STATS8: Introduction to Biostatistics

Overview

Babak Shahbaba
Department of Statistics, UCI

The role of statistical analysis in science

- This course discusses some biostatistical methods, which involve applying statistical methods to biological problems.

- We use **empirical evidence** to study **populations** and make informed **decisions**.

- To study a population, we measure a set of characteristics, which we refer to as **variables**.

- The objective of many scientific studies is to learn about the **variation** of a specific characteristic (e.g., BMI, disease status) 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 or **predictors** depending on the role of these variables.
- Statistical analysis begins with a scientific problem usually presented in the form of a **hypothesis testing** or a **prediction** problem.

Sampling

- To answer our scientific questions, we would, ideally, observe or perform an experiment on all members of the *population* of interest.
- However, this is usually impossible either physically, ethically, or economically.
- Instead, we select a **sample** of representative members from the population.
- Then with the methods of **statistical inference**, the conclusions based on the sample can cautiously be attributed to the whole population.

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 (e.g., people, households, cells) 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.

Sampling

- Some sampling schemes:
    - Simple random sampling
    - Stratified sampling
    - Cluster sampling

Observational studies and experiments

- After obtaining the sample, the next step is gathering the relevant information from the selected members.
- 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**.
- It is usually easier to establish causality by using **experiments**.
- In **experiments**, researchers attempt to control the process as much as possible.

Observational studies and experiments

- Retrospective and prospective observational studies
- Case-control studies
- Randomization, replication, and blocking in experiments
- Cross-Sectional, Longitudinal, and Time Series data

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.

Data exploration

- Using data exploration techniques, we can learn about the **distribution** of a variable.
- Informally, the distribution of a variable tells us the possible values it can take, the chance of observing those values, and 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**.

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.

Statistical inference

- In Statistics, the mathematical tool to address uncertainty is **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.

Computation

- We usually use computer programs to perform most of our statistical analysis and inference.

- The computer programs commonly used for this purpose are SAS, STATA, SPSS, MINITAB, MATLAB, and R.

- R is free and arguably the most common software among statisticians.

- For the purpose of this course, we use R-Commander, which allows us to do basic statistical analysis without necessarily learning the programing language of R.

- You are however encouraged to learn R for additional flexibility in your data analysis.