Estadística descriptiva

Dr. Edgar Ramírez Galeano

¿Qué es la estadística descriptiva?

El término estadística descriptiva se refiere al análisis, el resumen y la presentación de los resultados relacionados con un conjunto de datos derivados de una muestra o de toda la población.

La estadística descriptiva comprende tres categorías principales: distribución de frecuencias, medidas de tendencia central y medidas de variabilidad.

Tabla de frecuencia de datos no agrupados

- Los datos no agrupados son observaciones en estudios estadísticos presentados en la forma original.
- Indican con qué frecuencia se muestran las estadísticas sin cambiar los tamaños de las unidades originales.
- Se utilizan cuando a variable tiene un número pequeño de valores o es discreta

Tabla de Frecuencias para Datos no Agrupados

- **Valor:** Los valores observados en el conjunto de datos.
- **Frecuencia absoluta:** La cantidad de veces que cada valor aparece en el conjunto de datos.
- **Frecuencia relativa:** Es la frecuencia dividida entre el total de los datos.
- **Frecuencia absoluta acumulada** Es la suma de las frecuencias absolutas de todos los valores inferiores o iguales al valor considerado.
- **Frecuencia relativa acumulada** Es el cociente entre la frecuencia acumulada de un determinado valor y el número total de datos.

Ejemplo

En el diseño de un plato dosificador para siembra de maíz con una sembradora de precisión, se observó la longitud, ancho y grosor de 27 granos de maíz oloton y se obtuvieron los siguientes datos (en mm):

Longitud	Ancho	Grosor	Longitud	Ancho	Grosor	Longitud	Ancho	Grosor
11	9	4	10	11	6	11	11	6
11	11	4	11	11	4	12	12	4
11	9	5	10	11	3	10	11	4
11	10	3	11	11	5	9	10	5
12	10	5	10	10	5	11	10	6
10	12	6	10	9	11	9	10	5
11	11	3	10	11	4	11	10	4
11	10	4	12	10	3	12	11	3
8	9	3	10	11	5	11	10	4

Longitud maíz oloton

8	9	9	10	10	10	10	10	10	10	10	11	11
11	11	11	11	11	11	11	11	11	11	12	12	12
12												

Tabla de Frecuencias

Longitud	Frecuencia Absoluta	Frec. Abs. Acum.	Frecuencia Relativa	Frec. Rel. Acum.
8	1	1	3.70%	3.70%
9	2	3	7.41%	11.11%
10	8	11	29.63%	40.74%
11	12	23	44.44%	85.19%
12	4	27	14.81%	100%

Ancho maíz oloton

9	9	9	9	10	10	10	10	10	10	10	10	10
10	11	11	11	11	11	11	11	11	11	11	11	12
12												

Tabla de Frecuencias

Ancho	Frecuencia Absoluta	Frec. Abs. Acum.	Frecuencia Relativa	Frec. Rel. Acum.
9	4	4		
11			0.41	0.56
10		25		
12	2		0.07	

Tablas de Frecuencias para Datos Agrupados

Máquina seleccionadora de duraznos para almíbar bajo la norma NMX-F-034-1982

Perímetro duraznos

15.2	15.3	15.4	15.5	15.5	15.5	15.6	15.6	15.6	15.7
15.8	15.8	15.9	16.1	16.1	16.1	16.2	16.3	16.3	16.4
16.4	16.4	16.5	16.5	16.6	16.6	16.6	16.8	16.9	17.2
17.3									

Intervalos	Valor Medio	f_i	F_i	h_i	H_i
[15.2 , 15.9]	15.55	13	13	41.94%	41.94%
(15.9 , 16.6]	16.25	14	27	45.16%	87.10%
(16.6 , 17.3]	16.95	4	31	12.90%	100%

- Intervalos = 3
- Rango = Valor Máx - Valor Mín = 17.3 - 15.2 = 2.1
- Amplitud = Rango / Int = 2.1 / 3 = 0.7

Regla de Sturges

$$k = 1 + 3.322 \cdot log(N)$$

- k es el número de clases.
- N es el número total de observaciones de la muestra.
- Log es el logaritmo base 10.

Regla de Scott

$$k = 3.5 \cdot S \cdot n^{-1/3}$$

- k es el número de clases.
- S es la desviación estándar.
- n el total de elementos.

Regla de Freedman & Diaconis

$$k = 2 \cdot IQ \cdot n^{-1/3}$$

- k es el número de clases.
- IQ es el el rango intercuartílico.
- n el total de elementos.

REGLAS GENERALES PARA FORMAR LAS DISTRIBUCIONES DE FRECUENCIA

- Determine los números mayores y menores en los datos primarios y entonces halle el recorrido o amplitud.
- Divida el recorrido por un número conveniente de intervalos de clase que tengan el mismo tamaño. El número de los intervalos de clase usualmente se toma entre 5 y 20, dependiendo de los datos.
- Determine el número de observaciones que caen dentro de cada intervalo de clase, es decir halle las frecuencias de clase.

HISTOGRAMA

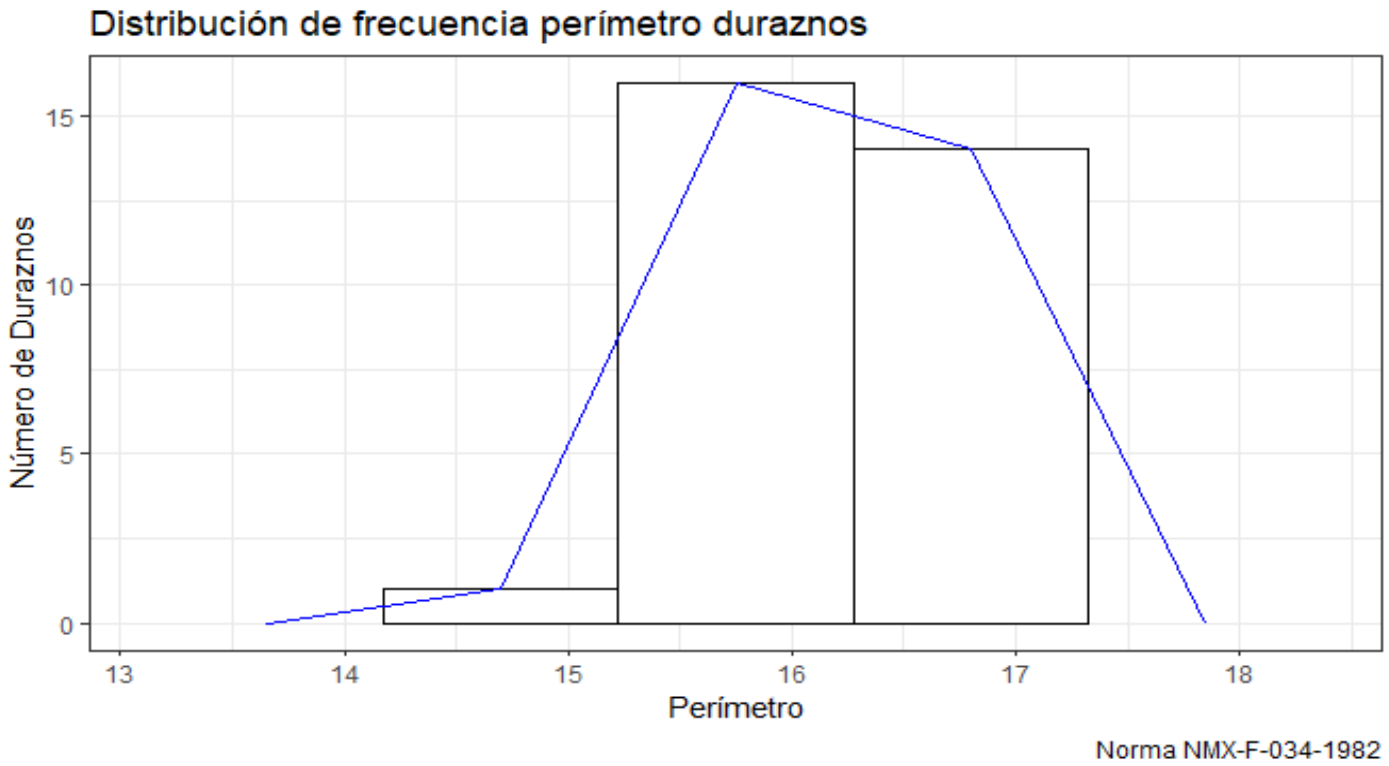
Representaciones gráficas de las distribuciones de frecuencia:

Un histograma consiste en un conjunto de rectángulos (gráfica de barras) que tienen:

- Base sobre un eje horizontal (el eje X) con los centros en las marcas de clase o puntos medios y las longitudes iguales a los tamaños de los intervalos de clase.
- El eje vertical (el eje Y) representa las frecuencias con que se repiten las mediciones en un intervalo de clase.

POLÍGONOS DE FRECUENCIA

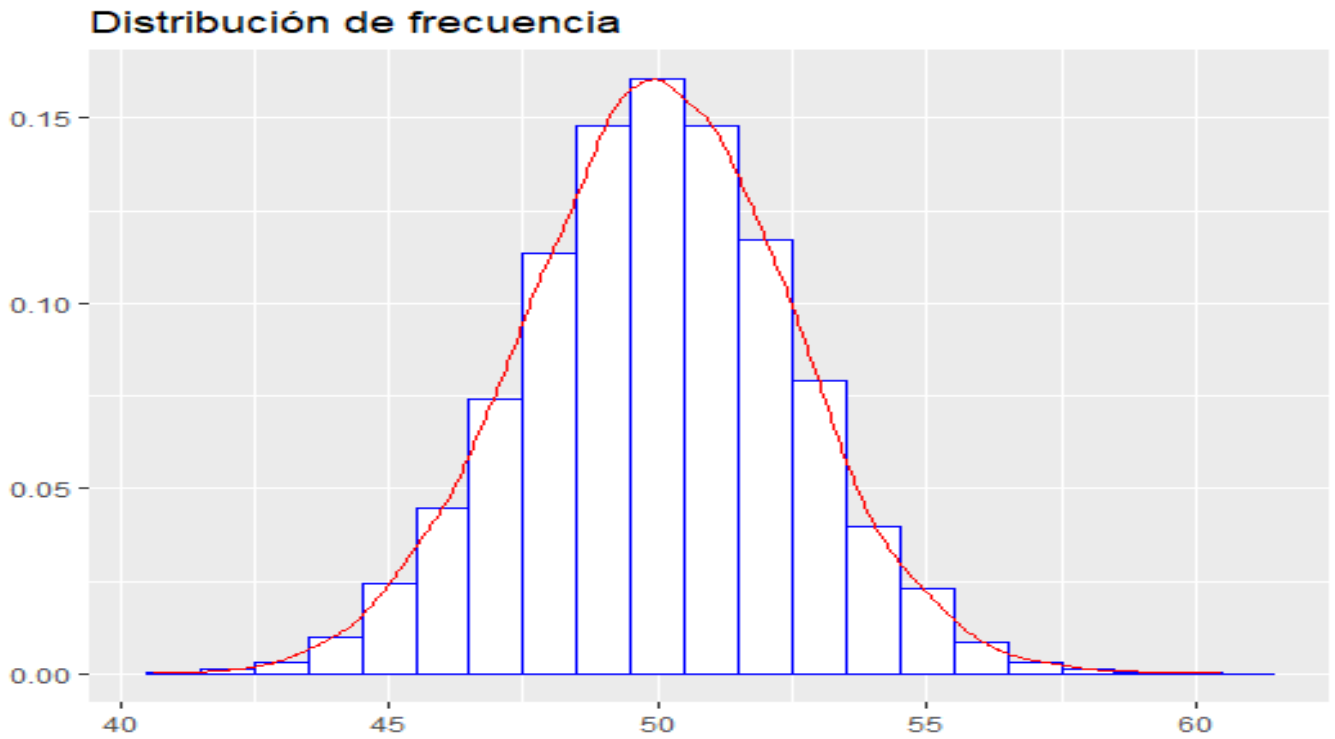
Es un gráfico de líneas de frecuencia de clase trazado en función del punto medio.



TIPOS DE CURVAS DE FRECUENCIA

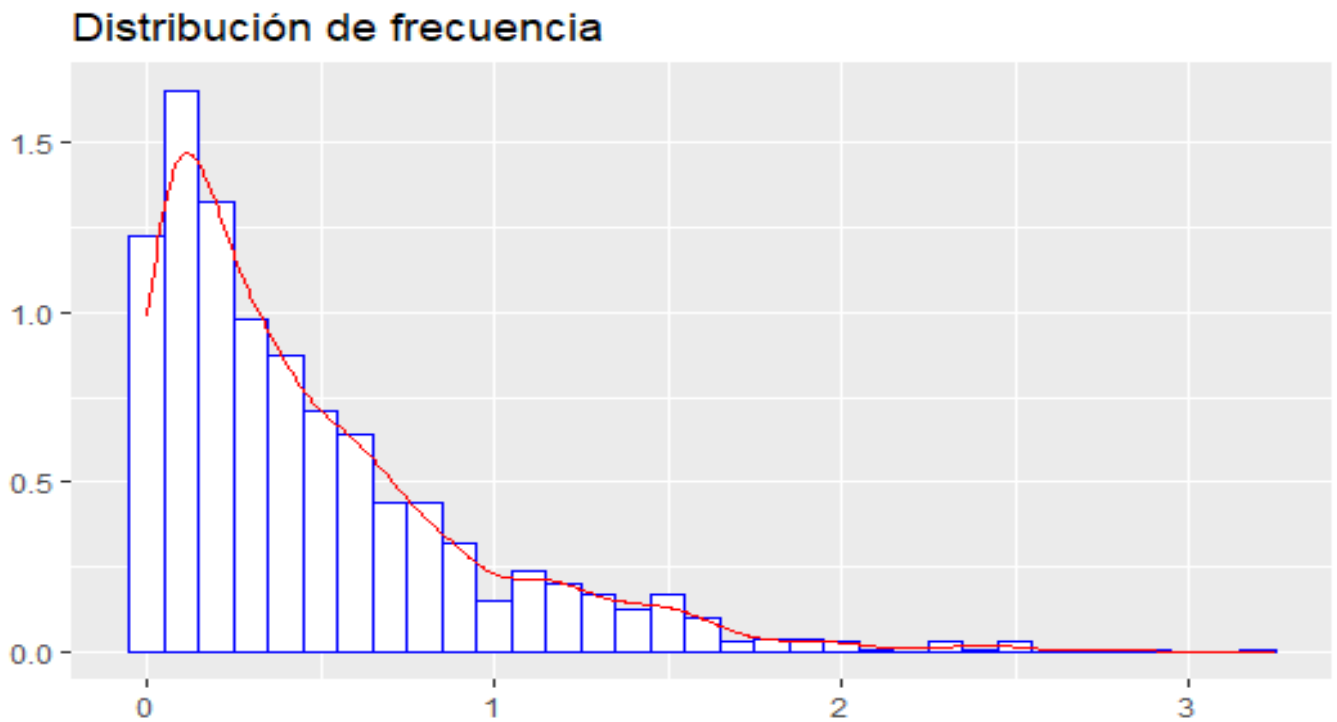
Curvas simétricas o acampanadas

Se caracterizan por el hecho de que las observaciones equidistantes del máximo central tienen la misma frecuencia.



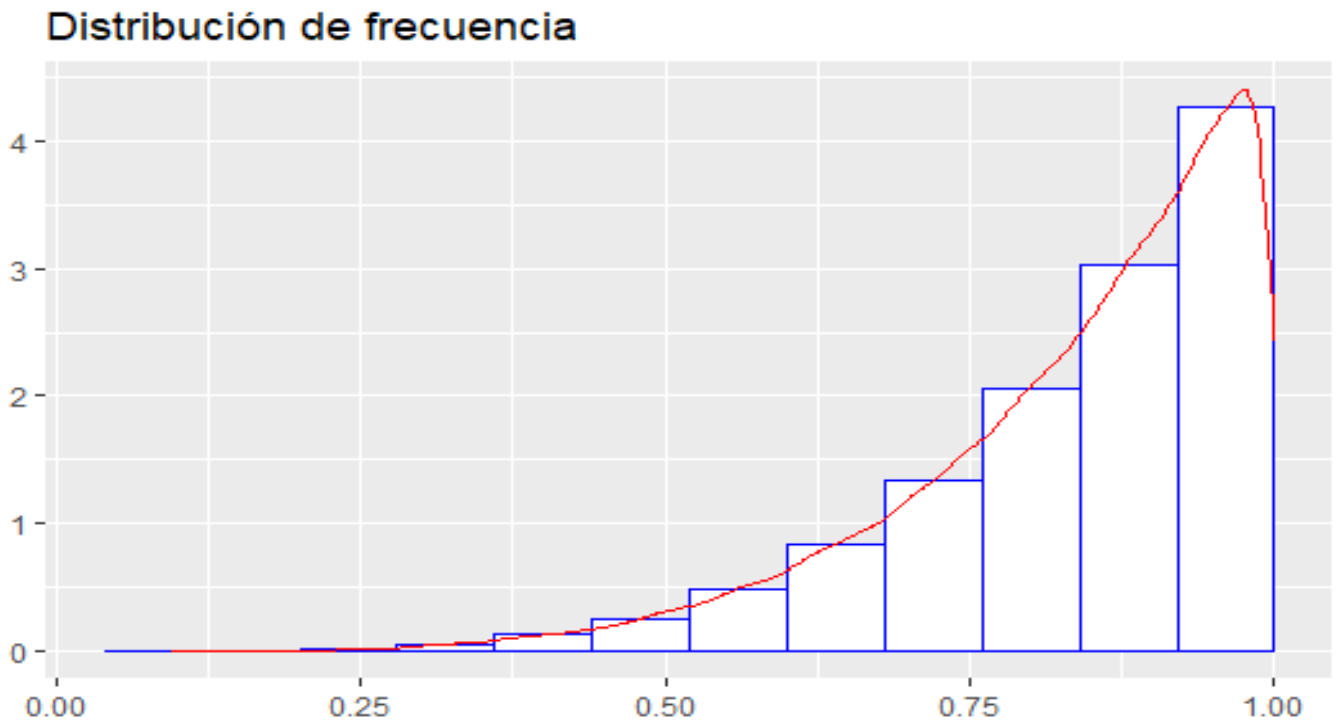
Curva asimétrica a la derecha (asimetría positiva)

Decimos que hay asimetría positiva (o a la derecha) si la "cola" a la derecha de la media es más larga que la de la izquierda, es decir, si hay valores más separados de la media a la derecha.



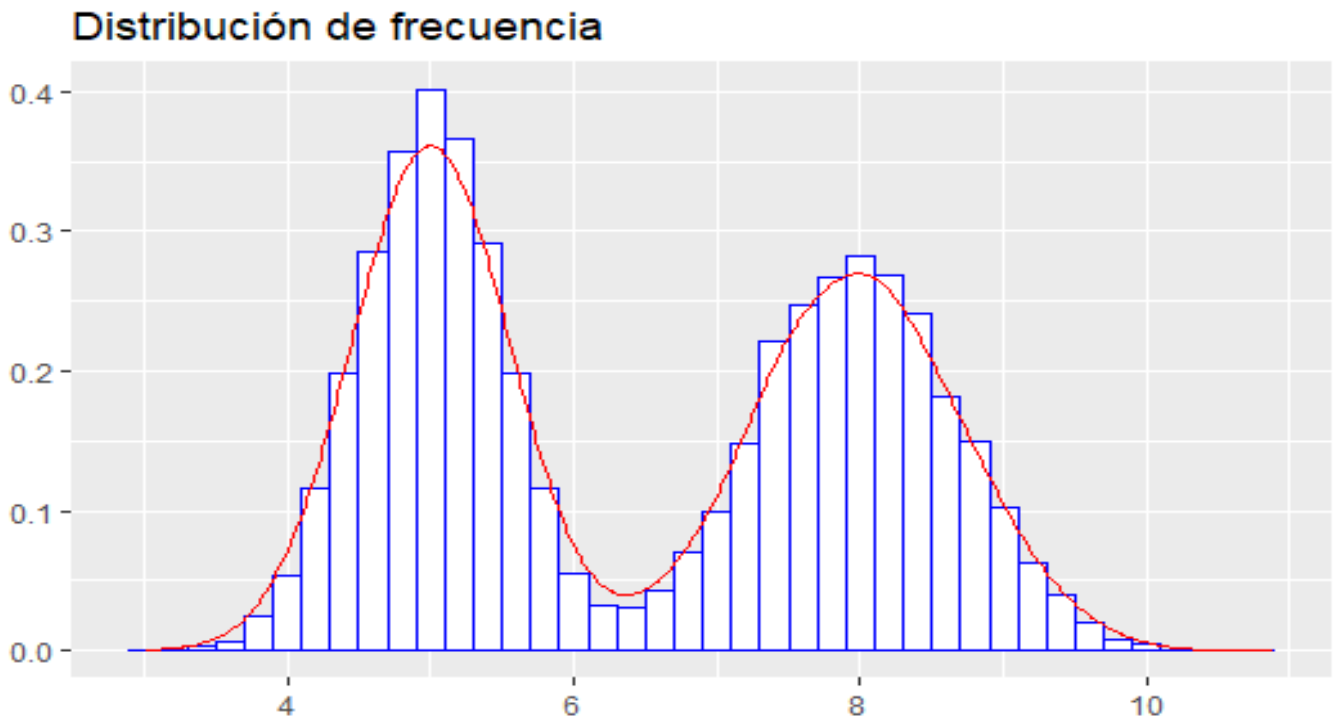
Curva asimétrica a la izquierda (asimetría negativa)

Si la "cola" a la izquierda de la media es más larga que la de la derecha, es decir, si hay valores más separados de la media a la izquierda.



Curva Bimodal

Una curva de frecuencia bimodal (g) tiene dos máximos.



Parámetros estadísticos

Los parámetros estadísticos de una muestra son valores que describen ciertas características de los datos recogidos en la muestra. Estos parámetros se utilizan para hacer inferencias sobre la población de la cual se ha extraído la muestra.

Estos parámetros estadísticos permiten resumir y entender mejor las características principales de una muestra, facilitando así el análisis y la interpretación de los datos.

Parámetros estadísticos de centralidad

Son medidas que describen el punto central o típico de una distribución de datos. Estos parámetros son fundamentales para entender la tendencia central de los datos, es decir, la localización de su "centro". Los principales parámetros de centralidad son:

- Media (Media aritmética).
- Mediana.
- Moda.

Media aritmética

Se utiliza para representar el valor promedio de un conjunto de datos. Es una medida que proporciona una idea general del "centro" de los datos.

- Para una población:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} X_i$$

- Para una muestra:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

La media es muy sensible a los valores extremos (outliers).

Para los datos 1, 2, 3, 4, 5, la media es $\bar{x} = \frac{1+2+3+4+5}{5} = 3$

Para 1, 2, 3, 4, 100, la media es $\bar{x} = \frac{1+2+3+4+5+100}{6} = 22$

Mediana

Es el valor que divide la muestra en dos partes iguales cuando los datos están ordenados de menor a mayor. Si el número de observaciones es impar, la mediana es el valor central. Si es par, la mediana es el promedio de los dos valores centrales.

Para los datos 1, 2, 3, 4, 5, la mediana es 3.

La mediana es una medida robusta de la tendencia central porque no se ve afectada por los valores extremos.

Para los datos 1, 2, 3, 4, 100, la mediana es 3.

Moda

Es el valor que se repite con mayor frecuencia en la muestra. Una distribución de datos puede tener más de una moda (bimodal, multimodal) o no tener ninguna moda si todos los valores son únicos.

Para los datos 1, 2, 2, 3, 4, la moda es 2.

Para los datos 1, 1, 2, 3, 3, las modas son 1 y 3 (bimodal).

Utilización

- **Media:** Se usa cuando se desea una medida que considere todos los valores de la muestra y los datos no contienen valores extremos significativos.
- **Mediana:** Es útil cuando los datos contienen outliers o están sesgados, ya que proporciona una representación más robusta de la tendencia central.
- **Moda:** Es especialmente útil para datos categóricos y para identificar el valor más común en un conjunto de datos.

Peso de 7 individuos

Individuo	1	2	3	4	5	6	7	8
Peso(kg)	63	52	78	49	71	62	68	52

Calcule los siguientes parámetros estadísticos de centralidad:

- Media
- Moda
- Mediana

Parámetros estadísticos de dispersión

- Son medidas que describen cómo se distribuyen los datos alrededor de una medida central (como la media).

- Estos parámetros nos proporcionan información sobre la variabilidad o la dispersión de los datos en un conjunto

Varianza

Es la media de los cuadrados de las diferencias de cada dato respecto a la media del conjunto.

- Para una población: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2$
- Para una muestra: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

Dada la muestra $\{2, 4, 6\}$

$\bar{X} = \frac{2+4+6}{3} = 4$

$$s^2 = \frac{1}{2}\left[(2-4)^2 + (4-4)^2 + (6-4)^2\right] = \frac{1}{2}[4+0+4] = 4.$$

Interpretación de la Varianza

- Un valor de varianza grande indica que los datos están muy dispersos alrededor de la media. Un valor de varianza pequeño indica que los datos están más concentrados cerca de la media.

- La varianza se expresa en las unidades al cuadrado de los datos originales, lo que puede hacerla difícil de interpretar en comparación directa con los datos originales.

Desviación estándar

Es la raíz cuadrada de la varianza, proporcionando una medida de dispersión en las mismas unidades que los datos originales.

- Para una población:

$$\sigma = \sqrt{\sigma^2}$$

- Para una muestra:

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

Más fácil de interpretar que la varianza porque está en las mismas unidades que los datos originales.

Coeficiente de variación (CV)

la relación entre la desviación estándar y la media, expresada como un porcentaje. Mide la dispersión relativa de los datos.

- Para una población:

$$CV = \frac{\sigma}{\mu} \times 100$$

- Para una muestra:

$$CV = \frac{\sigma}{\bar{x}} \times 100$$

Permite comparar la dispersión entre diferentes conjuntos de datos con distintas unidades o medias.

Peso de 7 individuos

Individuo	1	2	3	4	5	6	7	8
Peso(kg)	63	52	78	49	71	62	68	52

Calcule los siguientes parámetros estadísticos de dispersión:

- Varianza
- Desviación Estándar
- CV

Parámetros estadísticos de posición

Son valores que indican la ubicación relativa de un dato dentro de un conjunto de datos. Estas medidas permiten describir la posición de una observación en relación con el resto de los datos y se utilizan para hacer comparaciones y análisis detallados de la distribución.

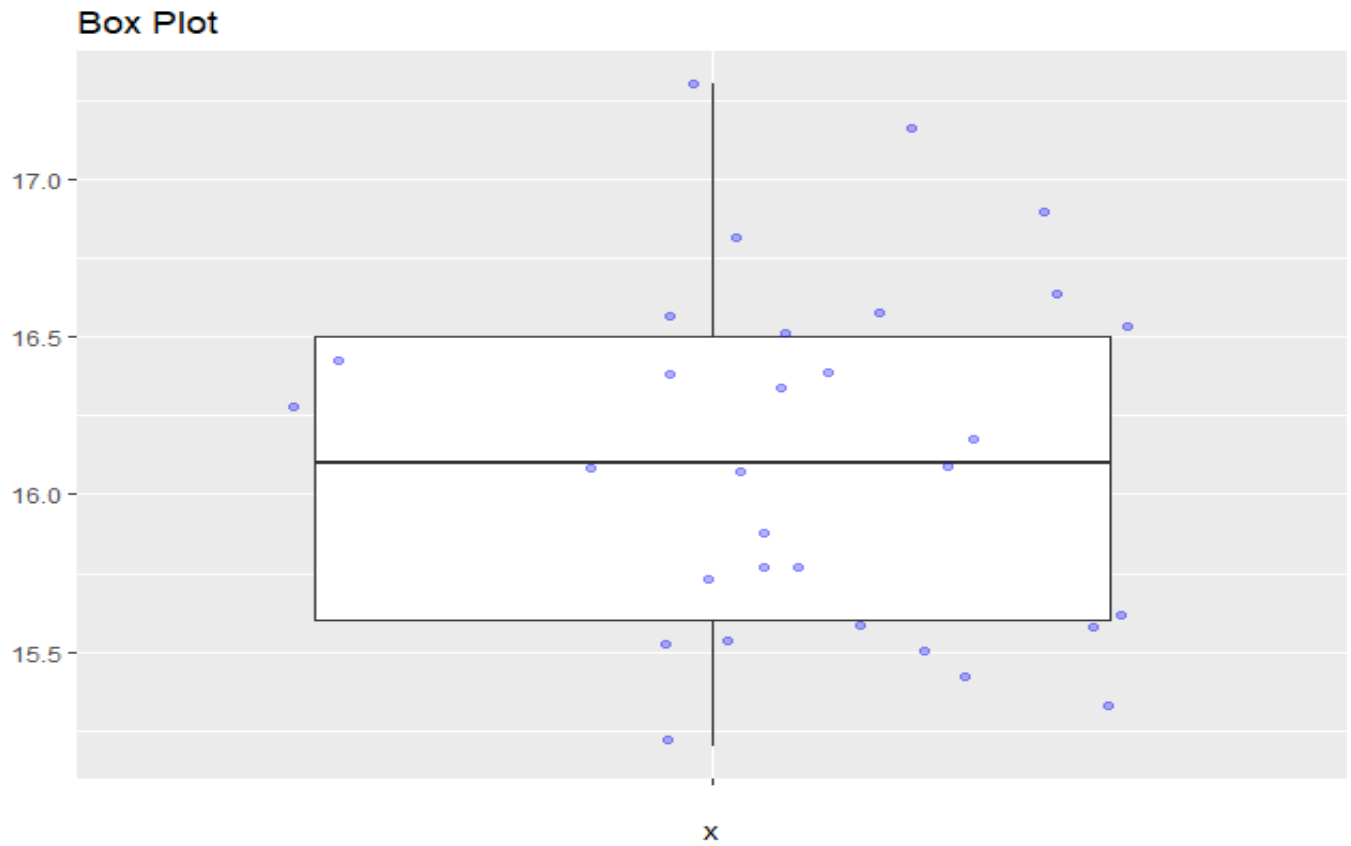
Cuartiles

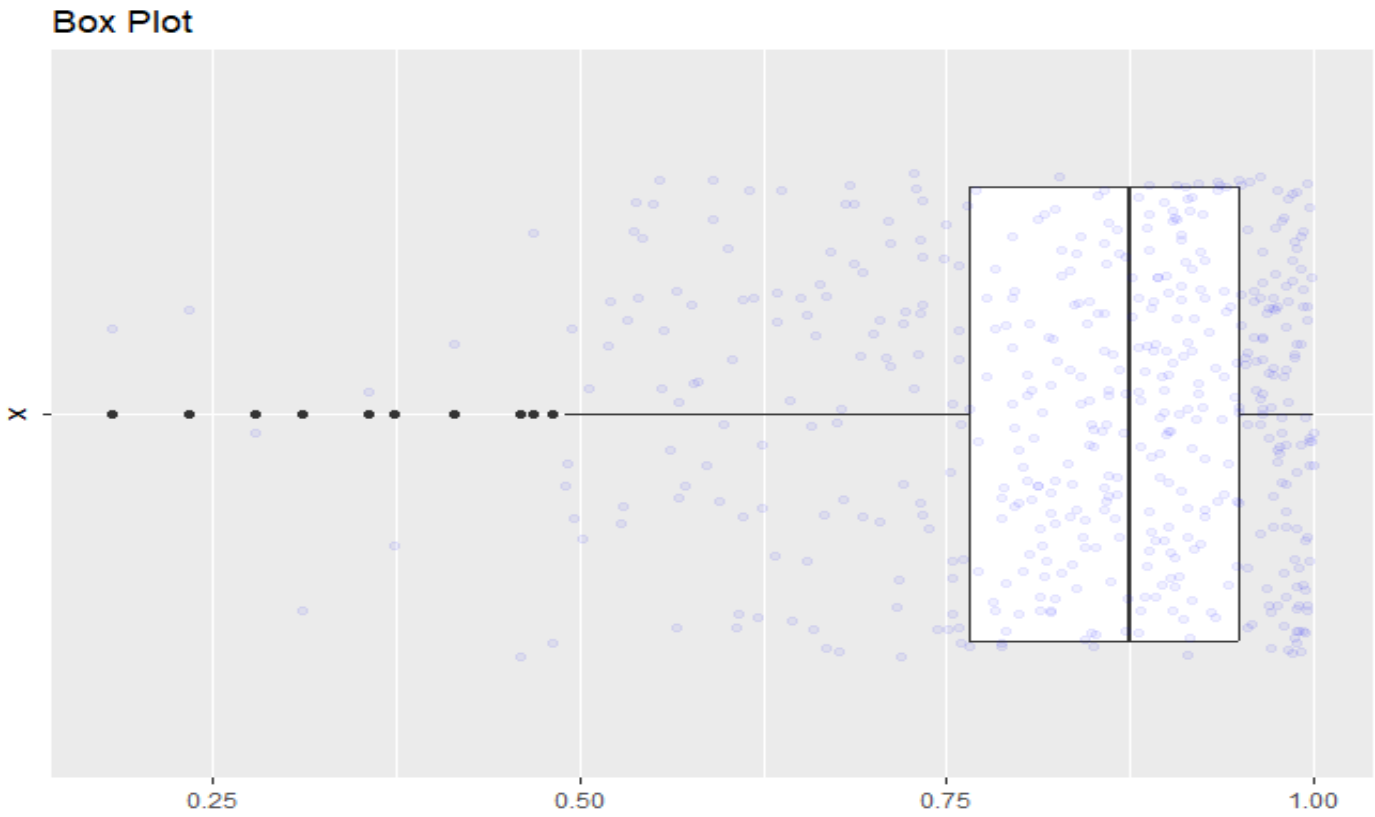
Dividen un conjunto de datos ordenado en cuatro partes iguales.

- **Primer cuartil (Q1):** Es el valor por debajo del cual se encuentra el 25% de los datos.
- **Segundo cuartil (Q2):** Es la mediana.
- **Tercer cuartil (Q3):** Es el valor por debajo del cual se encuentra el 75% de los datos.

Cuartiles para el perímetro de los duraznos

0%	25%	50%	75%	100%
15.2	15.6	16.1	16.5	17.3




Box Plot
x
0.25
0.50
0.75
1.00

Rango intercuartílico (IQR)

Es la diferencia entre el tercer cuartil (Q3) y el primer cuartil (Q1). Mide la dispersión de la mitad central de los datos.

$$IQR = Q3 - Q1$$

$$IQR = 16.5 - 15.6 = 0.9$$

No se ve afectado por valores atípicos y proporciona una mejor medida de dispersión para datos no simétricos.

Parámetros estadísticos de forma

Medidas que describen la distribución de un conjunto de datos, más allá de sus medidas de tendencia central y dispersión.

- Asimetría (Skewness).
- Curtosis (Kurtosis):

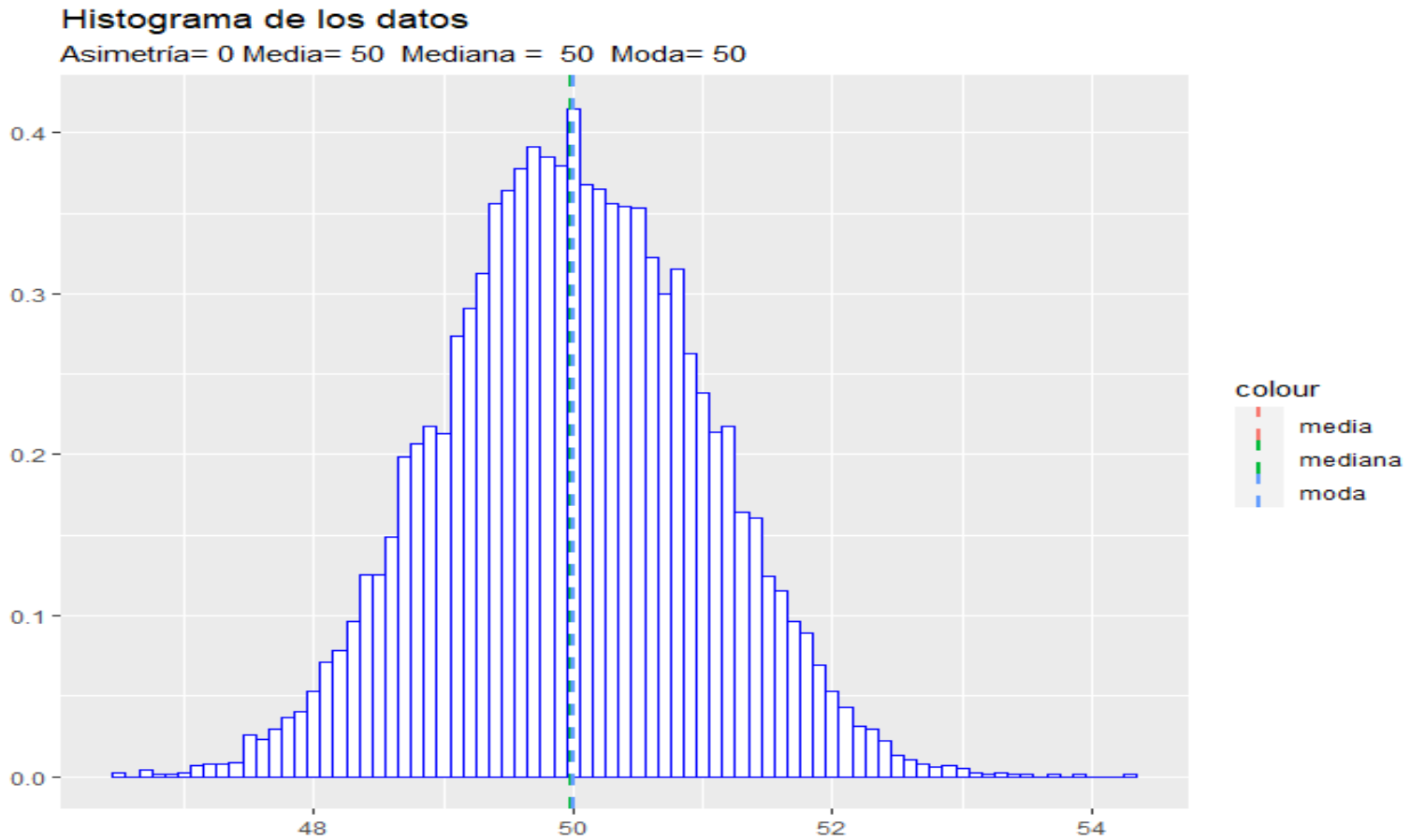
Asimetría (Skewness)

La asimetría mide la simetría de la distribución de los datos.

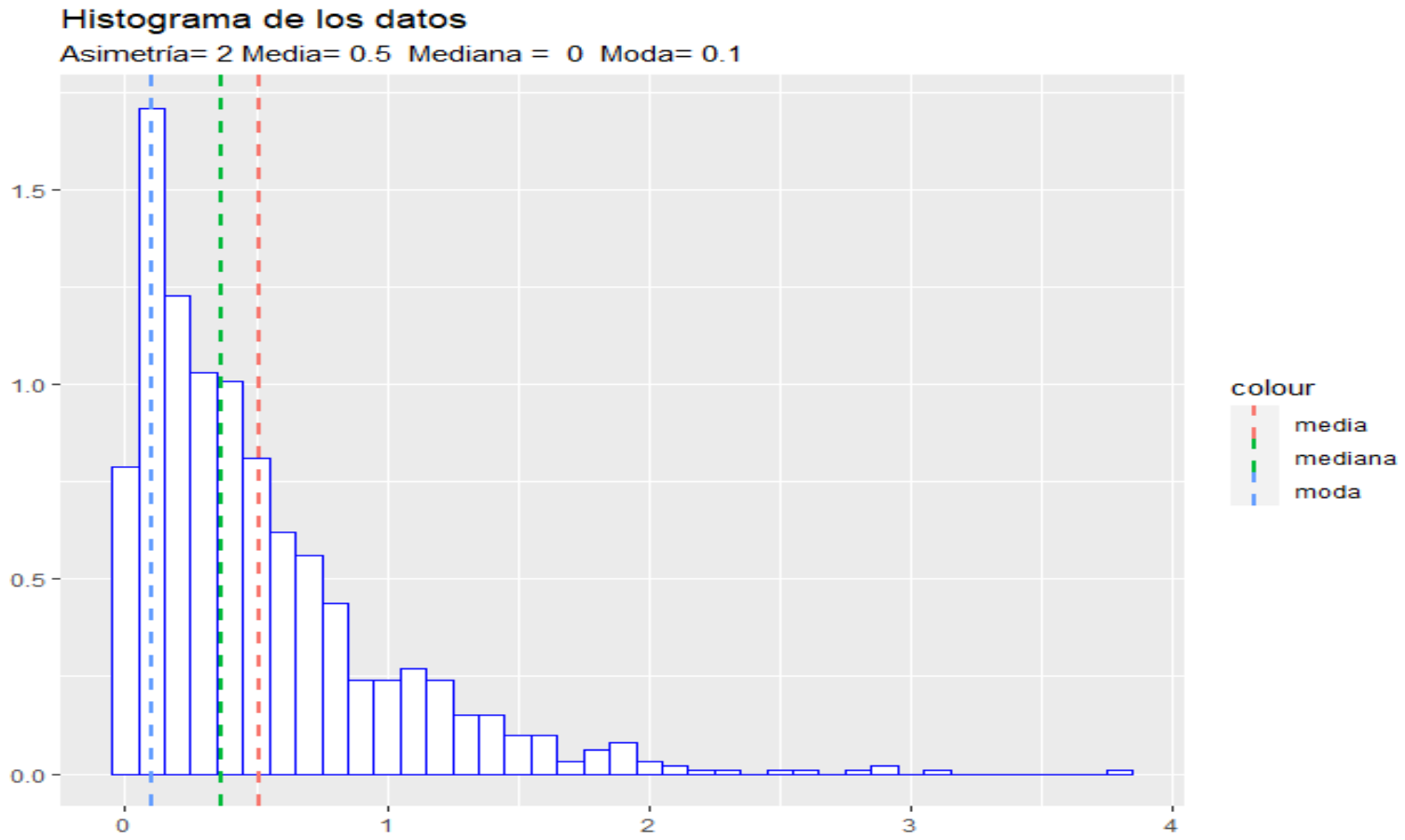
$$\text{Asimetría} = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3$$

- Asimetría $= 0$: Distribución simétrica
- Asimetría > 0 : Distribución asimétrica a la derecha
- Asimetría < 0 : Distribución asimétrica a la izquierda

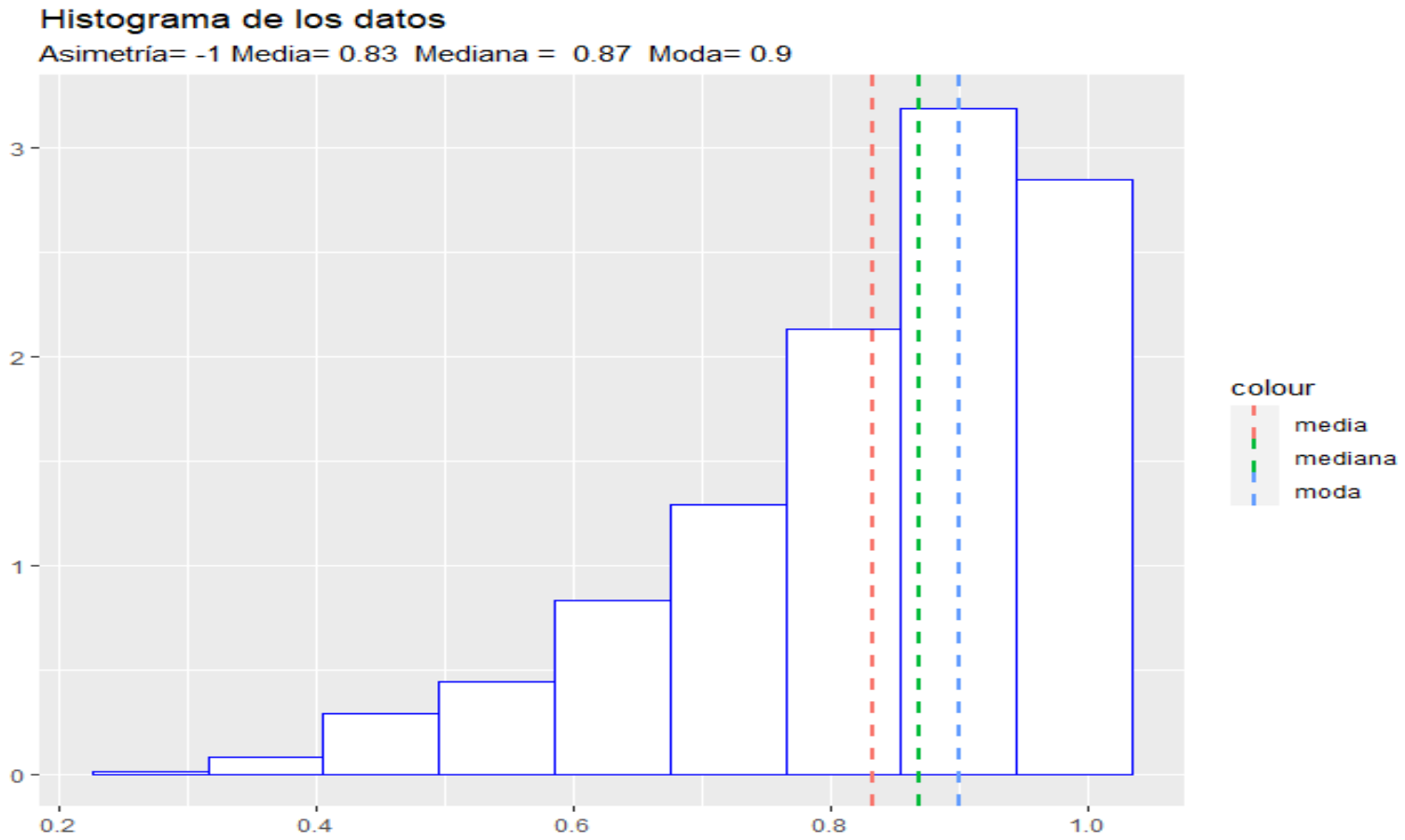
Distribución Simétrica



Distribución Asimétrica a la Derecha



Distribución asimétrica a la izquierda



Curtosis (Kurtosis)

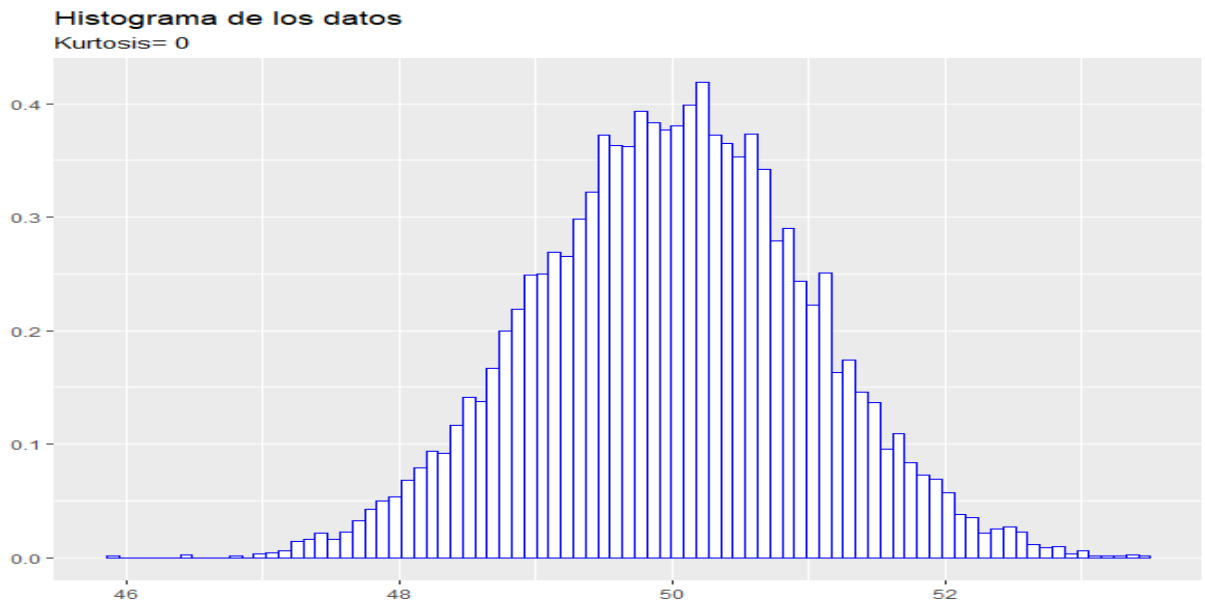
La curtosis mide el "apuntamiento" de la distribución de los datos. Indica la concentración de valores alrededor de la media y la presencia de valores extremos en la distribución de datos.

$$\text{Curtosis} = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

- Kurtosis $= 0$: Mesocúrtica
- Kurtosis > 0 : Leptocúrtica
- Kurtosis < 0 : Platicúrtica

Mesocúrtica

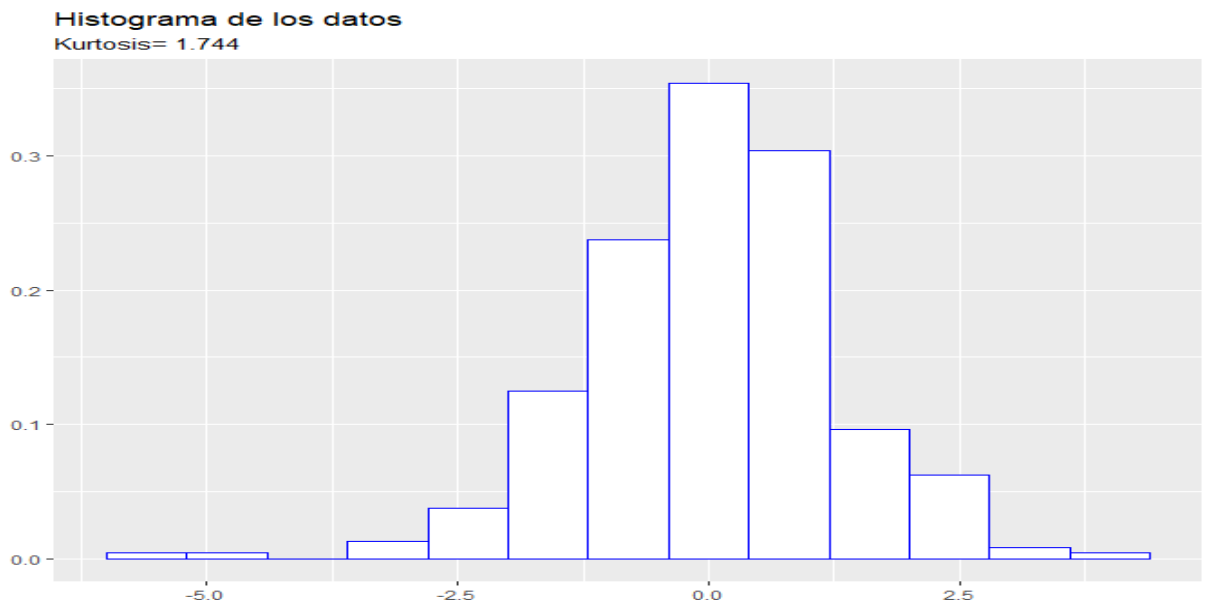
Presenta un grado de concentración medio alrededor de los valores centrales de la variable.



La curtosis se aproxima a cero para una distribución mesocúrtica

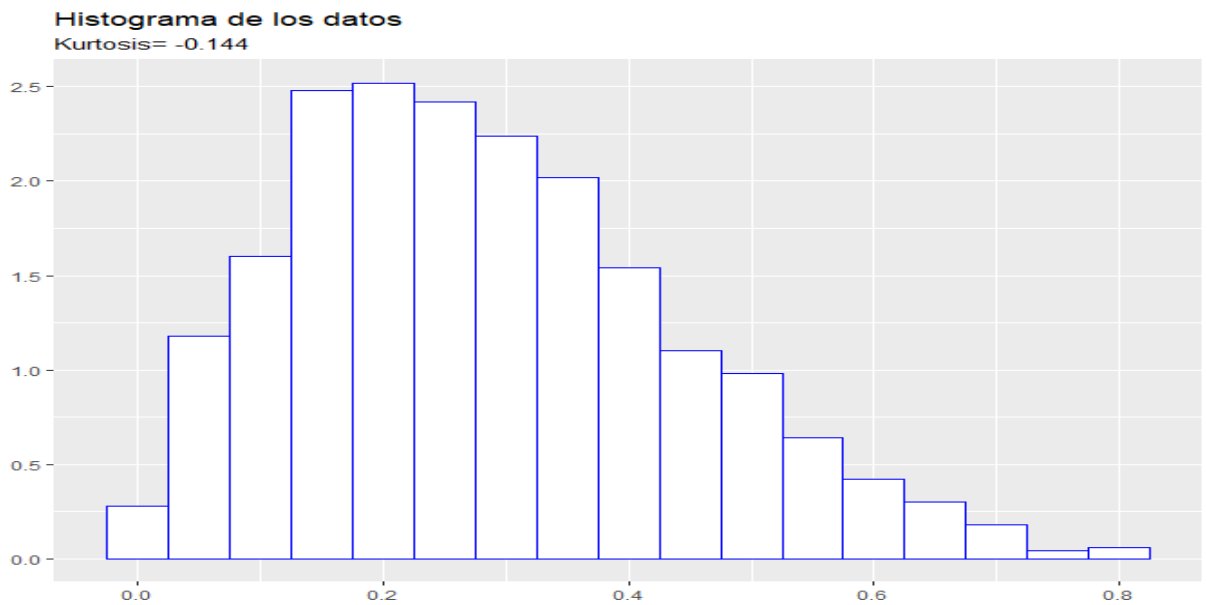
Leptocúrtica

Presenta un elevado grado de concentración alrededor de los valores centrales de la variable.Las colas de la distribución son más pesadas, lo que significa que los valores extremos son más probables.



Platicúrtica

Presenta un reducido grado de concentración alrededor de los valores centrales de la variable y sus valores están más dispersos. Tienen una forma achatada con un pico más bajo en el centro y colas menos pronunciadas.

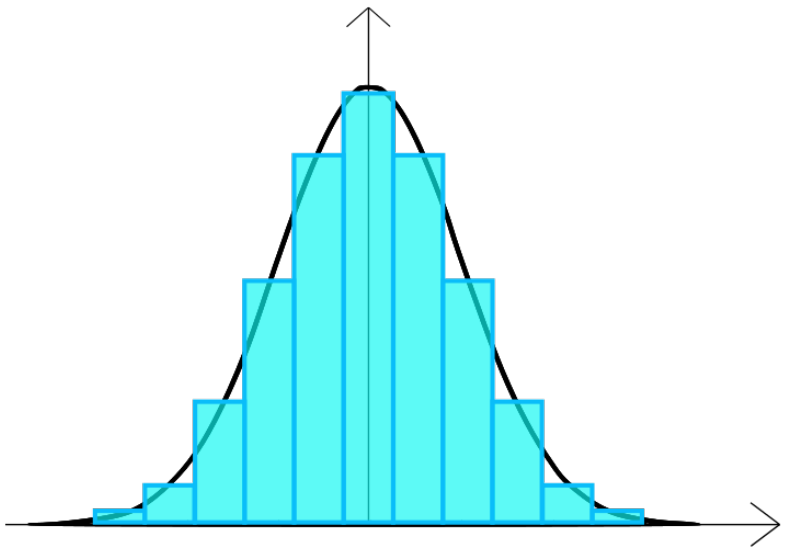


¿Qué es una Distribución de Probabilidades?

- Una distribución de probabilidades describe cómo se distribuyen los valores de una variable aleatoria.
- Existen dos tipos principales:
  - **Distribuciones Discretas**
  - **Distribuciones Continuas**
- Las distribuciones de probabilidades se utilizan para modelar fenómenos y realizar predicciones.

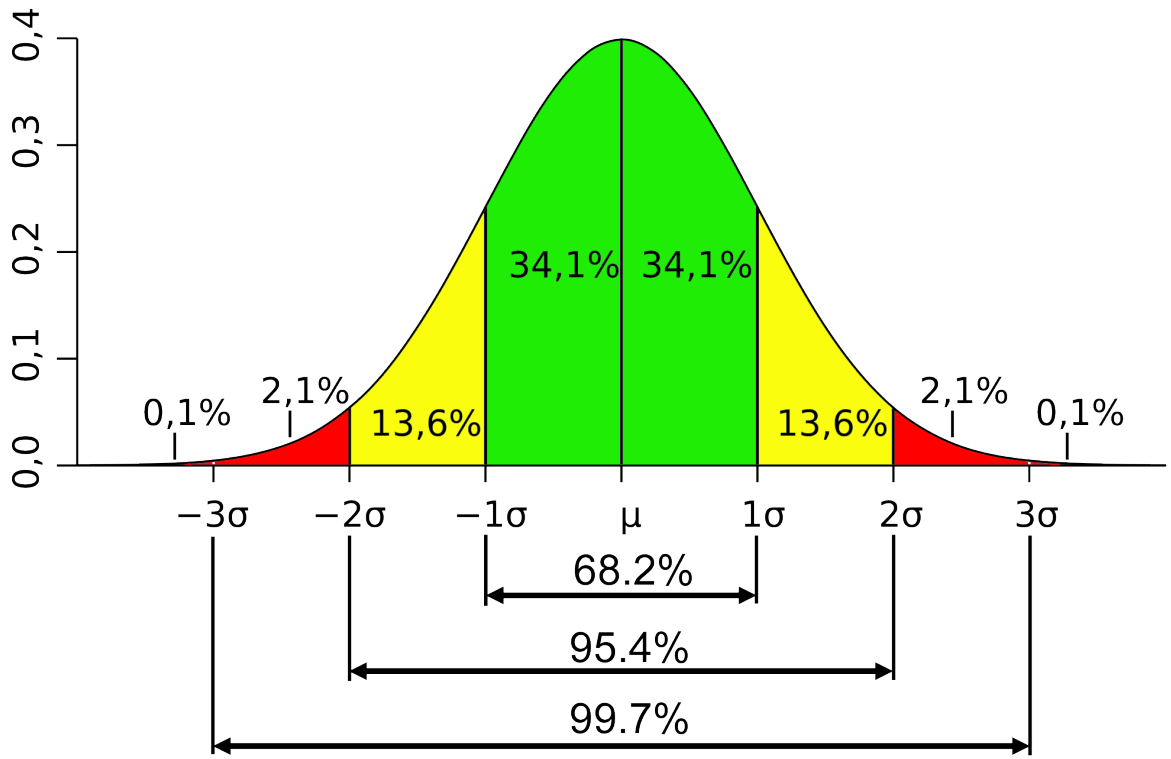
¿Qué es una Distribución Normal?

- También conocida como distribución de Gauss.
- Es una de las distribuciones de probabilidad más importantes en estadística.
- Se utiliza para modelar fenómenos naturales y sociales.



Características de la Distribución Normal

- **Simetría**: La curva es simétrica respecto a la media.
- **Campana**: Tiene forma de campana.
- **Media, mediana y moda**: Son iguales y se encuentran en el centro de la distribución.
- **Desviación estándar**: Determina el ancho de la curva.



Función de Densidad de Probabilidad

La función de densidad de probabilidad (pdf) de una distribución normal se define como:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- μ es la media.
- σ es la desviación estándar.
- La curva se integra a 1, es decir, $\int_{-\infty}^{\infty} f(x)\, dx = 1$.