Estadística I

Tema 2: Análisis de datos univariantes

Tema 2: Análisis de datos univariantes

Contenidos

- **Gráficas** para **datos categóricos** (diagrama de barras, diagrama de sectores).
- **Gráficas** para **datos numéricos** (histograma, polígono de frecuencias, diagrama de cajas).
- **Medidas numéricas** para describir:
  - tendencia central (media, mediana, moda)
  - variación (varianza, desviación típica, cuasi-varianza y cuasi-desviación típica, rango, RIC, coeficiente de variación)
  - otros (cuartiles, percentiles)

Capítulo 2: Análisis de datos univariantes

Lecturas recomendadas

- Peña, D., Romo, J., *Introducción a la Estadística para las Ciencias Sociales*.
  - Capítulos 4, 5.
- Newbold, P. *Estadística para los Negocios y la Economía* (2009).
  - Capítulo 2.

Representación gráfica de datos

Una vez obtenida la distribución de frecuencias de los datos, se pueden determinar las siguientes representaciones gráficas:

Categórico ⇓

- diagrama de sectores
- diagrama de barras

Numérico ⇓

- histograma
- polígono de frecuencias
- diagrama de caja

Gráficos para datos cualitativos: diagrama de sectores

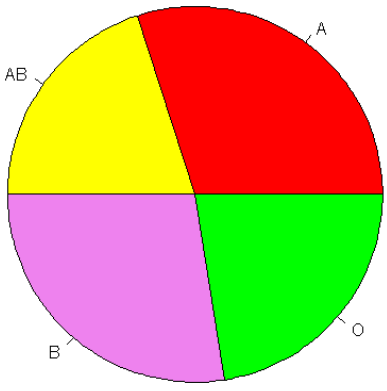
Ejemplo 1: La siguiente tabla de frecuencias corresponde a los datos de grupos sanguíneos obtenidos de una muestra de 40 individuos.

Clase	Frecuencia Absoluta	Frecuencia Relativa
A	12	0.300
B	11	0.275
AB	8	0.200
O	9	0.225
Total	40	1

Diagrama de sectores

Ejemplo 1 cont.:

- Cada sector es una fracción del total del círculo.
- Los sectores están etiquetados con los nombres de las clases.
- Muchos programas ordenan las clases en orden alfabético.
- Aunque es *vistoso*, es más complejo de leer que el diagrama de barras.
- Evitar los diagramas de sectores en 3D, ya que los sectores traseros tienden a parecer menores que los sectores delanteros.



Gráficos para datos cualitativos: diagrama de barras

Ejemplo 2: La tabla inferior muestra diferentes niveles de satisfacción en relación a 901 empleados.

Clase	Frecuencia Absoluta	Frecuencia Relativa	Frecuencia Absoluta Acumulada	Frecuencia Relativa Acumulada
MI	62	0.07	62	0.07
I	108	0.12	170	0.19
S	319	0.35	489	0.54
MS	412	0.46	901	1
Total	901	1		

Diagrama de barras

Ejemplo 2 cont.:

- Las barras tienen la misma amplitud y son equidistantes, con alturas correspondientes a las frecuencias (absolutas).
- Existen huecos entre las barras.
- Las barras están etiquetadas con los nombres de las clases.
- Muchos programas ordenan las clases en orden alfabético.

.

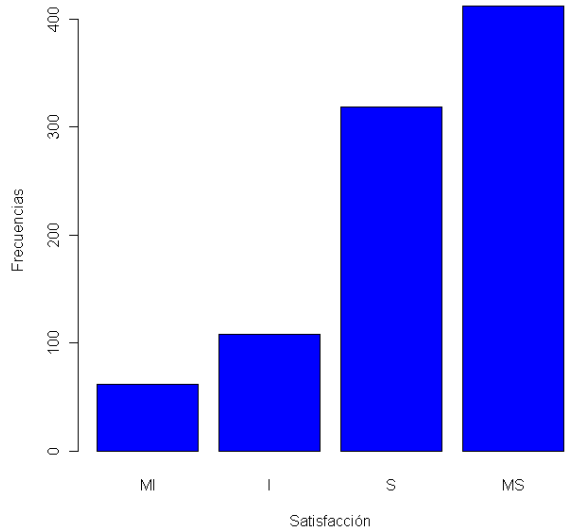


Diagrama de barras

- Los diagramas de barras pueden construirse también para datos discretos si no existen demasiados valores diferentes.
- Este es el diagrama de barras para el **Ejemplo 3** del Tema 1, donde se consideraba el número de hojas infectadas por un hongo en una muestra de 50 plantas.

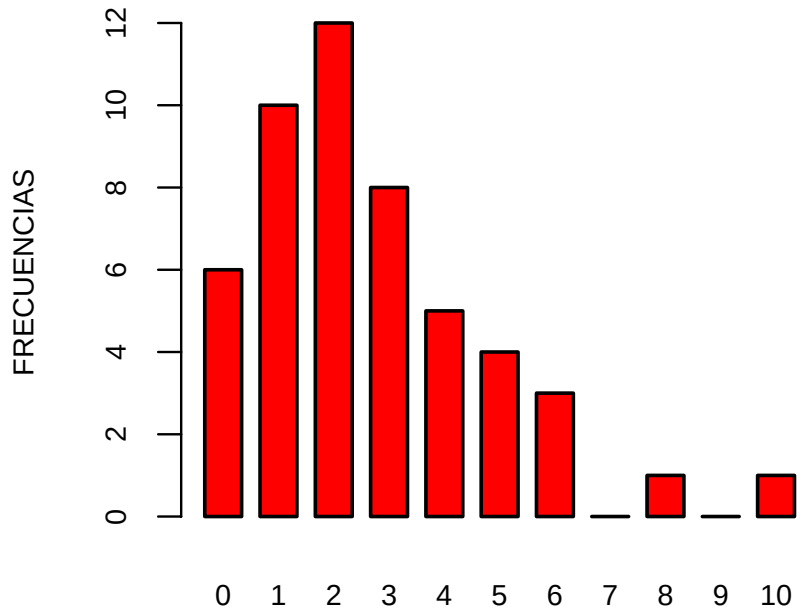
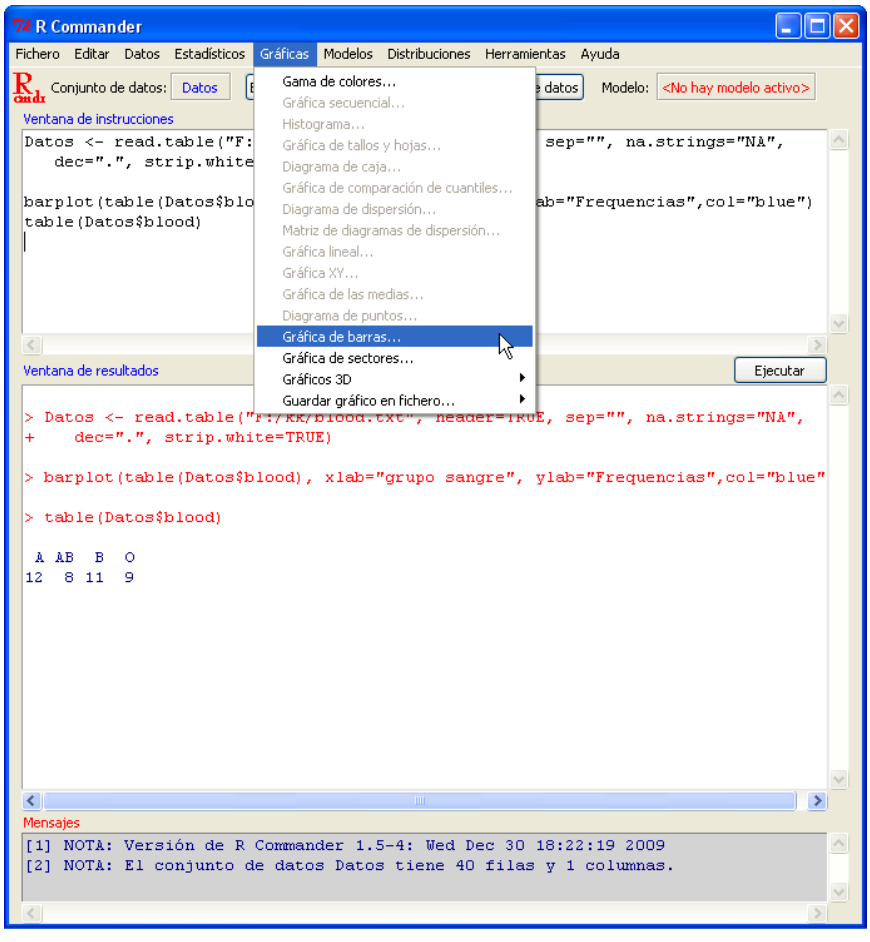


Diagrama de barras en RCommander



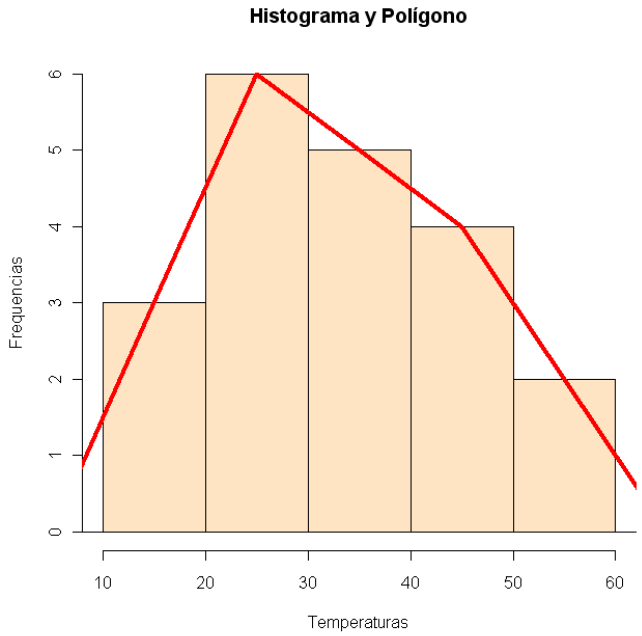
Gráficos para datos cuantitativos: histograma y polígono de frecuencias

Ejemplo 4: La distribución de frecuencias de la temperatura más alta del día (en grados °F) tomada en 20 días de invierno es como sigue:

Intervalo	Marca de clase	n_i	f_i	N_i	F_i
$[10, 20)$	15	3	0.15	3	0.15
$[20, 30)$	25	6	0.30	9	0.45
$[30, 40)$	35	5	0.25	14	0.70
$[40, 50)$	45	4	0.20	18	0.90
$[50, 60)$	15	2	0.10	20	1
Total		20	1		

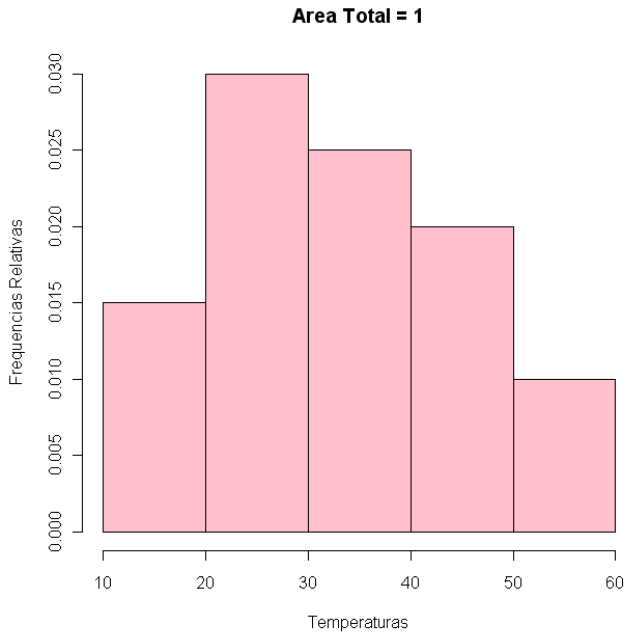
Histograma y polígono de frecuencias

- No hay *huecos* entre las barras/cajas.
- Amplitud de cajas $\equiv$ amplitud de intervalos (idénticos) y los límites de las clases se marcan en el eje horizontal.
- Alturas de cajas $\equiv$ frecuencias (aquí, absoluta).
- Las áreas de cajas son proporcionales a las frecuencias.

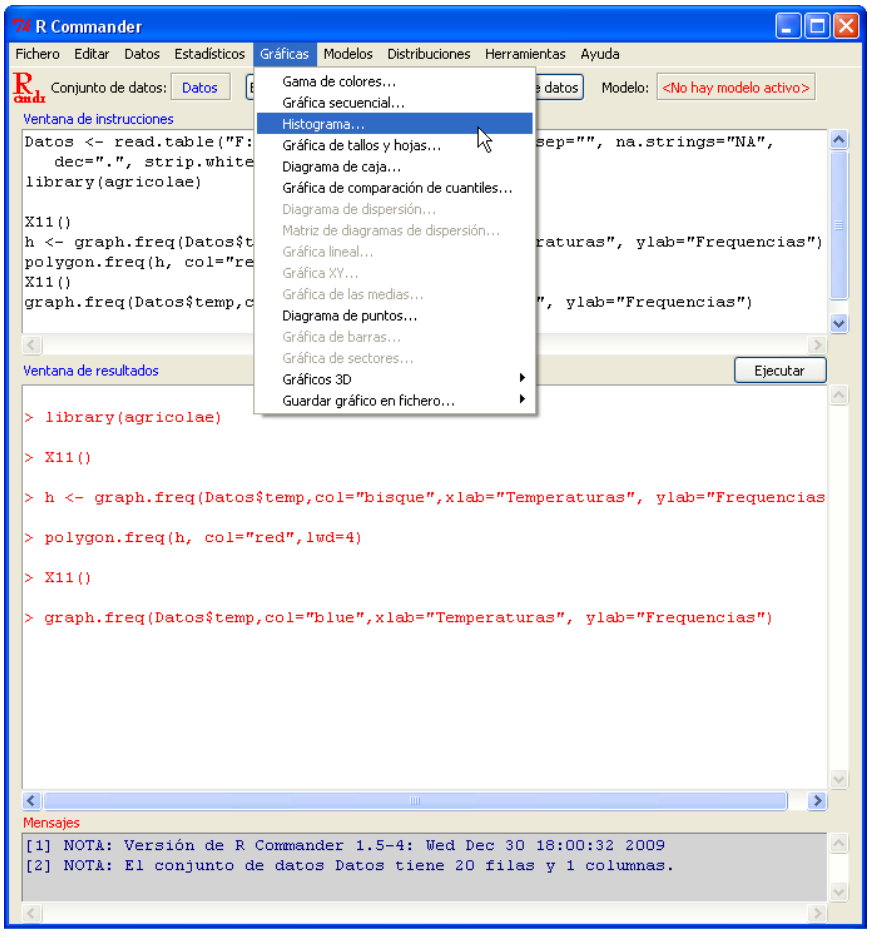


Histogramas de área 1 (sobre una escala de densidad)

- Amplitud de cajas $\equiv$ amplitud de intervalos (no necesariamente idénticos).
- alturas de cajas $= \frac{f_i}{l_i - l_{i-1}}$
- áreas de cajas $= f_i$



Histograma en RCommander



Descripción numérica de datos

Centro	Variación	Otros
⇓	⇓	⇓
• media	• rango	• cuartiles
• mediana	• rango intercuartílico	• percentiles
• moda	• varianza	
	• desviación típica	
	• coef. de variación	

Nueva notación:

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n$$

($\sum$: suma, $i = 1$: el límite inferior, n: el límite superior, x_i: ejemplo de fórmula dependiente de i)

Ejemplo:

$$\sum_{i=-1}^{3} i^2 = (-1)^2 + 0^2 + 1^2 + 2^2 + 3^2 = 15$$

Tendencia central: media (aritmética)

- La medida de tendencia central más común.
- Media poblacional.

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + \ldots + x_N}{N}$$

- Media muestral

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + \ldots + x_n}{n}$$

- Si a, b $(b \neq 0)$ son números reales e $y = a + bx$, se tiene

$$\bar{y} = a + b\bar{x}$$

- Afectado por valores extremos (observaciones atípicas (outliers)).

Ejemplo: X: $3, 1, 5, 4, 2$, Y: $3, 1, 5, 4, 200$

$$\bar{x} = \frac{3+1+5+4+2}{5} = 3 \quad \bar{y} = \frac{3+1+5+4+200}{5} = 42.6!$$

Tendencia central: mediana

- En la lista de observaciones ordenada, la mediana M es el número que está en la mitad de la lista.

$$M = \begin{cases} x_{((n+1)/2)} & \text{si } n \text{ impar (número en la mitad)} \\ \frac{x_{(n/2)}+x_{(n/2+1)}}{2} & \text{if } n \text{ par (promedio de los dos números en la mitad)} \end{cases}$$

  ($x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ significa que las observaciones están ordenadas en orden creciente, ej. $x_{(1)} = x_{\text{mín}}$, $x_{(n)} = x_{\text{máx}}$)

- No afectado por observaciones atípicas (outliers)

Ejemplo: Dadas las observaciones $3, 1, 5, 4, 2$ ($n = 5$), ordenar los datos 1,2,3,4,5, e identificar el/los números situados en la mitad de la lista

$$M = x_{((5+1)/2)} = \overbrace{x_{(3)}}^{3^{\text{o}} \text{ menor}} = 3$$

Ejemplo: Dadas las observaciones $3, 1, 5, 4, 2, 0$ ($n = 6$), ordenar los datos 0,1,2,3,4,5, e identificar el/los números en la mitad de la lista

$$M = \frac{x_{(6/2)} + x_{(6/2+1)}}{2} = \overbrace{\frac{x_{(3)} + x_{(4)}}{2}}^{\text{el promedio del } 3^{\text{o}} \text{ y el } 4^{\text{o}}} = \frac{2+3}{2} = 2.5$$

Tendencia central: moda

- El valor que aparece más a menudo.
- No afectado por valores atípicos=outliers.
- Utilizado tanto para datos numéricos como categóricos.
- Puede no haber moda o puede haber más de una moda.

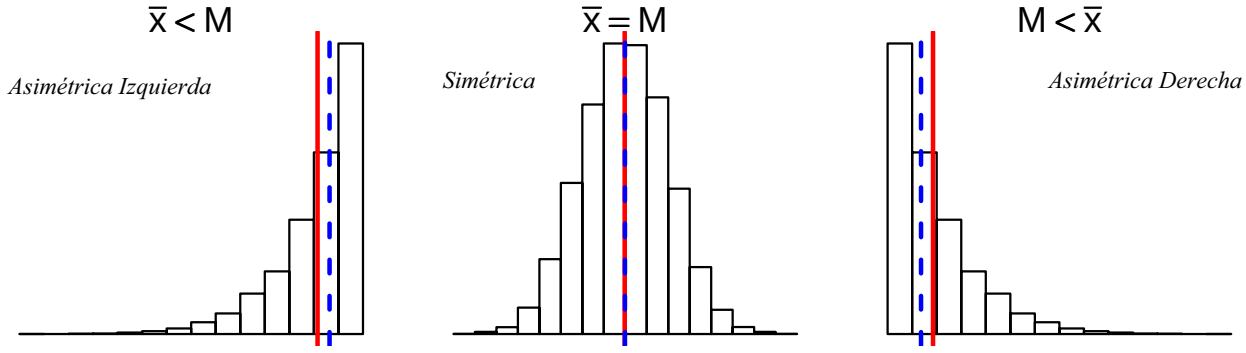
Ejemplo: Dadas las observaciones $3, 1, 5, 4, 2$, **no** hay moda

Ejemplo: Dadas las observaciones $3, 1, 5, 4, 2, 1$, la moda es 1

Forma: comparación de la media y la mediana

Tres tipos de distribuciones:

- Asimétrica a la izquierda Media $<$ Mediana.
- Simétrica Media $=$ Mediana.
- Asimétrica a la derecha Mediana $<$ Media.



Nota: La distribución en que está en el centro se conoce como normal o acampanada (ver figuras)

Cuartiles y percentiles

- Los cuartiles dividen los datos ordenados en cuatro segmentos que recogen la misma cantidad de observaciones.
- El primer cuartil Q_1 ocupa la posición $\frac{1}{4}(n+1)$.
- El segundo cuartil Q_2 ($=$ mediana) ocupa la posición $\frac{1}{2}(n+1)$.
- El tercer cuartil Q_3 ocupa la posición $\frac{3}{4}(n+1)$.

Ejemplo: Dadas las observaciones $22, 18, 17, 16, 16, 13, 12, 21, 11$ ($n = 9$), se ordenan los datos 11, 12, 13, 16, 16, 17, 18, 21, 22, a continuación de identifican las posiciones

$$Q_1 = x_{(2.5)} = 12.5 \quad Q_3 = 16 \quad Q_3 = x_{(7.5)} = 19.5$$

- El p % de los datos ($0 < p < 100$) se encuentran por debajo o sobre el p-ésimo percentil.

Ejemplo cont.: 33-ésimo percentil $= 13$

Variación: rango y rango intercuartílico (RIC)

- El rango es la medida de variación más simple

$$R = x_{\text{máx}} - x_{\text{mín}}$$

- Ignora la manera en que se distribuyen los datos.
- Sensible a observaciones atípicas (outliers).

Ejemplo: Dadas las observaciones $3, 1, 5, 4, 2$, $R = 5 - 1 = 4$
Ejemplo: Dadas las observaciones $3, 1, 5, 4, 100$, $R = 100 - 1 = 99$

- El rango intercuartílico (RIC) puede eliminar ciertos problemas con los datos atípicos (outliers). Se eliminan las observaciones de mayor valor y las de menor valor y se calcula el rango de los 50 % de los datos que se encuentran en la mitad.

$$RIC = \text{3er cuartil} - \text{1er cuartil} = Q_3 - Q_1$$

Variación: Rango intercuartílico y diagrama de cajas

- Las observaciones atípicas (outliers) se encuentran
  - por debajo de $Q_1 - 1.5 \cdot RIC$
  - por encima de $Q_3 + 1.5 \cdot RIC$
- Para observaciones atípicas (outliers) extremos, reemplazar 1.5 por 3 en la definición anterior

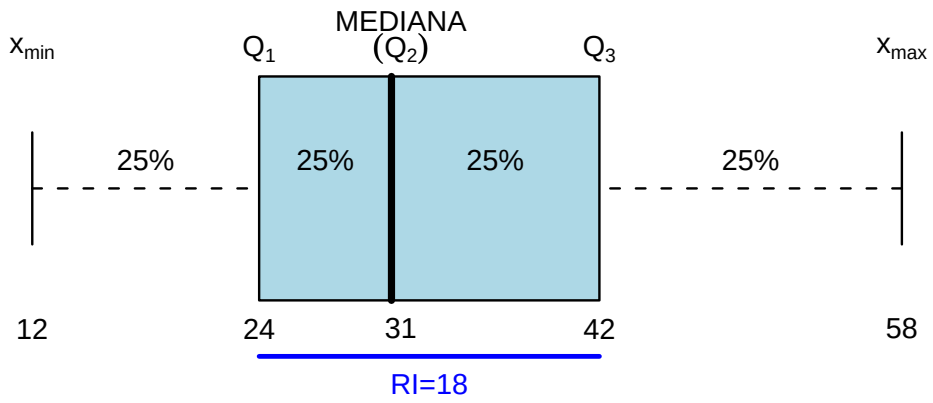
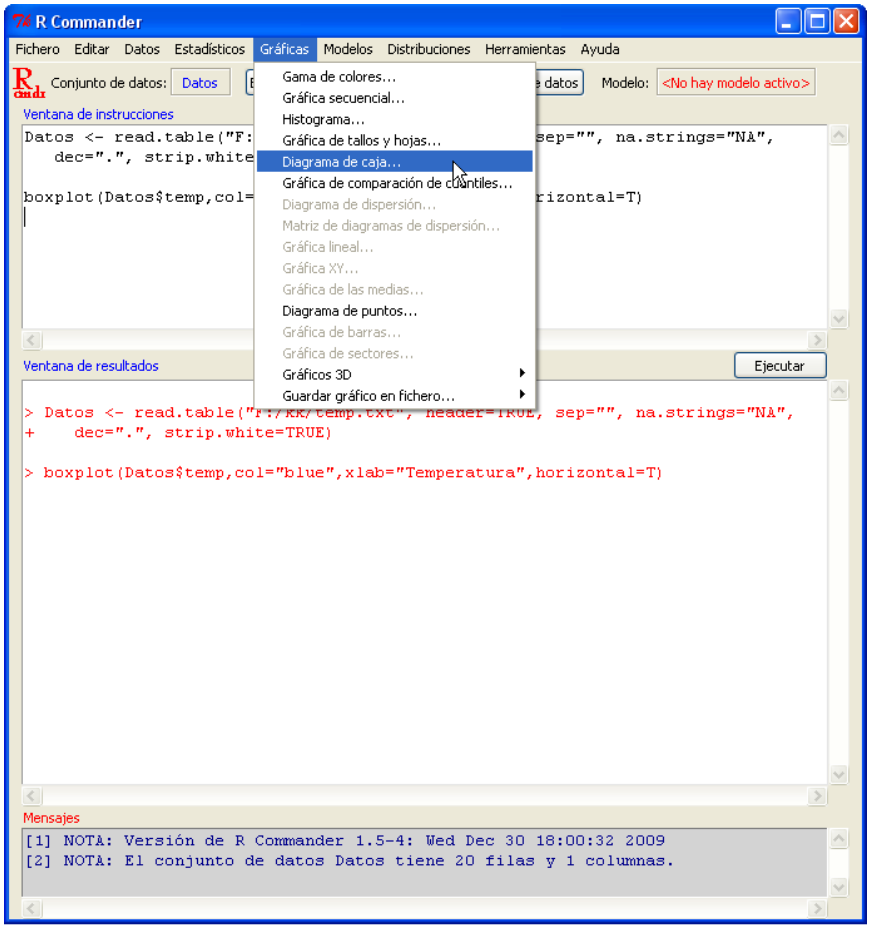


Diagrama de cajas en RCommander



Medida de variación: varianza

- Promedio de cuadrados de las desviaciones de valores a la media.
- Varianza poblacional.

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

- Varianza muestral

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n} = \overbrace{\frac{\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2}{n}}^{\text{más rápido de calcular}}$$

$\Leftarrow$ dividido por n

- Cuasi-varianza muestral (varianza muestral corregida)

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2}{n-1}$$

$\Leftarrow$ dividido por $n-1$

- $\hat{\sigma}^2$ es sesgado, mientras s^2 es insesgado (Tema 5). Su relación es

$$\hat{\sigma}^2 = \frac{n-1}{n} s^2$$

- Si a, b $(b \neq 0)$ son números reales e $y = a + bx$, se tiene $s_y^2 = b^2 s_x^2$

Medida de variación: desviación típica (DT)

- La medida de dispersión más comúnmente utilizada.
- La desviación típica poblacional, la desviación típica muestral y la cuasi-desviación típica muestral son respectivamente

$$\sigma = \sqrt{\sigma^2} \qquad \hat{\sigma} = \sqrt{\hat{\sigma}^2} \qquad s = \sqrt{s^2}$$

- Muestra la variación sobre la media.
- Posee las misma unidades que los datos, mientras que para la varianza se tienen unidades^2
- Varianza y DT se encuentran ambos afectados por observaciones atípicas (outliers).

Cálculo de la varianza y la desviación típica

Ejemplo: X : $11, 12, 13, 16, 16, 17, 18, 21$,
Y : $14, 15, 15, 15, 16, 16, 16, 17$, Z : $11, 11, 11, 12, 19, 20, 20, 20$

$$\bar{x} = \frac{124}{8} = 15.5 \qquad \bar{y} = \frac{124}{8} = 15.5 \qquad \bar{z} = \frac{124}{8} = 15.5$$

$$\sum_{i=1}^{n} x_i^2 = 11^2 + 12^2 + \ldots + 21^2 = 2000$$

$$\sum_{i=1}^{n} y_i^2 = 14^2 + 15^2 + \ldots + 17^2 = 1928$$

$$\sum_{i=1}^{n} z_i^2 = 11^2 + 11^2 + \ldots + 20^2 = 2068$$

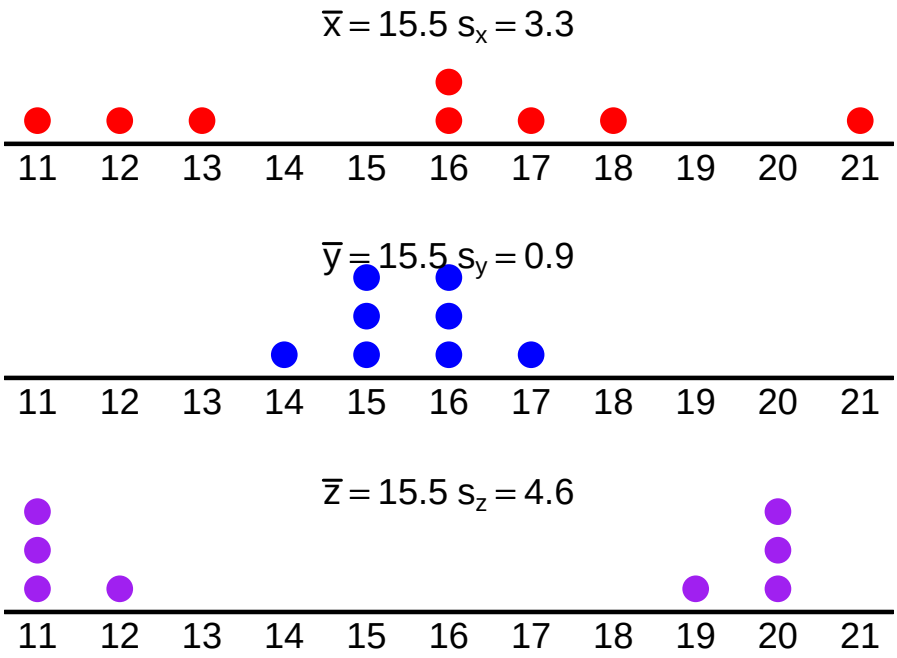
$$s_x^2 = \frac{\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2}{n-1} = \frac{2000 - 8(15.5)^2}{8-1} = \frac{78}{7} = 11.1429 \Rightarrow s_x = 3.3381$$

$$s_y^2 = \frac{1928 - 8(15.5)^2}{8-1} = \frac{6}{7} = 0.8571 \Rightarrow s_y = 0.9258$$

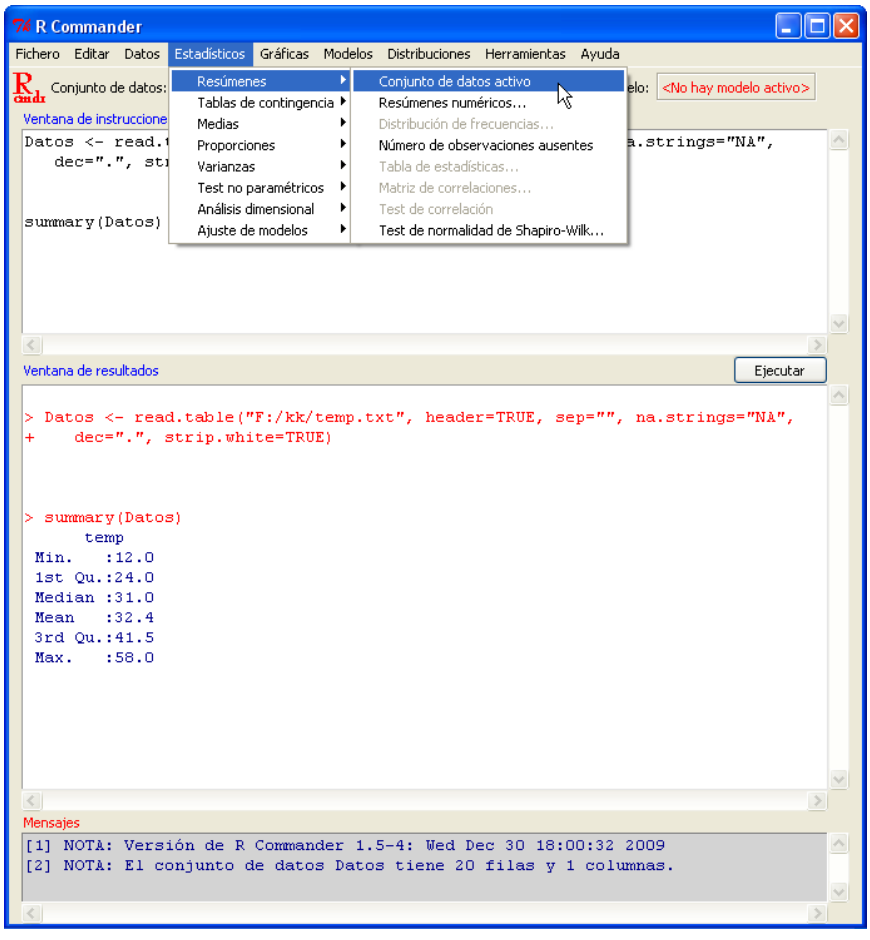
$$s_z^2 = \frac{2068 - 8(15.5)^2}{8-1} = \frac{146}{7} = 20.8571 \Rightarrow s_z = 4.5670$$

Comparación de desviaciones típicas

Ejemplo cont.: X : 11, 12, 13, 16, 16, 17, 18, 21,
Y : 14, 15, 15, 15, 16, 16, 16, 17, Z : 11, 11, 11, 12, 19, 20, 20, 20



Resúmenes numéricos en RCommander



Regla empírica

Si la distribución de los datos es acampanada (normal), es decir, simétrica y con colas suaves, se verifica:

- 68 % de los datos en $(\bar{x} - 1s, \bar{x} + 1s)$
- 95 % de los datos en $(\bar{x} - 2s, \bar{x} + 2s)$
- 99.7 % de los datos en $(\bar{x} - 3s, \bar{x} + 3s)$

Nota: Esta regla se conoce también como la regla del 68-95-99.7

Ejemplo: Sabemos que para una muestra de 100 observaciones, la media es 40 y la cuasi-desviación típica es 5. Asumiendo que los datos poseen distribución acampanada, proporciona los límites del intervalo que captura el 95 % de las observaciones.

$$95\,\%\text{ de } x_i\text{'s están en: } (\bar{x} \pm 2s) = (40 \pm 2(5)) = (30, 50)$$

Medidas de variación: coeficiente de variación (CV)

- Es una medida relativa de variación que se define como

$$CV = \frac{s}{|\bar{x}|}$$

- Es un número sin unidad (se expresa a veces en %'s).
- Muestra la variación con respecto a la media.

Ejemplo: Stock A: Precio promedio el año anterior = 50, Desviación típica = 5
Stock B: Precio promedio el año anterior = 100, Desviación típica = 5

$$CV_A = \frac{5}{50} = 0.10 \quad CV_B = \frac{5}{100} = 0.05$$

Ambos stocks poseen la misma DT, pero el stock B es menos variable en relación a la media de su precio.