Statistical Data Analysis

Prof. Dr. Nizamettin AYDIN

naydin@yildiz.edu.tr

http://www3.yildiz.edu.tr/~naydin

Statistical Data Analysis

Introduction

The role of statistical analysis in science

- This course discusses some statistical methods,
  - which involve applying statistical methods to various problems such as biological, economics, social, health, etc.
- We use empirical evidence to study populations and make informed decisions
- To study a population, we measure a set of characteristics,
  - which are referred to as variables
- The objective of many scientific studies is to learn about the variation of a specific characteristic in the population of interest

The role of statistical analysis in science

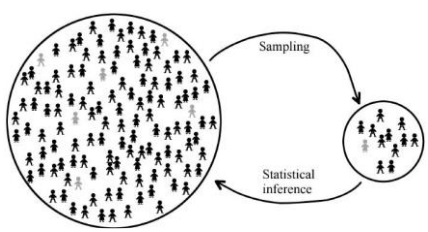
- In many studies, we are interested in possible relationships among different variables.
- We refer to the variables that are the main focus of our study as
  - the response (or target) variables.
- In contrast, we call variables that explain or predict the variation in the response variable as
  - explanatory variables
  - predictors

  depending on the role of these variables.
- Statistical analysis begins with a scientific problem usually presented in the form of
  - a hypothesis testing
  - a prediction problem.

Description of samples and populations

- Statistics is about making statements about a population from data observed from a representative sample of the population.
- A population
  - a collection of subjects whose properties are to be analyzed.
  - contains all subjects of interest.
- A sample
  - a part of the population of interest
  - a subset selected by some means from the population.

Description of samples and populations



- we sample subjects from a large population and use the information obtained from the sample to infer characteristics about the general population.

Description of samples and populations

- A parameter
  - a numerical value that describes a characteristic of a population
- A statistic
  - a numerical measurement that describes a characteristic of a sample
- We use a statistic to infer something about a parameter.

7

Description of samples and populations

- {For example, we are interested in the average height of a population of individuals.
  - The average height of the population, *m*, is a parameter,
    - but it would be too expensive and/or time-consuming to measure the height of all individuals in the population.
  - Instead we draw a random sample of, say, 12 individuals and measure the height of each of them.
    - The average of those 12 individuals in the sample is our statistic,
      - if the sample is representative of the population and the sample is sufficiently large, we have confidence in using the statistic as an estimate or guess of the true population parameter *m*. }

8

Description of samples and populations

- The distinction between population and sample depends on the context and the type of inference that you wish to perform.
  - If we were to deduce the average height of the total population, then the 12 individuals are indeed a sample.
  - If for some reason we were only interested in the height of these 12 individuals, and had no intention to make further inferences beyond the 12,
    - then the 12 individuals themselves would constitute the population.

9

Sampling

- The samples are selected randomly
  - i.e., with some probability from the population.
- Unless stated otherwise, these randomly selected members of populations are assumed to be independent.
- The selected members are called sampling units.
- The individual entities from which we collect information are called observation units, or simply observations.
- Our sample must be representative of the population, and their environments should be comparable to that of the whole population.

10

Sampling

- Some of the most widely used sampling designs
  - Simple Random Sampling
    - the chance of being selected is the same for any group of *n* members in the population
  - Stratified Sampling
    - The population is first partitioned into subpopulation and sampling is performed separately within each subpopulation
    - a.k.a. strata
  - Cluster Sampling
    - Group observations units into clusters and then sample from these clusters

11

Designing Studies

- Once a research question is defined, the next step is designing a study in order to answer that question.
- This amounts to figuring out what process you will use to get the data you need.
- After obtaining the sample, the next step is gathering the relevant information from the selected members.
- There are two major types of studies
  - observational studies
  - experiments

12

Observational studies and experiments

- In observational studies, researchers are passive examiners,
  - trying to have the least impact on the data collection process.
- Observational studies are quite helpful in detecting relationships among characteristics.
- When studying the relationships between characteristics, it is important to distinguish between association and causality.
  - The realationship is casual if one characteristic influences the other one.
- It is usually easier to establish causality by using experiments.
  - In experiments, researchers attempt to control the process as much as possible.
  - An experiment imposes one or more treatments on the participants in such a way that clear comparisons can be made.

13

Data exploration

- After collecting data, the next step towards statistical inference and decision making is to perform data exploration,
  - which involves visualizing and summarizing the data.
    - The objective of data visualization is to obtain a high level understanding of the sample and their observed (measured) characteristics.
- To make the data more manageable, we need to further reduce the amount of information in some meaningful ways so that we can focus on the key aspects of the data.
  - Summary statistics are used for this purpose.

14

Data exploration

- Using data exploration techniques, we can learn about the distribution of a variable.
  - The distribution of a variable tells us
    - the possible values it can take,
    - the chance of observing those values,
    - how often we expect to see them in a random sample from the population.
- Through data exploration, we might detect previously unknown patterns and relationships that are worth further investigation.
  - We can also identify possible data issues, such as unexpected or unusual measurements, known as outliers.

15

Statistical inference

- We collect data on a sample from the population in order to learn about the whole population.
  - {For example, Mackowiak, et al. (1992) measure the normal body temperature for 148 people to learn about the normal body temperature for the entire population.
    - In this case, we say we are estimating the unknown population average.
      - However, the characteristics and relationships in the whole population remain unknown.
    - Therefore, there is always some uncertainty associated with our estimations.}

16

Statistical inference

- The mathematical tool to address uncertainty in Statistics
  - probability.
- The process of using the data to draw conclusions about the whole population, while acknowledging the extent of our uncertainty about our findings, is called statistical inference.
- The knowledge we acquire from data through statistical inference allows us to make decisions with respect to the scientific problem that motivated our study and our data analysis.

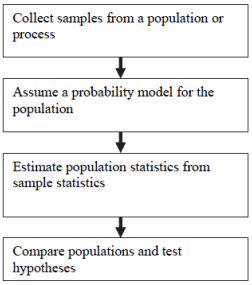
17

Computation

- We usually use computer programs to perform most of our statistical analysis and inference.
- The computer programs commonly used for this purpose
  - SAS,
  - STATA,
  - SPSS,
  - MINITAB,
  - MATLAB,
  - R,
  - Pyton,
  - …
- R is free and the most common software among statisticians
- You are encouraged to learn R for additional flexibility in your data analysis.

Summary

- The steps for performing statistical analysis of data.



19

Why statistics?

- Reasons for using statistical data summary and analysis:
  - The real world is full of random events that cannot be described by exact mathematical expressions
  - Variability is a natural and normal characteristic of the natural world
  - We like to make decisions with some confidence.
    - This means that we need to find trends within the variability

20

Questions to address

- There are several basic questions we hope to address when using numerical and graphical summary of data:
  - Can we differentiate between groups or populations?
    - probably the most frequent aim of biomedical research
  - Are there correlations between variables or populations?
  - Are processes under control?
    - Such a question may arise if there are tight controls on the manufacturing specifications for a medical device

21