Introduction to statistics in R

José Burgos

2023-11-15

Course statistics

2 branch have of statistics

Descriptive statistics: Describe and summarize data

Inferential statistics * Use a sample of data to make inferences abour a larger population

Types of data 1. Numeric (Quantitative) * Continuous (Measured) ** Airplane speed ** Time spent waiting in line

- Discrete (Counted) ** Number of pets ** Number of packages shipped

2. Categorical (Qualitative)

- Nominal (Unordered) ** Married/Unmarried ** Country of residence
- Ordinal (Ordered) _____Which measure to use???_____ Median and mean

Cuando los datos estan sesgados a la izquierda la media es menor que la mediana y mayor que la mediana en los datos sesgados a la derecha.

Se recomienda usar la media en datos simétricos y la mediana en datos asimétricos.

Debido a que la media es arrastrada por los valores extremos, es mejor usar la medianan ya que se ve menos afectada por los valores atípicos.

Measures of spread

Medidas de propagación.

- Varianza: Mide la distancia promedio desde cada punto de datos hasta la media de los datos. Cuanto mayor sea la varianza, más dospersos están los datos.

```
var(iris$Sepal.Length)
```

```
## [1] 0.6856935
```

- Desviación estándar SD es una medida de dispersión, calculada tomando la raís cuadrada de la varianza.
- Media absoluta de la desviación MAD: Calculada como la media del valor abosoluto de las diferencias de los valores menos el promedio de los valores.

```r
sd(iris$Sepal.Length)
```

```
## [1] 0.8280661
```

```r
mad(iris$Sepal.Length)
```

```
## [1] 1.03782
```

La deviación estándar y la desviación media absoluta, son similares pero no son lo mismo. La desviación estándar eleva al cuadrado las distancias, por lo que las distancias más largas se penalizan más que las más cortas, miestras que la deviación media absoluta penaliza cada distancia por igual.

Quartiles — Cuartiles

Los cuartiles dividen los datos en cuatros partes iguales.

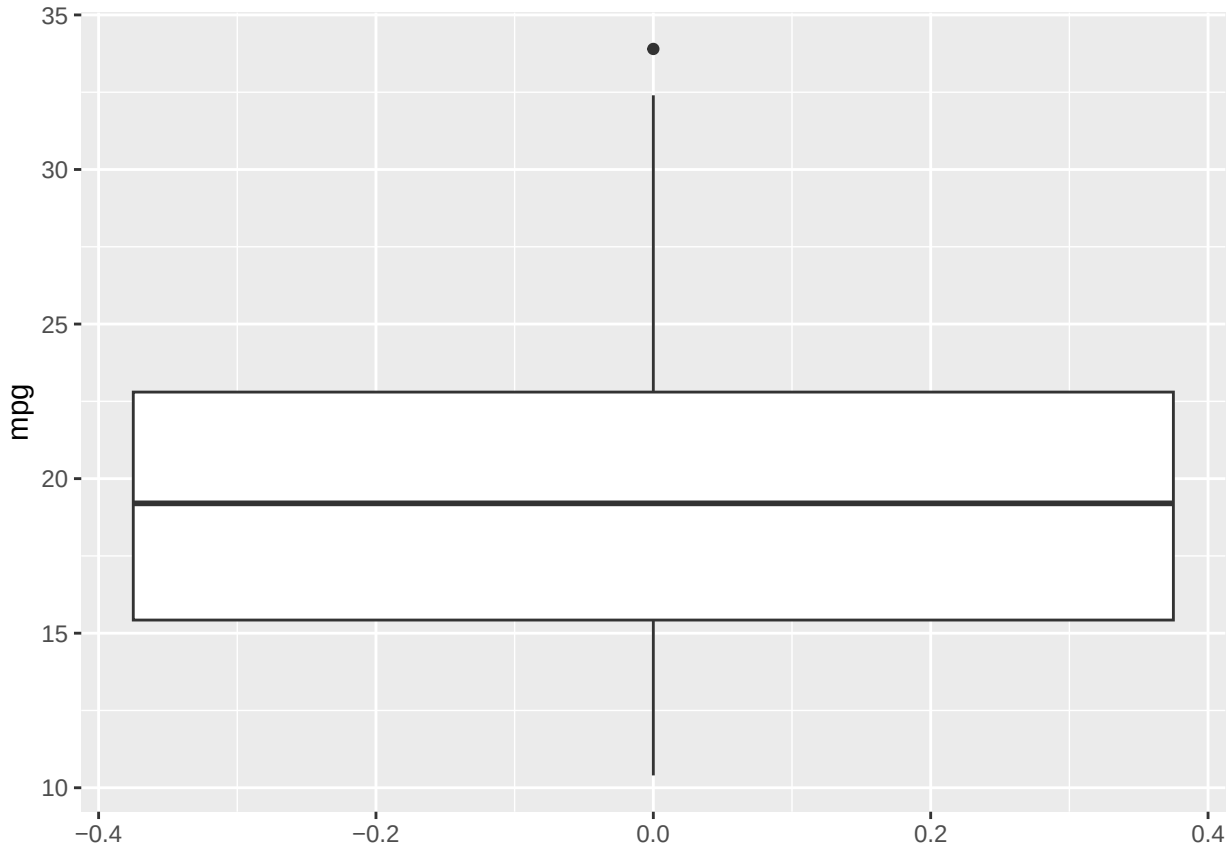
```r
quantile(iris$Sepal.Length)
```

```
##   0%  25%  50%  75% 100%
##  4.3  5.1  5.8  6.4  7.9
```

Boxplots use quartiles

Las cajas de diagramas de cajas representan cuartiles

```r
library(ggplot2)
ggplot(mtcars, aes(y = mpg)) +
  geom_boxplot()
```



Quantiles cuartiles o percertiles:

Con la misma función agregandole el argumento de probs que toma un vector de las proporciones, podemos dividir los datos en 5 partes:

```
quantile(mtcars$cyl, probs = c(0,0.2,0.4,0.6,0.8,1))
```

```
##   0%  20%  40%  60%  80% 100%
##    4    4    6    8    8    8
```

Otra menera es, usando la función `seq`:

seq(from, to, by)

Donde, from es el numero menor, to el mayor y by es el numero de salto o escala que tendrán.

```
quantile(mtcars$cyl, probs = seq(0,1,0.2))
```

```
##   0%  20%  40%  60%  80% 100%
##    4    4    6    8    8    8
```

Rango intercuartilico **IQR**

Es la diferencia entre los percertiles 25 y 75, lo que es la misma altura de la caja en un gráfico de boxplot diagrama de caja.

Outliers

Valores atípicos, son los valores de una base de datos que son sustancialmente diferentes de los demás. Para considerar un valor atípico, usualmente se utiliza la regla general, que dice que los datos son atípico cuando son menor que el primer cuartil menos **1.5** el IQR, así como cualquier punto mayor que el tercer cuartil más **1.5** el IQR.

Regla para considerar un valor atípico:

- $data < Q1 - 1.5 * IQR$
- $data > Q1 + 1.5 * IQR$

```
iqr <- quantile(mtcars$mpg, 0.75) - quantile(mtcars$mpg, 0.25)
lower_threshold <- quantile(mtcars$mpg, 0.25) - 1.5 * iqr
upper_threshold <- quantile(mtcars$mpg, 0.75) - 1.5 * iqr

mtcars |>
  filter(mpg < lower_threshold | mpg > upper_threshold)
```

Encontrar valores atípicos

```
##                    mpg cyl  disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
## Duster 360        14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
## Merc 240D         24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
## Merc 230          22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
## Merc 280          19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
## Merc 280C         17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
## Merc 450SE        16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
## Merc 450SL        17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
## Merc 450SLC       15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
## Chrysler Imperial 14.7   8 440.0 230 3.23 5.345 17.42  0  0    3    4
## Fiat 128          32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
## Honda Civic       30.4   4  75.7  52 4.93 1.615 18.52  1  1    4    2
## Toyota Corolla    33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1
## Toyota Corona     21.5   4 120.1  97 3.70 2.465 20.01  1  0    3    1
## Dodge Challenger  15.5   8 318.0 150 2.76 3.520 16.87  0  0    3    2
## AMC Javelin       15.2   8 304.0 150 3.15 3.435 17.30  0  0    3    2
## Camaro Z28        13.3   8 350.0 245 3.73 3.840 15.41  0  0    3    4
## Pontiac Firebird  19.2   8 400.0 175 3.08 3.845 17.05  0  0    3    2
## Fiat X1-9         27.3   4  79.0  66 4.08 1.935 18.90  1  1    4    1
## Porsche 914-2     26.0   4 120.3  91 4.43 2.140 16.70  0  1    5    2
## Lotus Europa      30.4   4  95.1 113 3.77 1.513 16.90  1  1    5    2
## Ford Pantera L    15.8   8 351.0 264 4.22 3.170 14.50  0  1    5    4
## Ferrari Dino      19.7   6 145.0 175 3.62 2.770 15.50  0  1    5    6
## Maserati Bora     15.0   8 301.0 335 3.54 3.570 14.60  0  1    5    8
## Volvo 142E        21.4   4 121.0 109 4.11 2.780 18.60  1  1    4    2
```