Univariate Statistics: Theoretical aspects and practical applications.

Universidad Nacional de Colombia

Andrés Felipe Beltrán

May 5, 2020

UNIVERSIDAD NACIONAL DE COLOMBIA

Univariate Statistics: a reminder

The **distribution** of data plays an important role in **statitistics**.

- Number of data often unknown
- type of distribution unknown

Univariate Statistics: a reminder

The **distribution** of data plays an important role in **statitistics**.

- Number of data often unknown
- type of distribution unknown

Univariate Statistics: a reminder

- The values of a variable **x** (say the concentration of a chemical compound in a set with n samples) have an `empirical distribution`.

Univariate Statistics: a reminder

- The values of a variable **x** (say the concentration of a chemical compound in a set with n samples) have an `empirical distribution`.
- whenever possible it should be visually inspected to obtain a better insight into the data.

Univariate Statistics: a reminder

Descriptive measures of a distribution are:

- Minimum
- mean
- median
- maximum
- quantiles

Quantiles

- A quantile is defined for a fraction α (between 0 and 1); it is the value when a fraction α of the data is below this value, and a fraction $\mathbf{1}$-α is above this value.

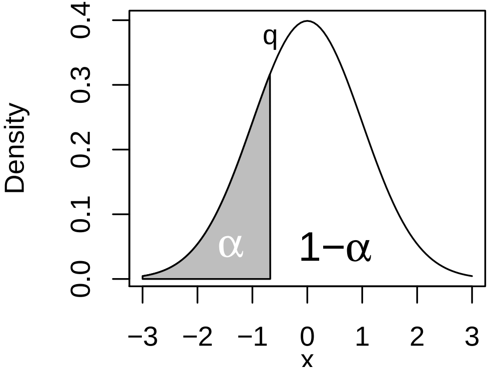


Figure 1: quantile graphical demonstration in standard normal distribution

Quantiles

- A quantile is defined for a fraction α (between 0 and 1); it is the value when a fraction α of the data is below this value, and a fraction $\mathbf{1}$-α is above this value.
- For **precentiles**, α is expressed in percent (%).

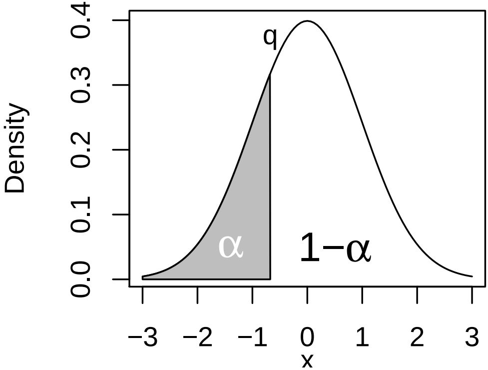


Figure 1: quantile graphical demonstration in standard normal distribution

Quartiles

- **quartiles** divide the data distribution into four parts:

1. First quartile $\boldsymbol{(Q_1)} = 25\ \%$
2. Second quartile $\boldsymbol{(Q_2)} = 50\ \% = \text{median}$
3. Third quartile $\boldsymbol{(Q_3)} = 75\ \%$

Quartiles

- **quartiles** divide the data distribution into four parts:

1. First quartile $(\boldsymbol{Q_1}) = 25\ \%$
2. Second quartile $(\boldsymbol{Q_2}) = 50\ \% =$ median
3. Third quartile $(\boldsymbol{Q_3}) = 75\ \%$

- The **Interquartile Range** $\mathbf{(IQR} = \boldsymbol{Q_3} - \boldsymbol{Q_1})$ is the difference between the first and third quartile.

Quartiles

- **quartiles** divide the data distribution into four parts:

1. First quartile $\boldsymbol{(Q_1)}$ = 25 %
2. Second quartile $\boldsymbol{(Q_2)}$ = 50 % = median
3. Third quartile $\boldsymbol{(Q_3)}$ = 75 %

- The **Interquartile Range** $\boldsymbol{(\mathrm{IQR} = Q_3 - Q_1)}$ is the difference between the first and third quartile.

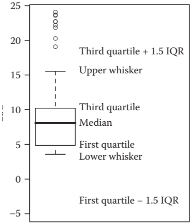
- 

Figure 2: Boxplot ilustration.

Boxplot

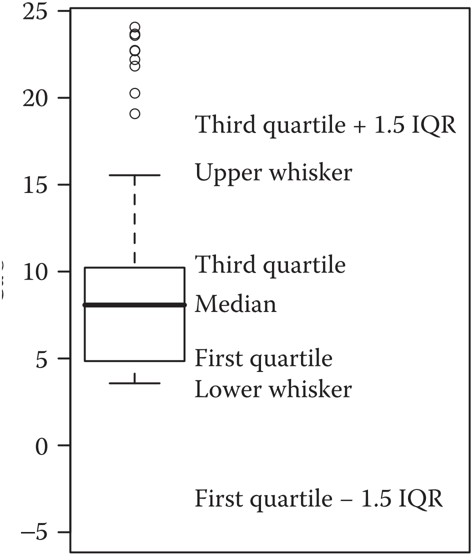


Figure 3: Boxplot ilustration. (zoom)

Theoretical distributions

- Theoretical distributions
  1. Normal distribution:

$$N(\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{1}$$

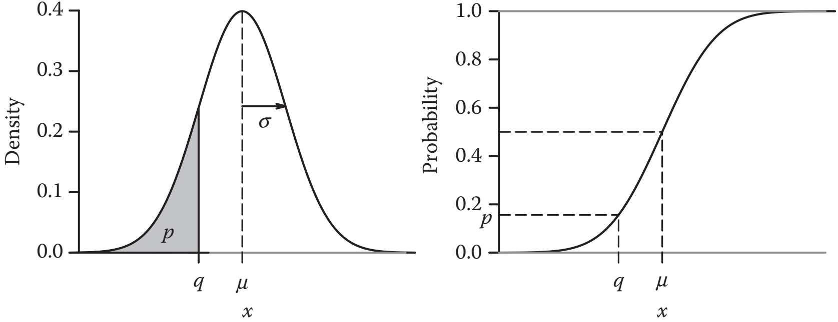


Figure 4: Probability density function (PDF) (left) and cumulative distribution function (right) of the normal distribution.

The probability density, d, at value x is defined by

- For a standard normal distribution:

$$N(0.1) : d(x) = \frac{1}{\sqrt{2\pi}} exp\left(\frac{-x^2}{2}\right) \tag{2}$$

$\mathbb{R}$: `d <- dnorm(x, mean=0, sd=1)`

- For a normal distribution:

$$N(\mu, \sigma^2) : d(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{3}$$

$\mathbb{R}$: `d <- dnorm(x, mean=mu, sd=sigma)`

Standard normal distribution

- Data values x following a normal distribution $N(\mu, \sigma^2)$ can be transformed to a standard normal distribution by the so called z-transformation:

$$z = \frac{(x - \mu)}{\sigma} \tag{4}$$

Normal distribution

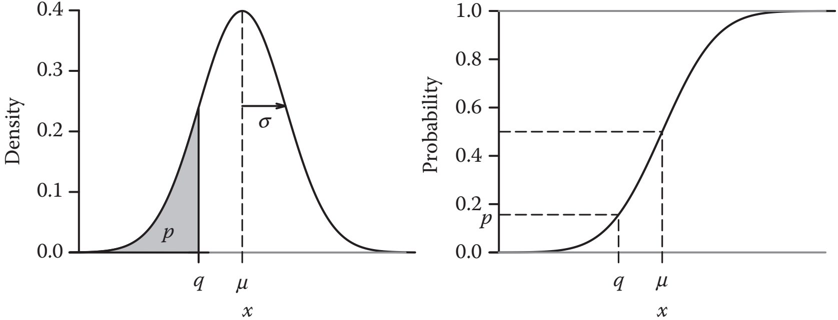


Figure 5: Probability density function (PDF) (left) and cumulative distribution function (right) of the normal distribution.

Other distributions

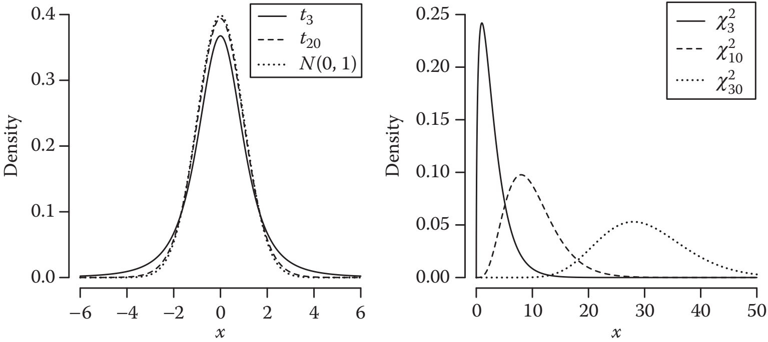


Figure 6: t- distribution and standard normal distribution with different degrees of freedom (left), chi-squared distribution with different degrees of freedom (right).

Other distributions

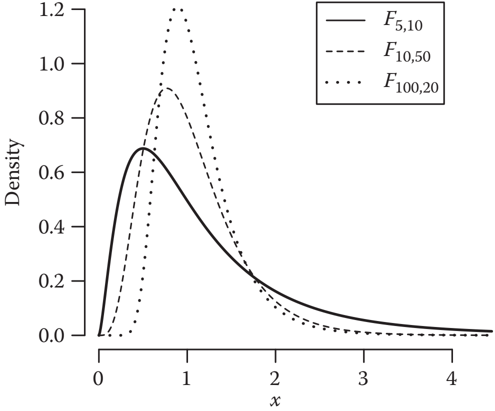


Figure 7: F distribution determined by two parameters.

Quantile Quantile plots

To corroborate that a theoretical distribution, for example the normal distribution, is in fact a good aproximation, we can use quantile quantile plots (qq-plots).

Quantile Quantile plots

To corroborate that a theoretical distribution, for example the normal distribution, is in fact a good aproximation, we can use quantile quantile plots (qq-plots).

- we can generate a vector of random numbers following a normal distribution by:
  $\mathbb{R}$:
  ```
  x.sim <- rnorm(50, mean=0, sd=1) %>% jitter
  ps <- (seq(0,99)+0.5)/100
  qs <- quantile(x.sim, ps)
  normalqs <- qnorm(ps, mean(x.sim), popsd(x.sim))
  ```

Bibliography I

Varmuza, K., & Filzmoser, P.
Introduction to multivariate statistical analysis in chemometrics.
CRC press, 2016.

Massart, D. L., Vandeginste, B. G.N., et al.
Handbook of chemometrics and qualimetrics, part A.
ElSevier, Amsterdam, The netherlands, 1997.

Meichenbächer, M., & Einax, J. W.
Challenges in analytical quality assurance.
Springer Science & Business Media.