Estadística

Pablo de la Cuesta García

2024-01-08

Contents

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 valores posibles son infinitos. Por ejemplo, la altura.

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

```r
# 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.

```r
# 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

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.

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

```r
# 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
```

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

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

```r
# 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.

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

```r
# 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.

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

```r
# 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.

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

```
# 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.

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

```
# 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.

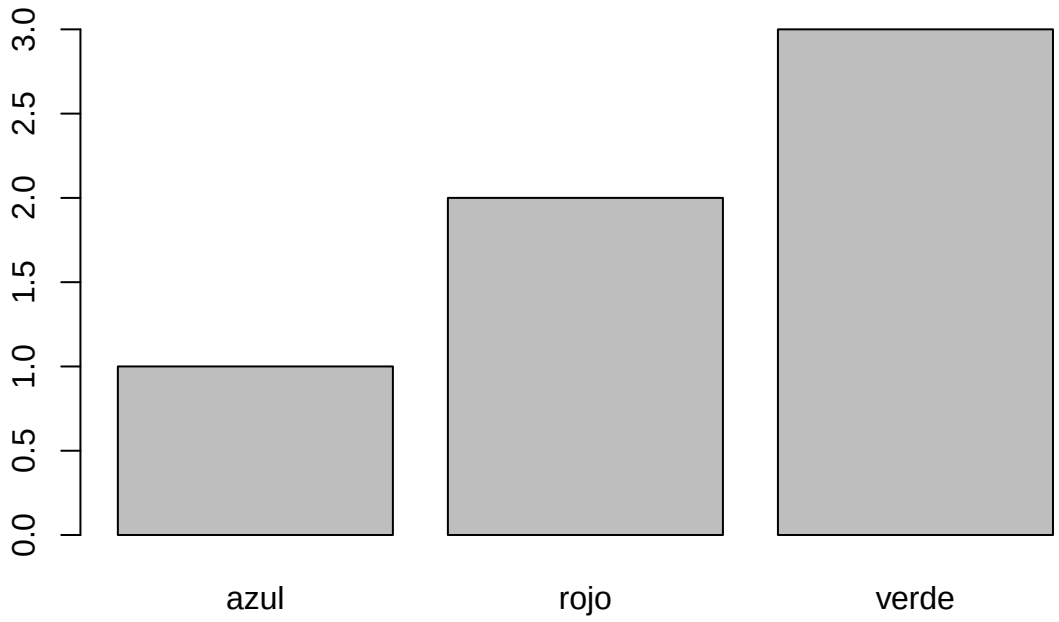
Para la visualización de datos en R podemos utilizar las bibliotecas nativas o bien alternativas de terceros. Una alternativa popular es la biblioteca `ggplot2`, que se basa en la gramática de gráficos de Wilkinson.

2.3.1 Visualización de datos cualitativos

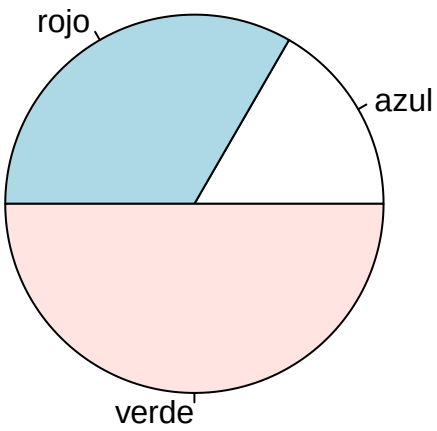
La visualización de datos cualitativos se utiliza para visualizar datos categóricos. Los gráficos más comunes para visualizar datos cualitativos son los gráficos de barras y los gráficos circulares.

```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))
```



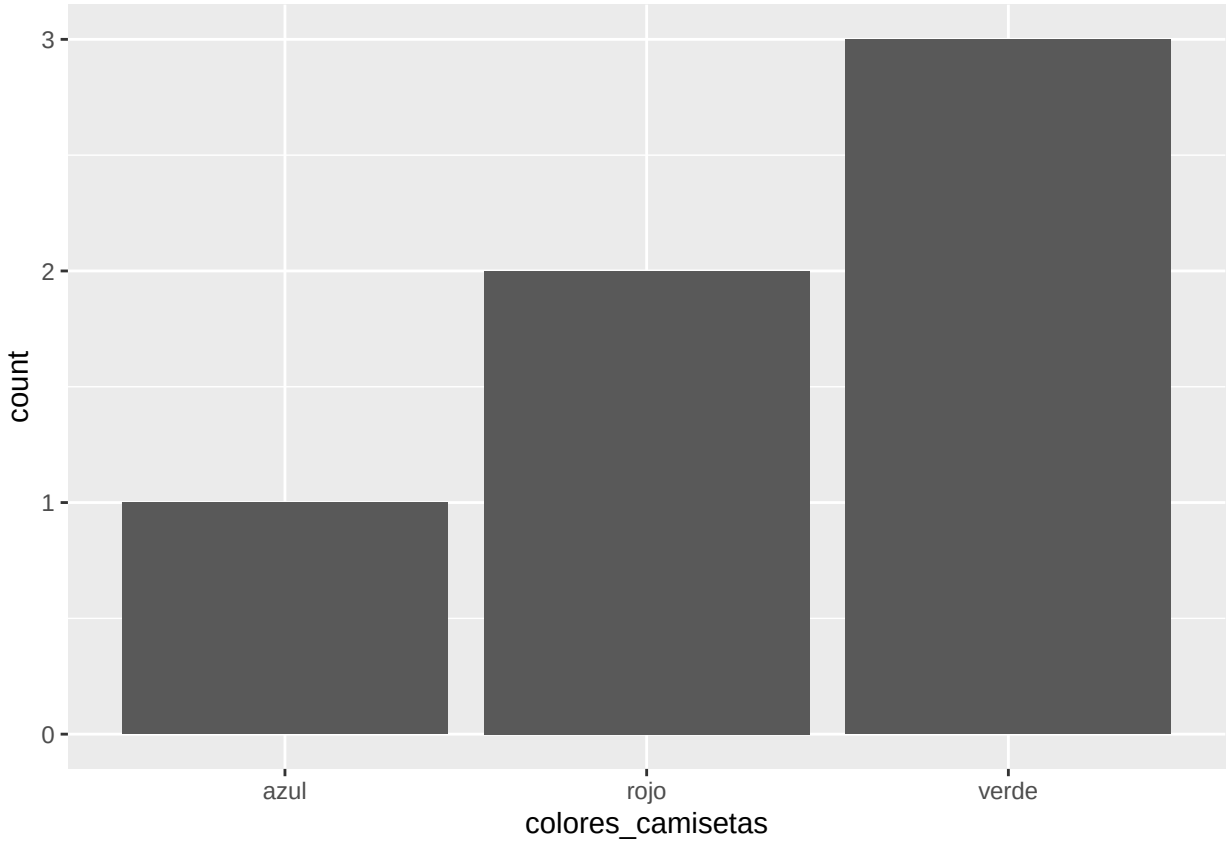

```r
# Crear un gráfico circular
pie(table(colores_camisetas))
```



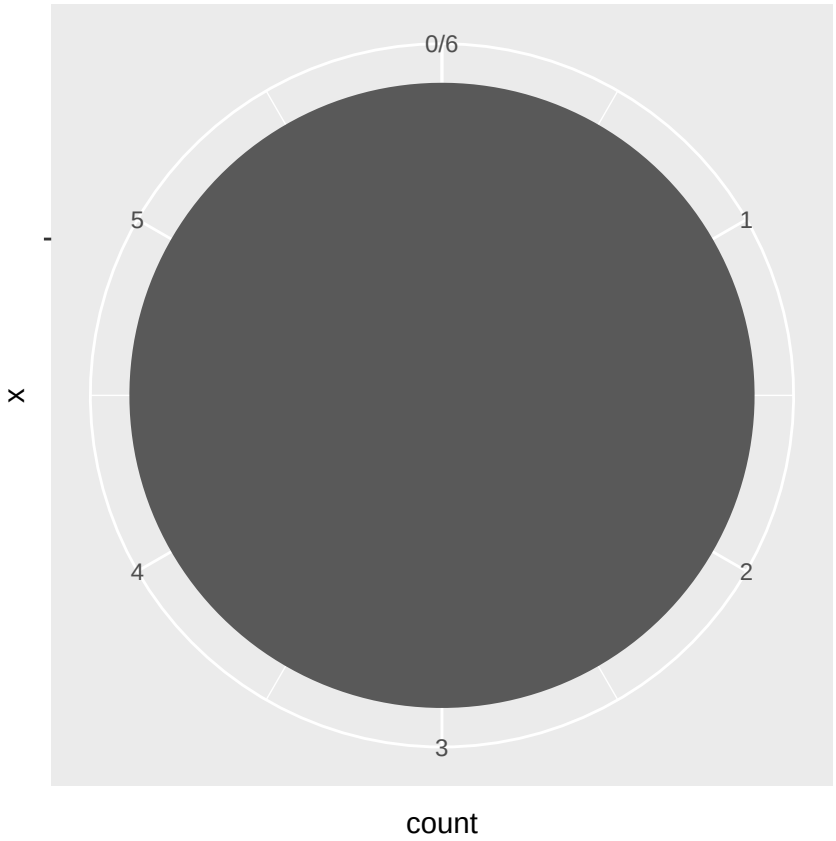
Los mismos gráficos pero utilizando la biblioteca `ggplot2`:

```
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")
```

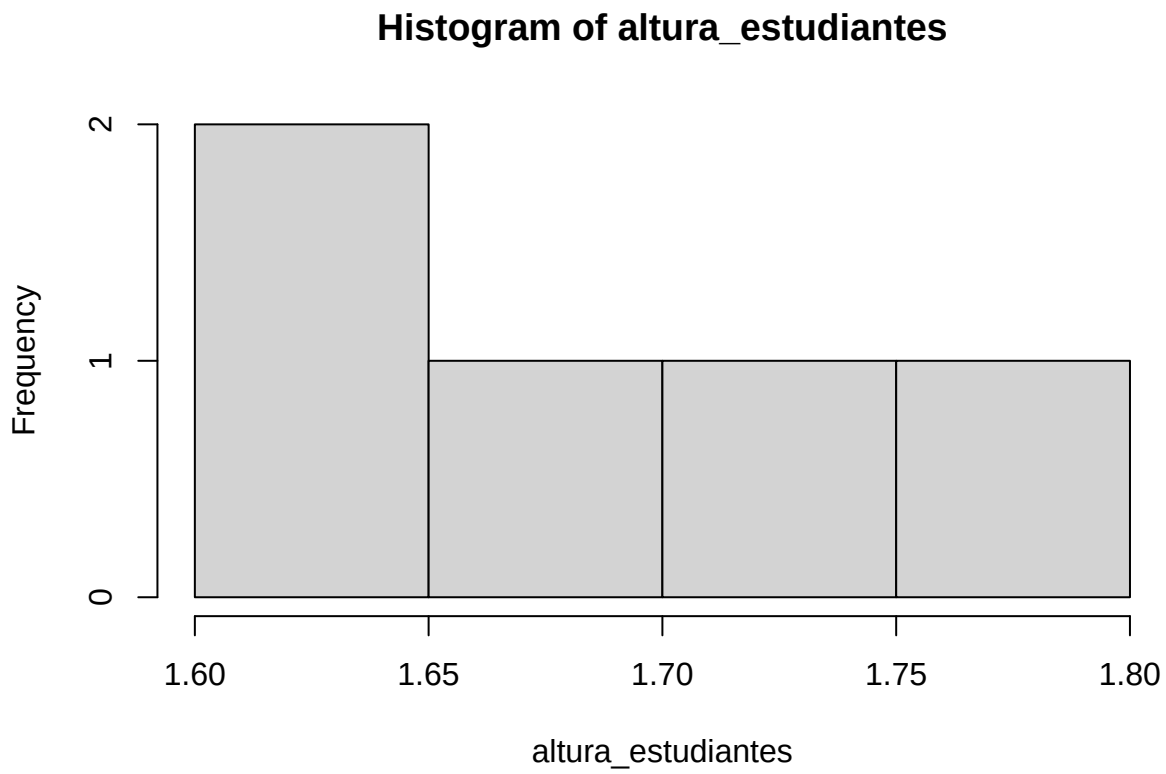


2.3.2 Visualización de datos cuantitativos

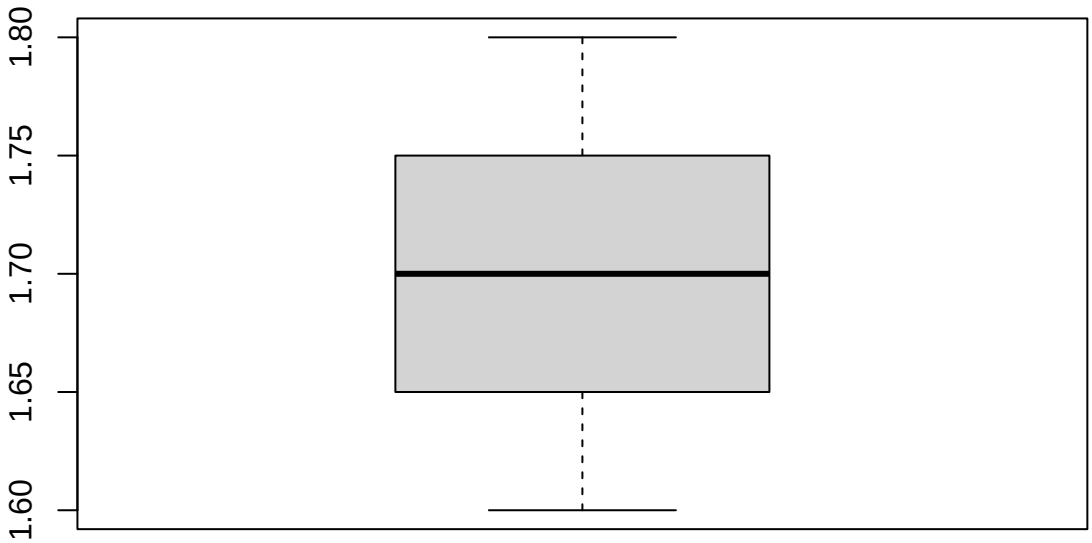
La visualización de datos cuantitativos se utiliza para visualizar datos continuos o discretos. Los gráficos más comunes para visualizar datos cuantitativos son los histogramas y los diagramas de caja.

```r
# 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)
```


1.80
1.75
1.70
1.65
1.60