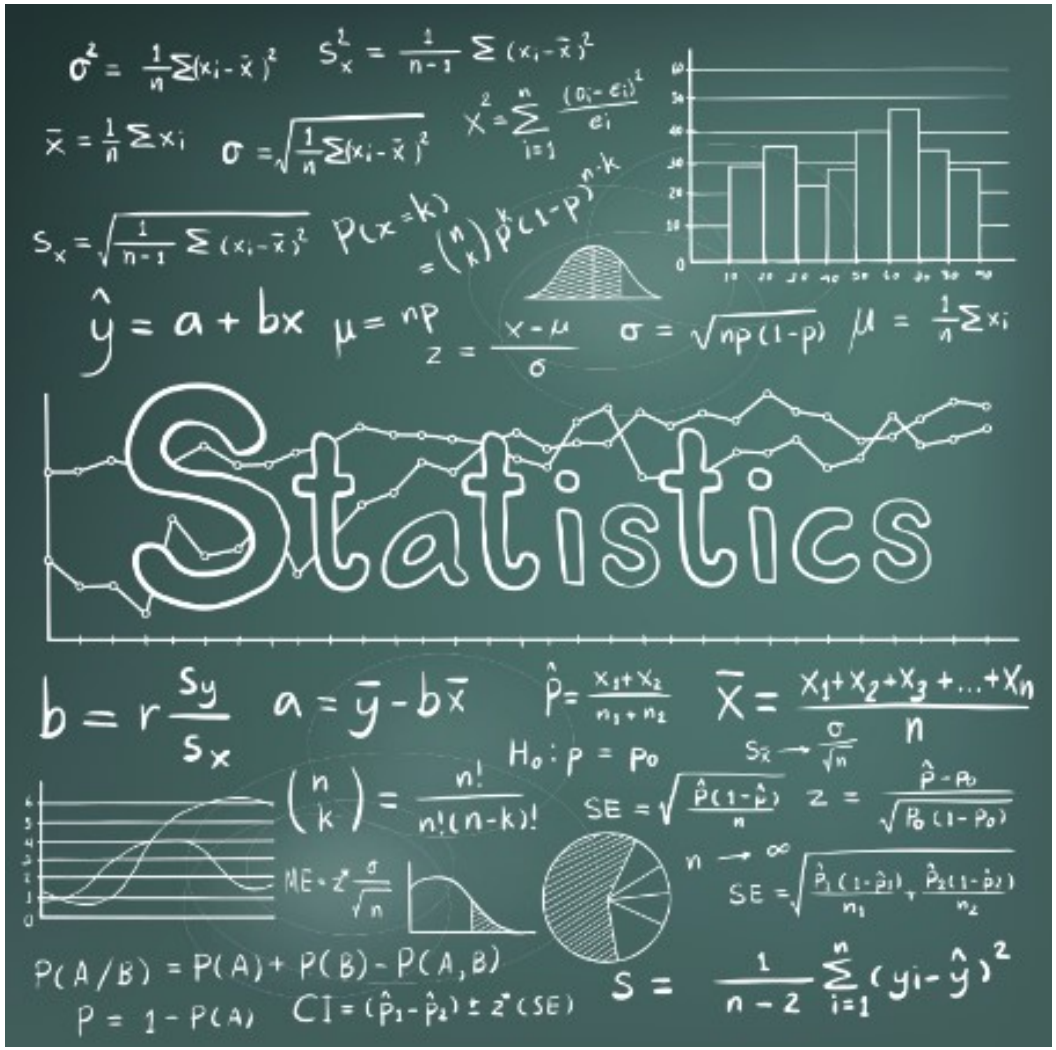
Applied Computer Science and Artificial Intelligence



Monia Ranalli

monia.ranalli@uniroma1.it

Outline

- **What is statistics**
- **Syllabus**
- **Material**
- **Exams**
- **Organization**

Statistics: Using Data to Answer Statistical Questions

Examples of statistical questions

1. What is the selling price (price) of an apartment in USA cities?
2. What is the square feet (sqft) of living space?
3. Is there any relation between price and sqft?

In order to answer the previous questions, we collect the following **DATA**.

The data are a random sample of records of resales of homes from Feb 15 to Apr 30, 2013 from the files maintained by the Albuquerque Board of Realtors. It is collected by multiple listing agencies in many cities and is used by realtors as an information base.

Number of cases (observations): 117

Column Names:

1. **PRICE** = Selling price ($ hundreds)
2. **SQFT** = Square feet of living space
3. **AGE** = Age of home (years)
4. **FEATS** = Number out of 11 features (dishwasher, refrigerator, microwave, disposer, washer, intercom, skylight(s), compactor, dryer, handicap fit, cable TV access)
5. **NE** = Located in northeast sector of city (1) or not (0)
6. **COR** = Corner location (1) or not (0)
7. **TAX** = Annual taxes ($)
8. **FLO** = Main flooring (carpet, hardwood, ceramic tile)

Data

PRICE	SQFT	AGE	FEATS	NE	COR	TAX	FLO
2050	2650	13	7	1	0	1639	carpet
2080	2600	4	4	1	0	1088	carpet
2150	2921	3	6	1	0	1635	hardwood
1999	2580	4	4	1	0	1732	ceramic tile

...

Examples of statistical answers

1. What is the selling price (price) of an apartment in USA cities?
   mean price = 1161.463
2. What is the square feet (sqft) of living space?
   median sqft = 1556.5 (50% of apartments have a sqft less than 1556.5)
3. Is there any relation between price and sqft?

The price increases when the sqft increases

Data

The information we gather with experiments and surveys is collectively called **data**

Examples of Collecting Data

Example: Experiment on low carbohydrate diet

- Data could be measurements on subjects before and after the experiment

Example: Survey on effectiveness of a TV ad

- Data could be percentage of people who went to Starbucks before and after the ad aired

Define Statistics

Statistics is the art and science of:

- Designing studies
- Analyzing the data produced by these studies
- Translating data into knowledge and understanding of the world around us

Statistical Methods

The three **main components** of statistics for answering a statistical question are:

1. **Design**: Planning how to obtain data
   - How to conduct the experiment, or
   - How to select people for the survey to ensure trustworthy results

   Examples:
   - Planning the methods for data collection to study the effects of Vitamin C.
   - For a marketing study, how do you select people for your survey so you will get data that provide accurate predictions about future sales?

2. **Description**: Summarizing the data
   - Summarize the raw data and present it in a useful format (e.g., average, charts or graphs)

   Examples:
   - It is more informative to use a few numbers or a graph to summarize the data, such as an average amount of TV watched, or
   - a graph displaying how number of hours of TV watched per day relates to number of hours per week exercising.
3. **Inference**: Making decisions and predictions
   - Are the conclusions that we draw from the data general or limited to the particular situation (sample)?

Why Study Statistics?

- **To evaluate numerical facts**: ...the annual report of a company printed that the sales next year are expected to be 11.50 million
- **To perform statistical data analysis or to interpret the results of sampling**: ...it has been asked to project the sales of a company for next year.
- **To make inference about the population through the sample**: ...a survey on the drinking habits of Italians estimated the percentage of adults across the country who drink beer, wine, or hard liquor, at least occasionally. Of the 1516 adults interviewed, 985 said that they drank at least occasionally. What can we say about the proportion of Italians that drink at least occasionally? Due to various constraints, for example, time or budget, one can only sample from the population instead of take a census of the population. We need a sample that closely represents the population. One way is to obtain a random sample.

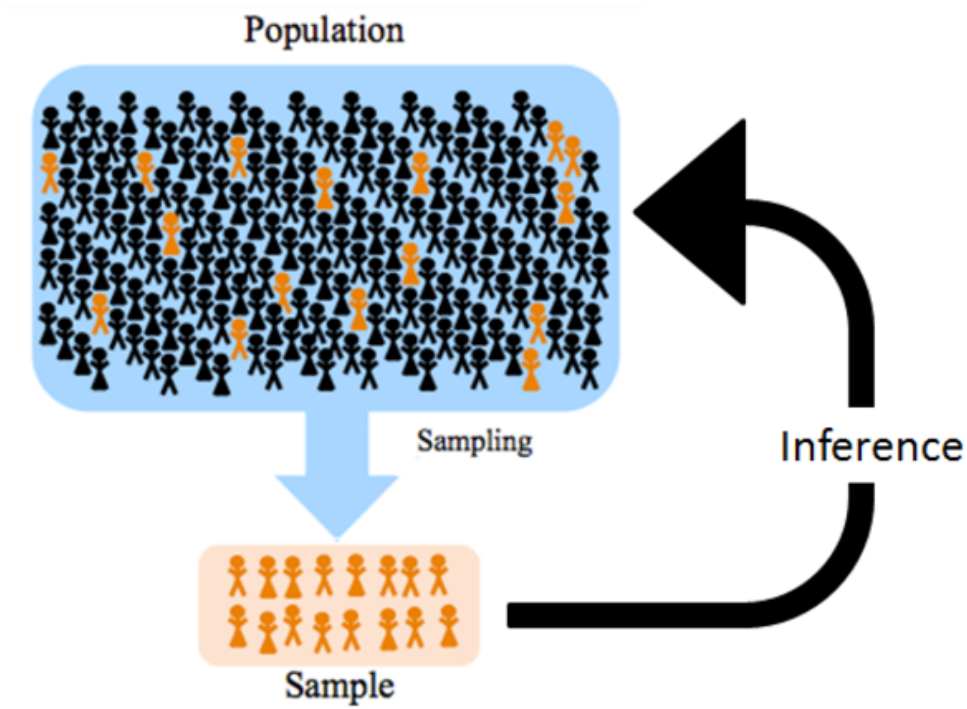
What Do Statisticians Do?

- **Gather data** $\rightarrow$ Draw a random sample of students, for example. The sample size depends on how accurate you need your inference to be and the margin of error you can tolerate.
- **Summarize data** $\rightarrow$ Summarize data from the sample (e.g. sample mean and sample standard deviation).
- **Analyze data** $\rightarrow$ Analyze the data through statistical techniques and make inference through confidence interval and hypothesis testing.
- **Draw conclusions and report the results of their analysis** $\rightarrow$ Write reports and support your conclusions including plots, tabular and numerical displays.

Sample & Population

- **Population:** the entire set of possible observations (individuals, schools, rats, countries, days, or widgets) in which we are interested
- **Sample:** a subset of the population from which information is actually collected

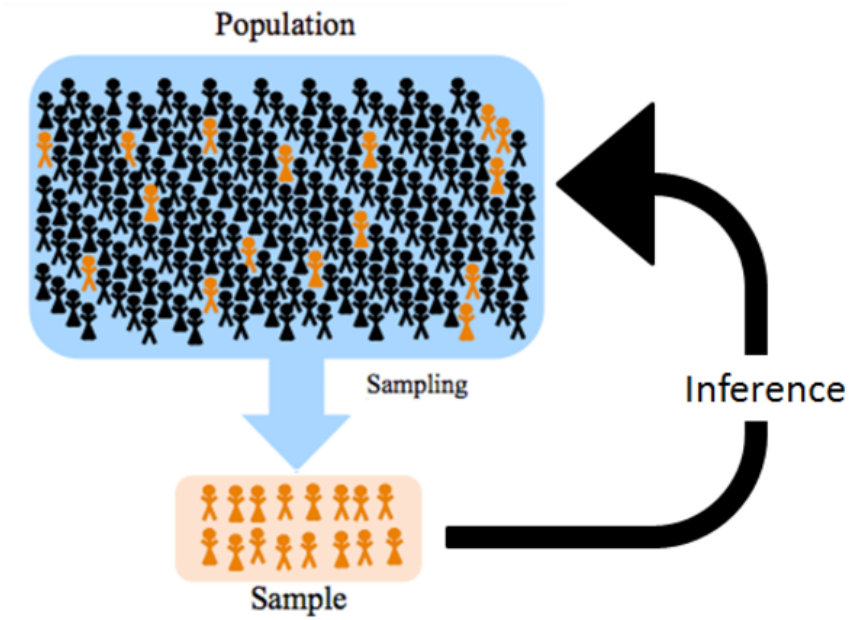
We observe samples but are interested in populations



Example: An Exit Poll

The purpose was to predict the outcome of the next gubernatorial election in California. An exit poll sampled 3889 of the 9.5 million people who voted. Define the sample and the population for this exit poll.

- The **Population** was the 9.5 million people who voted in the election.
- The **Sample** was the 3889 voters interviewed in the exit poll.



Sample Statistics & Population Parameters

- A **parameter** is a numerical summary of the population $\rightarrow$ proportion of all teenagers in the United States who have smoked in the last month
- A **statistic** is a numerical summary of a sample taken from the population $\rightarrow$ proportion of teenagers who have smoked in the last month out of a sample of 200 randomly selected teenagers in the United States

Statistics are used to make inference about population parameters

Descriptive Statistics and Inferential Statistics

- **Descriptive statistics** $\rightarrow$ Techniques of describing data in ways to capture the essence of the information. Summaries consist, for example, of graphs and numbers such as averages and percentages.

  Exit poll: In that poll, 24.0% of the sampled subjects voted for candidate A.

- **Inferential statistics** $\rightarrow$ to draw conclusions from data about the population.

  Exit poll: We are 95% confident that the percentage of all adult who voted candidate A falls between 22.7% and 25.3%.

Why only a sample?

We consider only a sample instead of the entire population because:

- **cost**;
- **time**;
- **destructive nature of the collecting process** (quality control: light bulbs, bullets,...);
- it is **impossible to collect the entire population** (infinite populations);
- **accuracy** (non-sampling errors due to mistakes made along the process of data acquisition are less frequent).

How do we choose the sample?

Let us consider the problem of to know which is the percentage of adult inhabitants of a particular region that in the last year have read at least one book. To give an answer to our problem we:

- mail a questionnaire to each family;
- take the respondents as our sample;
- compute the percentage of readers in the sample, this will be our estimate.

Problem: the selection criterion is strongly related with the variable of interest. People with a low cultural level usually do not respond to mail questionnaires and do not read books. Probably, we are overestimating the true value.

In order to avoid selection problems, the sample units are drawn at random from the population and each unit in the population is equally likely to be chosen. In this way we can use the probability calculus to control and, if possible, reduce the sampling error.

Randomness and Variability

- **Simple Random Sampling**: each subject in the population has the same chance of being included in the sample
- **Randomness** is crucial to insuring that the sample is representative of the population so that powerful inferences can be made

So we have to face with two different **sources of variability**:

- measurements may vary from subject to subject, and
- measurements may vary from sample to sample

Syllabus

- Descriptive statistics, Refresher in Probability calculus, Statistical inference.
- Simple Linear Regression and multiple linear regression - model specification, parameter estimation and statistical inference.
- Logistic Regression - model specification, parameter estimation and statistical inference.

Material

- **Textbook**. Lassy Wasserman. "All of Statistics A Concise Course in Statistical Inference" Springer
- **Further reading**. Alan Agresti, Christine Franklin, Bernhard Klingenberg "Statistics: The Art and Science of Learning From Data" Pearson; 4th International Edition, ISBN 9781292164779.

Attendance and expectations of students

It is **strongly recommended** the lesson attendance. The successful students are those who **do not wait until the last minute** to review and study the notes, solve the exercises and compare their solutions with those provided online, ask questions during the lesson, and prepare for the exam as follows: **read the slides and review the notes after each class**(have book available to use as a reference), **review the tutorials**, then **solve further exercises covering the topics learned during the class**.

Exams

- The **final exam** is a 90 minutes closed book written test consisting of theoretical questions and exercises.

Students are not allowed to consult any materials or other sources of information (including notebooks, laptops, dictionaries, mobile phones) during the exam.