Introduction to Advanced Statistical Methods

This lecture introduces “statistics” as a science that deals with describing data and making predictions that have a much wider scope than merely summarizing the collected data.

Data

- The observations gathered on the characteristics of interest are collectively called ***data***.
- Existing archived collections of data are called ***databases***.
  - Many databases are now available on the Internet.
  - See www.kaggle.com

What is Statistical Science?

- ***Statistics*** consists of a body of methods for obtaining and analyzing data.

Statistical Science

Statistical science provides methods for

1. ***Design***: Planning how to gather data for a research study to investigate questions of interest to us.
2. ***Description***: Summarizing the data obtained in the study.
3. ***Inference***: Making predictions based on the data, to help us deal with uncertainty in an objective manner.

Statistical Science

- Graphs, tables, and numerical summaries such as averages and percentages are called ***descriptive statistics***. We use descriptive statistics to reduce the data to a simpler and more understandable form without distorting or losing much information.
- Predictions made using data are called ***statistical inferences***.
- ***Description*** and ***inference*** are the two types of ways of analyzing the data. Social scientists use descriptive and inferential statistics to answer questions about social phenomena.

1.2 Descriptive Statistics and Inferential Statistics

A statistical analysis is classified as ***descriptive*** or ***inferential***, according to whether its main purpose is to describe the data or to make predictions.

Populations and Samples

- The entities on which a study makes observations are called the sample ***subjects*** for the study.
- The ***population*** is the total set of subjects of interest in a study. A ***sample*** is the subset of the population on which the study collects data.
- ***Descriptive statistics*** summarize the information in a collection of data.
- ***Inferential statistics*** provide predictions about a population, based on data from a sample of that population.

Parameters and Statistics

- A **descriptive statistic** is a numerical summary of the sample data.
- A ***parameter*** is a numerical summary of the population.

Defining Populations: Actual and Conceptual

Usually the population to which inferences apply is an actual set of subjects, such as all adult residents of the United States. Sometimes, though, the generalizations refer to a *conceptual* population—one that does not actually exist but is hypothetical.

Example: parameters / statistics

In practice, our main interest is in the values of the *parameters*, not merely the values of the *statistics* for the particular sample selected.

For example, in viewing results of a poll before an election, we're more interested in the *population* percentages favoring the various candidates than in the *sample* percentages for the people interviewed.

Defining Populations: Actual and Conceptual

Suppose a medical research team investigates a newly proposed drug for treating lung cancer by conducting a study at several medical centers. Such a medical study is called a *clinical trial*. The conditions compared in a clinical trial or other experiment are called *treatments*.

Defining Populations: Actual and Conceptual

Basic descriptive statistics compare lung cancer patients who are given the new treatment to other lung cancer patients who instead receive a standard treatment, using the percentages who respond positively to the two treatments. In applying inferential statistical methods, the researchers would ideally like inferences to refer to the conceptual population of *all* people suffering from lung cancer now or at some time in the future.

1.3 The Role of Computers and Software in Statistics

Statistical software packages include R, SPSS, SAS, STATA, Excel & ChatGPT

Software relieves us of computational drudgery and helps us focus on the proper application and interpretation of the statistical methods.

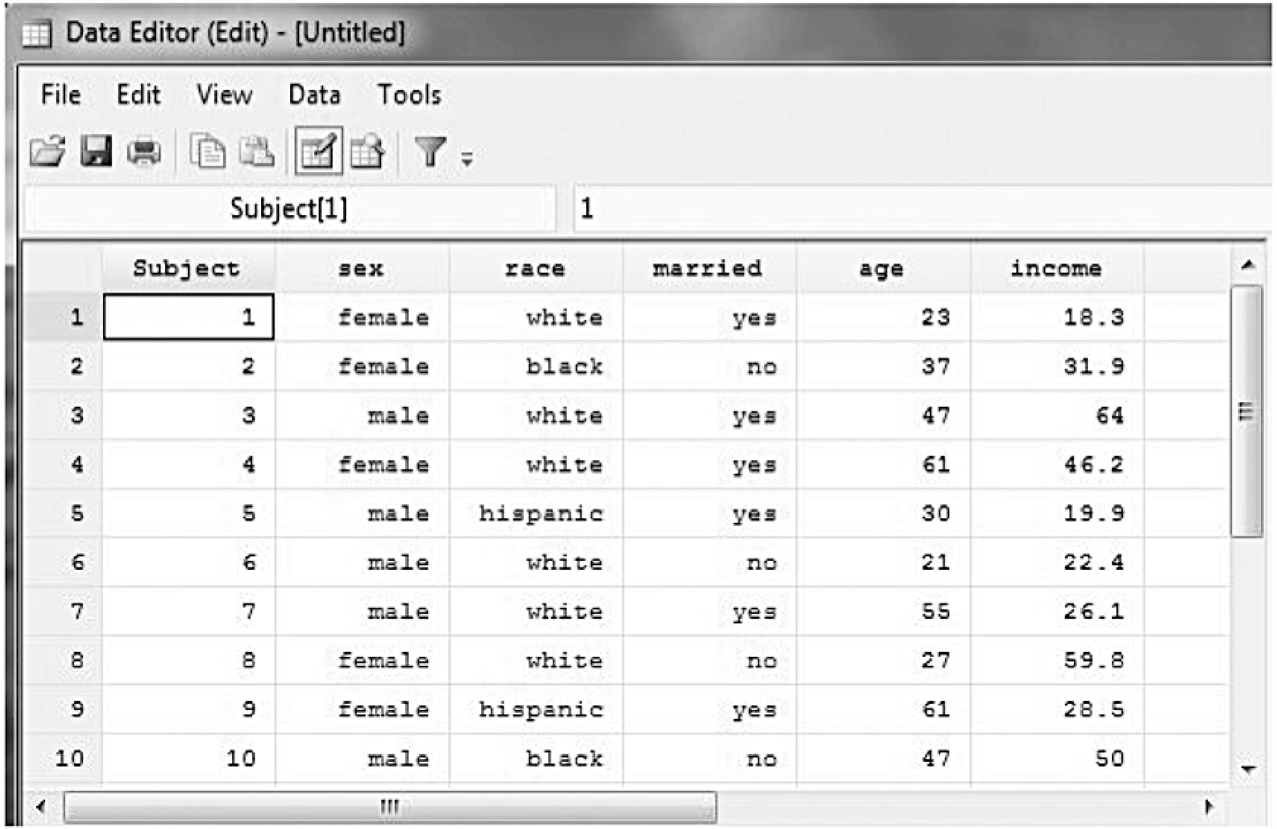
Data Files

Statistical software analyzes data organized in the spreadsheet form of a ***data file***:

- Any one row contains the observations for a particular subject (e.g., person) in the sample.
- Any one column contains the observations for a particular characteristic.

Data Files

The figure shows an example of a data file, in the form of a window for editing data using Stata software. Some of the data are numerical, and some consist of labels.



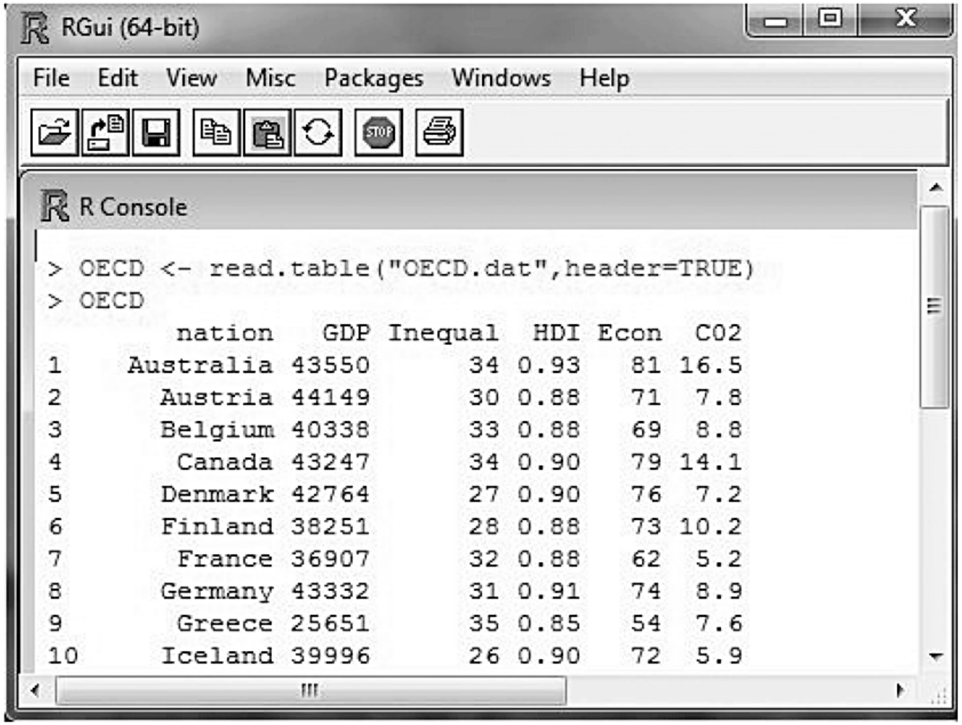
	Subject	sex	race	married	age	income
1	1	female	white	yes	23	18.3
2	2	female	black	no	37	31.9
3	3	male	white	yes	47	64
4	4	female	white	yes	61	46.2
5	5	male	hispanic	yes	30	19.9
6	6	male	white	no	21	22.4
7	7	male	white	yes	55	26.1
8	8	female	white	no	27	59.8
9	9	female	hispanic	yes	61	28.5
10	10	male	black	no	47	50

Uses and Misuses of Statistical Software

A note of caution: The easy access to statistical methods using software has dangers as well as benefits. It is simple to apply inappropriate methods. A computer performs the analysis requested whether or not the assumptions required for its proper use are satisfied.

Uses and Misuses of Statistical Software

Example of Part of an R Session for Loading and Displaying Data



```
> OECD <- read.table("OECD.dat",header=TRUE)
> OECD
      nation   GDP Inequal  HDI Econ  C02
1  Australia 43550      34 0.93   81 16.5
2    Austria 44149      30 0.88   71  7.8
3    Belgium 40338      33 0.88   69  8.8
4     Canada 43247      34 0.90   79 14.1
5    Denmark 42764      27 0.90   76  7.2
6    Finland 38251      28 0.88   73 10.2
7     France 36907      32 0.88   62  5.2
8    Germany 43332      31 0.91   74  8.9
9     Greece 25651      35 0.85   54  7.6
10   Iceland 39996      26 0.90   72  5.9
```

It is vital to understand the method before using it. Just knowing how to use statistical software does not guarantee a proper analysis.

Summary

The field of statistical science includes methods for

- designing research studies,
- describing the data (***descriptive statistics***), and
- making predictions using the data (***inferential statistics***).

Summary

Statistical methods apply to observations in a ***sample*** taken from a ***population***. ***Statistics*** summarize sample data, while ***parameters*** summarize entire populations.

- ***Descriptive statistics*** summarize sample or population data with numbers, tables, and graphs.
- ***Inferential statistics*** use sample data to make predictions about population parameters.

Summary

A ***data file*** has a separate row of data for each subject and a separate column for each characteristic. Software applies statistical methods to data files.