Conceptos básicos de la inferencia estadística

Unidad 2. Estadística descriptiva

Javier Santibáñez

IIMAS, UNAM

jsantibanez@sigma.iimas.unam.mx

Semestre 2020-1

Contenido

1. Tipos de variables y escalas de medición
2. Medidas de tendencia central
3. Medidas de dispersión
4. Medidas de forma
5. Medidas de asociación
6. Representaciones gráficas

Tipos de variables

Las variables se clasifican según los valores que pueden tomar. Una de las clasificaciones más utilizadas es la siguiente:

- Categóricas: corresponden a mediciones no cuantificables.
  - Nominales: Su rango está compuesto de categorías sin un orden evidente (nacionalidad, género, código postal).
  - Ordinales: Su rango está compuesto de categorías ordenadas (nivel de escolaridad, dominio de un idioma, preferencias en escala Likert).
- Numéricas: corresponden a mediciones cuantificables.
  - Discretas: Tienen rango numerable (edad en años, número de hijos, número de cuartos en la vivienda).
  - Continuas: Tienen rango no numerable (virtualmente cualquier magnitud física como tiempo, masa, temperatura).

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 de medir es poder deducir información acerca de los entes medidos a partir de operar con sus mediciones.

- En el caso de las magnitudes físicas hay una correspondencia entre manipular físicamente a los objetos medidos y manipular sus mediciones con operaciones matemáticas.
- 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

Las escalas de medición se clasifican según las operaciones permitidas con las mediciones. mediciones.

- Nominal: Comparaciones de igualdad.
- Ordinal: Comparaciones de igualdad y comparaciones de orden relacionadas con la *intensidad* del atributo medido.
- De intervalos. Comparaciones de igualdad y orden, además de operaciones aritméticas con las diferencias, aunque el cero y las unidades de medida son arbitrarias.
- De razón. Comparaciones de igualdad y orden, además de operaciones aritméticas con las mediciones, en este caso el cero es absoluto e indica ausencia del atributo pero las unidades de medida son arbitrarias.

Ejemplos de variables y sus escalas

- Nominal: nacionalidad, género, código postal.
- Ordinal: nivel de escolaridad, dominio de un idioma, preferencias en escala Likert.
- De intervalos: año calendario, escalas Celsius y Farenheit de temperatura.
- De razón: edad, escala Kelvin de temperatura, escalas para medir otras magnitudes físicas.

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.

Tipos de estadísticas descriptivas

- Medidas de tendencia central: se utilizan para describir el comportamiento típico de las observaciones. En general, son estadísticos que permiten conocer algún aspecto de la localización de las mediciones.
- Medidas de dispersión: se utilizan para describir la variabilidad en las observaciones.
- Medidas de forma: se utilizan para describir la forma en cómo se distribuyen las observaciones. Generalmente cuantifican desviaciones a la forma de campana de la distribución normal.
- Medidas de asociación: se utilizan para cuantificar asociaciones entre pares de variables. Las medidas más utilizadas cuantifican asociaciones lineales.

Medidas de tendencia central

Suponer que se tienen las observaciones: $x_1, x_2, \ldots, x_n$.

- **Promedio o media aritmética**:

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

- **Estadísticos de orden**:

$$x_{(1)} = \text{mín} \{x_1, x_2, \ldots, x_n\}$$
$$x_{(n)} = \text{máx} \{x_1, x_2, \ldots, x_n\}$$

En un conjunto con n observaciones hay n-estadísticos de orden. El i-ésimo estadístico de orden $x_{(i)}$ es la observación que ocupa la i-ésima posición cuando las datos se ordenan de menor a mayor.

Medidas de tendencia central

- **Mediana**:

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

- **Cuartiles**:

$$Q_2 = \text{mediana}\left\{x_1, x_2, \ldots, x_n\right\}$$

$$Q_1 = \text{mediana}\left\{x_i \ : \ x_i \leq Q_2\right\}$$

$$Q_3 = \text{mediana}\left\{x_i \ : \ x_i \geq Q_2\right\}$$

Medidas de tendencia central

- La mediana divide la recta real en dos segmentos tales que en cada uno está el 50 % de las observaciones.
- Los cuartiles Q_1, Q_2 y Q_3, que dividen la recta real en cuatro segmentos tales que cada uno contiene el 25 % de las observaciones.
- Q_1 corresponde a la mediana de las observaciones menores al la mediana de todas las observaciones.
- Q_3 corresponde a la mediana de las observaciones mayores a la mediana de todas las observaciones.

Medidas de tendencia central

- **Cuantiles**: Para cada $\alpha \in (0,1)$, el cuantil α , denotado por q_α, como el número tal que $100\alpha\,\%$ de las observaciones son menores a q_α.
- **Deciles**: son los nueve cuantiles que dividen a la recta real en diez segmentos, cada uno con igual número de observaciones, es decir:

$$q_{0.1}, q_{0.2}, q_{0.3}, \ldots, q_{0.9}.$$

- **Porcentiles o percentiles**: son los 99 cuantiles que dividen a la recta real en 100 segmentos, cada uno con igual número de observaciones, es decir:

$$q_{0.01}, q_{0.02}, q_{0.03}, \ldots, q_{0.97}, q_{0.98}, q_{0.99}.$$

Cuantiles

- Existen distintas formas de calcular los cuantiles muestrales, que se basan en distintos supuestos sobre la distribución poblacional subyacente que generó a las observaciones.
- La forma o método que se usa para calcular los cuantiles es relevante si se tienen pocas observaciones o sí existen empates.
- En `R` se utiliza la función `quantile` para calcular los cuantiles. En el argumento `prob` se introduce el valor de α.
- Para tener una idea de la cantidad de *algoritmos* disponibles para calcular los cuantiles podemos introducir revisar la ayuda de la función `quantile` y revisar el apartado del argumento `type`.

Medidas de dispersión

- **Varianza y desviación estándar**:

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

- **Desviación absoluta media**:

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

- **Rango, rango intercuartil y rango interdecil**:

$$R = x_{(n)} - x_{(1)}, \quad IQR = Q_3 - Q_1 \quad \text{e} \quad IDR = q_{0.9} - q_{0.1}$$

Medidas de dispersión

- La varianza es el promedio de las desviaciones cuadráticas de las observaciones con respecto a al promedio de los datos. Se eleva al cuadrado para eliminar el efecto de los signos.
- En algunos casos se prefiere utilizar s en lugar de s^2, ya que s está expresada en las mismas unidades que las observaciones originales.
- Hay un efecto negativo de elevar al cuadrado las desviaciones para calcular la varianza y es que las desviaciones pequeñas son reducidas y las desviaciones grandes son amplificadas.
- Los distintos rangos ofrecen distintos grados de robustez ante la presencia de observaciones extremas, inusualmente grandes o pequeñas.

Medidas de forma

- **Coeficiente de simetría**:

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

La simetría se refiere a la comparación de las frecuencias con las que se observan valores grandes y pequeños

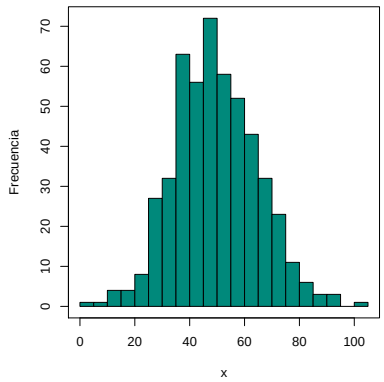
- **Coeficiente de curtosis**:

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

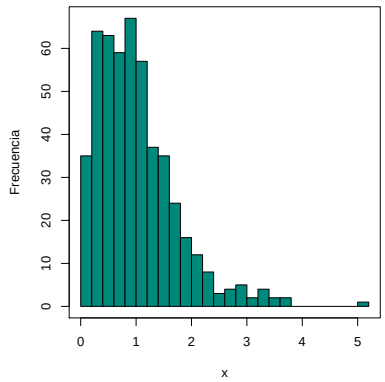
La curtosis se refiere a la forma en qué tan concentrados están los datos alrededor de la media.

Interpretación de γ

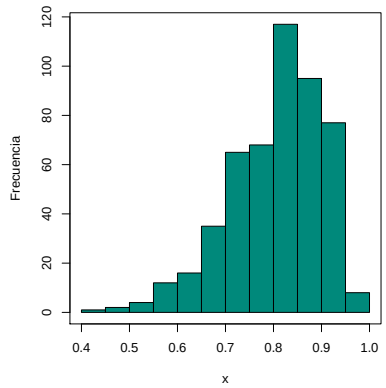
- Si $\lambda \approx 0$ la distribución es simétrica alrededor de su media.
- Si $\lambda > 0$ la distribución tiene sesgo positivo. La mayoría de las observaciones son pequeñas (menores a $\bar{x}$).
- Si $\lambda < 0$ la distribución tiene sesgo negativo. La mayoría de las observaciones son grandes (mayores a $\bar{x}$).



$\gamma = 0.17$



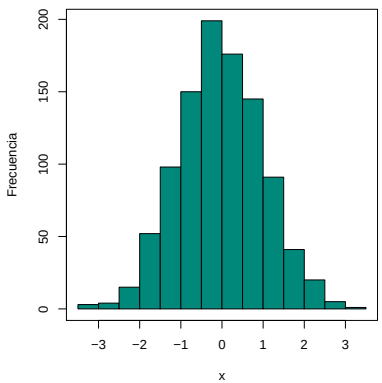
$\gamma = 1.44$



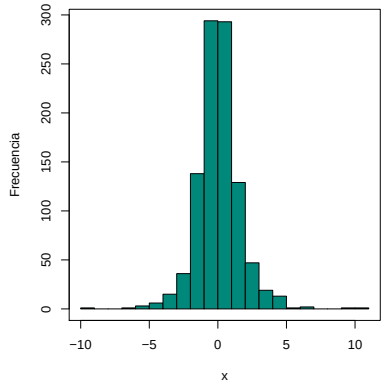
$\gamma = -0.79.$

Interpretación de κ

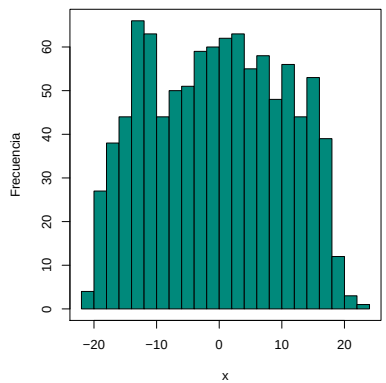
- Si $\kappa \approx 0$ la distribución es *mesocúrtica*. La concentración de los datos alrededor de la media es similar a la de la distribución normal.
- Si $\kappa > 0$, la distribución es *leptocúrtica*. La concentración de los datos alrededor de la media es mayor a la de la distribución normal.
- Si $\kappa < 0$, la distribución es *platicúrtica*. La concentración de los datos alrededor de la media es menor que en la distribución normal.



$\kappa = 0.07$



$\kappa = 5.75$



$\kappa = -1.09.$

Medidas de asociación

Suponer que se tienen observaciones de dos variables X y Y:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n).$$

- **Covarianza**:

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

donde $\bar{x}$ es el promedio de las X y $\bar{y}$ es el promedio de las Y.

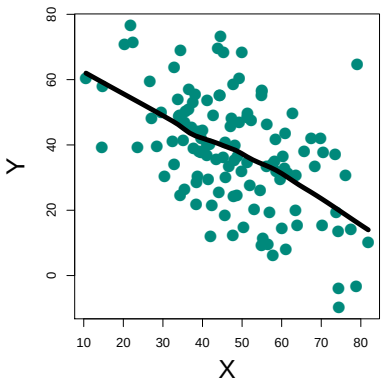
- **Correlación**:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

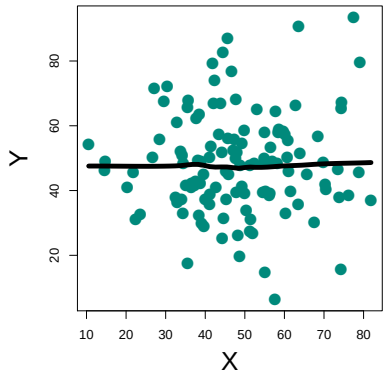
donde s_x es la desviación estándar de las X y s_y es la desviación estándar de las Y.

Interpretación de s_{xy} y r_{xy}

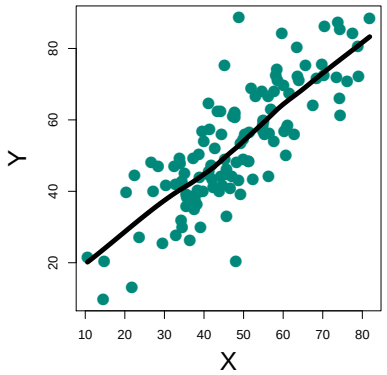
- s_{xy} y r_{xy} son medidas de asociación lineal, si la asociación entre X y Y es no lineal, estas medidas no son útiles.
- r_{xy} es una versión *estandarizada* de s_{xy}, que toma valores en el intervalo $(-1, 1)$, por lo que es más fácil de interpretar:
  - Si $r_{xy} \approx 1$, la relación lineal es directa.
  - Si $r_{xy} \approx 0$, no hay relación **lineal**.
  - Si $r_{xy} \approx -1$, la relación lineal es inversa.



$r_{xy} = -0.51$



$r_{xy} = 0.04$



$r_{xy} = 0.80$

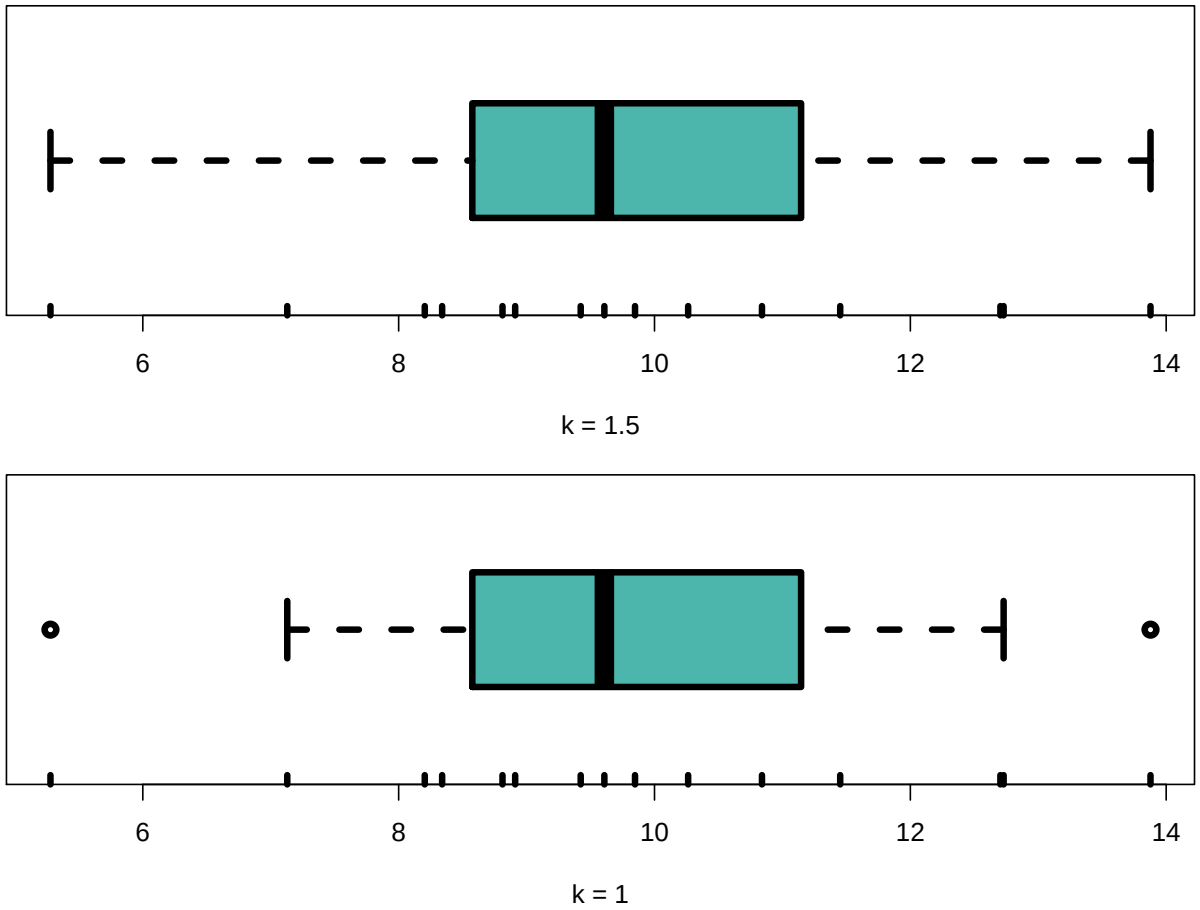
Boxplots o gráficos de caja (y bigotes)

Los *boxplots* permiten analizar fácilmente la distribución una serie de observaciones, incluso comparar series diferentes, a partir de los cuartiles y el IQR.

- Los lados de la caja representan Q_1 y Q_3, de manera que la base del rectángulo tiene una longitud igual al IQR.
- El segmento al interior de la caja representa Q_2 y su posición relativa se utiliza para estudiar la simetría de la distribución de los datos.
- Bigotes tienen como longitud máxima un múltiplo del IQR pero se ajustan para coincidir con una observación.
- Llos puntos representan observaciones extremas, es decir, aquellas fuera del intervalo $(Q_1 - kIQR, Q_3 + kIQR)$.

Ejemplo: boxplot

En `R` se utiliza el comando `boxplot` para hacer gráficos de caja y bigotes.



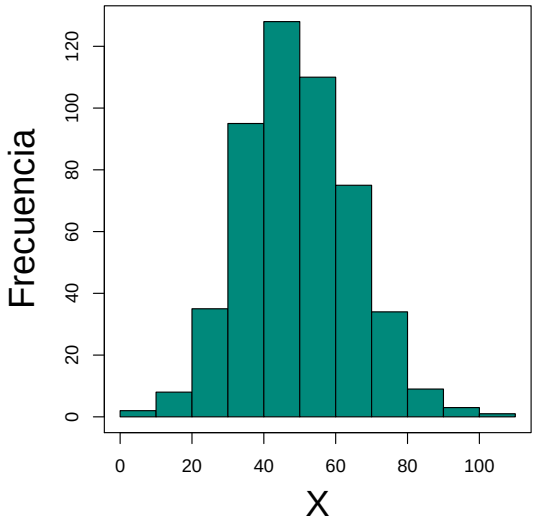
Histograma

Los histogramas son útiles para analizar la forma de la distribución de una serie de datos e incluso compararla con un modelo teórico.

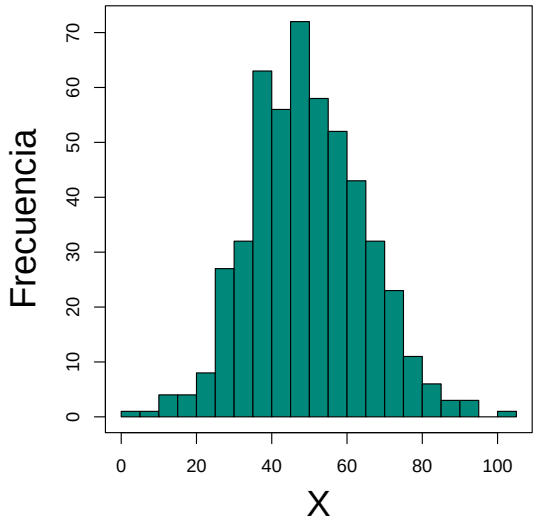
- La altura de las barras representa la frecuencia con la que se observa un determinado valor o un intervalo de valores de x.
- Usualmente se hacen histogramas de variables continuas, por lo que se agrupan las observaciones en intervalos igualmente espaciados.
- El número de intervalos a usar se debe determinar de acuerdo al número de observaciones y a su distribución.
- El objetivo es elegir un número de intervalos de manera que el histograma sea informativo. Se recomienda utilizar $\sqrt{n}$ intervalos.

Ejemplos: histograma

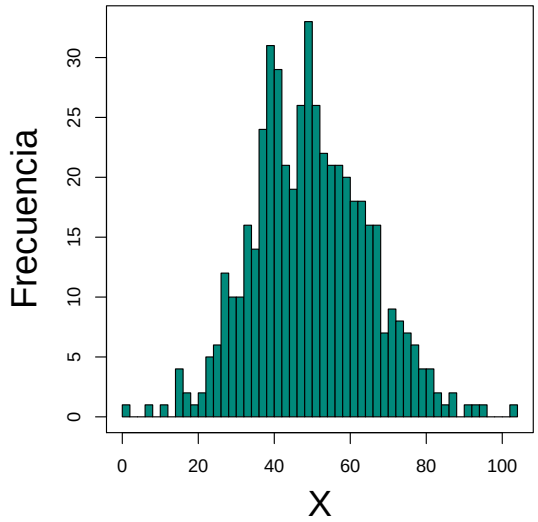
En `R`, se utiliza el comando `hist` para graficar histogramas. La opción `breaks` se utiliza para especificar el número de *cortes*, por lo que el número de barras es `breaks + 1`.



$b = 11$



$b = 21$

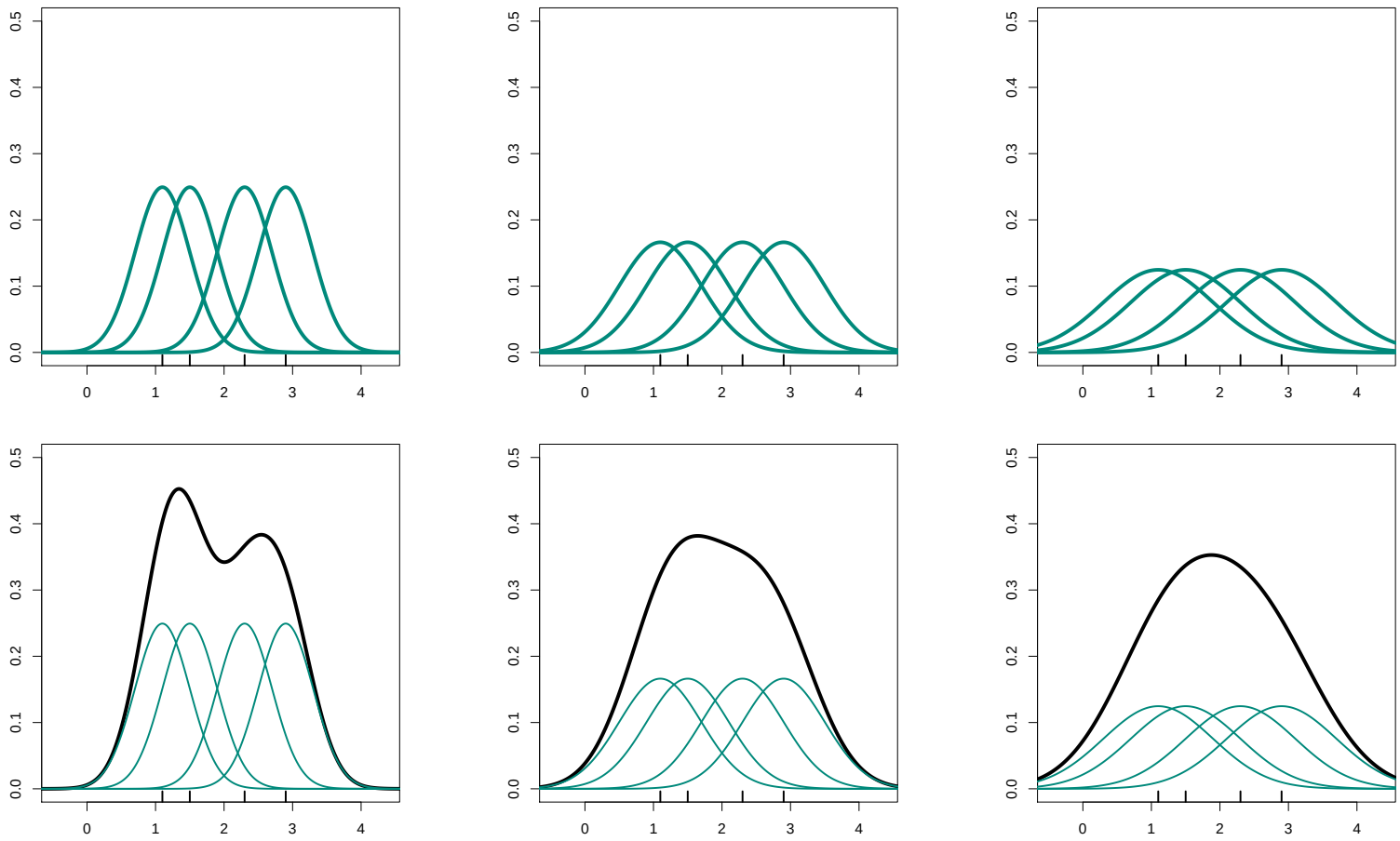


$b = 51$

Gráfico de densidad estimada

- Los gráficos de densidad estimada también se utilizan para estudiar la distribución una serie de 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...
- La densidad estimada es el resultado de sumar las contribuciones de cada observación, calculadas según el kernel seleccionado, que escalado por el número de observaciones y un parámetro de amplitud.
- En `R` se utiliza las funciones `density` y `plot` para calcular y gráficar la densidad estimada, respectivamente. Por defecto se utiliza el kernel gaussiano.

Kernel gaussiano



$h = 0.4$ $h = 0.6$ $h = 0.8.$

Gráfico de dispersión

Los gráficos de dispersión son gráficos que se utilizan para explorar posibles relaciones entre pares de variables. Son muy útiles en el análisis de regresión.

- Los ejes del gráfico de dispersión corresponden a las variables. Las coordenadas de los puntos del gráfico corresponden a los valores de las variables.
- Los gráficos de dispersión se utilizan para identificar dependencias entre pares de variables, en cuando a su naturaleza e intensidad.
- En el análisis de regresión se utilizan este tipo de gráficos para identificar patrones tales como tendencias no lineales, así como para identificar posibles valores atípicos.

Ejemplos de gráficos de dispersión