Introduction to Statistics summary

A. Tsanakas, Prof. Ioannis Pitas
Aristotle University of Thessaloniki
pitas@csd.auth.gr
www.aiia.csd.auth.gr
Version 2.7

Artificial Intelligence & Information Analysis Lab

Introduction to Statistics

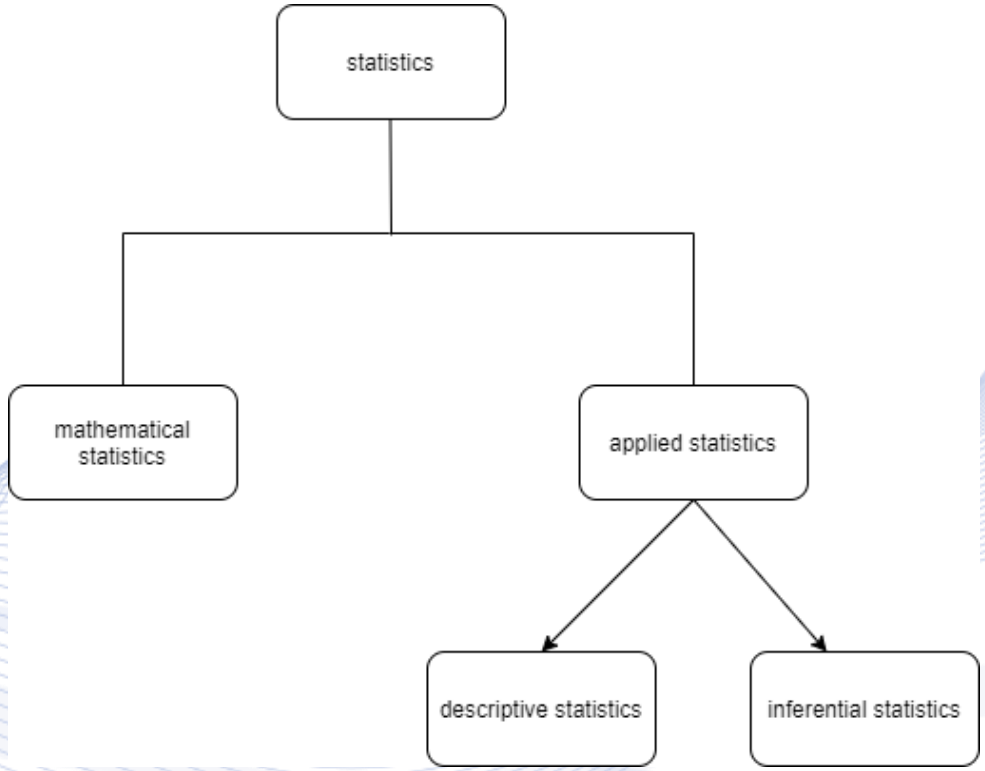
- Introduction
- Random Variable
- Types of Data
- Sampling
- Descriptive statistics
  - Graphs
  - Measures

Introduction

- ***Statistics*** is a mathematical science that aims to help in the study of the phenomena or properties of a population using a selected part of the population or the phenomenon.

Introduction

- Mathematical and theoretical study of statistics, which is based on probability theory and mathematical analysis, is called theoretical statistics or mathematical statistics.
- On the other hand, applied statistics is divided into 2 major categories: descriptive statistics and statistical inference.



Introduction to Statistics

- Introduction
- Random Variable
- Types of Data
- Sampling
- Descriptive statistics
  - Graphs
  - Measures

Random Variable

Let a Borel field $\mathcal{F}$, $\mathcal{S} = \mathbb{R} = \{z\}$ be used to define an ***event***:

$$\{z : X(z) \leq x\} \in \mathcal{F}.$$

- $X(z)$: ***random variable*** having probability $P\{z: X(z) \leq x\}$ for every real x.

Random Variable

Function $F_X(x) = P\{z : X(z) \leq x\}$
is called ***probability distribution*** of X.

- Also called ***cumulative distribution function*** (***cdf***) of X.

Propability Density Fuction

The ***probability density function*** (***pdf***) $f_X(x)$ of variable X is the derivative of the probability distribution function $F_X(x)$:

$$f_X(x) = \frac{d}{dx} F_X(x).$$

Pdf properties:

- Since $F_X(x)$ is non-decreasing, then $f_X(x) \geq 0$.

$$F_X(x) = \int_{-\infty}^{x} f_X(x)dx,$$

$$\int_{x_1}^{x_2} f_X(x)dx = F_X(x_2) - F_X(x_1).$$

Introduction to Statistics

- Introduction
- Random Variable
- Types of Data
- Sampling
- Descriptive statistics
  - Graphs
  - Measures

Types of Data

- ***Categorical Data***: This type of data is attributes treated as distinct symbols or just names. The types of categorical data is:

1. Nominal Data: This is a type of data used to name variables without providing any numerical value.
2. Ordinal Data: This is a data type with a set order or scale to it.

- Examples of categorical variables are ***race***, ***sex***, ***age group***, and ***educational level***.

Types of Data

- ***Numeric data***: It is a data type expressed only ordinal numbers.

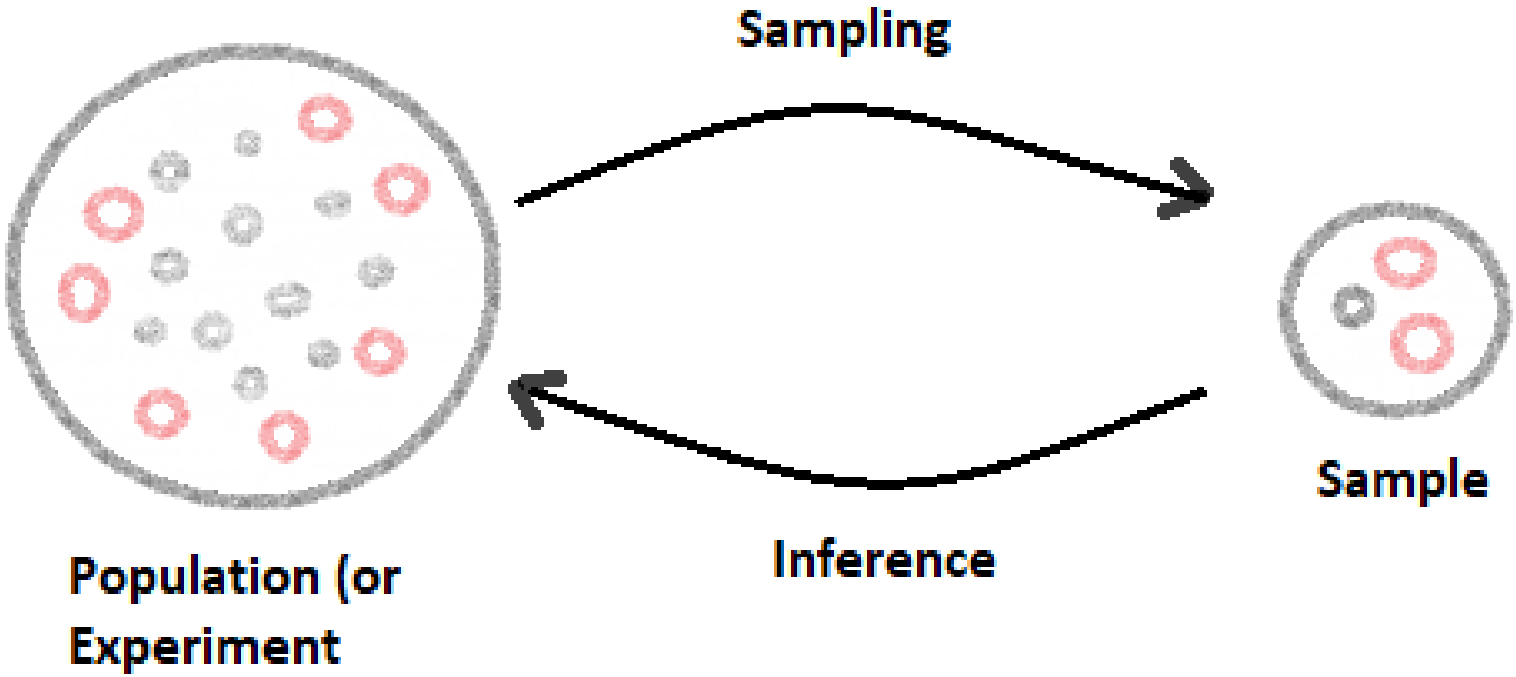
- Examples of numerical data are the number of ***IQ***, ***movies watched***, ***age***, ***height***, ***weight***, etc.

- To graph numerical data, one uses ogive graphs, dot plots, stem, histograms, box plots, leaf graphs, and scatter plots.

Introduction to Statistics

- Introduction
- Random Variable
- Types of Data
- Sampling
- Descriptive statistics
  - Graphs
  - Measures

Sampling



Sampling - Types

- There are four main types of probability sample:
  - ***Simple random*** sampling
  - ***Systematic*** sampling
  - ***Stratified*** sampling
  - ***Cluster*** sampling

Introduction to Statistics

- Introduction
- Random Variable
- Types of Data
- Sampling
- Descriptive statistics
  - Graphs
  - Measures

Descriptive statistics

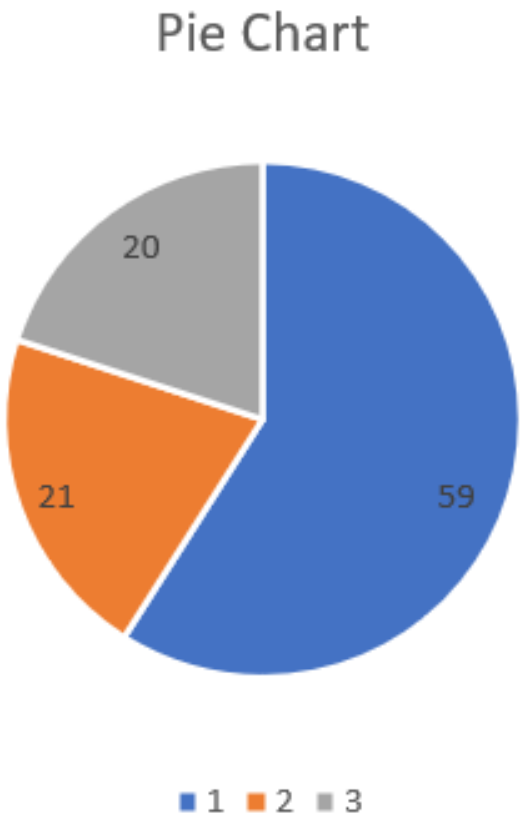
- ***Descriptive statistics*** are useful to describe the basic features of the data in a statistics study.

- The basis of every quantitative analysis of data is descriptive statistics and graphics analysis.

Introduction to Statistics

- Introduction
- Random Variable
- Types of Data
- Sampling
- Descriptive statistics
  - Graphs
  - Measures

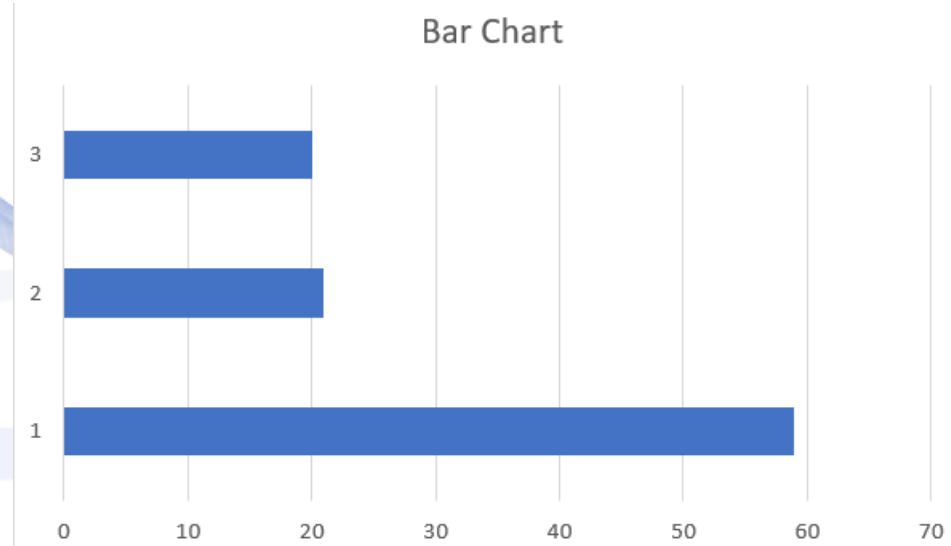
Statistical Graphs - Pie Chart

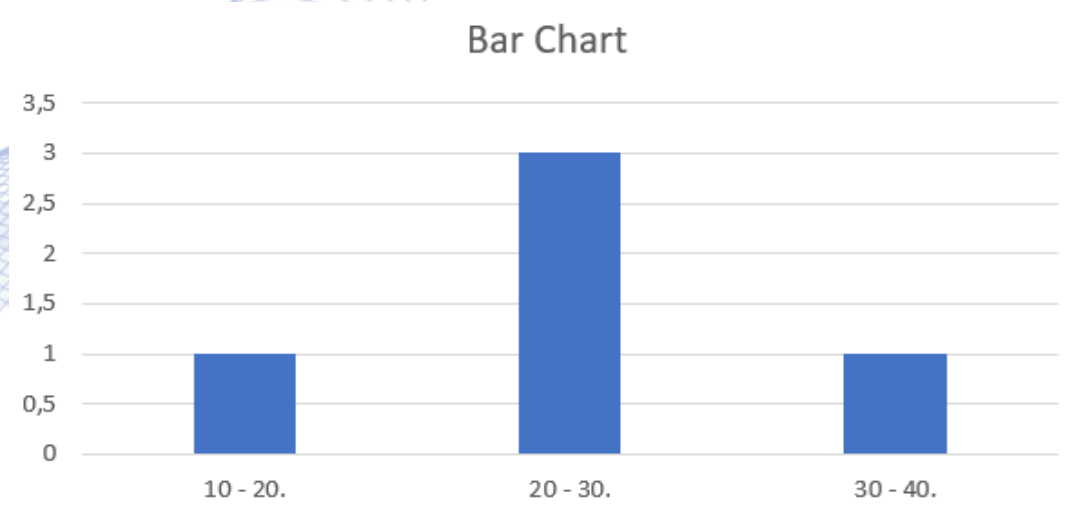


An important tool to graphically represent categorical data is the ***pie chart***. It is a circular chart divided into sectors, illustrating relative magnitudes in frequencies or precents. In a pie chart, the area is proportional to the quantity it represents.

Statistical Graphs - Bar Charts

Bar graph is often preferable to pie charts s as a way to display categorical data, because the human eye is good at judging linear measures and poor at judging relative areas. The bars can be plotted vertically or horizontally.

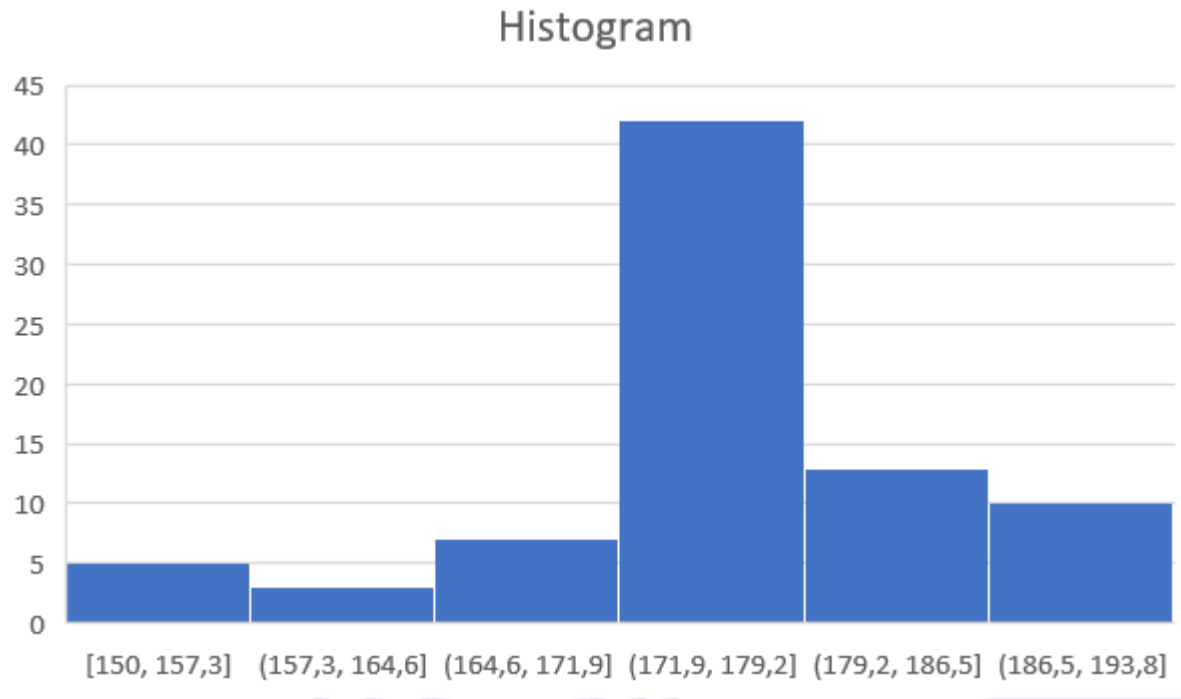




Statistical Graphs - Histogram

- A ***histogram*** is an approximate representation of the distribution of numerical data.

- For example, the histogram of the height of 80 people in centimeters:



Introduction to Statistics

- Introduction
- Random Variable
- Types of Data
- Sampling
- Descriptive statistics
  - Graphs
  - Measures

Measures of Central Tendency

- ***Measures of central tendency*** help you find the middle, or the average, of a data set. The 2 most common measures of central tendency are the mean and median.

Median

The median take the middle value for $x_1, x_2, \ldots, x_n$ after the data has been sorted from smallest to largest: $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$.

$(1), x(2), \ldots, x(n)$.

- If n is odd: $median(x) = x_{\left(\frac{(n+1)}{2}\right)}$.
- If n is even : $median(x) = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}$.

Mean

For a collection of numeric data, $x_1, x_2, \ldots, x_n$, the sample mean is the numerical average:

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = \frac{1}{n}\sum_{1}^{n} x_i$$

Or, if the value x occurs $n(x)$ times in the data:

$$\bar{x} = \frac{1}{n}\sum_x xp(x)$$

Measures of Variation

- The measures of ***central tendency*** fail to find variability within a data set.

- Measures of ***dispersion/variation*** describe the similarity of a variable's values

- This type of measure only applies to ***interval***, ***ordinal***, and ***ratio data.***

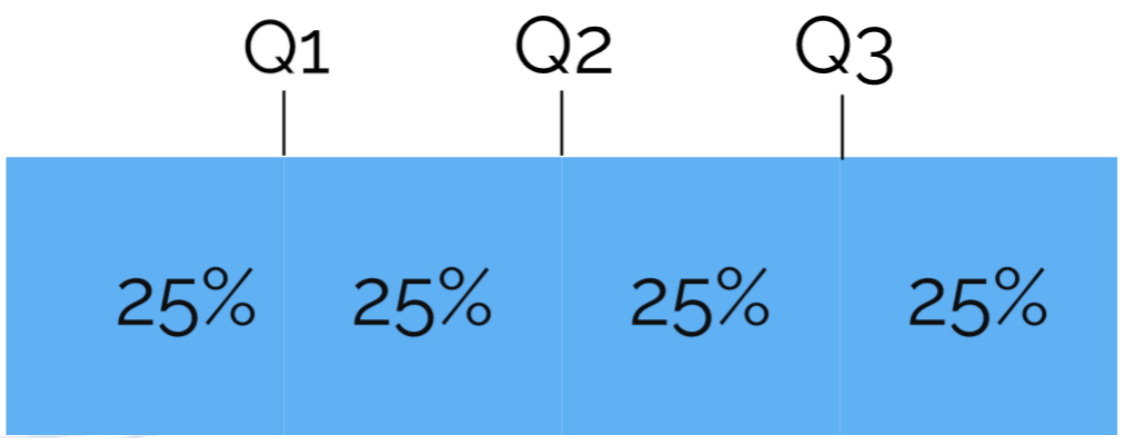
Variance and Standard Deviation

- The ***sample variance*** averages the square of the differences from the mean:
  - $Var(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})$
- Also, the ***sample standard deviation***, s_x, is the square root of the sample variance.
  - $s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})}$

Quartiles

- ***Quartiles*** divide a dataset that is rank-ordered into four equal parts.

- The first, second, and third quartiles are the values that come out after the division. They are denoted by Q_1, Q_2, and Q_3, respectively.



Bibliography

[FIS09] Fisher, Murray J., and Andrea P. Marshall. "Understanding descriptive statistics." *Australian Critical Care* 22.2 (2009): 93-97.

[EVJ04] Evans, Michael J., and Jeffrey S. Rosenthal. *Probability and statistics: The science of uncertainty*. Macmillan, 2004.

[LAM05] Larsen, Richard J., and Morris L. Marx. *An introduction to mathematical statistics*. Prentice Hall, 2005.

[LJK99] Liu, Regina Y., Jesse M. Parelius, and Kesar Singh. "Multivariate analysis by data depth: descriptive statistics, graphics and inference,(with discussion and a rejoinder by liu and singh)." *The annals of statistics* 27.3 (1999): 783-858.

[PCJ15] Peck, Roxy, Chris Olsen, and Jay L. Devore. *Introduction to statistics and data analysis*. Cengage Learning, 2015.

[LAN17] Lane, David M., et al. *Introduction to statistics*. Houston, TX: Rice University, 2017.

Q & A

Thank you very much for your attention!

More material in
http://icarus.csd.auth.gr/cvml-web-lecture-series/

Contact: Prof. I. Pitas
pitas@csd.auth.gr