Introduction to Statistics

Mahbub Latif, PhD

October 2024

Introduction

- One needs to collect data to learn something and *statistics* is the art of learning from data
- Statistics is concerned with collection of data, its description and analysis, which leads to the drawing of conclusions

Introduction

- Statistics is the science to make inferences about specific random phenomena on the basis of relatively limited sample material
- We use statistics to make decisions in different areas of our life
  - E.g. health, sports, business, etc.

Data collection and descriptive statistics

- Data could be either available or nor available at the beginning of a statistical analysis
- Governmental organization regularly collect data
  - Statistics can be used to describe, summarize, and analyse these data
- If data are not available, statistical theory can be used to design an appropriate experiment to generate data

A hypothetical example

- Suppose an instructor is interested in determining which of *two teaching methods* for a computer programming for beginners is most effective
  - The instructor might divide the students into two groups, and use a different teaching method for each group
  - At the end of the class the students can be tested and the scores of the members of the different groups compared
- To draw a valid conclusion from the data, students should be divided into two groups "randomly"

Inferential statistics and probability models

- *Descriptive statistics* is a part of statistics that is concerned with the description and summarization of the data
- The part of the statistics that is concerned with drawing conclusion is known as *inferential statistics*
- Inferential statistics takes into account the possibility of chance

Inferential statistics and probability models

- E.g. in the example of comparing two teaching method suppose it is found that average score of the first group is higher than that of the second group
  - Can we conclude that this increase is due to the teaching method used?
  - Or is the increase just a chance occurrence?

Inferential statistics and probability models

- Suppose we get 7 heads in 10 flips
  - Does it indicate that it is more likely to get heads in the future flips? (it can happen by chance)
  - What about getting 47 heads in 50 flips?

Inferential statistics and probability models

- To be able to draw logical conclusions from data, we usually make some assumptions about the chances (probabilities) of obtaining different data values
    - E.g. what is the probability that the difference between two mean scores is greater than 5
- The totality of all such assumptions are known as probability models for the data

Inferential statistics and probability models

- Sometimes the nature of the data suggests the form of the probability model that is assumed
    - E.g., suppose that an engineer wants to find out what proportion of computer chips, produced by a new method, will be defective
- In other situations, the appropriate probability model for a given data set will not be readily apparent

Inferential statistics and probability models

- The basis of statistical inference is the formulation of probability models to describe the data
  - It requires some knowledge of the theory of probability
- Statistical inference starts with the assumption that the important aspects of the phenomenon under study can be described by probabilities
- Conclusions are drawn by using data to make inferences about these probabilities

Populations and samples

- In statistics, population is referred as the total collection of all elements given the *objective* of the study
- Often the population is too large to examine all of its members
  - all the residents of a given state
  - all the television sets produced in the last year by a particular manufacturer
  - all the households in a given community

Populations and samples

- A subgroup of members (known as the sample) is selected to learn about the population
- The sample should be representative of the population
- We are interested in learning about the age distribution of people residing in a given city
    - "We obtain the ages of the first 100 people to enter the town library?" Is it representative of the total city population?
- A representative sample must be randomly selected

Changing definitions of statistics

- Statistics has then for its object that of presenting a faithful representation of a state at a determined epoch. (Quetelet, 1849)

- Statistics may be regarded as the study of *populations*, *variation*, and *methods of the reduction of data* (Fisher, 1925)

Changing definitions of statistics

- Statistics is the name for that science and art which deals with uncertain inferences - which uses numbers to find out something about nature and experience. (Weaver, 1952)
- Statistics has become known in the 20th century as the mathematical tool for analyzing experimental and observational data. (Porter, 1986)
- Statistics is the art of learning from data. (this book, 2009)

Problems

- An election will be held next week and, by polling a sample of the voting population, we are trying to predict whether the Republican or Democratic candidate will prevail. Which of the following methods of selection is likely to yield a representative sample?
  - Poll all people of voting age attending a college basketball game.
  - Obtain a copy of the voter registration list, randomly choose 100 names, and question them.
  - Use the results of a television call-in poll, in which the station asked its listeners to call in and name their choice

Problems

- A university plans on conducting a survey of its recent graduates to determine information on their yearly salaries. It randomly selected 200 recent graduates and sent them questionnaires dealing with their present jobs. Of these 200, however, only 86 were returned. Suppose that the average of the yearly salaries reported was $75,000.
    - Would the university be correct in thinking that $75,000 was a good approximation to the average salary level of all of its graduates? Explain the reasoning behind your answer.
    - If your answer was no, can you think of any set of conditions relating to the group that returned questionnaires for which it would be a good approximation?