Unidad 2. Estadística descriptiva

Tipos de variables y escalas de medición

Se puede clasificar a las variables según los valores que pueden tomar y sus propiedades. Una de las clasificaciones más utilizadas es la siguiente.

Tipos de variables

- Categóricas: corresponden a mediciones no cuantificables.
  - Nominales. Su rango está compuesto de categorías.
  - Ordinales. Su rango está compuesto de categorías ordenadas.
- Numéricas: corresponden a mediciones cuantificables.
  - Discretas. Tienen rango numerable.
  - Continuas. Tienen rango no numerable.

- Los modelos para variables categóricas y numéricas discretas, se ven en el curso *análisis de datos categóricos*.
- Los modelos para variables numéricas continuas se ven en el resto de los cursos de estadística (inferencia estadística, diseño de experimentos, métodos multivariados, análisis de regresión, entre otros).

Escalas de medición

Medir (DRAE, 2014)

1. Comparar una cantidad con su respectiva unidad, con el fin de averiguar cuántas veces la segunda está contenida en la primera.
3. Comparar algo no material con otra cosa.

- Cuando se habla de medición de magnitudes físicas se hace referencia a la asignación de cantidades numéricas.
- El propósito es poder deducir información acerca de los entes medidos a partir de manipular/operar sus mediciones.
- De nuevo, en el caso de las magnitudes físicas hay una correspondencia entre manipular físicamente a los objetos medidos y manipular sus mediciones con operaciones aritméticas.
- Si tenemos dos objetos de masas m_1 y m_2, ambas en las mismas unidades, sabemos que $m_1 + m_2$ corresponde a la masa de los dos objetos combinados y que $m_1/2$ corresponde a la masa de cada una de las partes de resulta de dividir al objeto 1 exactamente a la mitad.

Escalas de medición

- Cuando lo que se *mide* no son magnitudes físicas, puede no existir una correspondencia entre la manipulación física o abstracta de los objetos medidos y la aplicación de operaciones aritméticas a sus mediciones.
- El mayor logro que se puede tener en medición es conseguir una escala que permita tal relación. Aunque es posible lograr escalas de medición intermedias que permitan obtener conclusiones relevantes.

Escalas de medición

- Nominal. Sólo admiten comparaciones de igualdad.
- Ordinal. Igualdad + comparaciones de orden (relacionadas con la *intensidad* del atributo)
- De intervalos. Igualdad + orden + operaciones con las diferencias (el cero y las unidades de medida son arbitrarias)
- De razón. Todo lo anterior + operaciones con los atributos (el cero es absoluto e indica ausencia del atributo pero las unidades de medida son arbitrarias)

Estadística descriptiva

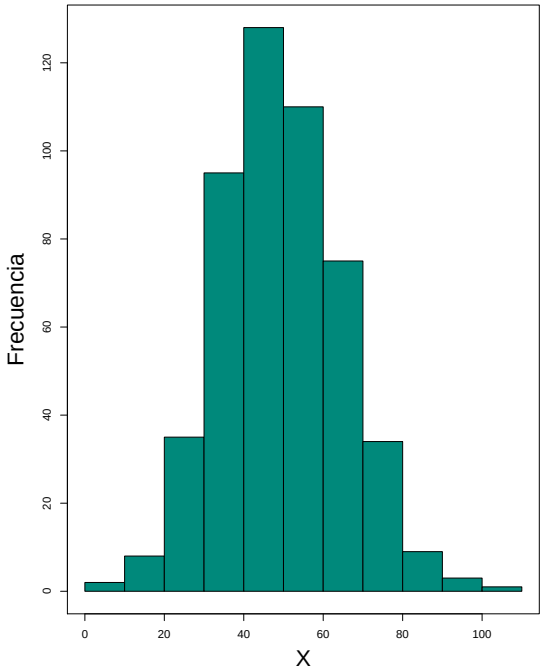
- La *estadística descriptiva* es el conjunto de técnicas analíticas y gráficas que se utilizan para describir un conjuntos de datos.
- Generalmente el interés se centra en estudiar la distribución de una cierta variable en el conjunto de datos o estudiar asociaciones entre pares de variables.
- Hay dos tipos de características que son importantes para describir la *distribución* de una variable: tendencia central, dispersión y forma.
- Nos enfocaremos solamente en variables numéricas.

- Medidas de tendencia central: se utilizan para identificar el valor que *mejor* describe al conjunto de observaciones.
- Medidas de dispersión: se utilizan para medir qué tanto difieren las observaciones con respecto a algún valor *central*.
- Medidas de forma: se utilizan para cuantificar desviaciones a la forma de campana de la distribución normal.
- Medidas de asociación: se utilizan para cuantificar asociaciones entre pares de variables.

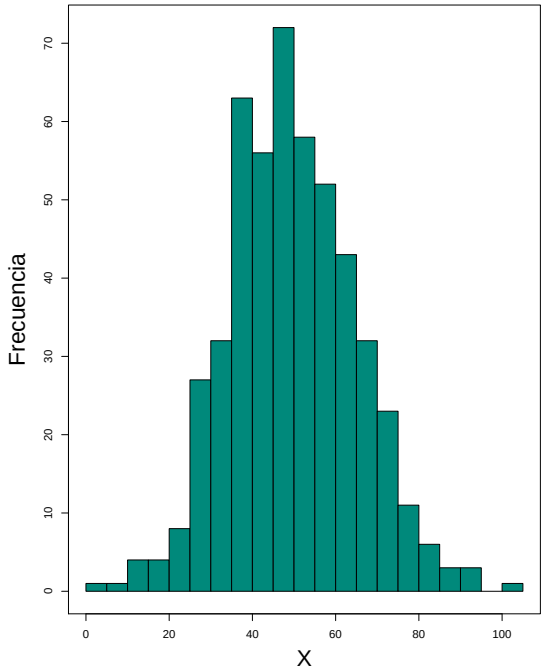
Histograma

- Los histogramas son gráficas de barras en las que en el eje x se representa la variable de interés y la altura de las barras es la frecuencia (absoluta o relativa) con la que se observa un determinado valor o un intervalo de valores de x.
- Los histogramas se utilizan para estudiar la distribución de los dato, esto es, la frecuencia con la que se observan determinados valores.
- Usualmente se hacen histogramas de variables continuas, por lo que se agrupan las observaciones en intervalos igualmente espaciados.
- El número de intervalos a usar se debe determinar de acuerdo al número de observaciones y a su distribución. Generalmente se especifica el número de intervalos k y éstos se construyen de longitud $(\max x_i - \min x_i)/k$.
- El objetivo es elegir k de manera que el histograma proporcione información valiosa acerca de la distribución de las observaciones.

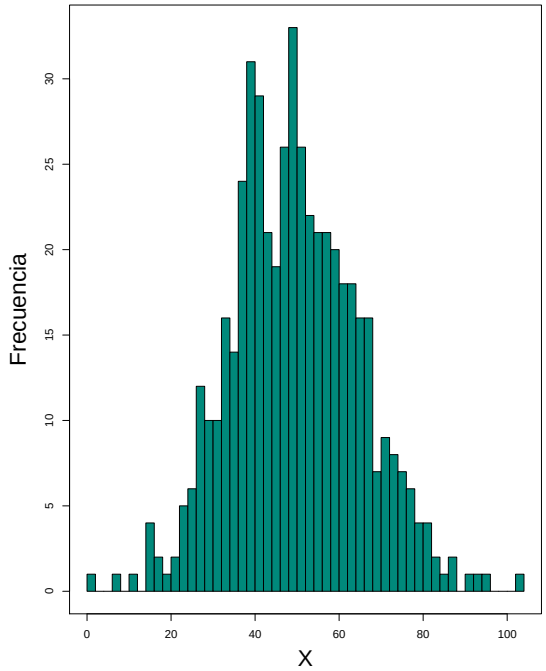
Ejemplos: histograma



$b = 11$



$b = 21$



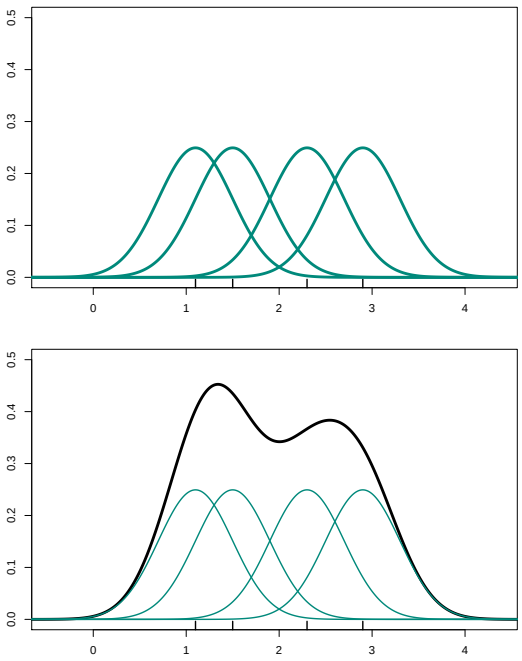
$b = 51.$

En `R`, se utiliza el comando `hist` para graficar histogramas. La opción `breaks` se utiliza para especificar el número de *cortes*, por lo que el número de barras es `breaks + 1`.

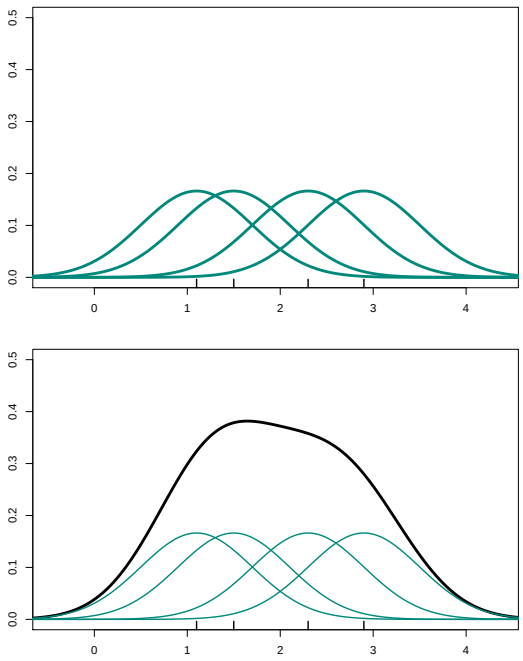
Gráfico de densidad estimada

- Los gráficos de densidad estimada se utilizan para estudiar la distribución de las observaciones. Se puede considerar a estos gráficos como histogramas *suavizados*.
- Una de las formas de estimar densidades es usando funciones *núcleo* o *kernel* y hay toda una teoría al respecto...
- A grandes rasgos se calcula la contribución de cada observación según el kernel seleccionado, escalado por el número de observaciones y un parámetro de amplitud. La densidad estimada es el resultado de sumar las contribuciones de cada observación.

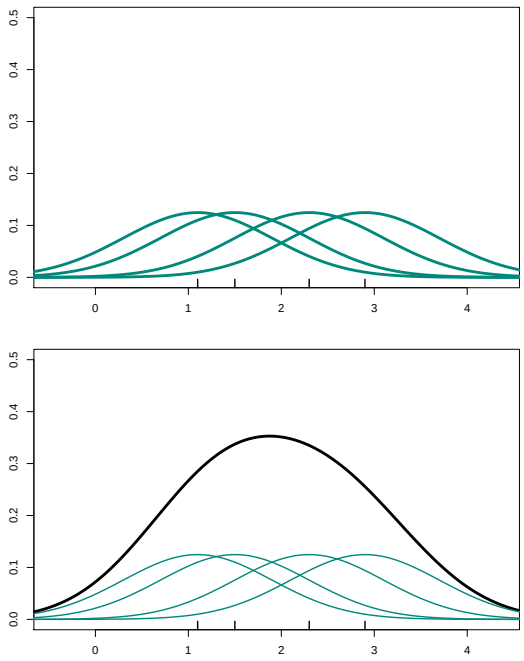
Ejemplos: densidad estimada (kernel gaussiano)



$h = 0.4$

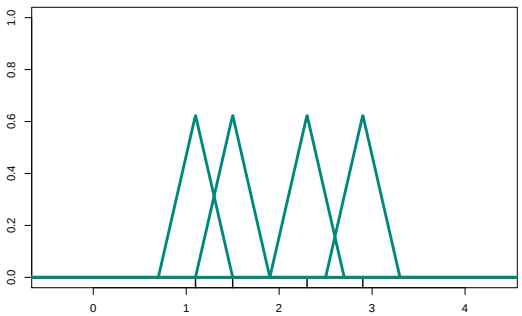


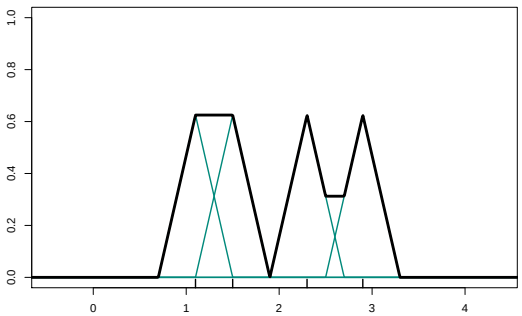
$h = 0.6$



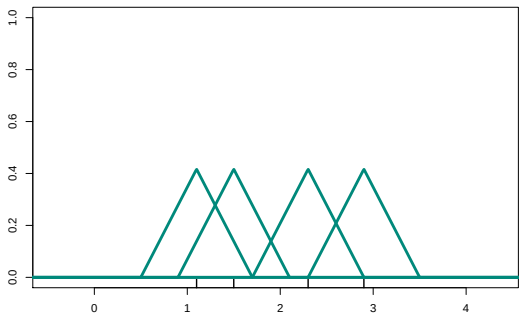
$h = 0.8.$

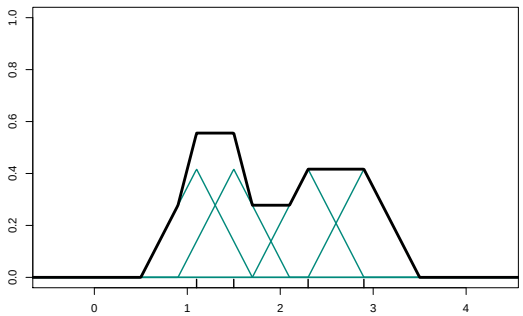
Ejemplos: densidad estimada (kernel triangular)



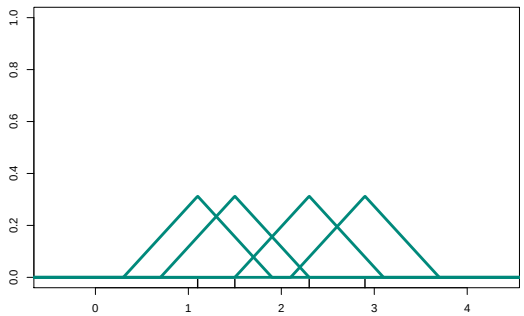


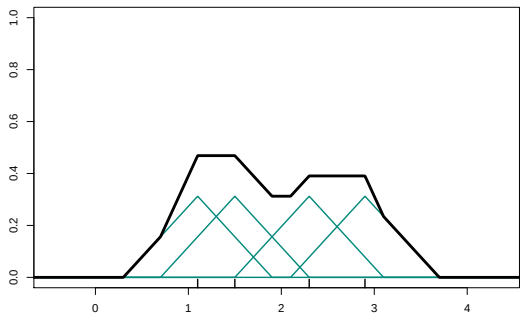
$h = 0.4$





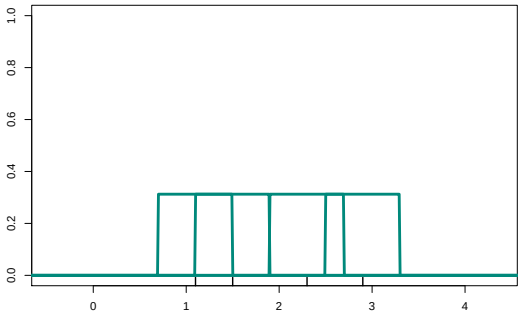
$h = 0.6$

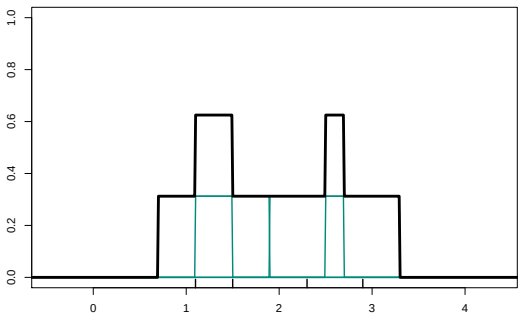




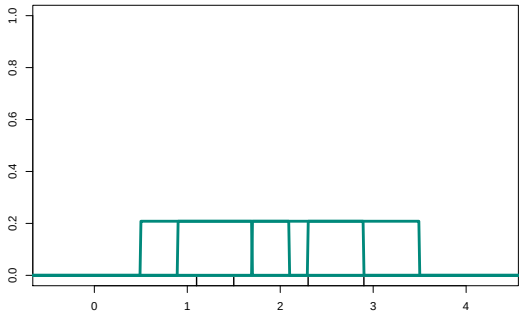
$h = 0.8.$

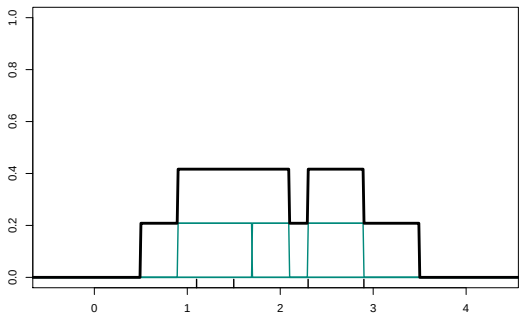
Ejemplos: densidad estimada (kernel rectangular)



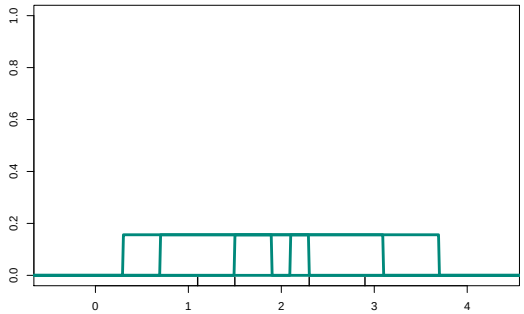


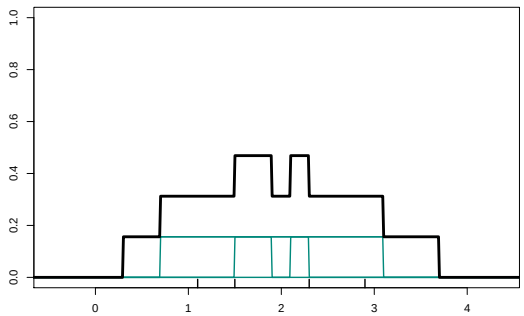
$h = 0.4$





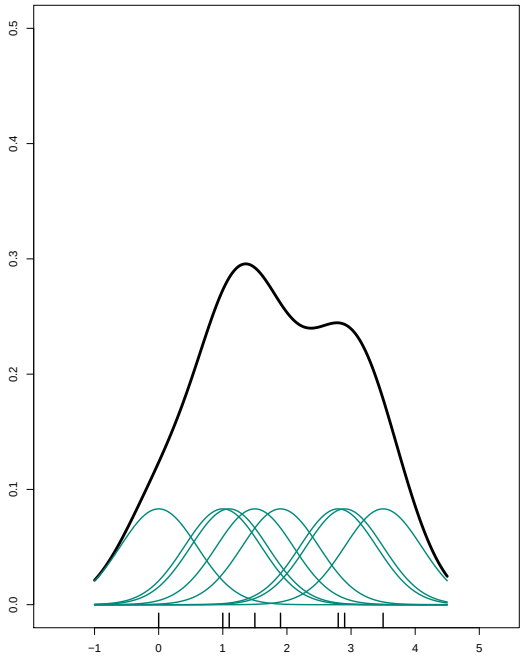
$h = 0.6$



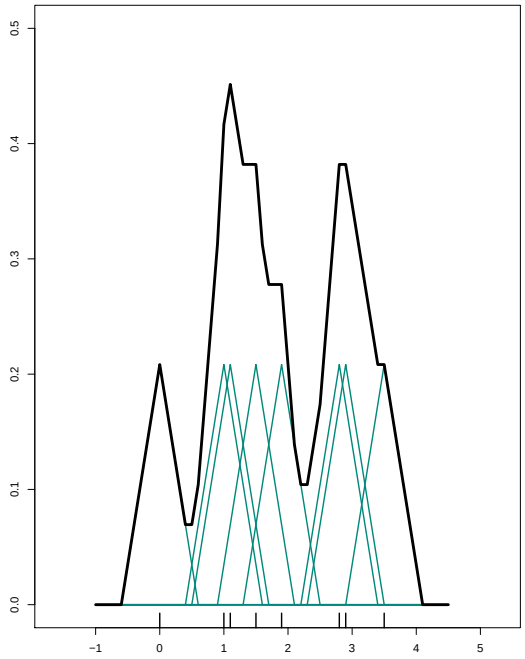


$h = 0.8.$

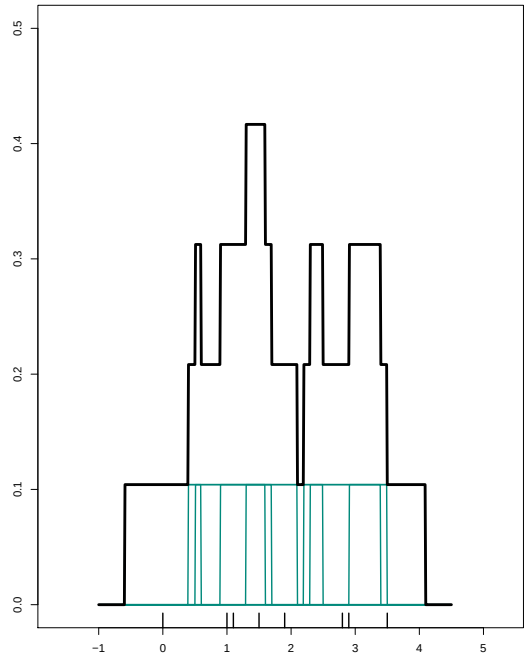
Ejemplos: densidad estimada (comparación)



Gaussiano



Triangular



Rectangular

Estadísticos de orden

- Los estadísticos de orden de una variable numérica sirven para identificar la posición de una medición con respecto al orden de las mediciones. Suponer que tenemos una muestra aleatoria $S = \{X_1, X_2, \ldots, X_n\}$, de alguna distribución.
- El primer estadístico de orden se denota por $X_{(1)}$, corresponde a la menor observación de toda la muestra y se le llama *mínimo*.
- El último (n-ésimo) estadístico de orden se denota por $X_{(n)}$, corresponde a la mayor observación de la muestra y se le llama *máximo*.
- El i-ésimo estadístico de orden se denota por $X_{(i)}$ y corresponde a la observación de la muestra que ocupa de i-ésima posición cuando ésta se ordena de menor a mayor.
- La distribución de los estadísticos de orden $X_{(i)}$ no necesariamente corresponde a la distribución de las variables originales X_i.
- Una vez que se ha observado la muestra los estadísticos de orden se pueden *calcular* y se denotan por $x_{(1)}, \ldots, x_{(n)}$.

Medidas de tendencia central

Suponer que se tienen las siguientes observaciones de una variable: $x_1, x_2, \ldots, x_n$. Se definen las siguientes medidas de tendencia central.

- **Promedio**: Se calcula como la media aritmética de las observaciones y se denota por $\bar{x}$.

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

- **Mediana**: Se define la mediana como la observación que divide a la población en dos partes iguales, el 50 % con valores mayores y el 50 % con valores menores. Se calcula como sigue

$$\text{Mediana} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{si } n \text{ es impar} \\ \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} & \text{si } n \text{ es par} \end{cases}$$

Medidas de dispersión

- **Varianza**: Es el promedio de los cuadrados de las desviaciones de los valores observados con respecto a su media y se calcula como:

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

- **Desviación estándar**: es la raíz cuadrada de la varianza, tiene la ventaja que está expresada en las mismas unidades que las mediciones originales.
- **Desviación absoluta**: Es el promedio de las desviaciones absolutas de las observaciones con respecto a su media y se calcula como

$$\text{Desv. Abs.} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|.$$

  Tiene la ventaja que está expresada en las mismas unidades que las variables originales.
- **Rango**: Es la longitud del menor intervalo que contiene a todos las observaciones y se calcula como

$$\text{Rango} = x_{(n)} - x_{(1)}.$$

Medidas de forma

Definición (coeficiente de asimetría)

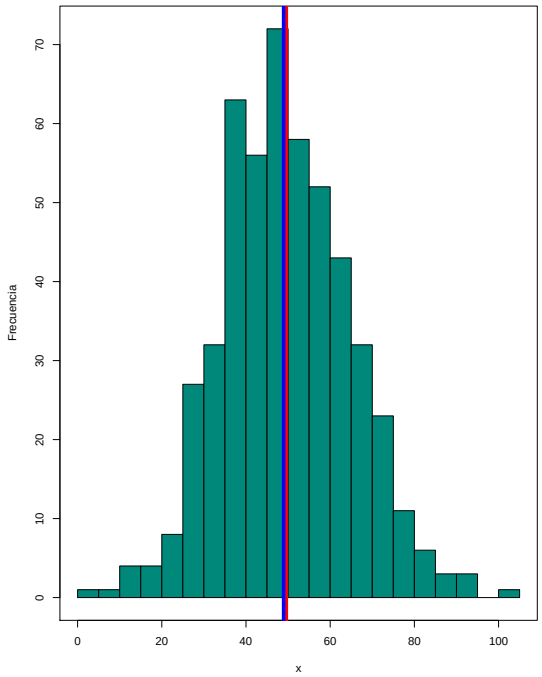
Se define el coeficiente de simetría de un conjunto de observaciones $x_1, \ldots, x_n$ como:

$$\gamma = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}$$

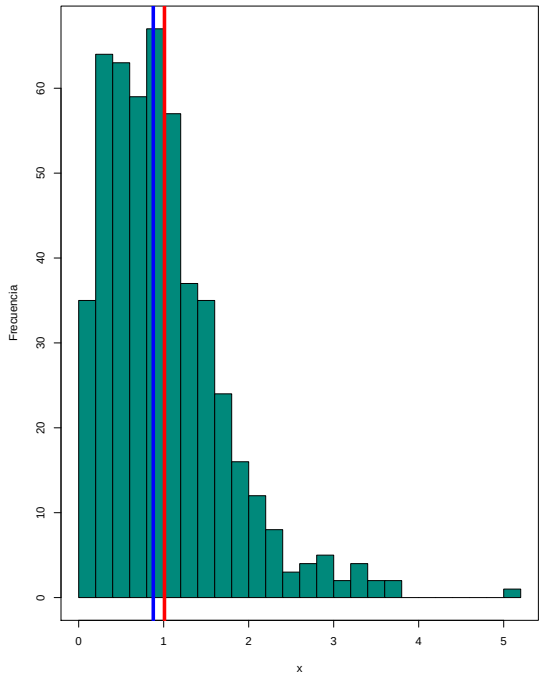
Interpretación

- Si $\lambda \approx 0$ la distribución es simétrica alrededor de su media.
- Si $\lambda > 0$ la distribución tiene sesgo positivo. La mayoría de las observaciones son menores a $\bar{x}$ y el histograma o la densidad estimada tienen una *joroba* a la izquierda.
- Si $\lambda < 0$ la distribución tiene sesgo negativo. La mayoría de las observaciones son mayores a $\bar{x}$ y el histograma o la densidad estimada tienen una *joroba* a la derecha.

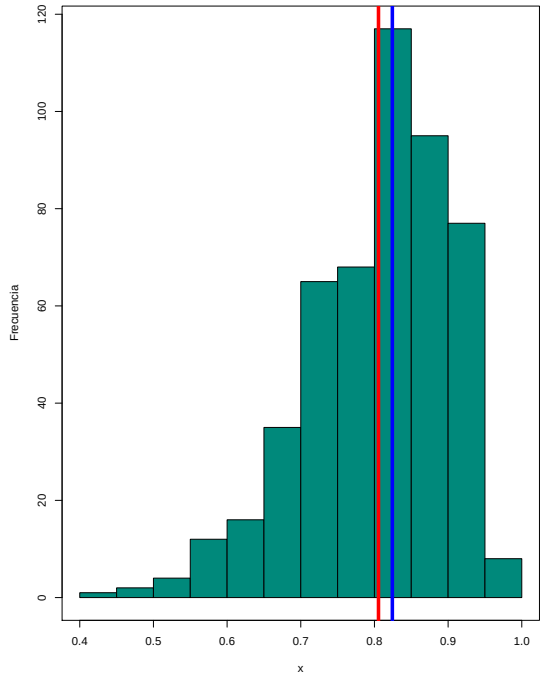
Ejemplos: curtosis



$\gamma = 0.17$



$\gamma = 1.44$



$\gamma = -0.79.$

Las líneas representan la media (rojo) y la mediana (azul) de las observaciones.

Medidas de forma

Definición (curtosis)

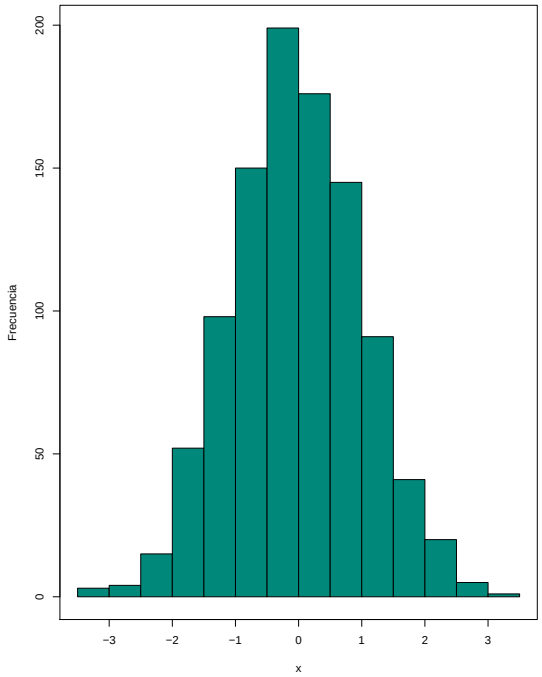
Dado un conjunto de observaciones $x_1, \ldots, x_n$, se define la curtosis como

$$\kappa = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{s^4} - 3$$

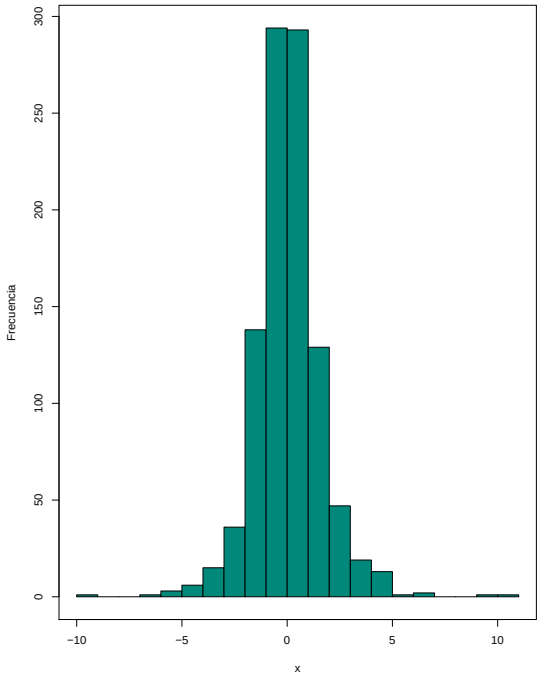
Interpretación

- Si $\kappa \approx 0$, la distribución de los datos es *mesocúrtica*. La concentración alrededor de la media y las colas tienen la misma forma que en la distribución normal.
- Si $\kappa > 0$, la distribución de los datos es *leptocúrtica*. La forma alrededor de la media es más afilada y las colas son más pesadas que las de una distribución normal.
- Si $\kappa < 0$, la distribución de los datos es *platicúrtica*. La forma alrededor de la media es más plana y las colas son más ligeras que las de una distribución normal.

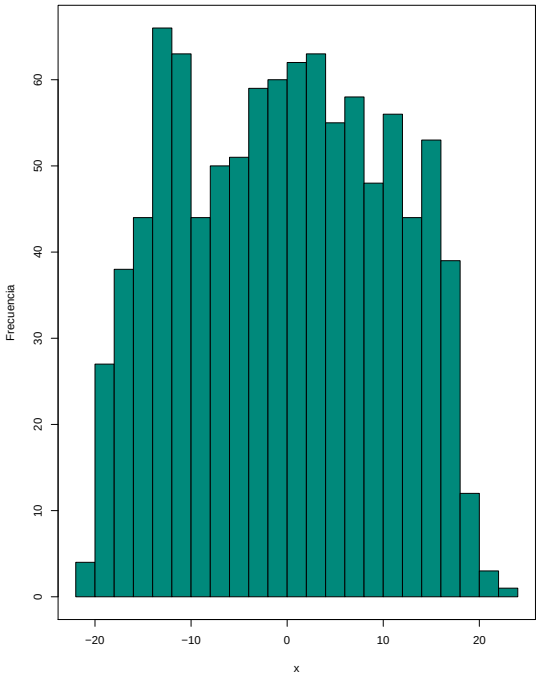
Ejemplos: curtosis



$\kappa = 0.07$



$\kappa = 5.75$



$\kappa = -1.09.$

Medidas de asociación

Definición (covarianza)

Se define la covarianza entre un conjunto de observaciones $x_1, \ldots, x_n$ y $y_1, \ldots, y_n$ como

$$s_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

Definición (correlación)

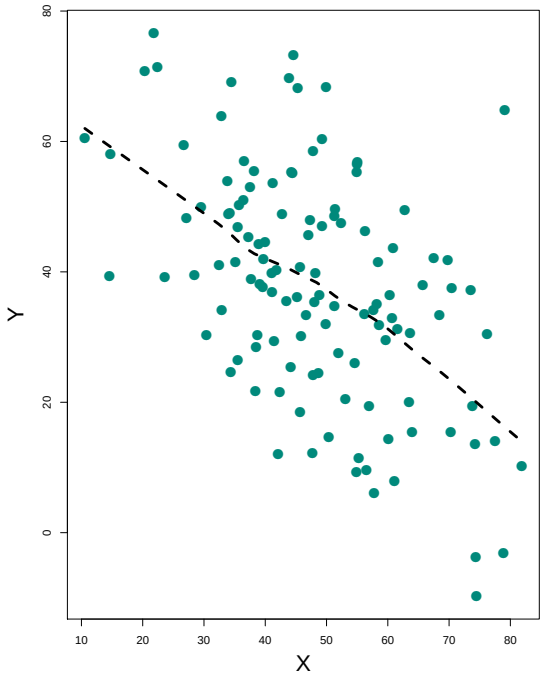
Se define la covarianza entre un conjunto de observaciones $x_1, \ldots, x_n$ y $y_1, \ldots, y_n$ como

$$r_{xy} = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}}.$$

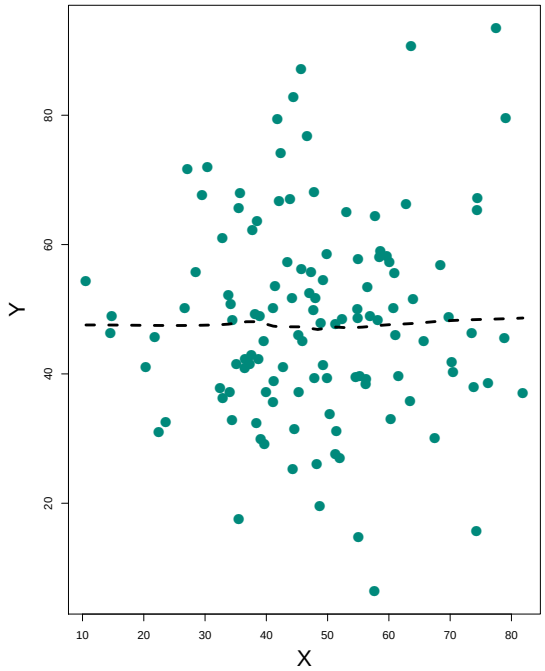
Medidas de asociación: interpretación

- s_{xy} y r_{xy} son medidas de asociación *lineal*. Por lo que no deben ser utilizadas si se sospecha algún tipo de relación no lineal entre las variables.
- r_{xy} es una versión *estandarizada* de s_{xy}, que toma valores en el intervalo $(-1, 1)$, por lo que se prefiere para estudiar asociación lineal entre pares de variables.
- Valores de r_{xy} cercanos a -1 indican fuerte asociación lineal inversa entre X y Y, esto significa que valores grandes de X están relacionados con valores pequeños de Y y viceversa.
- Valores de r_{xy} cercanos a 0 indican ausencia de asociación linean entre X y Y, aunque es posible que exista otro tipo de asociación no lineal.
- Valores de r_{xy} cercanos a 1 indican fuerte asociación lineal directa entre X y Y, esto significa que valores grandes de X están relacionados con valores grandes de Y y viceversa.
- $r_{xy} = \pm 1$ si y sólo si existen constantes a y b, tales que $y_i = a + bx_i$, $i = 1, \ldots, n$.

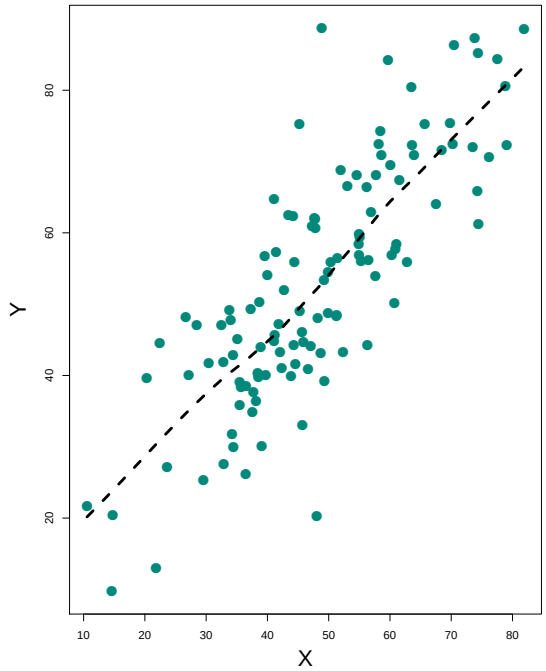
Ejemplos: medidas de asociación



$r_{xy} = -0.51$



$r_{xy} = 0.04$



$r_{xy} = 0.80.$