Estadística

Pablo de la Cuesta García

2024-01-08

Contents

1 Introducción a la Probabilidad **2**

1.1 Probabilidad . . . 2

1.1.1 Definición clásica . . . 3

1.1.2 Definición frecuentista . . . 3

1.1.3 Definición axiomática . . . 3

1.2 Eventos . . . 3

1.3 Operaciones con eventos . . . 4

1.4 Leyes de Conjuntos en Probabilidad . . . 4

1.4.1 Ejemplo . . . 4

1.5 Tipos de datos . . . 5

1.5.1 Datos Cualitativos . . . 5

1.5.2 Datos Cuantitativos . . . 6

1.6 Tablas de frecuencias . . . 6

1.6.1 Componentes de una tabla de frecuencias . . . 7

1.6.2 Ejemplo . . . 7

1.7 Probabilidad condicionada . . . 7

1.7.1 Definición formal . . . 8

1.7.2 Carecterísticas . . . 8

1.7.3 Propiedades . . . 8

1.7.4 Ejemplo . . . 8

1.8 Teorema de Bayes . . . 9

1.8.1 Definición formal . . . 9

1.8.2 Ejemplos . . . 9

1.9 Variables aleatorias discretas . . . 10

1.9.1 Función de probabilidad . . . 10

1.9.2 Función de distribución acumulada . . . 10

1.9.3 Ejemplo 11
1.10 Variables aleatorias continuas 11
1.10.1 Función de Densidad de Probabilidad 11

2 Estadística Descriptiva **11**
2.1 Medidas de tendencia central 12
2.1.1 Media 12
2.1.2 Mediana 12
2.1.3 Moda 13
2.2 Medidas de dispersión 13
2.2.1 Rango 13
2.2.2 Varianza 13
2.2.3 Desviación estándar 14
2.3 Visualización de datos 14
2.3.1 Visualización de datos cualitativos (una variable) 14
2.3.2 Visualización de datos cualitativos (dos variables) 18
2.3.3 Visualización de datos cuantitativos 19
2.3.4 Visualización de datos cuantitativos (dos variables) 21

1 Introducción a la Probabilidad

Para comenzar nuestro estudio de la estadística es necesario introducir y entender algunos conceptos básicos de probabilidad. La relacion entre probabilidad y estadística es muy estrecha, ya que la estadística se basa en la probabilidad para realizar inferencias sobre los datos. Es decir, la probabilidad proporciona el marco teórico para la estadística. La estadística, en esencia, es la aplicación de la probabilidad a los datos.

Mientras que el objetivo de la estadística descriptiva es resumir y describir los datos, el objetivo de la estadística inferencial es realizar inferencias sobre los datos. Por ejemplo, estimar la media de una población a partir de una muestra. Por tanto, la estadística inferencial utiliza modelos de probabilidad para realizar inferencias sobre los datos.

1.1 Probabilidad

La probabilidad es una medida que cuantifica la incertidumbre asociada a un evento o fenómeno. Se expresa como un número entre 0 y 1, donde 0 indica imposibilidad y 1 indica certeza absoluta.

La probabilidad se puede definir de tres formas diferentes:

- Definición clásica
- Definición frecuentista
- Definición axiomática

1.1.1 Definición clásica

La probabilidad clásica se define como la razón entre el número de resultados favorables y el número total de resultados posibles en un experimento aleatorio, asumiendo que todos los resultados son igualmente probables.

Características:

- Ω es finito: $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ y equiprobable $P(\omega_i) = 1/n$.
- $A \subset \Omega$ es un suceso.
- $P(A) = \frac{\text{casos favorables}}{\text{casos posibles}} = \frac{\text{casos favorables}}{\text{casos favorables+casos desfavorables}}$.

1.1.2 Definición frecuentista

La interpretación frecuentista de la probabilidad define la probabilidad de un evento como el límite de su frecuencia relativa en una serie de ensayos que se repiten infinitamente.

Características:

- Ω es infinito: $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ y equiprobable $P(\omega_i) = 1/n$.
- $A \subset \Omega$ es un suceso.
- $P(A) = \lim_{n\to\infty} \frac{\text{casos favorables}}{\text{casos posibles}} = \lim_{n\to\infty} \frac{\text{casos favorables}}{\text{casos favorables+casos desfavorables}}$.

1.1.3 Definición axiomática

La probabilidad axiomática se basa en un conjunto de axiomas propuestos por Andrey Kolmogorov en 1933. Estos axiomas establecen las propiedades fundamentales que debe cumplir cualquier medida de probabilidad.

- **Axioma 1**: La probabilidad de cualquier evento es un número no negativo $0 \leq P(A) \leq 1$.
- **Axioma 2**: La probabilidad del espacio muestral completo es 1. $P(\Omega) = 1$.
- **Axioma 3**: Para cualquier secuencia de eventos mutuamente excluyentes (no pueden ocurrir al mismo tiempo), la probabilidad de que ocurra alguno de los eventos es igual a la suma de sus probabilidades individuales. $P(A \cup B) = P(A) + P(B)$. O expresado de otra forma, $P(\cup A_i) = \Sigma P(A_i)$.

Estos axiomas permiten construir la teoría de la probabilidad de manera rigurosa y coherente, y son la base para la mayoría de las teorías modernas de probabilidad.

1.2 Eventos

Un evento es una colección de posibles resultados de un experimento aleatorio. En otras palabras, es un subconjunto del espacio muestral, que es el conjunto de todos los posibles resultados del experimento. Para poder medir la probabilidad de un evento, es necesario definir varios conceptos básicos:

- **Experimento aleatorio**. Es un experimento que se puede repetir en las mismas condiciones y que puede tener varios resultados posibles. Por ejemplo, lanzar un dado.
- **Espacio muestral**. Es el conjunto de todos los posibles resultados de un experimento aleatorio. Se denota por Ω.
- **Evento o suceso**. Es un subconjunto del espacio muestral. Se denota por A.
- **Evento o suceso elemental**. Es un elemento del espacio muestral. Se denota por ω.
- **Evento o suceso seguro**. Es el suceso que contiene todos los elementos del espacio muestral. Se denota por Ω.
- **Evento o suceso imposible**. Es el suceso que no contiene ningún elemento del espacio muestral. Se denota por $\emptyset$.

1.3 Operaciones con eventos

- **Unión**. Se denota por $A \cup B$ y es el suceso que contiene todos los elementos de A y B.
- **Intersección**. Se denota por $A \cap B$ y es el suceso que contiene los elementos comunes de A y B.
- **Complementario**. Se denota por A^c y es el suceso que contiene todos los elementos del espacio muestral que no están en A.
- **Diferencia**. Se denota por $A \setminus B$ y es el suceso que contiene los elementos de A que no están en B.

1.4 Leyes de Conjuntos en Probabilidad

- **Unión**. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- **Intersección**. $P(A \cap B) = P(A) + P(B) - P(A \cap B)$
- **Complementario**. $P(A^c) = 1 - P(A)$
- **Probabilidad total**. $P(A) = P(A \cap B) + P(A \cap B^c)$

1.4.1 Ejemplo

Imagina que lanzamos dos monedas. El espacio muestral Ω es {CC, CS, SC, SS}, donde C representa cara y S sello.

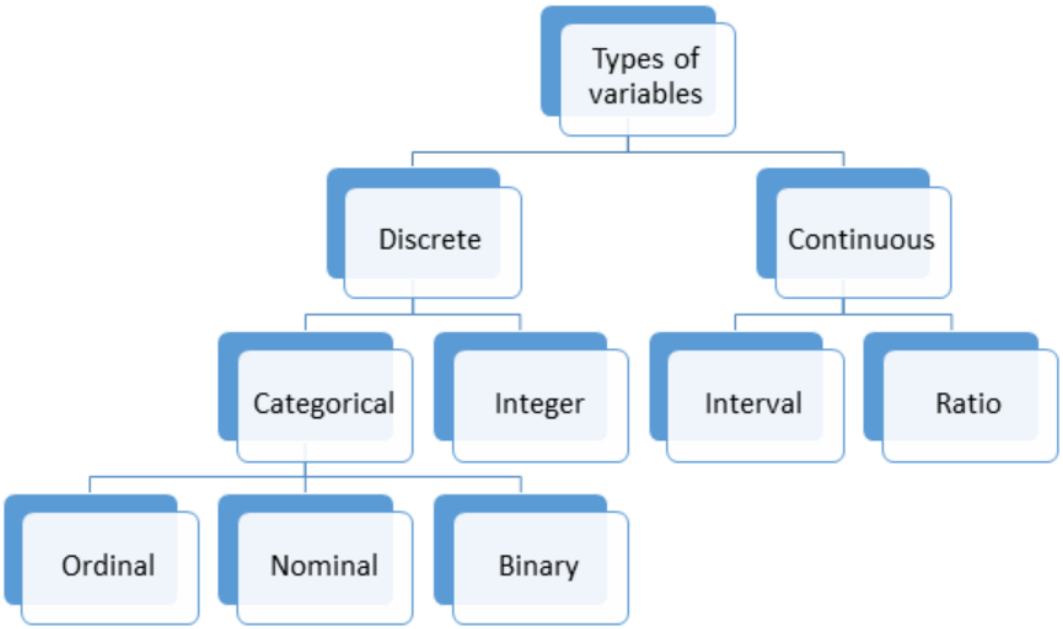
- Evento A: al menos obtenemos una cara $A = CC, CS, SC$.
- Evento B: al menos obtenemos un sello $B = CS, SC, SS$.

Por tanto:

- $P(A) = 3/4$
- $P(B) = 3/4$
- $P(A \cap B) = 1/2$
- $P(A \cup B) = 1$

Usando la ley de la suma: $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 3/4 + 3/4 - 1/2 = 1$.

1.5 Tipos de datos



1.5.1 Datos Cualitativos

Los datos cualitativos o categóricos describen cualidades o características. No se expresan en números y por lo general se clasifica en categorías. Por ejemplo, el color de los ojos, el género, el estado civil, etc.

Dentro de los datos cualitativos, podemos distinguir entre:

- **Nominales**. No existe un orden entre las categorías. Por ejemplo, el color de los ojos.
- **Ordinales**. Existe un orden entre las categorías. Por ejemplo, el nivel de satisfacción de un cliente.

Este código crea un vector con colores de camisetas, un ejemplo típico de datos nominales donde cada color es una categoría sin un orden inherente.

```r
# Crear un vector con datos nominales
colores_camisetas <- c("rojo", "azul", "verde", "azul", "rojo")

# Tabla de frecuencias
table(colores_camisetas)
```

```
## colores_camisetas
##  azul  rojo verde
##     2     2     1
```

Aquí, `niveles_educacion` es un factor con orden. Los niveles de educación tienen un orden inherente: primaria, secundaria y universitaria.

```r
# Crear un factor con datos ordinales
niveles_educacion <- factor(c("primaria", "universitaria", "secundaria", "secundaria", "universitaria")
                            levels = c("primaria", "secundaria", "universitaria"),
```

```
                            ordered = TRUE)

# Tabla de frecuencia
table(niveles_educacion)
```

```
## niveles_educacion
##      primaria    secundaria universitaria
##             1             2             2
```

1.5.2 Datos Cuantitativos

Los datos cuantitativos describen cantidades. Se expresan en números y se pueden realizar operaciones aritméticas con ellos. Por ejemplo, la edad, el peso, la altura, etc.

Dentro de los datos cuantitativos, podemos distinguir entre:

- **Discretos**. Los valores posibles son numerables. Por ejemplo, el número de hijos.
- **Continuos**. Los datos continuos pueden tomar cualquier valor en un intervalo especifi

Este ejemplo muestra un vector con el número de estudiantes en diferentes clases, un claro ejemplo de datos discretos que son contables.

```
# Crear un vector con datos discretos
numero_estudiantes <- c(25, 30, 22, 28, 31)

# Calcular estadísticas básicas
summary(numero_estudiantes)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    22.0    25.0    28.0    27.2    30.0    31.0
```

En este caso, altura_estudiantes representa un conjunto de datos continuos, ya que la altura puede tomar cualquier valor en un rango y puede ser tan precisa como se desee.

```
# Crear un vector con datos continuos
altura_estudiantes <- c(1.70, 1.65, 1.80, 1.75, 1.60)

# Calcular estadísticas básicas
summary(altura_estudiantes)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.60    1.65    1.70    1.70    1.75    1.80
```

1.6 Tablas de frecuencias

Las tablas de frecuencias son una herramienta estadística fundamental para organizar y resumir datos. Son especialmente útiles para entender la distribución de los datos en un conjunto y para la visualización de datos categóricos o numéricos discretos.

Por tanto, una tabla de frecuencia es una representación organizada de los datos que muestra la frecuencia que cada valor de un conjunto de datos ocurre. Básicamente, clasifica los valores de un conjunto de datos en categorías y proporciona el número de observaciones que pertenecen a cada categoría.

1.6.1 Componentes de una tabla de frecuencias

- **Categorías**. Son los valores únicos de un conjunto de datos. Por ejemplo, los colores de las camisetas.
- **Frecuencia**. Es el número de observaciones que pertenecen a cada categoría. Por ejemplo, el número de camisetas de cada color.
  - **Frecuencia absoluta**. Es el número de observaciones que pertenecen a cada categoría.
  - **Frecuencia relativa**. Es la proporción de observaciones que pertenecen a cada categoría.
  - **Frecuencia acumulada**. Es el número de observaciones que pertenecen a cada categoría y a todas las categorías anteriores.

1.6.2 Ejemplo

Utilizaremos un ejemplo de calificaciones de estudiantes en R para ilustrar cómo crear una tabla de frecuencias. El conjunto de datos `calificaciones` contiene las calificaciones de 10 estudiantes en un examen de matemáticas.

```r
# Crear el vector con las calificaciones
calificaciones <- c(5, 7, 8, 5, 3, 7, 6, 5, 9, 8)

# Calcular la frecuencia absoluta
frecuencia_absoluta <- table(calificaciones)

# Calcular la frecuencia relativa
frecuencia_relativa <- prop.table(frecuencia_absoluta)

# Frecuencias acumuladas
frecuencia_acumulada <- cumsum(table(calificaciones))

# Representación de las frecuencias
tabla_frecuencias <- data.frame(
  "Calificación" = names(frecuencia_absoluta),
  "Frecuencia Absoluta" = as.integer(frecuencia_absoluta),
  "Frecuencia Relativa" = as.numeric(frecuencia_relativa),
  "Frecuencia Acumulada" = as.integer(frecuencia_acumulada)
)
colnames(tabla_frecuencias) <- c("Calificación", "Frecuencia Absoluta", "Frecuencia Relativa", "Frecuen
knitr::kable(tabla_frecuencias, caption = "Frecuencia Absoluta de Calificaciones")
```

Table 1: Frecuencia Absoluta de Calificaciones

Calificación	Frecuencia Absoluta	Frecuencia Relativa	Frecuencia Acumulada
3	1	0.1	1
5	3	0.3	4
6	1	0.1	5
7	2	0.2	7
8	2	0.2	9
9	1	0.1	10

1.7 Probabilidad condicionada

La probabilidad condicionada describe cómo la probabilidad de un evento A cambia cuando se conoce que otro evento B ha ocurrido.

1.7.1 Definición formal

La probabilidad condicionada de A dado B se denota como $P(A|B)$ y se define como:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

donde:

- $P(A \cap B)$ es la probabilidad de que ocurran ambos eventos A y B.
- $P(B)$ es la probabilidad de que ocurra el evento B.

1.7.2 Carecterísticas

- **Actualización de información**: la probabilidad condicionada actualiza nuestras creencias sobre la probabilidad de un evento cuando se conoce que otro evento ha ocurrido. Es una medida de cómo un evento influye en la probabilidad de otro evento.
- **Rango**: Al igual que las probabilidades ordinarias, las probabilidades condicionadas están en el rango de 0 a 1.
- **Dependencia e Independencia**:
  - Si $P(A|B) = P(A)$, entonces los eventos A y B son independientes.
  - Si dos eventos son independientes, entonces $P(A \cap B) = P(A) \cdot P(B)$.
  - Si $P(A|B) \neq P(A)$, entonces los eventos A y B son dependientes.

1.7.3 Propiedades

- **Probabilidad Total**: Si se tiene un conjunto de eventos mutuamente excluyentes y colectivamente exhaustivos $B_1, B_2, \ldots, B_n$, entonces la probabilidad de un evento A se puede calcular como: $P(A) = \sum_{i=1}^{n} P(A|B_i) \cdot P(B_i)$.
- **Teorema de Bayes**: Si se tiene un conjunto de eventos mutuamente excluyentes y colectivamente exhaustivos $B_1, B_2, \ldots, B_n$, entonces la probabilidad de un evento B_j dado un evento A se puede calcular como: $P(B_j|A) = \frac{P(A|B_j) \cdot P(B_j)}{\sum_{i=1}^{n} P(A|B_i) \cdot P(B_i)}$.

1.7.4 Ejemplo

Imaginemos que tenemos un mazo de cartas (52 cartas en total). Queremos calcular la probabilidad de sacar un as dado que ya sabemos que la carta es de corazones. En este caso, A es el evento "sacar un as" y B es "sacar una carta de corazones".

- $P(A)$, hay 4 ases en el mazo, entonces: $P(A) = \frac{4}{52}$
- $P(B)$, hay 13 cartas de corazones, entonces: $P(B) = \frac{13}{52}$
- $P(A \cap B)$, solo hay 1 as de corazones, por lo que: $P(A \cap B) = \frac{1}{52}$

Utilizando la fórmula de la probabilidad condicionada:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{52}}{\frac{13}{52}} = \frac{1}{13}$$

Esto significa que la probabilidad de sacar un as, sabiendo que la carta es de corazones, es $\frac{1}{13}$.

1.8 Teorema de Bayes

El Teorema de Bayes, nombrado así en honor al reverendo Thomas Bayes, es un principio fundamental en el campo de la probabilidad y la estadística. Este teorema proporciona una forma de actualizar nuestras probabilidades a priori sobre un evento, basándonos en nueva evidencia o información. Es especialmente útil en situaciones donde las probabilidades condicionales son conocidas o pueden ser estimadas.

1.8.1 Definición formal

El Teorema de Bayes se expresa matemáticamente como:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

donde:

- $P(A|B)$ es la probabilidad posterior de A después de observar B.
- $P(B|A)$ es la probabilidad de B dado A.
- $P(A)$ es la probabilidad previa de A.
- $P(B)$ es la probabilidad previa de B.

1.8.2 Ejemplos

- **Diagnóstico Médico**: Como se mencionó anteriormente, el Teorema de Bayes es muy útil en medicina para interpretar los resultados de las pruebas. Por ejemplo, si una prueba para una enfermedad tiene cierta tasa de falsos positivos y falsos negativos, y conocemos la prevalencia de la enfermedad, el Teorema de Bayes puede ayudarnos a calcular la probabilidad de que un paciente realmente tenga la enfermedad dado un resultado positivo en la prueba.
- **Filtro de Spam en el Email**: Los filtros de correo electrónico utilizan el Teorema de Bayes para clasificar los mensajes como "spam" o "no spam". El filtro analiza la probabilidad de que un mensaje sea spam basándose en la presencia de ciertas palabras. Si una palabra aparece más frecuentemente en los mensajes de spam que en los mensajes normales, aumenta la probabilidad de que el mensaje sea spam.
- **Inferencia en Ciencia de Datos**: En machine learning y estadística, el Teorema de Bayes se usa para actualizar los modelos a medida que se recopilan nuevos datos. Por ejemplo, en un sistema de recomendación, las preferencias de los usuarios se actualizan constantemente a medida que interactúan con diferentes productos.
- **Toma de Decisiones bajo Incertidumbre**: En finanzas y economía, el Teorema de Bayes se utiliza para actualizar las probabilidades de los resultados económicos basándose en nueva información, como cambios en las políticas económicas o indicadores del mercado.
- **Ciencias Forenses**: En el análisis forense, el Teorema de Bayes puede ayudar a evaluar la evidencia. Por ejemplo, si se encuentra una huella dactilar en una escena del crimen, el teorema puede utilizarse para calcular la probabilidad de que pertenezca a un sospechoso, dados los patrones de huellas dactilares de la población en general.

1.9 Variables aleatorias discretas

Son aquellas que toman valores específicos y contables. Por ejemplo, el número de carros que pasan por un punto en una hora, o el número de estudiantes que aprueban un examen. Los valores que puede tomar una variable aleatoria discreta son finitos o infinitos contables (como el conjunto de los números enteros).

```
# En R, podemos usar dbinom para la función de probabilidad de una distribución binomial
# Por ejemplo, la probabilidad de obtener 3 éxitos en 5 ensayos con p = 0.5
probabilidad <- dbinom(3, size = 5, prob = 0.5)
probabilidad
```

```
## [1] 0.3125
```

- **Valores Discretos y Contables**: las variables aleatorias discretas toman valores específicos y contables. Pueden ser finitas o infinitamente contables.

1.9.1 Función de probabilidad

La función de probabilidad es una función que asigna una probabilida a cada posible valor de una variable aleatoria discreta. Matemáticamente, para una variable aleatoria discreta X, la función de probabilidad se define como $P(X = x)$, que es la probabilidad de que X tome el valor x.

Características:

- **No negatividad**: $P(X = x) \geq 0$ para todo x.
- **Normalización**: $\sum_x P(X = x) = 1$.

1.9.2 Función de distribución acumulada

La función de distribución acumulada (FDA) es una función que asigna una probabilidad acumulada a cada posible valor de una variable aleatoria discreta. Matemáticamente, para una variable aleatoria discreta X, la función de distribución acumulada se define como $F(x) = P(X \leq x)$, que es la probabilidad de que X tome un valor menor o igual a x.

La FDA es una función escalonada para el caso discreto, que aumenta en los puntos donde la variable aleatoria tiene probabilidad positiva. Se define como la suma de las probabilidades de los resultados hasta un cierto punto.

$$F(x) = P(X \leq x) = \sum_{y \leq x} P(X = y)$$

Ejemplo

Consideremos una variable aleatoria discreta X que representa el número de caras obtenidas al lanzar un dado tres veces. La FDA de $X \leq 2$ es $F(2) = P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) = 1/8 + 3/8 + 3/8 = 7/8 = 0.875$. Es decir, la probabilidad de obtener 2 o menos caras se calcula sumando las probabilidades de obtener 0, 1 o 2 caras.

```
n <- 3
p <- 0.5

# Calculamos la FDA para x=2
fda <- pbinom(2, size = n, prob = p)
fda
```

```
## [1] 0.875
```

1.9.3 Ejemplo

Aquí ilustraremos como se pueden mostrar tanto la función de probabilidad como la función de distribución acumulada de una variable aleatoria discreta. Este ejemplo mostrará la probabilidad de obtener un cierto número de éxitos en una serie de ensayos, así como la probabilidad acumulada hasta ese número de éxitos.

Vamos a usar una distribución binomial con los siguientes parámetros:

- $n = 10$, el número de ensayos.
- $p = 0.5$, la probabilidad de éxito en cada ensayo.

```
n <- 10
p <- 0.5

valores <- 0:n
```

1.10 Variables aleatorias continuas

Son aquellas que pueden tomar cualquier valor dentro de un intervalo. Por ejemplo, la altura de una persona, la temperatura de un día, etc. Los valores que puede tomar una variable aleatoria continua son infinitos no contables (como el conjunto de los números reales).

1.10.1 Función de Densidad de Probabilidad

Para variables continuas no podemos hablar de probabilidades en puntos específicos, sino de densidades. La Función de Densidad de Probabilidad (PDF), describe la probabilidad relativa de que la variable aleatoria tome un valor dado. Esta función se denota como $f(x)$, no nos da las probabilidades directamente, pero el área bajo la curva de la función entre dos puntos nos da la probabilidad de que la variable caiga dentro de ese intervalo.

Características:

- **No negatividad**: $f(x) \geq 0$ para todo x.
- **Normalización**: $\int_{-\infty}^{\infty} f(x)dx = 1$.

```
# En R, podemos usar dnorm para la función de densidad de una distribución normal
# Por ejemplo, la probabilidad de obtener un valor entre 0 y 1 en una distribución normal con media 0 y
densidad <- dnorm(0, mean = 0, sd = 1)
densidad
```

```
## [1] 0.3989423
```

2 Estadística Descriptiva

La estadística descriptiva es una rama de la estadística que se centra en la descripción de los datos. Su objetivo es resumir y organizar los datos para que sean más fáciles de entender y visualizar. Por tanto, la estadística descriptiva se puede utilizar para describir las características de un conjunto de datos, pero no para llegar a conclusiones más allá de los datos que tenemos o para realizar inferencias sobre una población.

2.1 Medidas de tendencia central

Las medidas de tendencia central son estadísticas que resumen la posición central de un conjunto de datos. Son especialmente útiles para describir la distribución de datos numéricos continuos o discretos. Las medidas de tendencia central más comunes son la media, la mediana y la moda.

2.1.1 Media

La media es la medida de tendencia central más común. Se define como la suma de todos los valores dividida por el número de valores. La media es muy sensible a los valores atípicos, por lo que no es una buena medida de tendencia central cuando los datos tienen valores atípicos.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

```
# Crear un vector con datos continuos
altura_estudiantes <- c(1.70, 1.65, 1.80, 1.75, 1.60)

# Calcular la media
mean(altura_estudiantes)
```

```
## [1] 1.7
```

```
# Calcular la media con valores NA
mean(c(1.70, 1.65, 1.80, 1.75, 1.60, NA))
```

```
## [1] NA
```

```
# Calcular la media con valores NA y valores infinitos
mean(c(1.70, 1.65, 1.80, 1.75, 1.60, NA, Inf))
```

```
## [1] NA
```

2.1.2 Mediana

La mediana es el valor que separa la mitad superior de un conjunto de datos de la mitad inferior. Es menos sensible a los valores atípicos que la media, por lo que es una mejor medida de tendencia central cuando los datos tienen valores atípicos.

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

```
# Crear un vector con datos continuos
altura_estudiantes <- c(1.70, 1.65, 1.80, 1.75, 1.60)

# Calcular la mediana
median(altura_estudiantes)
```

```
## [1] 1.7
```

2.1.3 Moda

La moda es el valor que ocurre con mayor frecuencia en un conjunto de datos. Es la única medida de tendencia central que se puede utilizar con datos categóricos. Sin embargo, no es una buena medida de tendencia central cuando los datos tienen múltiples valores que ocurren con la misma frecuencia.

$$\text{Moda} = \underset{x}{\text{argmax}} \left(\text{Frecuencia Absoluta}(x) \right)$$

```
# Crear un vector con datos categóricos
colores_camisetas <- c("rojo", "azul", "verde", "rojo", "verde", "verde")

# Calcular la moda
Mode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}

Mode(colores_camisetas)
```

```
## [1] "verde"
```

2.2 Medidas de dispersión

Las medidas de dispersión son estadísticas que resumen la variabilidad de un conjunto de datos. Son especialmente útiles para describir la distribución de datos numéricos continuos o discretos. Las medidas de dispersión más comunes son el rango, la varianza y la desviación estándar.

2.2.1 Rango

El rango es la diferencia entre el valor máximo y el valor mínimo de un conjunto de datos. Es la medida de dispersión más simple y fácil de calcular, pero también la menos informativa.

$$\text{Rango} = \max(x) - \min(x)$$

```
# Crear un vector con datos continuos
altura_estudiantes <- c(1.70, 1.65, 1.80, 1.75, 1.60)

# Calcular el rango
max(altura_estudiantes) - min(altura_estudiantes)
```

```
## [1] 0.2
```

2.2.2 Varianza

La varianza es la media de las diferencias al cuadrado entre cada valor y la media. Es una medida de dispersión muy común, pero no es muy intuitiva porque está en unidades al cuadrado.

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

```r
# Crear un vector con datos continuos
altura_estudiantes <- c(1.70, 1.65, 1.80, 1.75, 1.60)

# Calcular la varianza
var(altura_estudiantes)
```

```
## [1] 0.00625
```

2.2.3 Desviación estándar

La desviación estándar es la raíz cuadrada de la varianza. Es una medida de dispersión muy común y más intuitiva que la varianza porque está en las mismas unidades que los datos.

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

```r
# Crear un vector con datos continuos
altura_estudiantes <- c(1.70, 1.65, 1.80, 1.75, 1.60)

# Calcular la desviación estándar
sd(altura_estudiantes)
```

```
## [1] 0.07905694
```

2.3 Visualización de datos

La visualización de datos es una rama de la estadística que se centra en la visualización de datos. Su objetivo es resumir y organizar los datos para que sean más fáciles de entender y visualizar. Por tanto, la visualización de datos se puede utilizar para describir las características de un conjunto de datos, pero no para llegar a conclusiones más allá de los datos que tenemos o para realizar inferencias sobre una población.

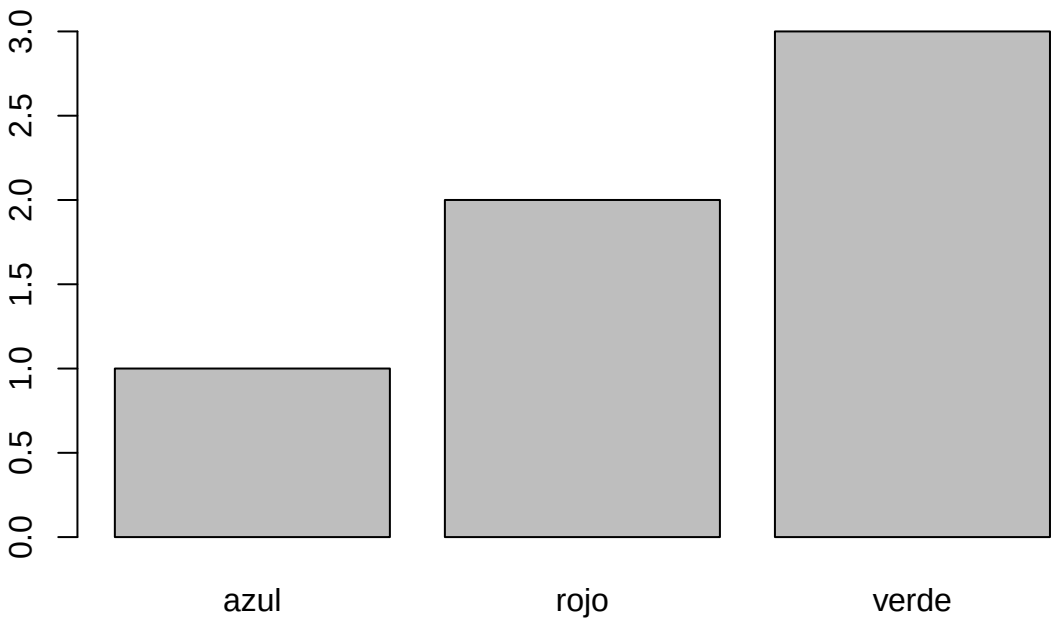
Para la visualización de datos en R podemos utilizar las bibliotecas nativas o bien alternativas de terceros. Una alternativa popular es la biblioteca `ggplot2`, que se basa en la gramática de gráficos de Wilkinson.

2.3.1 Visualización de datos cualitativos (una variable)

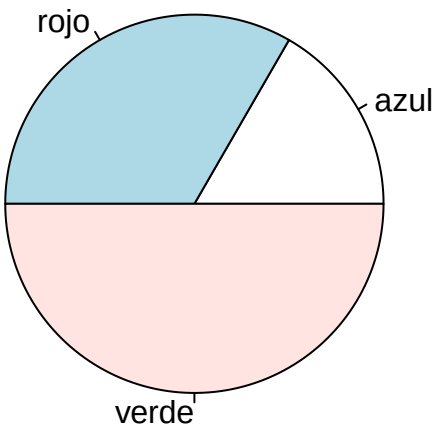
La visualización de datos cualitativos se utiliza para visualizar datos categóricos. Los gráficos más comunes para visualizar datos cualitativos son los gráficos de barras y los gráficos circulares.

```r
# Crear un vector con datos categóricos
colores_camisetas <- c("rojo", "azul", "verde", "rojo", "verde", "verde")

# Crear un gráfico de barras
barplot(table(colores_camisetas))
```



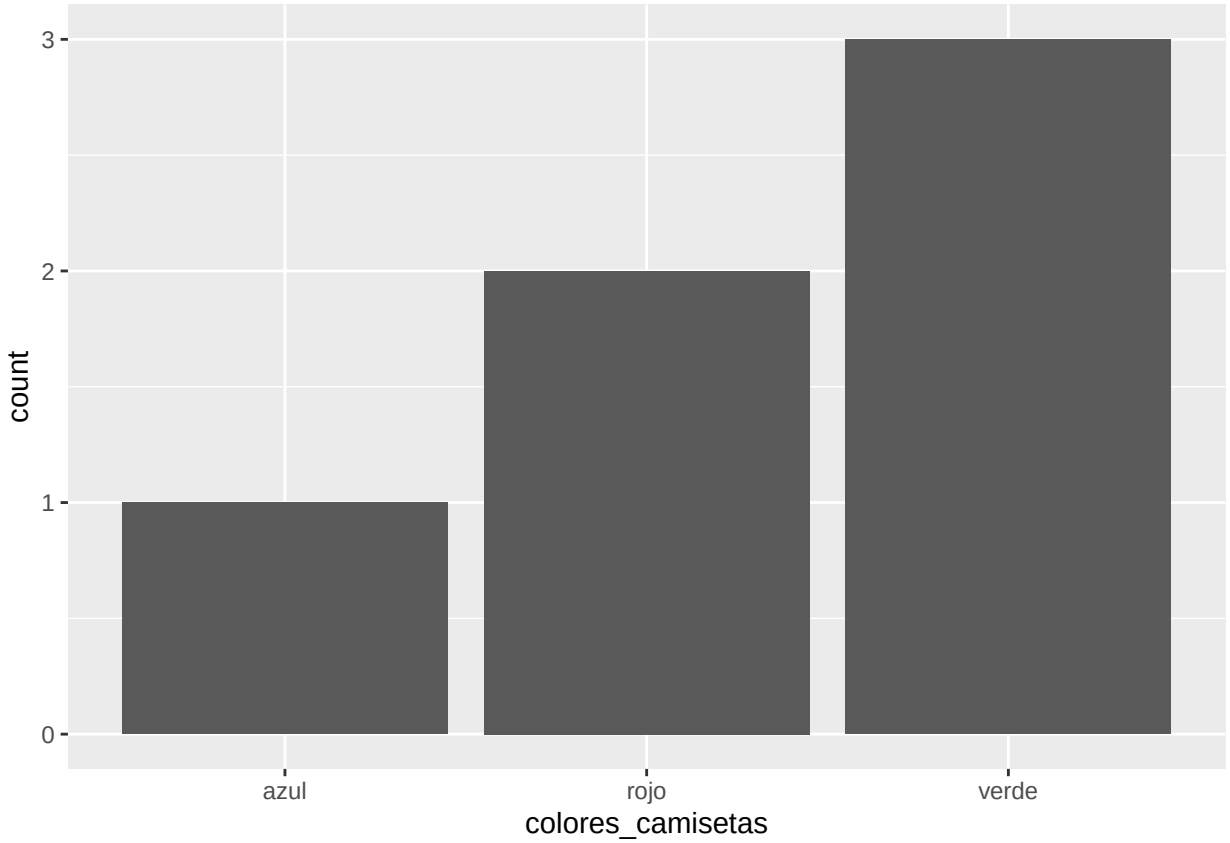

```r
# Crear un gráfico circular
pie(table(colores_camisetas))
```



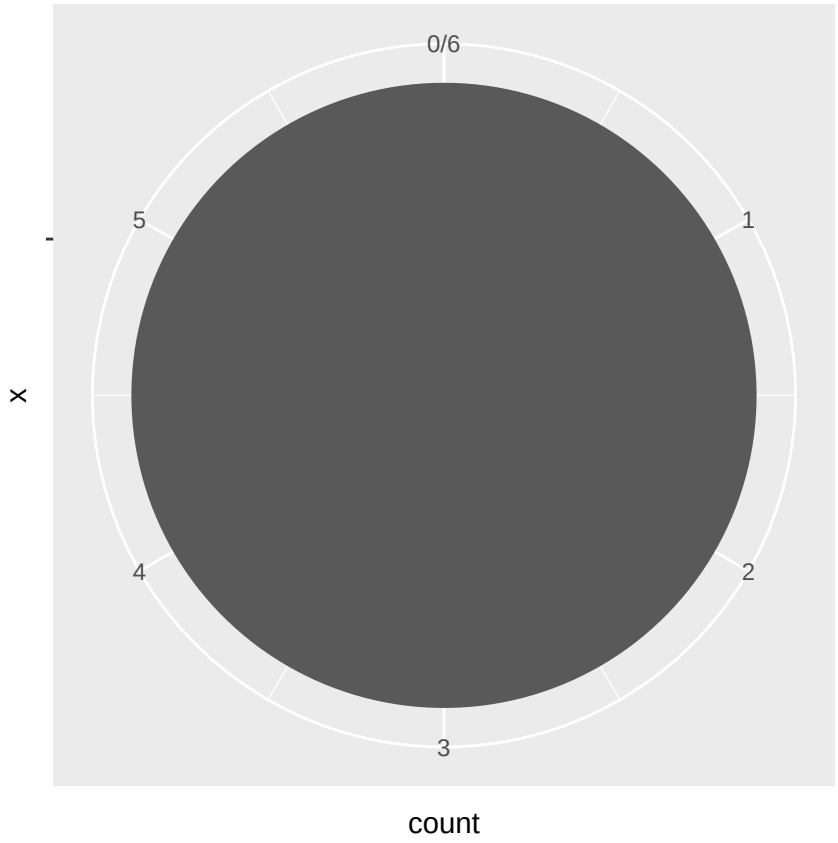
Los mismos gráficos pero utilizando la biblioteca `ggplot2`:

```
library(ggplot2)
# Crear un vector con datos categóricos
colores_camisetas <- c("rojo", "azul", "verde", "rojo", "verde", "verde")

# Crear un gráfico de barras
ggplot(data.frame(colores_camisetas), aes(x = colores_camisetas)) + geom_bar()
```




```
# Crear un gráfico circular
ggplot(data.frame(colores_camisetas), aes(x = "")) + geom_bar() + coord_polar(theta = "y")
```



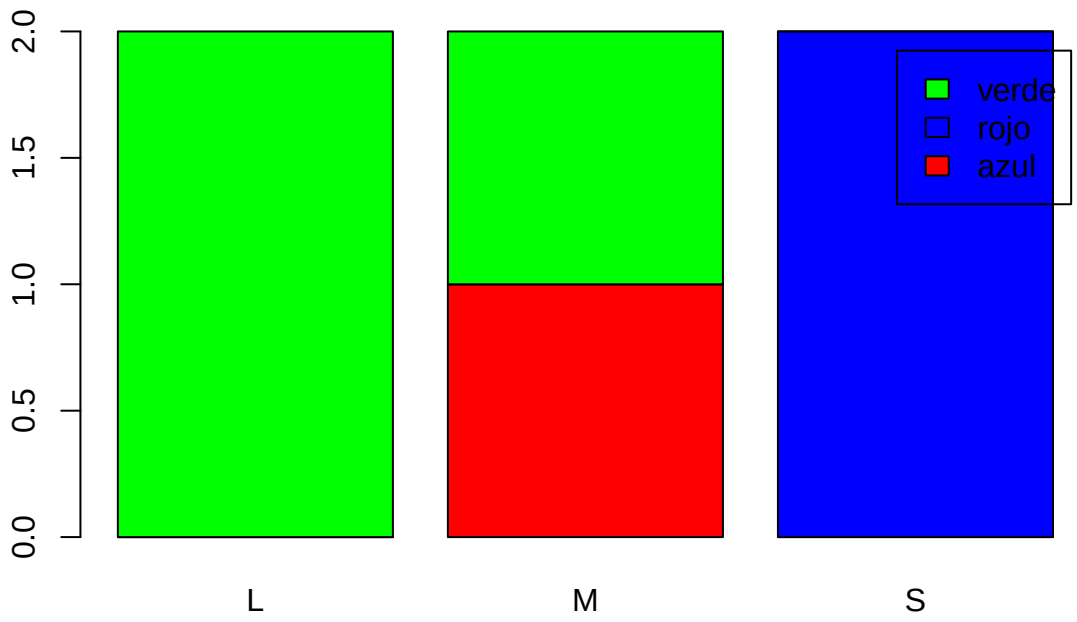
2.3.2 Visualización de datos cualitativos (dos variables)

Cuando dos variables se miden en una sola unidad experimental, los datos resultantes se denominan **datos bivariados**. La visualización de datos bivariados se utiliza para visualizar datos categóricos en dos variables. Los gráficos más comunes para visualizar datos bivariados son los gráficos de barras apiladas y los gráficos de barras agrupadas.

```r
# Crear un vector con datos categóricos
colores_camisetas <- c("rojo", "azul", "verde", "rojo", "verde", "verde")

# Crear un vector con datos categóricos
tallas_camisetas <- c("S", "M", "L", "S", "M", "L")

# Crear un gráfico de barras apiladas
barplot(table(colores_camisetas, tallas_camisetas), legend = TRUE, col = c("red", "blue", "green"))
```

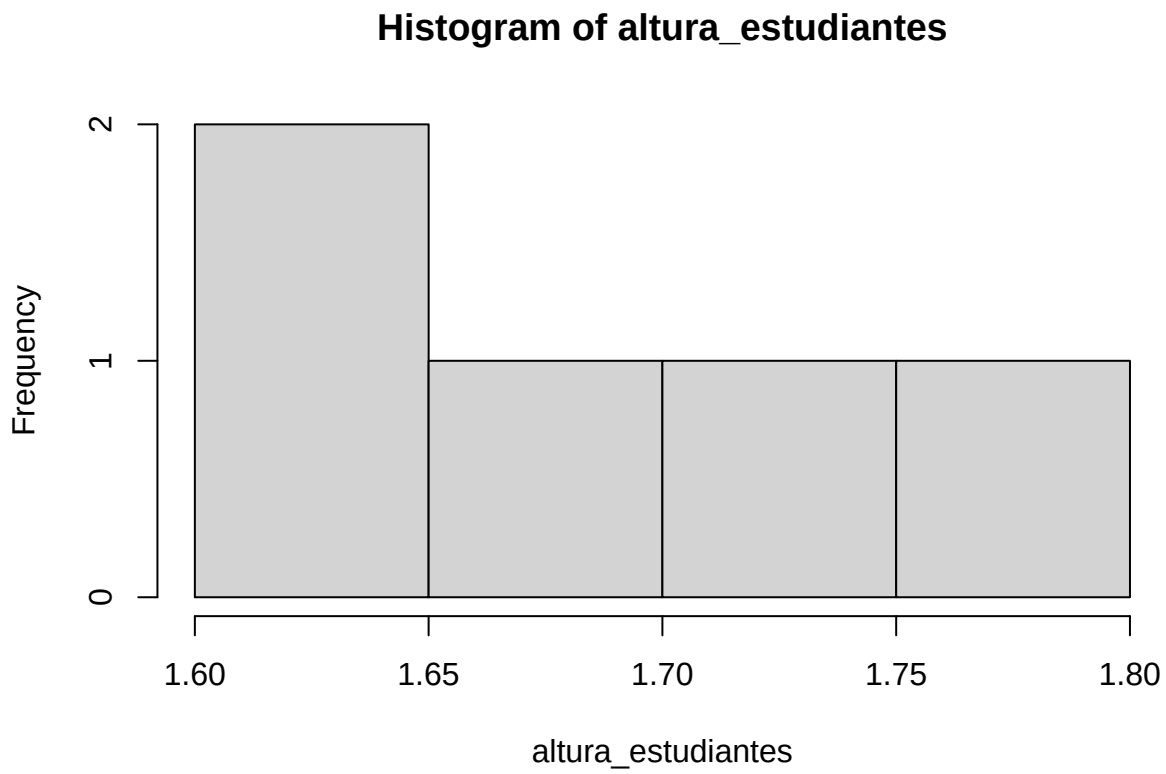


2.3.3 Visualización de datos cuantitativos

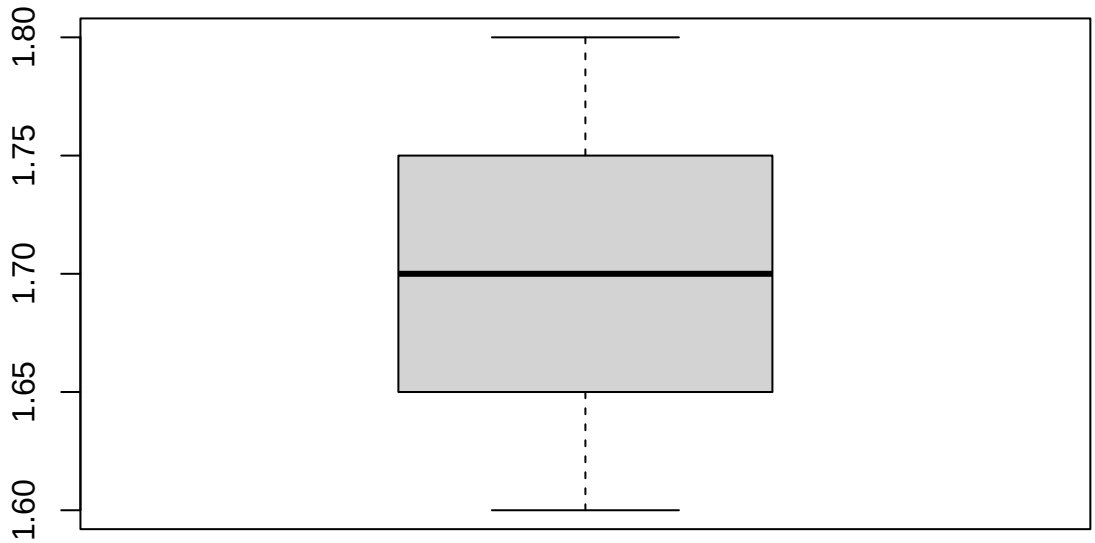
La visualización de datos cuantitativos se utiliza para visualizar datos continuos o discretos. Los gráficos más comunes para visualizar datos cuantitativos son los histogramas y los diagramas de caja.

```
# Crear un vector con datos continuos
altura_estudiantes <- c(1.70, 1.65, 1.80, 1.75, 1.60)

# Crear un histograma
hist(altura_estudiantes)
```




```r
# Crear un diagrama de caja
boxplot(altura_estudiantes)
```



Los diagramas de cajas son una forma de visualizar la distribución de un conjunto de datos. Se basan en el rango intercuartil, que es la diferencia entre el tercer y el primer cuartil. El primer cuartil es el valor que divide el conjunto de datos en dos partes iguales, y el tercer cuartil es el valor que divide el conjunto de datos en dos partes iguales. El diagrama de caja muestra el rango intercuartil como una caja, con el primer cuartil en la parte inferior y el tercer cuartil en la parte superior. La mediana se muestra como una línea dentro de la caja. Los valores atípicos se muestran como puntos fuera de la caja.

- Los lados inferior y superior de la caja representan el **primer** y el **tercer cuartil**. Por lo que la altura de la caja es igual al **rango intercuartil**. Así, la caja representa el 50% de los datos.
- La línea central de la caja representa la **mediana**.
- Los puntos fuera de la caja representan los **valores atípicos**. Estos son calculados de la siguiente manera:
  - **Límite inferior**: $Q_1 - 1.5 \times \text{RI}$
  - **Límite superior**: $Q_3 + 1.5 \times \text{RI}$

2.3.4 Visualización de datos cuantitativos (dos variables)

La representación gráfica más usual es el **diagrama de dispersión**. Este diagrama representa los valores de dos variables en un plano cartesiano. Cada punto representa un par de valores de las dos variables. El diagrama de dispersión se utiliza para visualizar la relación entre dos variables.

- **Correlación lineal**: cuando el diagrama muestra cierta **relación lineal** entre las variables, la medida resumen comúnmente utilizada es el **coeficiente de correlación de Pearson**. Este coeficiente toma valores entre -1 y 1. Un valor de 1 indica una correlación lineal positiva perfecta, un valor de -1

indica una correlación lineal negativa perfecta y un valor de 0 indica que no hay correlación lineal. Matemáticamente se expresa como: $r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$

  - Existe una correlación entre dos variables cuando los valores de una variable están de alguna manera asociados con los valores de la otra.
  - La correlación mide la *fuerza* y la *dirección* de la relación entre dos variables.

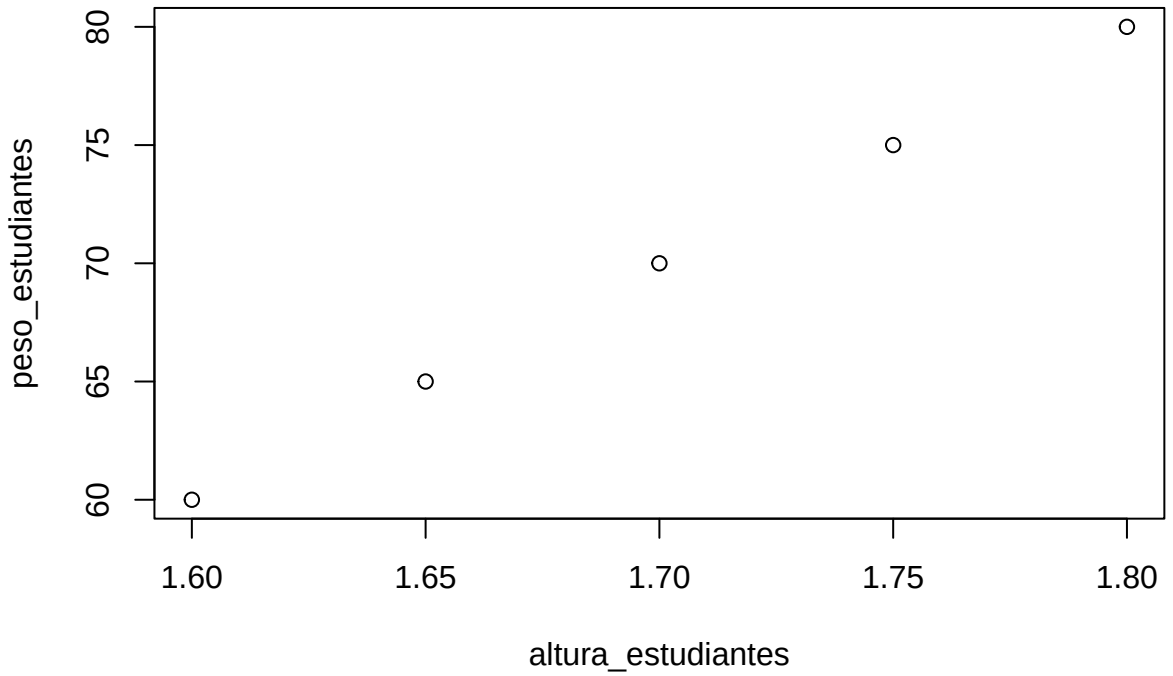
Algunas propiedades de la correlación son:

- El valor del coeficiente de correlación está entre -1 y 1. $-1 \leq r \leq 1$
- Si $r = 0$, no hay correlación lineal.
- Si $r > 0$, hay una correlación lineal positiva.
- Si $r < 0$, hay una correlación lineal negativa.
- La correlación r no tiene unidades y no depende de la escala de las variables.
- La correlación es simétrica. $r_{xy} = r_{yx}$

```r
# Crear un vector con datos continuos
altura_estudiantes <- c(1.70, 1.65, 1.80, 1.75, 1.60)

# Crear un vector con datos continuos
peso_estudiantes <- c(70, 65, 80, 75, 60)

# Crear un diagrama de dispersión
plot(altura_estudiantes, peso_estudiantes)
```



Podemos calcular la correlación mediante la función `cor()`:

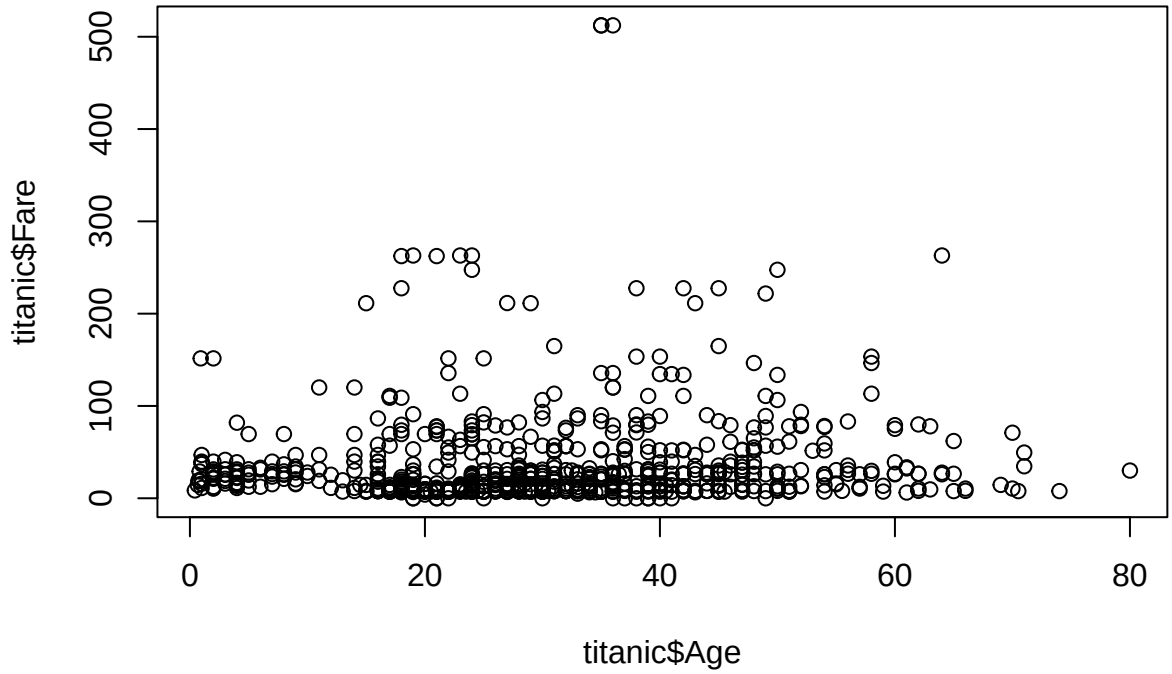
```r
cor(altura_estudiantes, peso_estudiantes)
```

```
## [1] 1
```

Si en lugar de datos inventados utilizamos los datos del dataset del **titanic** obtenemos el siguiente gráfico:

```r
# Cargar el dataset del titanic
titanic <- read.csv("datasets/titanic.csv")

# Crear un diagrama de dispersión
plot(titanic$Age, titanic$Fare)
```



Aquí el coeficiente de correlación es:

```r
cor(titanic$Age, titanic$Fare)
```

```
## [1] 0.1123286
```