1.2 Fundamentals of statistics: the first step towards the data mining

Prof. Massimiliano Grosso
University of Cagliari, Italy
massimiliano.grosso@dimcm.unica.it
GRICU PhD School 2021
Digitalization Tools for the Chemical and Process Industries
March 11, 2021

1

Motivations

- Experimental observations are affected by **uncertainty** deriving from:
  - measurement noise
  - fluctuations in the process that cannot be controlled
  - other …
- outcomes of the experiments cannot be predicted in a deterministic way, even after multiple replicate measurements
- **Goal: model uncertainty in the measurements**
- Uncertainty in the observation can be modelled as a **random process**

2

Basic definitions

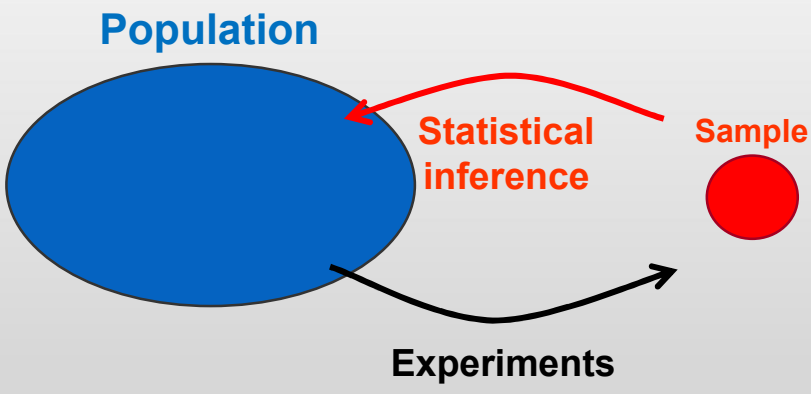
- **Observation**
- Single outcome of the process under investigation.

- **Population**
- Set of **all** the possible observations

- **Sample**
- Set of the observations at our disposal
  - The sample is a **subset** of the population

3

Goal of the statistical inference



The goal is to infer information on the **population** on the basis of the **sample** at our disposal

4

Population characterization

5

Population characterization

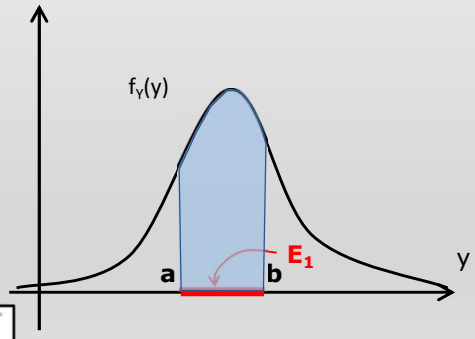
- The goal is to model the regularities that are present in the **population**
- The single experiment cannot be a-priori predicted.
- To the utmost, one can determine the probability that it assumes certain values

- Element of the population as an outcome of a **random variable**

6

Scalar random variables - Definitions

- A scalar random variable (RV) can be uniquely described in terms of its univariate **probability density function $f_Y(y)$**



$$\boldsymbol{P(Y \in E_1) = ?}$$

$$P(Y \in E_1) = P(a < Y < b) = \int_a^b f_Y(y)dy$$

7

Scalar random variables - Definitions

- Properties
- The probability density function is **normalized**

$$\int_{-\infty}^{+\infty} f_Y(y)dy = 1$$

- It assumes always positive values

$$f_Y(y) > 0 \qquad \forall\, y \in \mathbb{R}$$

8

Scalar random variables - Probability density function – Mean

- The **average value** μ or **mean** of a random variable Y is

$$\mu_Y = \int_{-\infty}^{\infty} y\, f_Y(y) dy$$

- It is also referred as the **Expected Value** of Y

$$\mu_Y = E_Y[Y]$$

- it represent the result that is expected, on the average, after infinite observations of the random variable

9

Scalar random variables - Probability density function – Variance

- **Variance** is a measure of the uncertainty of the random variable

$$\sigma_Y^2 = E[(Y - \mu_Y)^2] = \int_{-\infty}^{+\infty} (y - \mu_Y)^2 f_Y(y) dy$$

- It is a scalar always positive
- The higher the variance, the more uncertain is the process

10

The Gaussian random variable

- **Central limit theorem** (in the Lindenberg-Lévy form)
- Suppose $\{X_1, X_2, \ldots X_n\}$ is a sequence of i.i.d. random variables with $E[X_i] = \mu$ and $var[X_i] = \sigma^2 < \infty$.
- Then, as *n* approaches infinity, the random variable

$$\sqrt{n}(\bar{X}_n - \mu)$$

- converges in distribution to a Normal $\mathcal{N}(0, \sigma^2)$

11

The Gaussian random variable

- A random variable is said to be the Gaussian, with parameters μ and σ^2

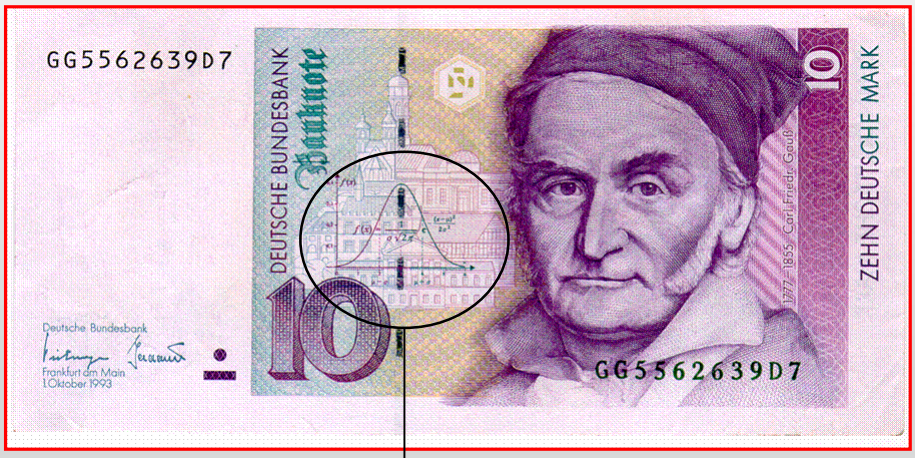
$$Y \sim \mathcal{N}(\mu, \sigma^2)$$

- if its pdf is given by:

$$f_y(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma_Y} exp\left[-\frac{1}{2}\frac{(y-\mu_y)^2}{\sigma_y^2}\right] \qquad \forall y \in \mathbb{R}$$

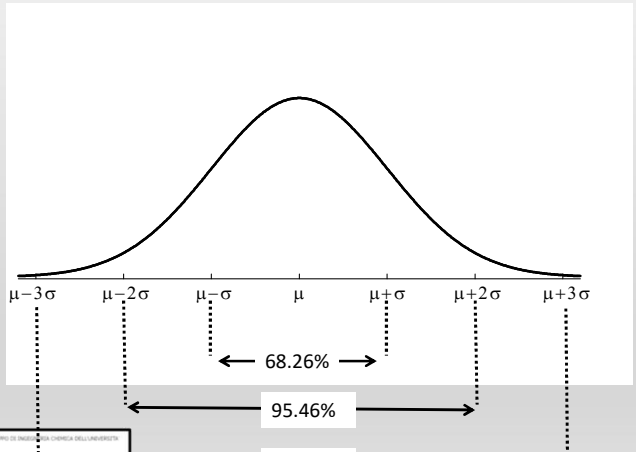
12

Carl Friederich Gauss



Gaussian

Gaussian random variables



$$P(\mu - \sigma < Y < \mu + \sigma) = 0.6827$$
$$P(\mu - 2\sigma < Y < \mu + 2\sigma) = 0.9545$$
$$P(\mu - 3\sigma < Y < \mu + 3\sigma) = 0.9973$$

- This is true for any value of μ and σ

Transformation of random variables

- Given a random variable Y with parameters mean μ_Y and variance σ^2_Y
- For a given transformation

$$Z = g(Y)$$

- one should be interested in the properties of the new random variable Z

15

Random variables –Expected value: definition

- Given a random variable Y and g(•) a measurable function, one can define the scalar

$$E[g(Y)] = \int_{-\infty}^{\infty} g(y)f(y)dy$$

- As the **expected value** of g(Y)

16

Random variables – Properties of the Expected value

- Useful properties of the expected value (it's a linear operator)

$$\boldsymbol{E_Y[c]} = \int_{-\infty}^{+\infty} c\, f_y(y)dy = c\int_{-\infty}^{+\infty} f_y(y)dy = \boldsymbol{c}$$

$$\boldsymbol{E_Y[c\, g(y)]} = \int_{-\infty}^{+\infty} c\, g(y)\, f_y(y)dy = c\int_{-\infty}^{+\infty} g(y) f_y(y)dy = \boldsymbol{cE_Y[g(Y)]}$$

$$\boldsymbol{E_Y[c_1\, g_1(Y) + c_2 g_2(Y)]} = \int_{-\infty}^{+\infty} \big(c_1\, g_1(y) + c_2 g_2(y)\big) f_Y(y)dy =$$

$$c_1\int_{-\infty}^{+\infty} g_1(y) f_y(y)dy + c_2\int_{-\infty}^{+\infty} g_2(y) f_y(y)dy = \boldsymbol{c_1 E_Y[g_1(Y)] + c_2 E_Y[g_2(Y)]}$$

17

Affine transformations of random variables

- Given an **affine** transformation

$$Z = a_0 + a_1 Y$$

- It is easy to demonstrate for the new random variable z

$$\boldsymbol{\mu_Z} = E_Z[Z] = E_Y[a_0 + a_1 Y] = E_Y[a_0] + E_Y[a_1 Y] = a_0 + a_1 E_Y[Y] = \boldsymbol{a_0 + a_1 \mu_Y}$$

$$\boldsymbol{\sigma_Z^2} = E_Z[(Z-\mu_z)^2] = E_Y\left[\big((a_0 + a_1 Y) - (a_0 + a_1\mu_Y)\big)^2\right] =$$
$$E_Y[a_1(Y-\mu_Y)^2] = \boldsymbol{a_1^2\, \sigma_Y^2}$$

18

Affine transformations of random variables

- Affine transformations do not change the nature of the random variable
- In particular, for a Gaussian random variable

$$Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2) \xrightarrow{Z = a_0 + a_1 Y} Z \sim \mathcal{N}(\mu_z, \sigma_Z^2)$$
$$\mu_z = a_0 + a_1 \mu_Y$$
$$\sigma_Z^2 = a_1^2 \sigma_Y^2$$

Affine transformations of random variables

- **Particular case:**
- Given a Gaussian random variable

$$Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$$

- let consider the following affine transformation

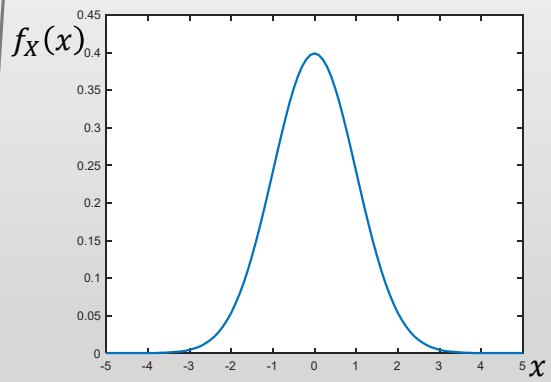
$$X = \frac{Y - \mu_Y}{\sigma_Y}$$

- The new RV X is Gaussian with zero mean and unity variance

$$X \sim \mathcal{N}(0,1)$$

Definition of **Standard** normal random variable

Standard normal distribution



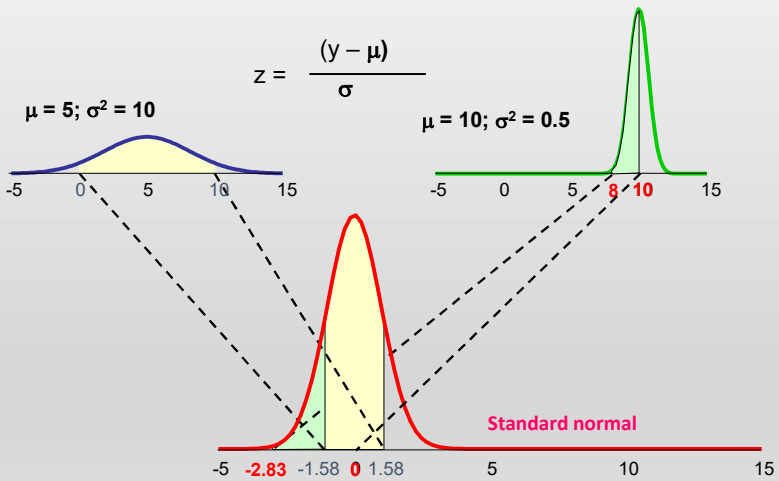
$$f_X(x) = \frac{1}{\sqrt{2\pi}} exp\left(-\frac{x^2}{2}\right)$$

$$P(-1 < X < +1) = 0.6827$$
$$P(-2 < X < +2) = 0.9545$$
$$P(-3 < X < +3) = 0.9973$$

21

Standard normal distribution

$$z = \frac{(y - \mu)}{\sigma}$$



22

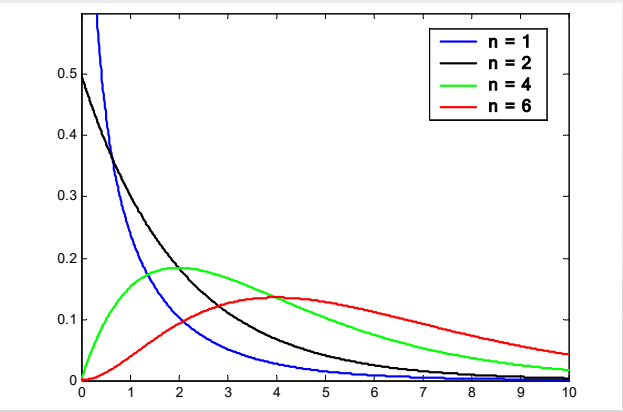
RVs derived by the Gaussian – Chi-square distribution χ^2

- The χ^2-distribution with *n* degrees of freedom is the sum of the squares of *n* independent standard normal random variables

$$Z = \sum_{i=1}^{n} X_i^2 \qquad X_i \sim \mathcal{N}(0,1)$$

- The RV depends on an **integer parameter**, the number of degrees of freedom, *n*

RVs derived by the Gaussian – Chi-square distribution χ^2



- Properties of the chi-square random variable

$$\mu = E[\chi_n^2] = n$$
$$\sigma^2 = 2n$$
$$\chi_n^2 = \begin{cases} A_n y^{\frac{n-2}{2}} e^{-\frac{y}{2}} & y \geq 0 \\ 0 & y < 0 \end{cases}$$

RVs derived by the Gaussian – T-student distribution

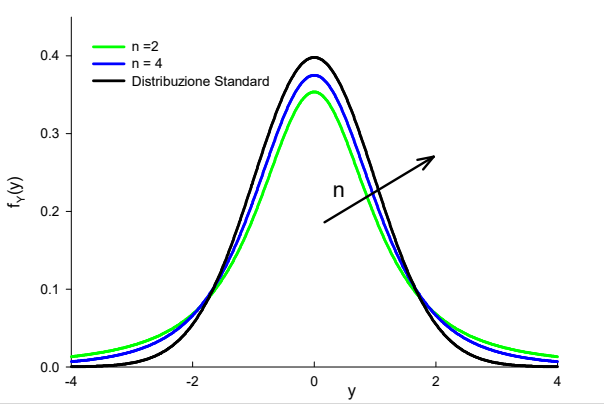
- The ratio between a standard normal RV and a χ_r^2 with *r* d.o.f is a **T-student** random variable **with *r* d.o.f.**

$$T = \frac{Y}{\sqrt{\frac{\chi_r^2}{r}}}$$

- The RV depends on an **integer parameter**, the number of degrees of freedom, *r*

25

RVs derived by the Gaussian – T-student distribution



- Properties of the T-student random variable

$$\mu = E[T_r] = 0$$

$$\sigma^2 = \frac{r}{r-2} \quad (r > 2)$$

$$f_{T_r}(y) = K \frac{1}{\left(1 + \frac{y^2}{r}\right)^{\frac{r+1}{2}}} \quad y \in \mathbb{R}$$

26

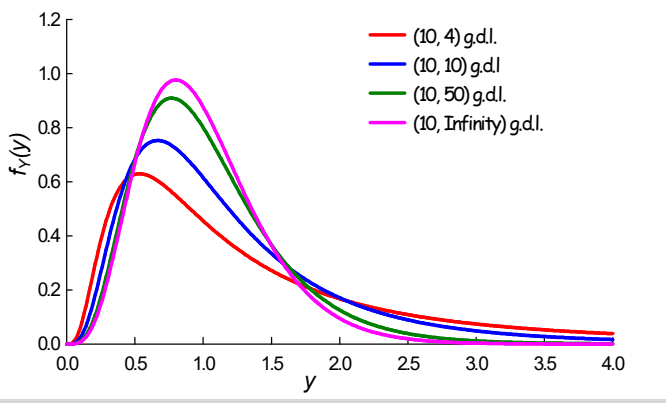
RVs derived by the Gaussian – Fisher distribution

- Given two chi-square random variables with a m and n degrees of freedom, respectively, then the ratio

$$Z = \frac{\frac{\chi_m^2}{m}}{\frac{\chi_n^2}{n}}$$

- is a Fisher random variable
- It has two parameters, m and n

RVs derived by the Gaussian – Fisher distribution



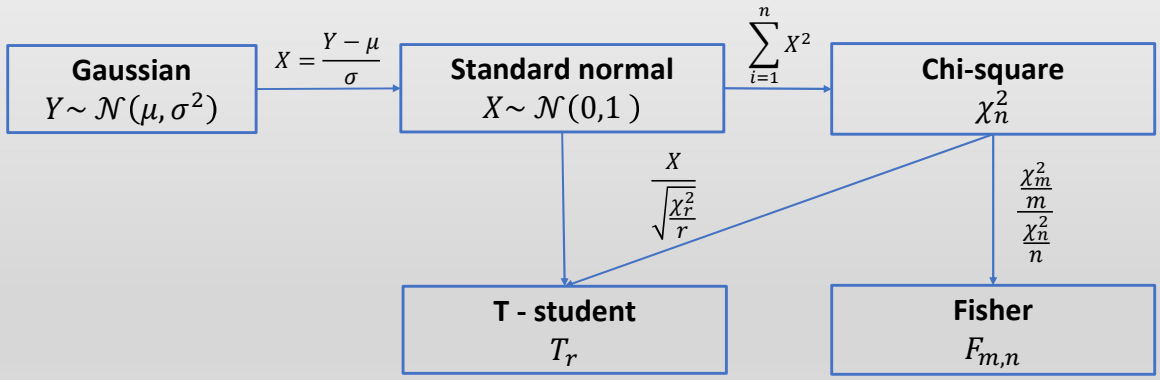
- Properties of the Fisher random variables

$$\mu_Y = \frac{n}{n-2}, \quad (n > 2)$$

$$\sigma_Y^2 = \frac{(m+n-2)2n^2}{m(n-4)(n-2)^2}$$

Scalar random variables - Summary

- Summary and relationship between random variables



Vector random variables

- The process to be investigated is often characterized by more objects that are jointly observed
  - (e.g., concentrations of different chemicals in a batch, wavelengths in IR spectra etc.)
- The concept of **vector random variable** is introduced

$$\underline{\mathbf{Y}} = \|Y_1, Y_2, \ldots, Y_N\|$$

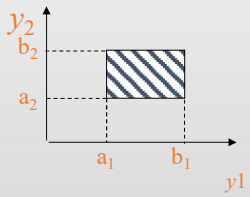
- If N=2, the events are subsets of the Euclidean plane

$$\underline{\mathbf{Y}} = \|Y_1, Y_2\|$$

For sake of simplicity we will mainly discuss 2D cases

Vector random variables – 2D example

- The observation is characterized by two different variables



- We are interested in evaluating the probability

$$P(a_1 < Y_1 < b_1; a_2 < Y_2 < b_2)$$

31

Vector random variable – 2D example

- One can introduce the **joint** probability density function

$$f_{\mathbf{Y}}(y_1, y_2)$$

- Properties

$$P(a_1 < Y_1 < b_1; a_2 < Y_2 < b_2) = \int_{a_2}^{b_2} \int_{a_1}^{b_1} f_{\mathbf{Y}}(y_1, y_2) dy_1\, dy_2$$

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_{\mathbf{Y}}(y_1, y_2) dy_1\, dy_2 = 1$$

32

Vector random variable – Gaussian case

- A vector Gaussian random variable is denoted by

$$\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$$

- The joint pdf can be written as

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{(2\,\pi)^{n/2} det(\mathbf{V})^{n/2}} exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T \mathbf{V}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right)$$

Gaussian vector random variable – 2D case

- A joint probability density function describing a Gaussian vector random variable is characterized by:
- A vector of means $\boldsymbol{\mu}$

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

- A covariance matrix

$$\boldsymbol{V} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$$

Gaussian vector random variable – 2D case

- Terms in the covariance matrix **V**

$\sigma_1^2 = V[Y_1] = E[(Y_1 - \mu_1)^2]$ variance of Y_1

$\sigma_2^2 = V[Y_1] = E[(Y_2 - \mu_2)^2]$ variance of Y_2

$\sigma_{12} = cov[Y_1, Y_2] = E[(Y_1 - \mu_1)(Y_2 - \mu_2)]$ **covariance** of Y_1 and Y_2

35

Gaussian vector random variables – Meaning of the covariance

- It measures the statistical **dependence** between two components
- N.B. The equivalence is rigorous only for Gaussian random variables

Y_1 and Y_2 are independent

$\sigma_{12} = 0$

36

Gaussian vector random variables – Definition of correlation

- The **correlation coefficient** between two components of the random variable is

$$\rho_{12} = \frac{\mathrm{cov}(Y_1, Y_2)}{\sqrt{V(Y_1)V(Y_2)}} = \frac{\sigma_{12}}{\sigma_1 \sigma_2}$$

Pure number

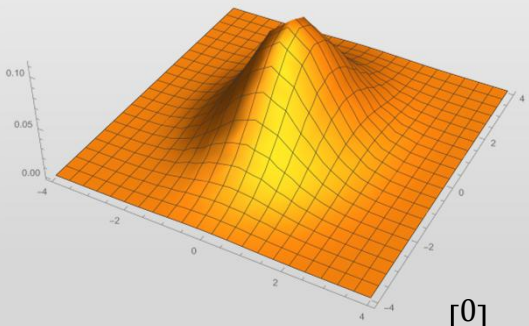
- By definition:

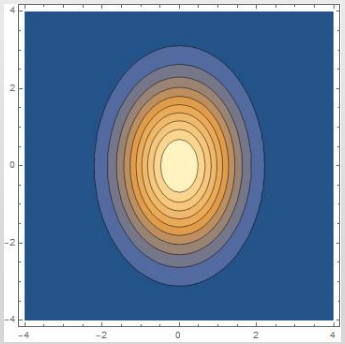
$$-1 \leq \rho_{12} \leq +1$$

37

Gaussian vector random variables – Independent components

- Case ($\sigma_{12} = 0$)

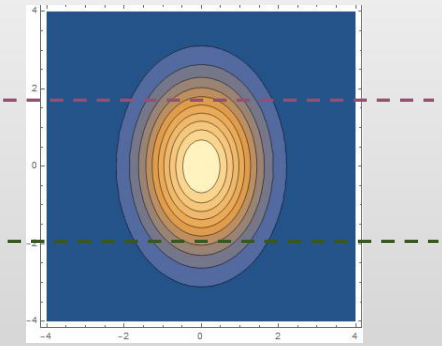




$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \mathbf{V} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$

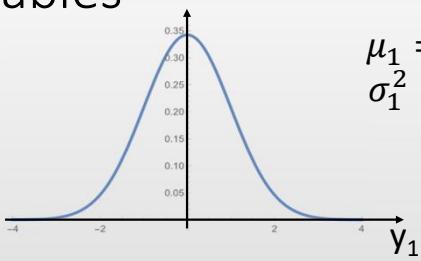
38

Gaussian vector random variables – Independent components



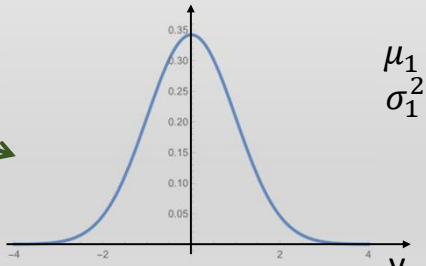
if y_2=2

if y_2=-2



$\mu_1 = 0$
$\sigma_1^2 = 1$

y_1



$\mu_1 = 0$
$\sigma_1^2 = 1$

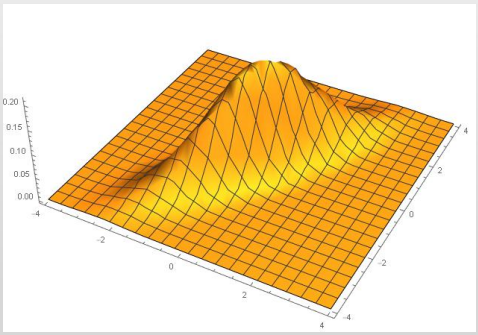
y_1

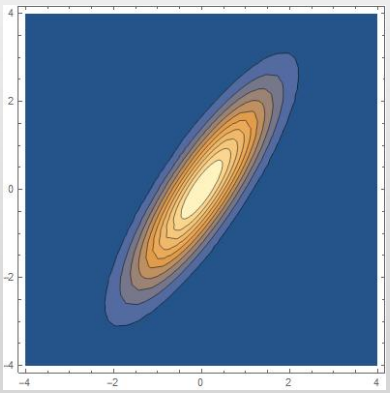
Probability of Y_1 does not depend on the value that Y_2 assumes

39

Gaussian vector random variables – Dependent components

- Case ($\sigma_{12} \neq 0$)

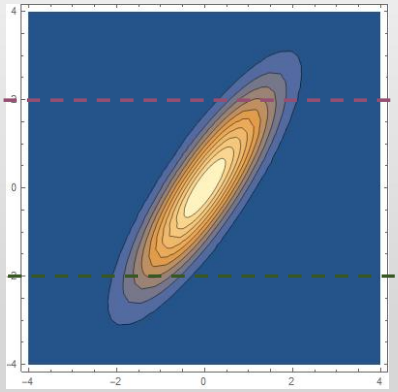




$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \mathbf{V} = \begin{bmatrix} 1 & \alpha \\ \alpha & 2 \end{bmatrix} \qquad \alpha > 0$$

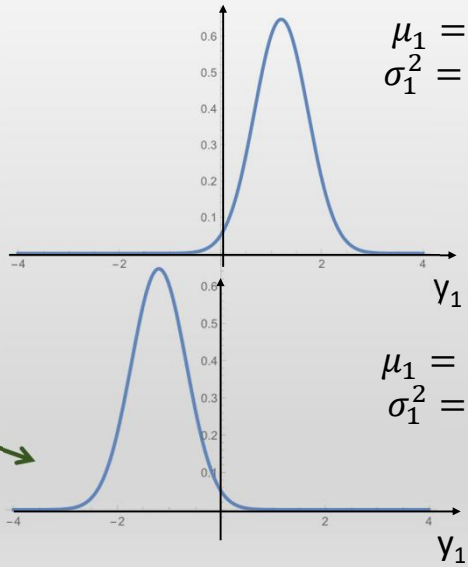
40

Gaussian vector random variables – Dependent components



if y_2=2

if y_2=-2



The value assumed by Y_2 affects the probability of Y_1

41

Vector random variables

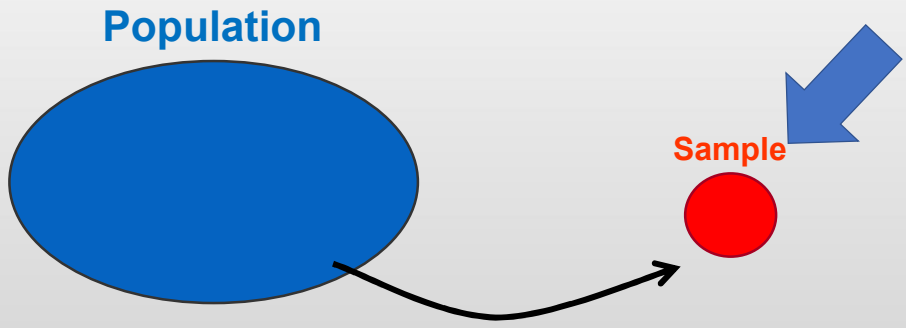
- **Remarks**
- Dependence on components of vector random variables leads to **redundancy** in the data
- Dealing with a parsimonious number of **independent variables** would be in general recommended

42

Sample characterization

43

Sample characterization



- The sample is constituted by a **finite** number of observation, whose **size** is equal to ***l***

44

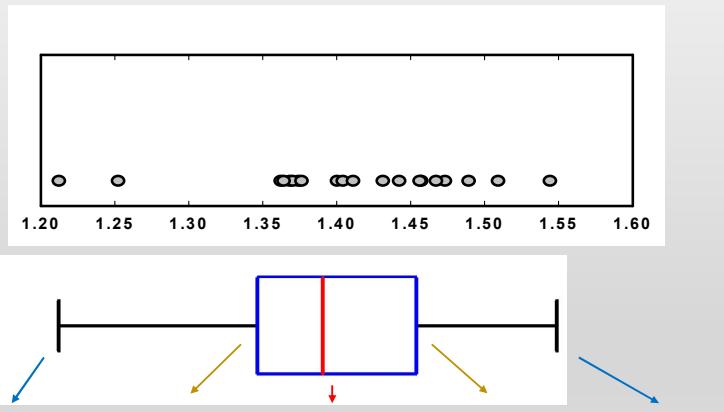
Descriptive statistics - Boxplot

- Given a sample of size *I* a boxplot is a method for graphically depicting groups of numerical data through their quartiles.
- A boxplot is a standardized way of displaying the dataset on the base of:
  - the minimum
  - the maximum
  - the sample median
  - the first and third quartiles
- Useful to give a synthetic though effective representation of the dataset distribution

45

Descriptive statistics - Boxplot

- Example



46

Descriptive statistics – Measures of central tendency

- **Arithmetic mean (or average value)**

$$\bar{y} = \frac{\sum_{i=1}^{I} y_i}{I}$$

- **Median**
  - The 50th percentile
- **Mode**
  - The most frequent value in the data set

47

Descriptive statistics – Measures of statistical dispersion

- **Sample variance**

$$s^2 = \frac{1}{I-1}\sum_{i=1}^{I}(y_i - \bar{y})^2$$

- **Sample standard deviation**

$$s = \sqrt{\frac{1}{I-1}\sum_{i=1}^{I}(y_i - \bar{y})^2}$$

- Useful since it has the same dimensions of the *y* variable

48

Multivariate descriptive statistics

- When we have a vector of measurements, other statistics are introduced to estimate the correlation among the components
- Consider a set of *I* experimental observations, the i-th observation is a vector of dimension J:

$$\mathbf{y}_i = \|y_{i1}, y_{i2}, \ldots, y_{iJ}\|$$

Multivariate descriptive statistics

- Data can be arranged in matrix form

y_{ij}

index ***i*** refers to the i-th experiment

index ***j*** refers to the j-th variable

$$\mathbf{Y} = \begin{bmatrix} y_{11} & \cdots & y_{1j} \\ \vdots & \ddots & \vdots \\ y_{I1} & \cdots & y_{IJ} \end{bmatrix}$$

- For each variable one can compute the (columnwise) mean

$$\bar{y}_{\bullet j} = \frac{1}{I}\sum_{i=1}^{I} y_{ij}$$

Multivariate descriptive statistics – Covariance matrix

- Mean centering allows to evaluate the covariance matrix $\mathbf{C} = \mathbf{X}^T\mathbf{X}$

$$\mathbf{Y} = \begin{bmatrix} y_{11}^{mc} & \cdots & y_{1j}^{mc} \\ \vdots & \ddots & \vdots \\ y_{I1}^{mc} & \cdots & y_{IJ}^{mc} \end{bmatrix} = \begin{bmatrix} y_{11} - \bar{y}_{\bullet 1} & \cdots & y_{1j} - \bar{y}_{\bullet j} \\ \vdots & \ddots & \vdots \\ y_{I1} - \bar{y}_{\bullet 1} & \cdots & y_{IJ} - \bar{y}_{\bullet J} \end{bmatrix}$$

- Indeed, for the element *kl* of **C**

$$\mathbf{C}_{kl} = \frac{1}{I-1}\left\|y_{k1} - \bar{y}_{\bullet 1}, y_{k2} - \bar{y}_{\bullet 2}, \ldots, y_{kJ} - \bar{y}_{\bullet J}\right\| \cdot \left\| \begin{matrix} y_{l1} - \bar{y}_{\bullet 1} \\ y_{l2} - \bar{y}_{\bullet 2} \\ \vdots \\ y_{lJ} - \bar{y}_{\bullet J} \end{matrix} \right\| = \frac{1}{I-1}\left\|y_{k1}^{mc}, y_{k2}^{mc}, \ldots, y_{kJ}^{mc}\right\| \cdot \left\| \begin{matrix} y_{l1}^{mc} \\ y_{l2}^{mc} \\ \vdots \\ y_{lJ}^{mc} \end{matrix} \right\|$$

$$\frac{1}{I-1}\sum_{m=1}^{I}(y_{mk} - \bar{y}_{\bullet k})(y_{ml} - \bar{y}_{\bullet l}) = \frac{1}{I-1}\sum_{m=1}^{I} y_{mk}^{mc} y_{ml}^{mc}$$

51

Multivariate descriptive statistics – Covariance matrix

- The diagonal elements of **C** are the *variance* related to the j-th variable

$$\mathbf{C}_{jj} = \frac{1}{I-1}(\mathbf{Y}^T\mathbf{Y})_{\boldsymbol{jj}} = \sum_{m=1}^{I}(y_{mj} - \bar{y}_{\bullet j})^2 = \sum_{m=1}^{I}(y_{mj}^{mc})^2$$

- The off-diagonal terms are the *covariance* between the k-th and the l-th variable

$$\mathbf{C}_{kl} = \frac{1}{I-1}(\mathbf{Y}^T\mathbf{Y})_{\boldsymbol{kl}} = \sum_{m=1}^{I}(y_{mk} - \bar{y}_{\bullet k})(y_{ml} - \bar{y}_{\bullet k}) = \sum_{m=1}^{I} y_{mk}^{mc} y_{ml}^{mc}$$

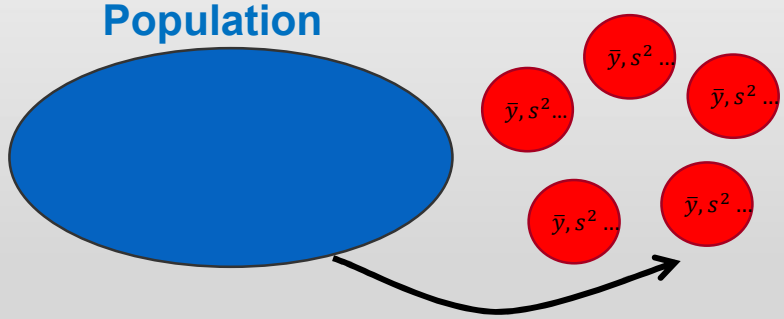
52

Statistical inference

53

Statistical inference

- The statistics previously introduced in the descriptive part refer to the current sample under investigation



- **In principle**, one can extract infinite samples, each of them with different statistics
- **Goal**:
- Characterize the «population» of the samples

54

Statistical inference

- The parameters associated to the sample
  - mean $\bar{y}$,
  - variance s^2

  differ as new samples are taken into account
- These statistics are in turn **random variables**

55

Statistical inference – Characterization of the estimator mean

- Given the random variable:

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} \qquad Y_i \sim \mathcal{N}(\mu, \sigma^2)$$

- the expected value of $\bar{Y}$ is computed as

$$\boldsymbol{E[\bar{Y}]} = E\left[\frac{\sum_{i=1}^{n} Y_i}{n}\right] = \frac{\sum_{i=1}^{n} E[Y_i]}{n} = \frac{\sum_{i=1}^{n} \mu}{n} = \frac{n\,\mu}{n} = \boldsymbol{\mu}$$

- As it regards the variance of $\bar{Y}$

$$\boldsymbol{var[\bar{Y}]} = var\left[\frac{\sum_{i=1}^{n} Y_i}{n}\right] = \frac{\sum_{i=1}^{n} var[Y_i]}{n^2} = \frac{\sum_{i=1}^{n} \sigma^2}{n^2} = \frac{n\,\sigma^2}{n^2} = \boldsymbol{\frac{\sigma^2}{n}}$$

56

Statistical inference – Characterization of the estimator mean

- Since the sum is a linear operator, the mean is **a Gaussian random variable**.
- Given *n* observations of the RV $Y_i \sim \mathcal{N}(\mu, \sigma^2)$ carried out at the same conditions,

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Statistical inference – Characterization of the estimator variance

- As it regards the **variance**, it follows a **chi-square distribution**

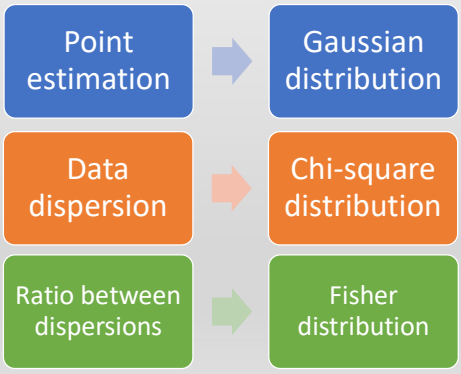
$$\frac{s^2}{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(Y_i - \bar{Y})^2}{\sigma^2}$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{Y_i - \bar{Y}}{\sigma}\right)^2 = \frac{1}{n-1}\chi^2_{n-1}$$

This may be approximated as a Standard random variable

Statistical inference – Concepts to remember

- **Analysis of a data set as an outcome of a random variable**



59

References

1. Bruce, P.C., Bruce A. *Practical statistics for data scientists : 50 essential concepts*. Sebastopol, CA: O'Reilly, 2017
2. Mendenhall, W. M., Sincich T. *Statistics for engineering and the sciences*. Boca Raton: CRC Press, Taylor & Francis Group, 2016.
3. Montgomery, D. C., Runger G.C. *Applied statistics and probability for engineers*. Hoboken, NJ: Wiley, 2018.
4. Papoulis, A., Pillai S.U. *Probability, random variables, and stochastic processes*. New Delhi New York: Tata McGraw-Hill, 2002.

60