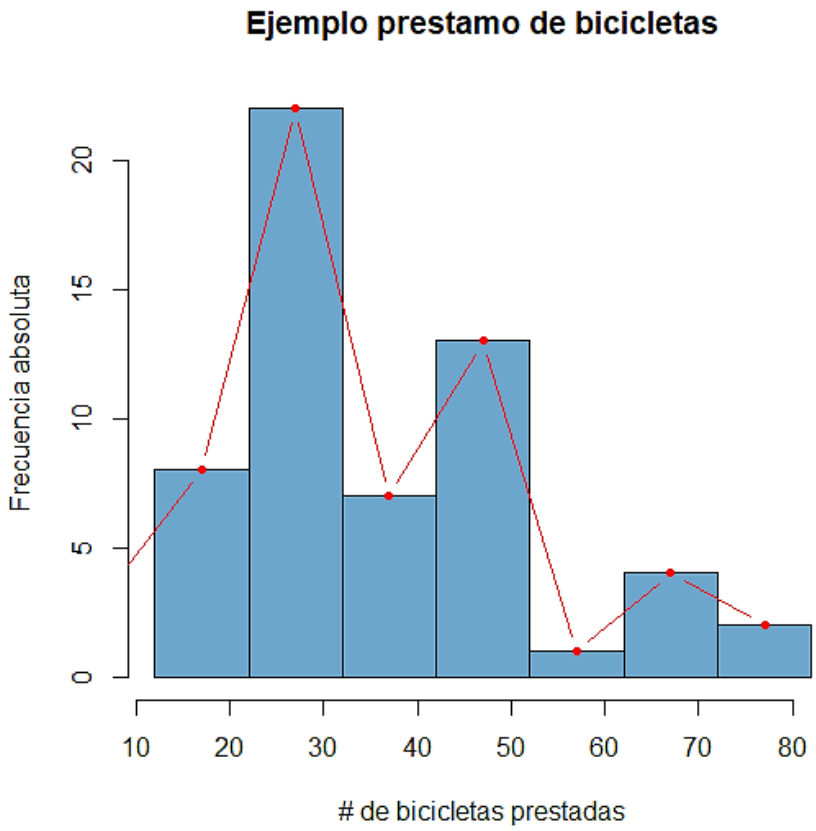
Ejemplo

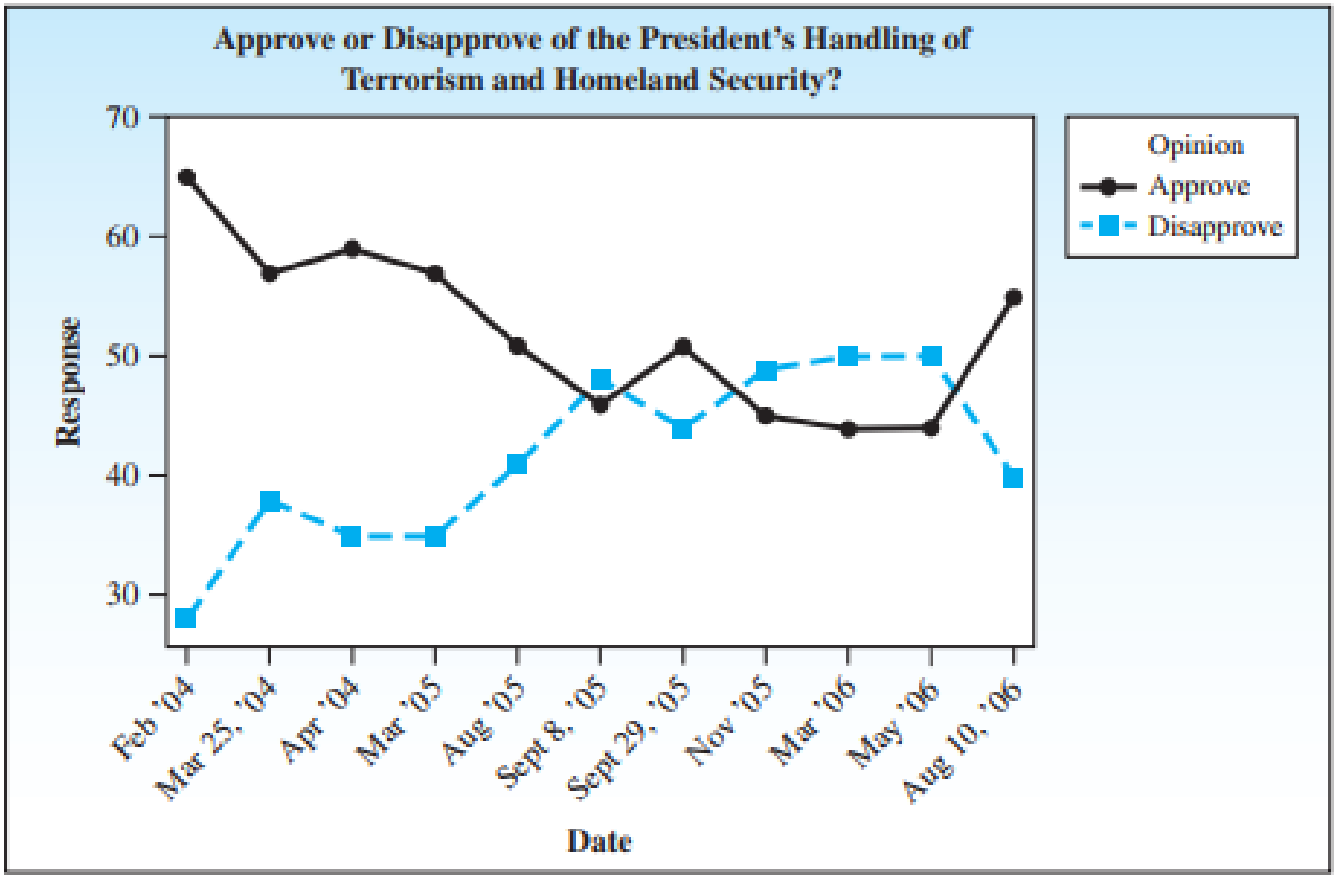


Intervalo de Clase	m_i	f_i	F_i	f_{Ri}	F_{Ri}	$f_{Ri}\%$	$F_{Ri}\%$
$[12,22)$	17	6	6	0,11	0,11	11%	11%
$[22,32)$	27	22	28	0,38	0,49	38%	49%
$[32,42)$	37	6	34	0,11	0,60	11%	60%
$[42,52)$	47	16	50	0,28	0,88	28%	88%
$[52,62)$	57	1	51	0,02	0,90	2%	90%
$[62,72)$	67	4	55	0,07	0,97	7%	97%
$[72,82)$	77	2	57	0,03	1	3%	100%

Polígono de frecuencia

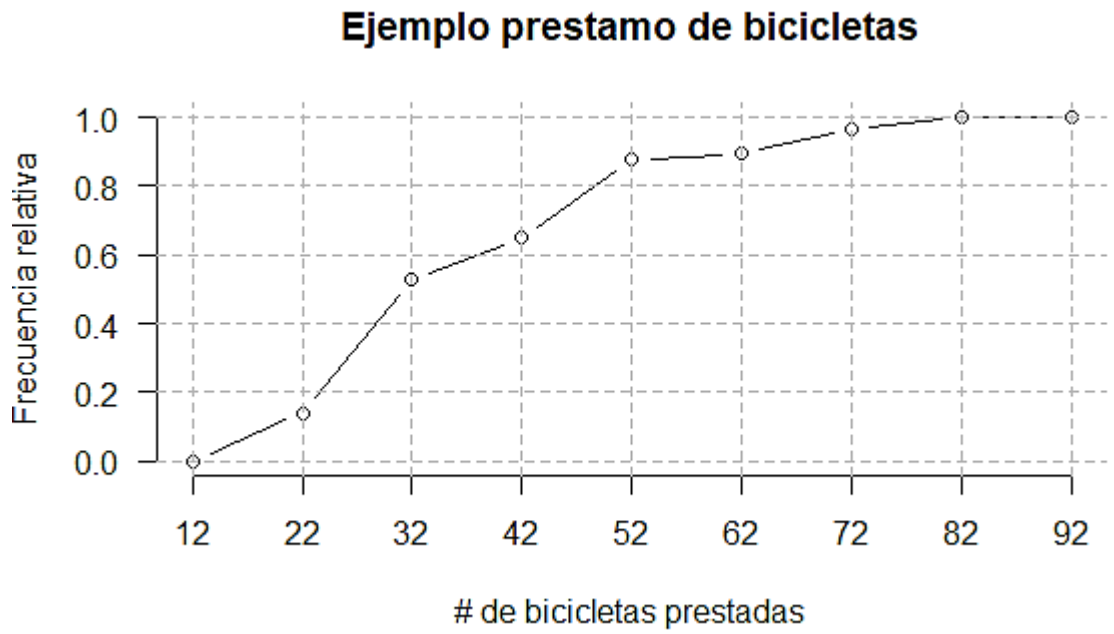
- Se forman puntos con los valores de la marca de clase y la frecuencia.
- Estos puntos se unen con segmentos de recta.




Approve or Disapprove of the President's Handling of Terrorism and Homeland Security?
Response
70
60
50
40
30
Opinion
Approve
Disapprove
Feb '04
Mar 25, '04
Apr '04
Mar '05
Aug '05
Sept 8, '05
Sept 29, '05
Nov '05
Mar '06
May '06
Aug 10, '06
Date

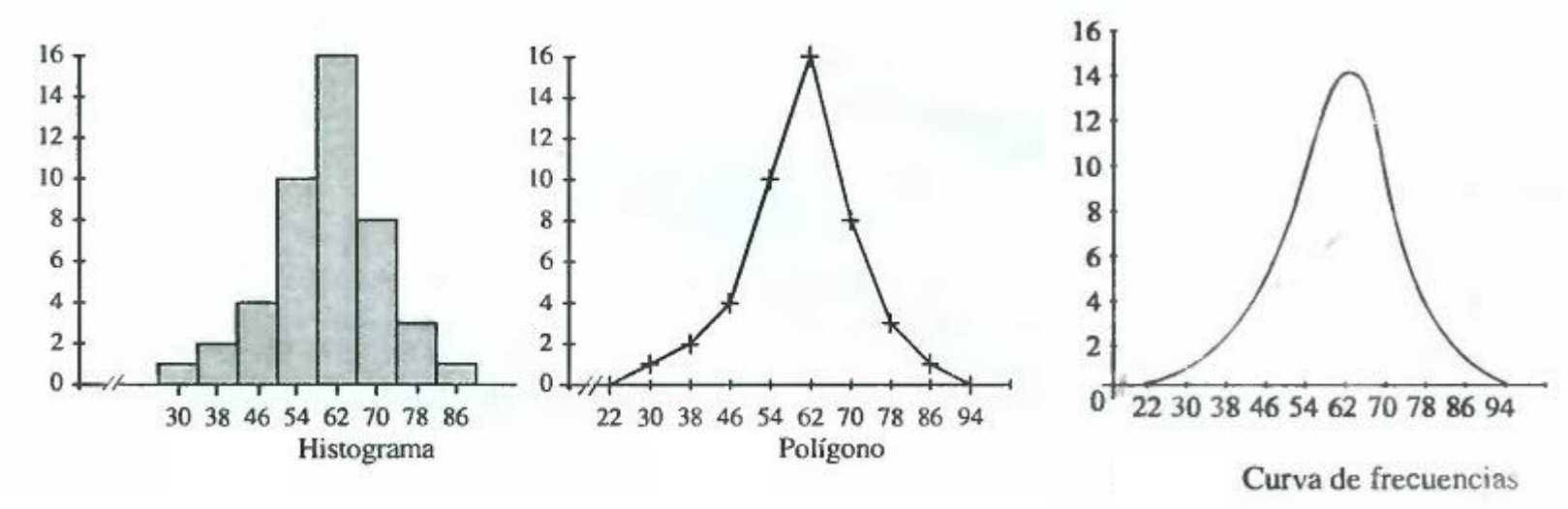
Ojiva

Es un polígono de frecuencias de la frecuencia acumulada (absoluta o relativa).



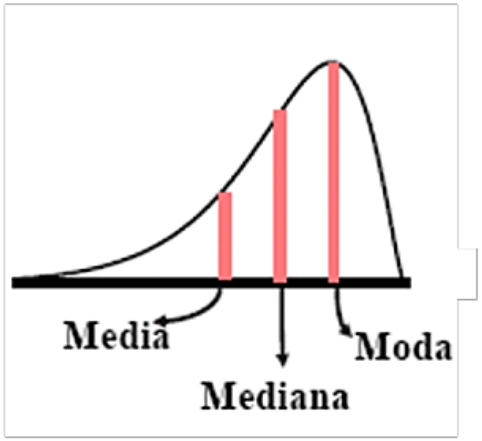
Curva de frecuencias

Una curva de frecuencias se obtiene del polígono de frecuencias "suavizando" sus puntos angulosos.
También llamada modelo de la población, y describe las características de la distribución de la población como: simetría, asimetría, tipos como: normal, bimodal, uniforme, etc..

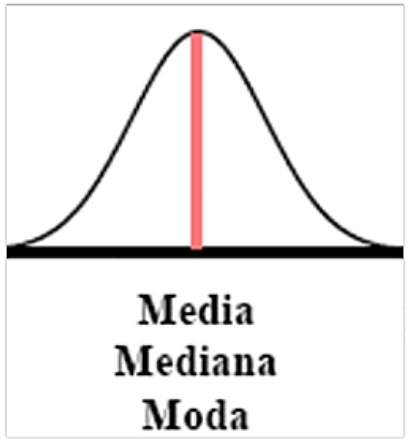


Simetría y sesgo

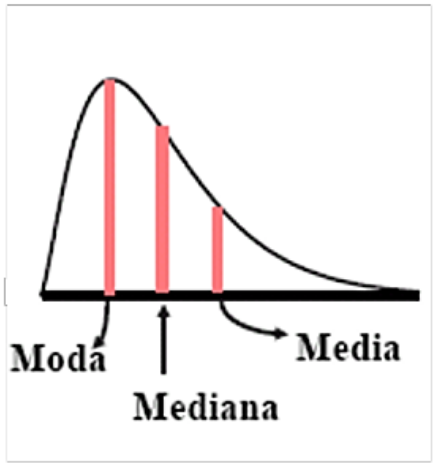
Comparación Media, Moda, Mediana



Sesgo a la Izquierda
o
Asimetría hacia la izquierda



Simétrica
(No sesgada)



Sesgo a la derecha
o
Asimetría hacia la derecha

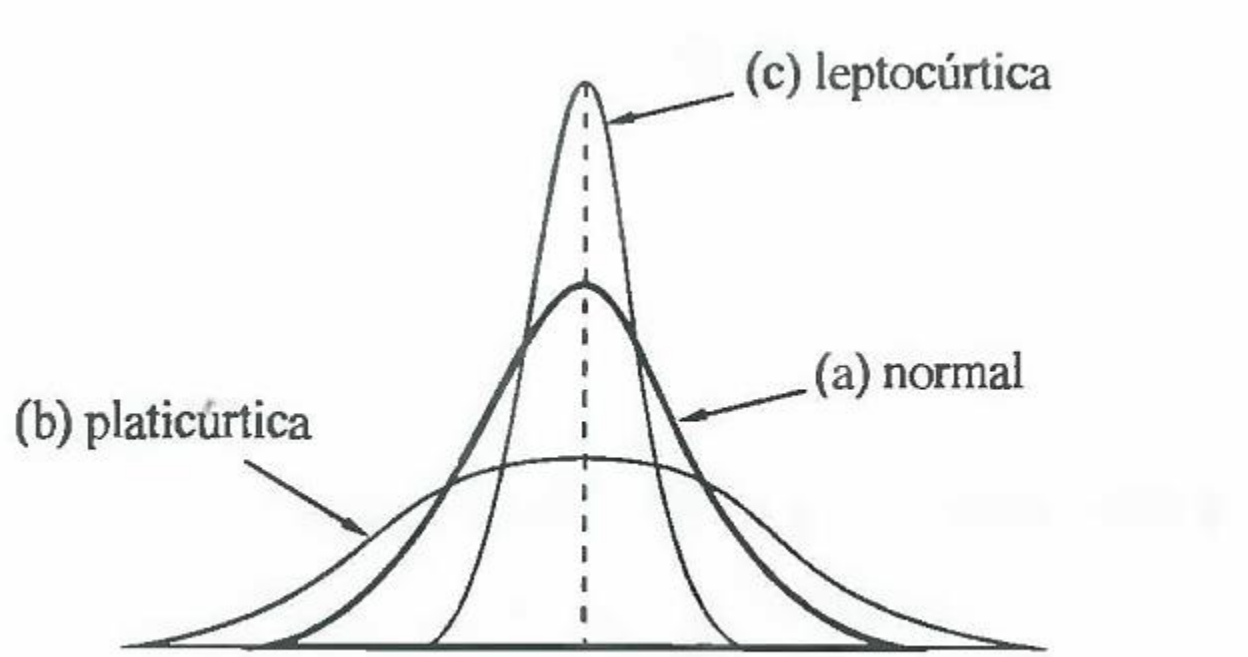
Índice de asimetría de Pearson

Es la medida de asimetría más utilizada, ya que no presenta ninguna condición previa y se aplica a cualquier tipo de distribución.

$$As = \frac{n\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{S}\right)^3}{(n-1)(n-2)}$$

- Si la distribución es simétrica $As = 0$.
- Si $As > 0$, es asimétrica positiva.
- Si $As < 0$, es asimétrica negativa.

Kurtosis



Curvas simétricas: (a) normal, (b) platicúrtica, (c) leptocúrtica

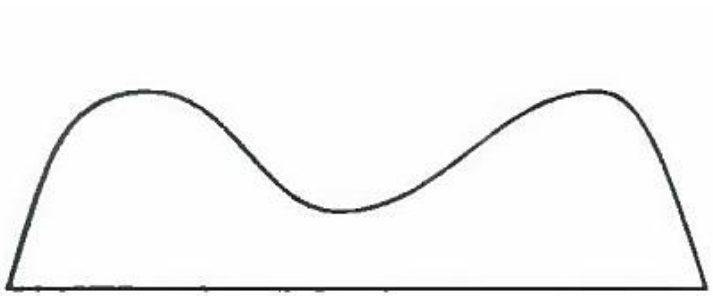
Coeficiente de curtosis

La curtosis se mide en comparación a la curva simétrica normal o mesocúrtica. Se puede calcular como:

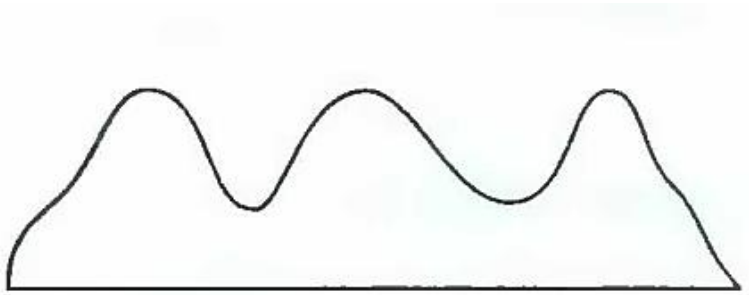
$$k = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{nS^4}$$

- Si la distribución es normal $K = 3$.
- Si $K > 3$, es leptocúrtica.
- Si $K < 3$ es platicúrtica.

Distribución multimodal



(a) Curva bimodal

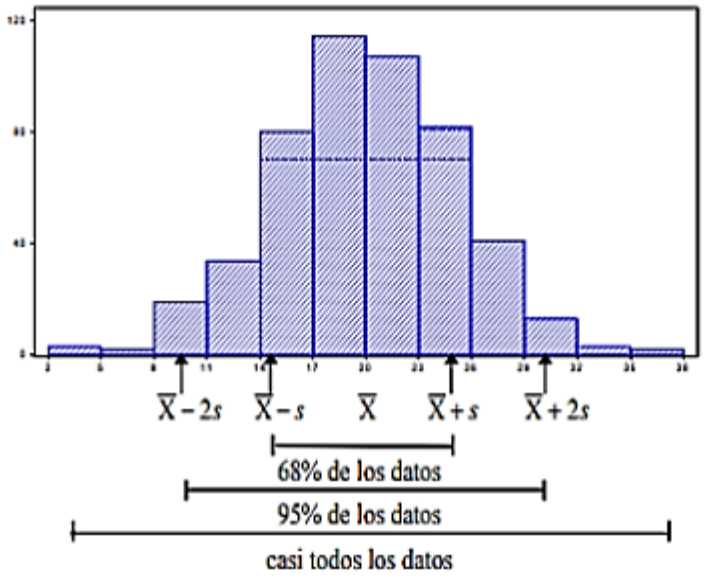


(b) Curva trimodal

Regla empírica

Si el histograma de los datos es aproximadamente simétrico y acampanado entonces:

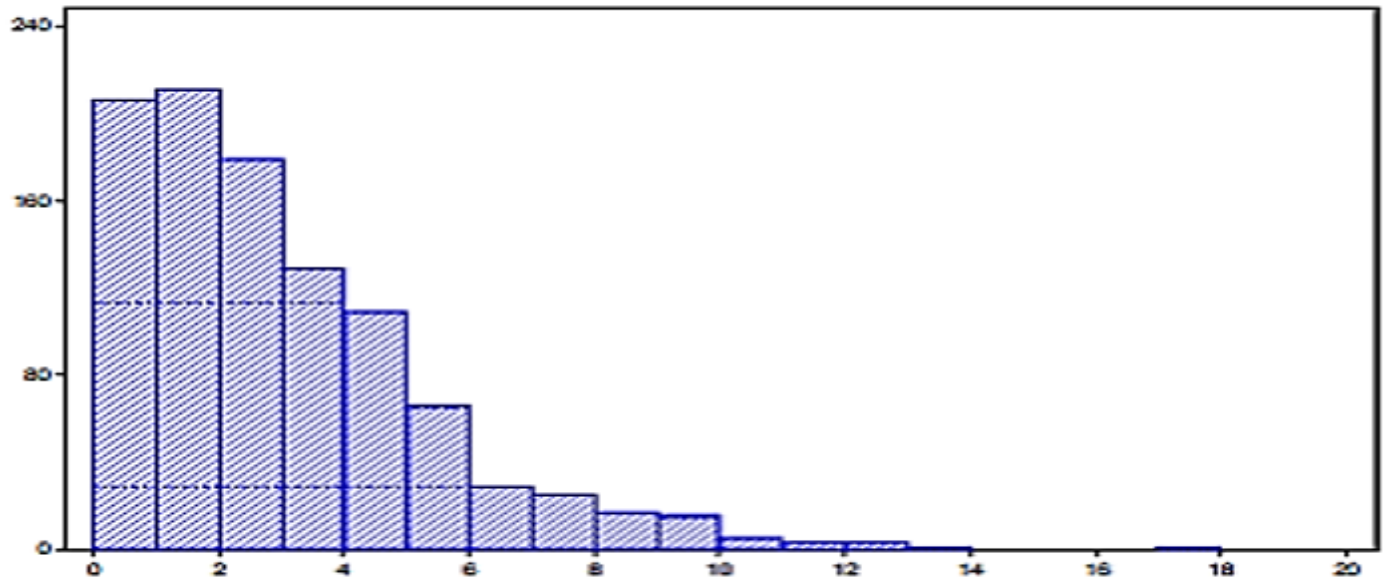
- Aproximadamente el 68 % de las observaciones caen en el intervalo $(\bar{X} - S,\ \bar{X} + S)$
- Aproximadamente el 95 % de las observaciones caen en el intervalo $(\bar{X} - 2S,\ \bar{X} + 2S)$
- Aproximadamente el 99 % de las observaciones caen en el intervalo $(\bar{X} - 3S,\ \bar{X} + 3S)$



Esta regla es válida para distribuciones no necesariamente acampanadas, pero puede ser errónea cuando se aplica a distribuciones fuertemente asimétricas.

Ejemplo

Consideremos la distribución de salario mensual (en millones de pesos) de una muestra empleados. Para estos datos, $\bar{X} = 3$ y $S = 2.45$.



En este caso, al restar $2S$ a la media, caemos fuera de la escala de la variable:

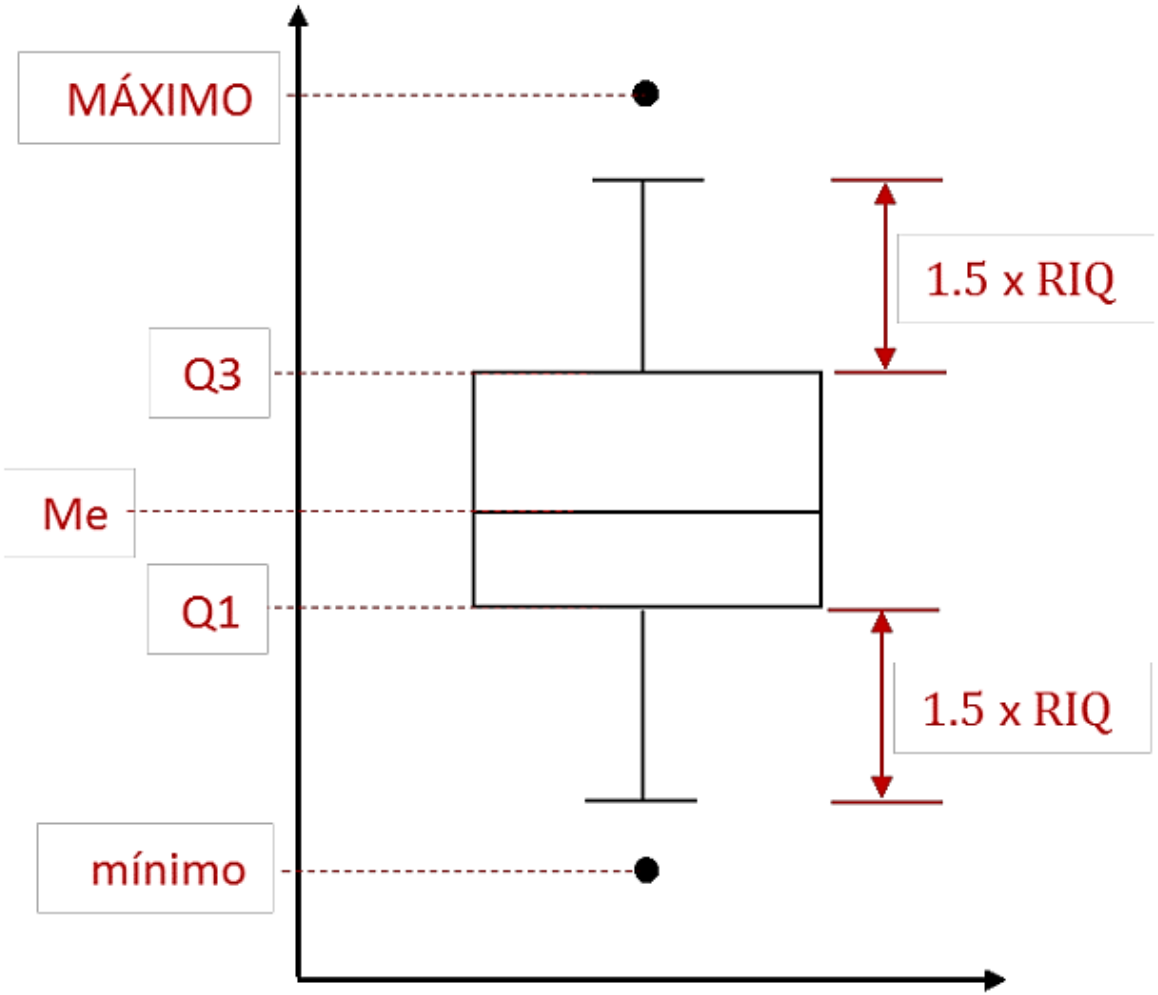
$$\bar{X} - 2S = 3 - (2)(2.45) = -1.9 \qquad (\textit{salarionegativo}?)$$

La interpretación que propusimos a través de la regla empírica resulta no ser apropiada.

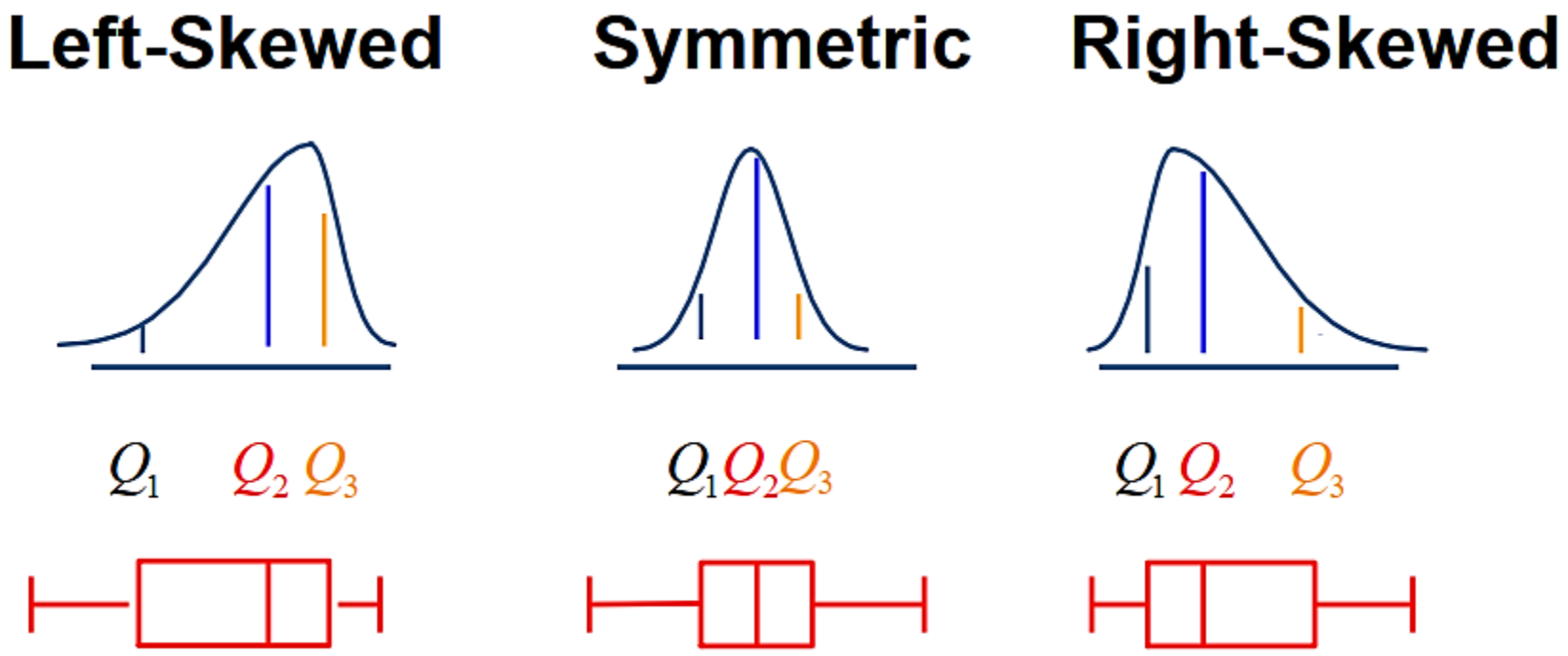
Boxplot

- Ilustra los cuartiles.
- Permite comparar una variable cuantitativa agrupada en un cualitativa.
- Eje x: Variable categórica (caso de 2 variables).
- Eje y: Variable numérica (variable de respuesta).
- Los ejes son intercambiables.

$$Me = \tilde{X}$$



Forma de la distribución y boxplot



The **five-number summary** is used to draw the graph.

- The minimum entry
- Q_1
- Q_2 (median)
- Q_3
- The maximum entry

Example:

Use the data from the 15 quiz scores to draw a box-and-whisker plot.

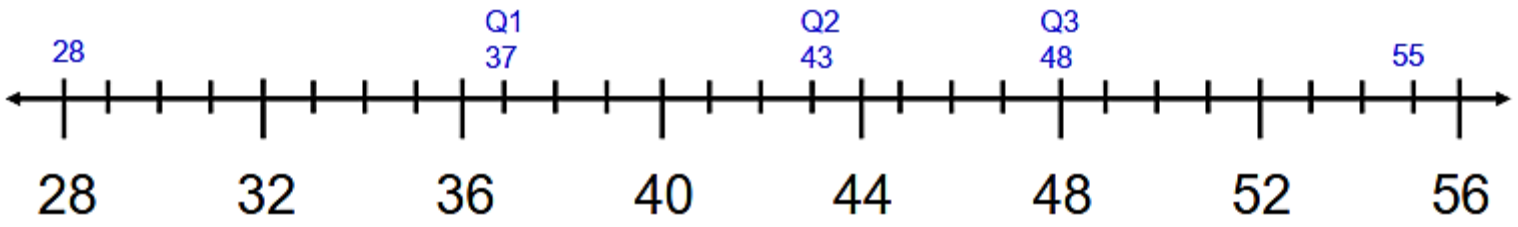
28 30 33 37 37 38 42 43 43 44 45 48 48 51 55

Five-number summary

- The minimum entry 28
- Q_1 37
- Q_2 (median) 43
- Q_3 48
- The maximum entry 55

$IQR = Q_3 - Q_1 = 11$

Max. length = 1.5 IQR = 16.5

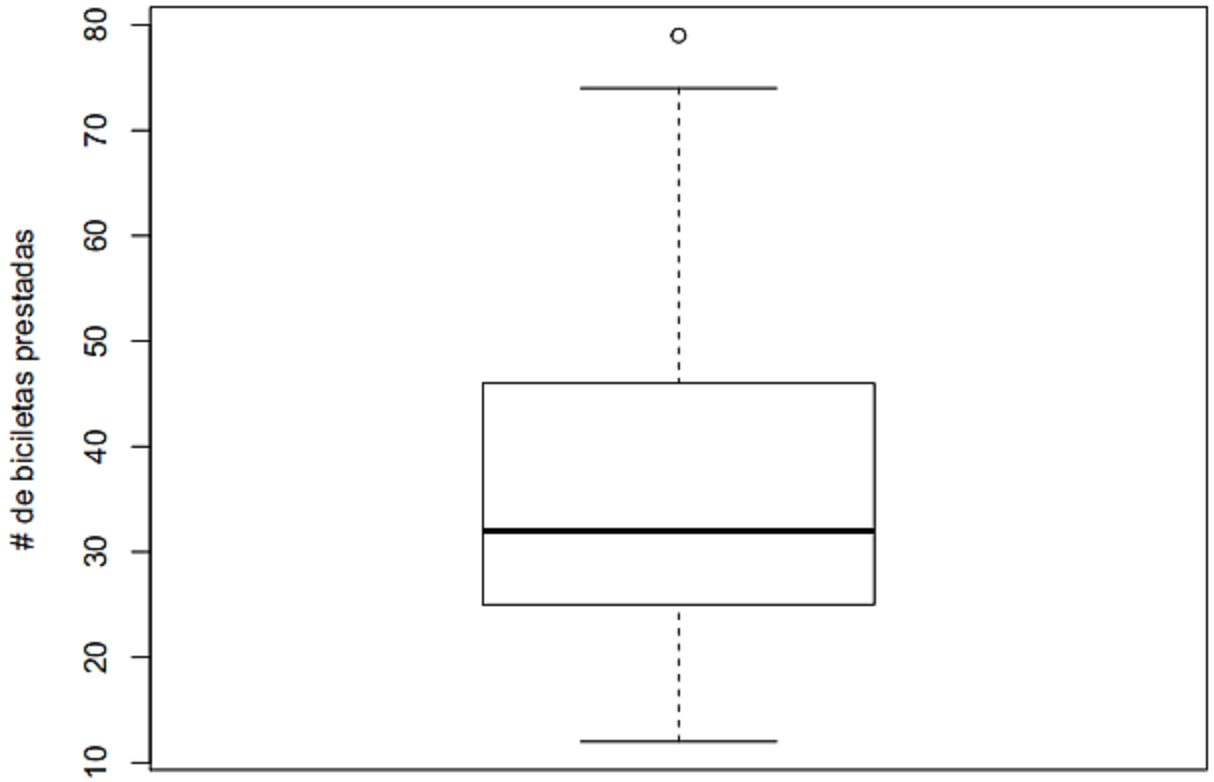


Ejemplo

Lun.	Mar.	Mier.	Jue.	Vier.	Sáb.
68	65	12	22	79	31
63	43	32	43	27	28
42	25	49	27	22	25
27	74	38	49	23	45
30	51	42	28	24	12
36	36	27	23	25	57
28	42	31	19	44	51
32	28	50	46	30	43
12	38	21	16	24	69
	47	23	49		

Ejemplo

Boxplot sin variable de agrupación:



Boxplot con variable de agrupación:

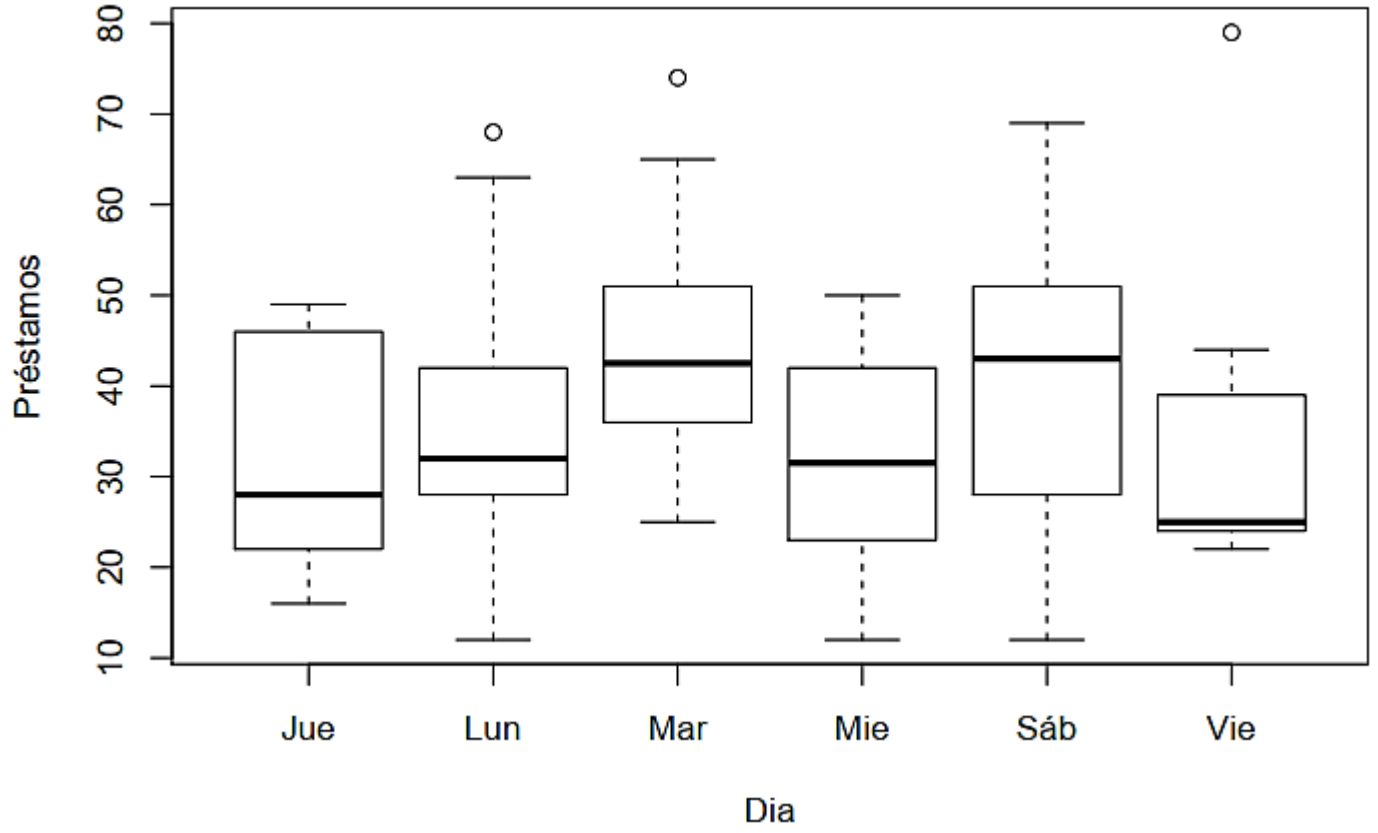


Grafico de medias

- Ilustra el promedio y la desviación estándar o el error estándar:

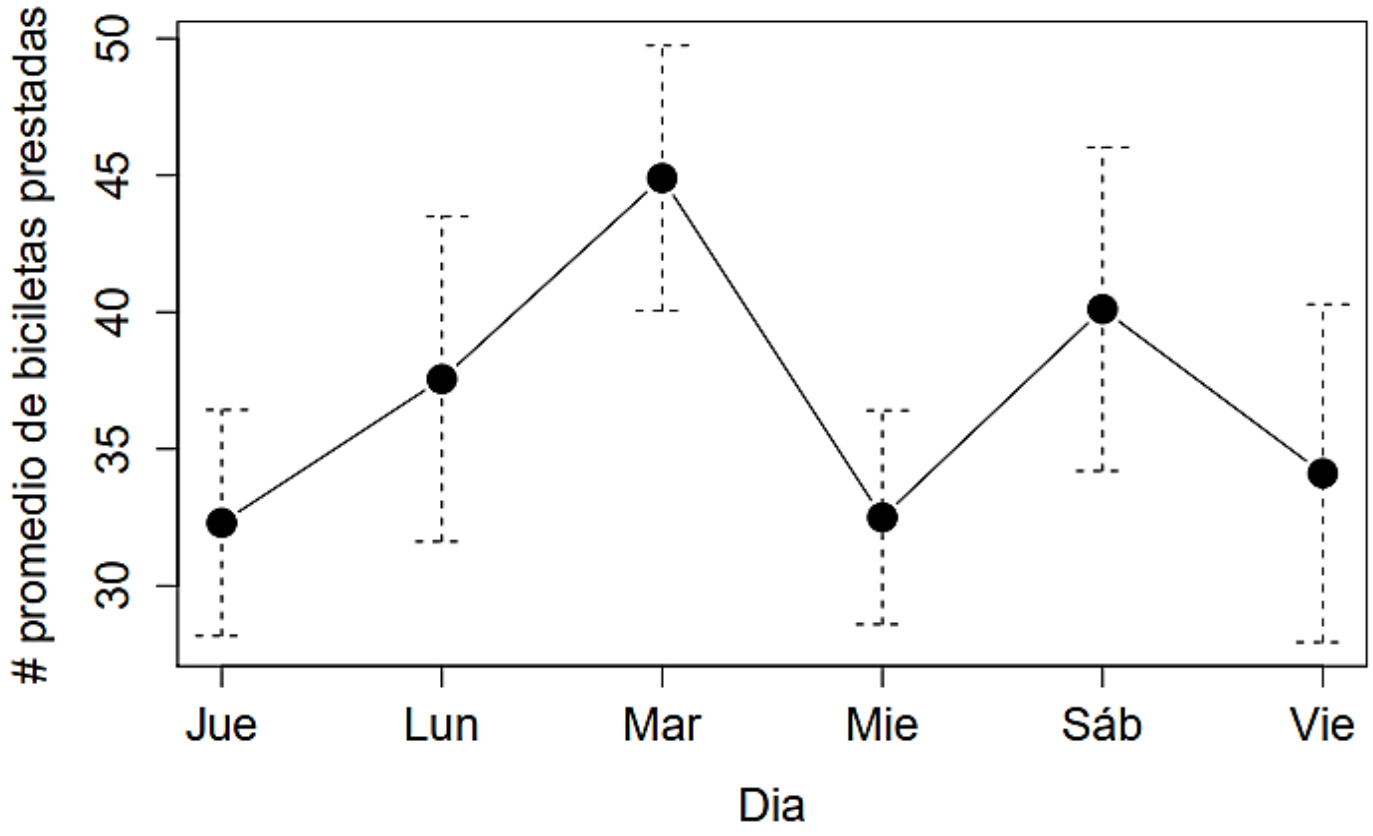
$$S.E = \frac{S}{\sqrt{n}} \qquad \text{(Error Estándar)}$$

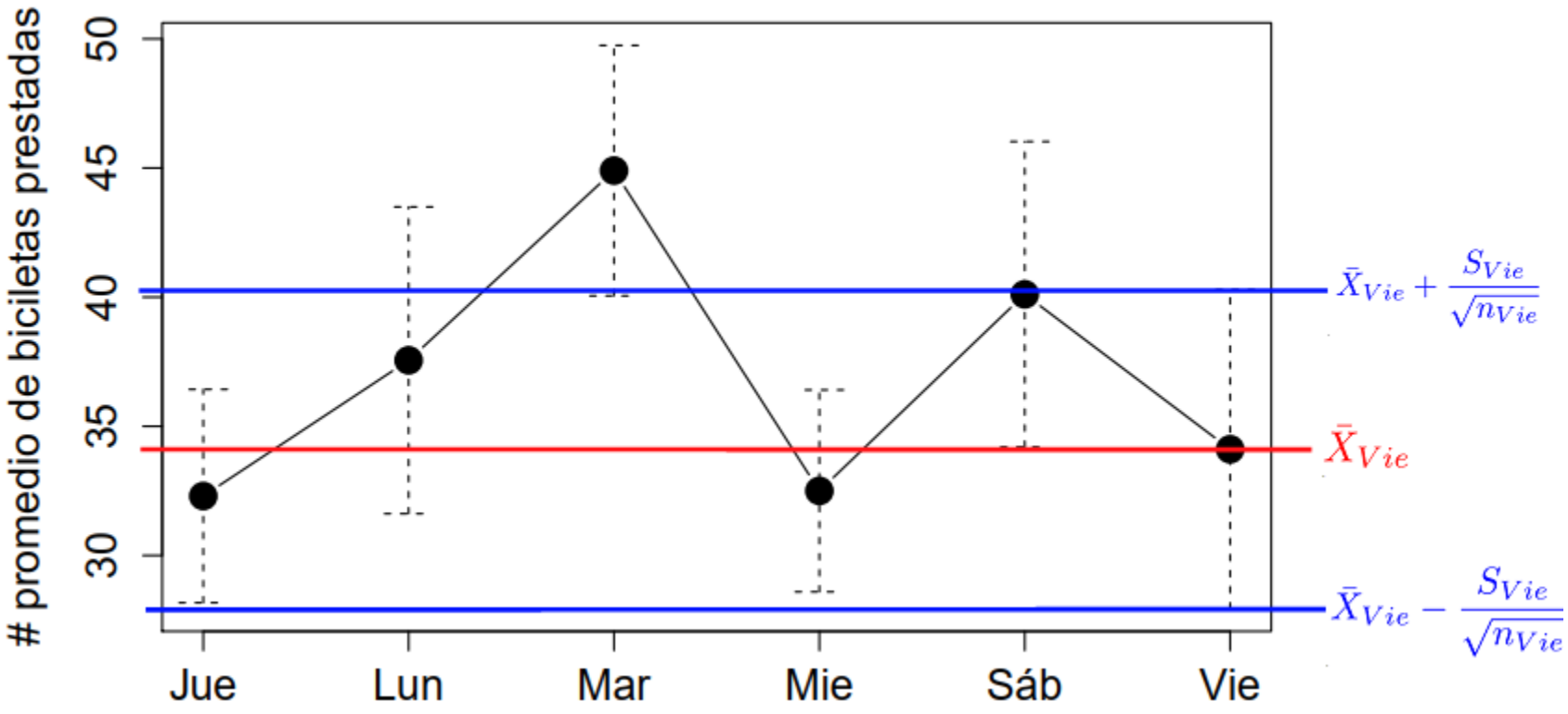
- Permite comparar una variable cuantitativa agrupada en un cualitativa.

- Eje x: Variable categórica (caso de 2 variables).

- Eje y: Variable numérica (variable de respuesta).

de Bicicletas promedio prestadas por día:





Recuerde, el promedio señalado por el punto negro, y el error estándar señalado por las barras de error (las líneas punteadas) se calculan con los datos correspondientes a cada día.