https://kesmarag.gitlab.io/courses/mem264

MEM-264 Applied Statistics

Department of Mathematics and Applied Mathematics, University of Crete

Costas Smaragdakis (kesmarag@gmail.com)

1st lecture : 08-02-2021

Definition : Statistics

Statistics is a group of methods used to collect, analyze, represent, interpreter data and to make decisions.

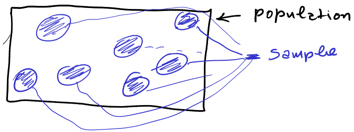
Aspects of Statistics

- Theoretical : proves the theorems. (Parametric Statistics - MEM 262)
- Applied : uses the theorems in real-world applications.

  Machine learning, data analysis, big data

Types of Applied Statistics

- Descriptive Statistics (Περιγραφική Στατιστική)
  - Descriptive statistics consists of methods for organizing, visualizing, and describing data by means of tables, graphs (plots) and summary measures.
- Inferencial Statistics (Επαγωγική Στατιστική)
  - Inferencial statistics consists of methods that help the scientists to make conclusions (or make decisions) about the statistical population using just a statistical sample.



.

Definition : Population

A statistical population consists of all elements (objects) whose characteristics are being studied.

Definition : Sample

A portion of the population selected for study is referred to as a statistical sample.

Definition : Representative sample

A statistical sample that represents the characteristics of the population as closely as possible is called a representative sample of the population.

Population and Sample

characteristic under study is the color
Color (magenta, blue, purple)

Population | the size of population is 26

16 – magenta
7 – blue
3 – purple

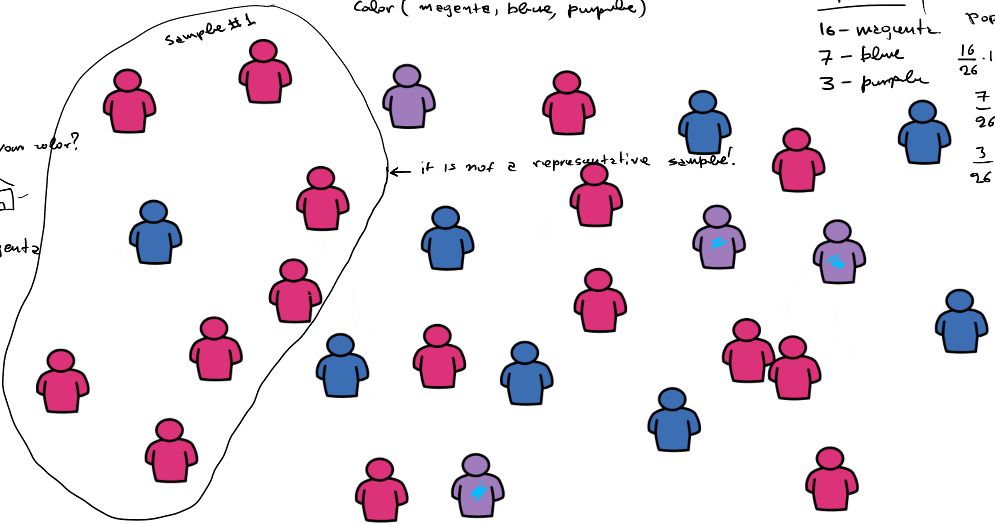
$\frac{16}{26} \cdot 100\% = 61.6\%$

$\frac{7}{26} \cdot 100\%$ 26.9%

$\frac{3}{26} \cdot 100\%$ 11.5%

question:
what is your color?

7 - magenta
1 - blue

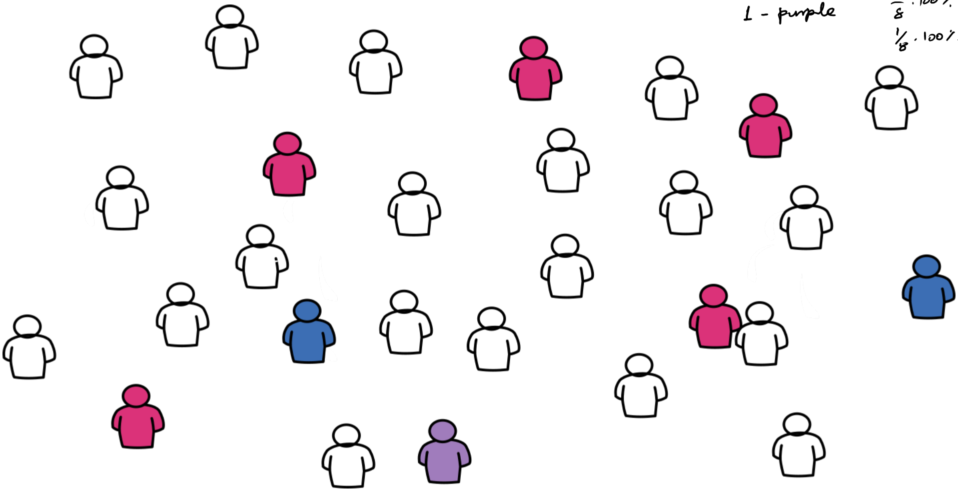


5 - magenta
2 - blue
1 - purple

$\frac{5}{8} \cdot 100\% = 62.5\%$

$\frac{2}{8} \cdot 100\% = 25\%$

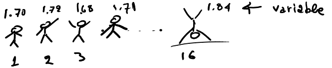
$\frac{1}{8} \cdot 100\% = 12.5\%$



Definition : Random sample

If each element of the population has a chance of being selected then the sample is called a random sample.

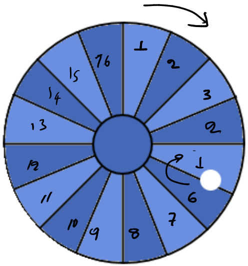
Example : Roulette's wheel sampling



selected twice

{1.71, 1.70, 1.84, 1.71}

Random Sample



Quatitative variables (Ποσοτικές μεταβλητές)

Variables that can be measured numerically.

- Discrete variables (eg: number of cars, houses, accidents)
- Continuous variables (eg: length, temperature)

Qualitative variables (Ποιοτικές μεταβλητές)

Variables whose values can be classified into two or more non-numerical categories.

- Ordinal (διατάξιμες) variables : we can consider an order relation among their values (eg: likert scale, level of education)
- Nominal (Ονομαστικές) variables : there is no any order relation among their values (eg: color, gender, brand).

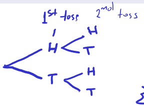
Definition : Sample space Ω

A sample space Ω is a set of all possible outcomes of a random experiment.

Definition : Event

An event A represents a subset of the Ω.

Example : tossing a coin twice.



$\Omega = \{HH, HT, TH, TT\}$ ← sample space

$\{HH\}$ is an event

$\{HH, HT\}$ is also an event!

Example : tossing a coin forever.

$\Omega = \{ (\omega_1, \omega_2, \ldots\ldots) \ , \ \omega_i \in \{H, T\} \}$

$A = \{ (H, H, \ldots\ldots) \}$ $\leftarrow$ In each throw we get Head!

Example : the average temperature of Heraklion tomorrow.

$\Omega = [-10, 50]\ °C$

- A sigma algebra $\mathcal{G}$ is a collection of subsets of the sample space Ω, which includes all the events of interest.

Mathematical definition

A collection $\mathcal{G}$ of subsets of Ω, it is a σ-algebra iff:

- $\Omega \in \mathcal{G}$
- If $A \in \mathcal{G}$, then $\underbrace{\Omega - A}_{A^c} \in \mathcal{G}$
- If $A_j \in \mathcal{G}$, then $\cup_j A_j \in \mathcal{G}$

These properties of a sigma algebra are inportant for defining a **probability space** $(\Omega, \mathcal{G}, P)$ where P is a measure of probability.

events space

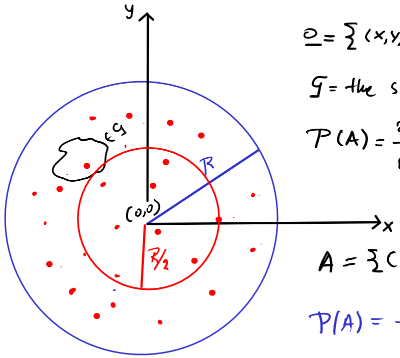
$$P : \mathcal{G} \to [0, 1]$$

- $P(\Omega) = 1.$
- $A_i \cap A_j = \emptyset$ for any i, j then $P(\cup_j A_j) = \sum_j P(A_j)$.

  $i \neq j$

Probability Space : $(\Omega, \mathcal{G}, P)$

Example : Uniformy distributed points inside a circle of radius R



$\Omega = \{(x,y) : x^2+y^2 < R^2\}$

$\mathcal{G}$ = the set of all subsets of Ω = 2^{Ω} ← power set of Ω

$P(A) = \dfrac{\text{area}(A)}{\text{area}(\Omega)}$, for all $A \in \mathcal{G}$

$A = \{(x,y) : x^2+y^2 < (\tfrac{1}{2}R)^2\}$

$P(A) = \dfrac{\pi(\frac{1}{2}R)^2}{\pi R^2} = \dfrac{1}{4}$

Definition : Random variable

Random variable is a (measurable) function from the sample space Ω to $\mathbb{R}$.

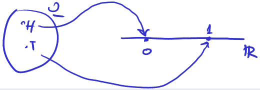
$$X : \Omega \to \mathbb{R}$$

Definition : Random vector

Random vector is called a vector whose elements are random variables.

Example : tossing a coin

$\Omega = \{H, T\}$



$$X(\omega) = \begin{cases} 0, & \text{if } \omega = H \\ 1, & \text{if } \omega = T \end{cases}$$

Example : tossing a coin twice

$\Omega = \{HH, HT, TH, TT\}$

$$X(\omega) = \begin{cases} 0, & \text{if } \omega = HH \\ 1, & \text{if } \omega = HT \\ 2, & \text{if } \omega = TH \\ 3, & \text{if } \omega = TT \end{cases}$$

if we are interested in studying the number of heads.

$Y(\omega) = \{\text{number of heads in } \omega\}$

for example if $\omega = HT$ then $Y(\omega) = 1$.

Definition : Discrete random variables

X is a discrete random variable if it takes countably many values

$$\{x_1, x_2, \dots\}$$

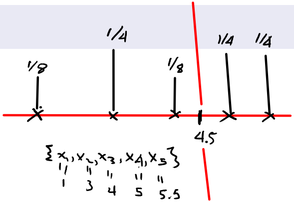
Definition : Probability mass function

$$f_X(x) := P\{X = x\},\ x \in \{x_1, x_2, \dots\}$$

Definition : Distribution function $F_X(x)$

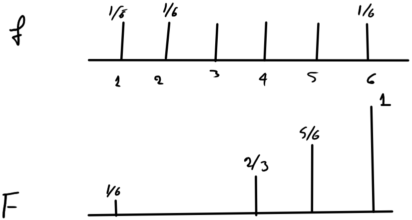
$$F_X(x) = P(X \le x) = \sum_{x_j \le x} f_X(x_j).$$

$$F_X(4.5) = 1/8 + 1/8 + 1/4 = 1/2$$



Example : Rolling a fair dice

$X(\omega) = \omega \quad , \quad \omega \in \Omega = \{1,2,3,4,5,6\}$

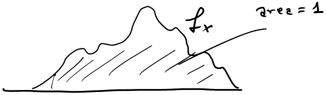


Continuous random variables

Definition : Continuous random variables and probability density function

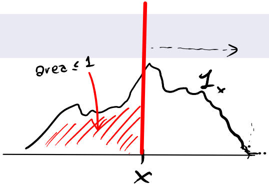
A random variable X is continuous if there exists a function f_X such that $f_X(x) \geq 0$, $\int_{-\infty}^{\infty} f_X(x)dx = 1$, and for every $a \leq b$,

$$P\{X \in (a,b)\} = \int_a^b f_X(x)dx.$$



Definition : Distribution function $F_X(x)$

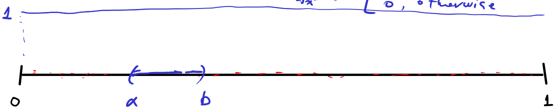
$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(x)dx.$$



Example : Let X be uniformly distributed on $(0, 1)$

$$\Omega = \{x : x \in (0,1)\}$$

$$f_X(x) = \begin{cases} 1, & x \in (0,1) \\ 0, & \text{otherwise} \end{cases}$$



$$X(\omega) = \omega, \quad \omega \in \Omega$$

$$0 < \alpha < b < 1$$

$$\mathbb{P}\{X \in (\alpha, b)\} = \int_\alpha^b f_X(x)\,dx = b - \alpha.$$

For the statistical analysis of the properties of a random sample, we wish to associate each variable under consideration with a random variable.

- For quantitative variables the conversion is straightforward !!

Example : number of customers in a store on a particular day

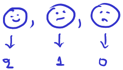
$\Omega = \{0, 1, \dots\}$ $\quad X(\omega) = \omega, \ \omega \in \Omega$ $\quad$ discrete random variable

Example : time between two successive events (eg: earthquakes of magitude m > 8)

$\Omega = [0, +\infty)$ $\quad X(\omega) = \omega, \ \omega \in \Omega$

- For ordinal qualitative variables is also simple.

Example : Customer satisfaction with a company's product

$\Omega = \{$  $\}$ $\quad X(\omega) \in \{0, 1, 2\}$

2 : 10 $\quad$ 0 : 20

1 : 20

the average level satisfaction

4/5

$\rightarrow \dfrac{2 \cdot 10 + 0 \cdot 20 + 1 \cdot 20}{50} = \dfrac{40}{50} = \dfrac{4}{5}$

Associating variables with random variables

- For nominal qualitative variables typically we employ binary random vectors

Example : What causes this sound?

observations

$$\underline{O} = \{ \underset{\substack{\downarrow \\ 0}}{\overset{20}{Car}},\ \underset{\substack{\downarrow \\ 1}}{spaceship},\ \underset{\substack{\downarrow \\ 2}}{\overset{20}{motorcycle}} \}$$

$$\frac{20 \cdot 0 + 20 \cdot 2}{40} = 1$$

expectation

Solution : we employ random vector

$$Car \to \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \qquad Spaceship \to \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \qquad moto \to \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

$$\frac{1}{40}\left(20 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + 20 \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right) = \begin{bmatrix} 1/2 \\ 0 \\ 1/2 \end{bmatrix}$$

express probabilities

Posterior

Dummy Variables : N categories $\longrightarrow$ N vectors of dimension N-1

$$\underset{Car}{\begin{bmatrix} 0 \\ 0 \end{bmatrix}} \qquad \underset{Spaceship}{\begin{bmatrix} 1 \\ 0 \end{bmatrix}} \qquad \underset{moto}{\begin{bmatrix} 0 \\ 1 \end{bmatrix}}$$