Preface

Table of contents

> "*Statistics is the grammar of science.*"
>
> — Karl Pearson

In this course, principles and techniques will be considered of collecting, displaying, and analysing data. The Study Guide covers the basic types of sets of numbers, namely single batches, two or more batches, related batches and data in tables. It contains Exploratory Data Analysis (EDA) techniques for displaying data, fitting a model, and considering transformations of variables so that the data better fit the assumptions of the models. The Study Guide also contains some methods for testing hypotheses, namely t, F and χ^2 tests; and some thoughts on the collection data.

The essence of Exploratory Data Analysis, or EDA, is to search for clues by whatever graphical or numerical means seem appropriate and even though these methods are often easy to understand, it is useful to have a computer at hand to perform the computations and draw graphs. Methods for EDA have developed very quickly over the past few decades. John Tukey [1] likened EDA to detective work where data are studied carefully to provide clues. Some of the techniques employed are quite old and date back to the nineteenth century and beyond when researchers collected information in an organised way and displayed it in tables and graphs. Other techniques are of more recent origin and have been developed to more easily discern peculiarities and trends in data and to provide robust methods of fitting models to data, and these methods not being overly sensitive to restrictive assumptions.

Inferential statistics[1] has also seen huge progress in the last 50 or so years since we've had fast computers. Inference uses probability models to test hypotheses and estimate parameters while quantifying uncertainty. These procedures lie at the heart of all quantitative science, and can be likened to the way a detective evaluates evidence for competing theories. Regression

[1] Inferential statistics is sometimes called "Confirmatory Analysis", but I'm not a fan of this term. Real scientists acknowledge that there is always uncertainty, and we can never really "confirm" anything!

and Analysis of Variance models are covered in the later part of the study guide, where we will use both EDA and inferential statistics. Topics such as time series analysis, nonparametric methods, and experimental design are presented only briefly. It is intended that these sections will give you an idea of how advanced modelling or analysis can be done.

Note that you are not expected to study all the topics in depth. Some sections of the study guide are intended for revision of the topics you studied in a first year statistics course. Some topics such as experimental design, correspondence analysis etc will not be examined in depth.

Data analysis is best learned by practice—exploring, summarising, plotting, and testing ideas with real data. Reading a textbook or study guide on data analysis is important, but this activity is more like reading a cook book. In the words of John Tukey:

> "*Yet more—to unlock the analysis of a body of data, to find the good way or ways to approach it, may require a key, whose finding is a creative act. Not everyone can be expected to create the key to any one situation. And, to continue to paraphrase Barnum, no one can be expected to create a key to each situation he or she meets.*
> "*To learn about data analysis, it is right that each of us try many things that do not work—that we tackle more problems than we make expert analyses of. We often learn less from an expertly done analysis than from one where, by not trying something, we missed—at least we were told about it—an opportunity to learn more.*" — Preface to *Exploratory Data Analysis*, John W. Tukey [1]

1 What is Data Analysis?

- **DATA can be defined as NUMBERS + STORY** - This reminds us that numbers have no meaning in themselves but depend on the reasons why they were collected, how they were collected and other associated information. We shall often concentrate on the numbers and their patterns but the story should determine how we examine the numbers and the conclusions we reach.
- **Analysis** - It would be better to use the plural, ANALYSES, for we rarely are interested in examining the data in only one way. Usually, there is no correct or even best answer but the analysis will depend on the way we interpret the story, the assumptions we make and other such considerations. We will take the view that the best approach to a problem is to view it in more than one way.

2 Approaches to Analyses

- **Describe the Data** - Our first approach to data will be to describe them pictorially or by summary statistics. At times, this will be all that is necessary and the reader will be left to draw conclusions.
- **Understand the data** We could use the analogy of getting to know a person. One way would be to see how they react in different situations. Another way would be to make certain assumptions (e.g. he is generous) and see if these are true (e.g. ask for a loan). The first approach is in keeping with the comment made above that the data should be examined in more than one way. In the second approach, we shall often assume that the data follows a normal distribution.
- **Generalise from the data** One way to generalise from the data is to be able to view them as a sample from a larger population so that estimates and confidence intervals can be formed. Another way is to predict the value of an estimate under different circumstances. In order to generalise, it is often necessary to make assumptions about the data.

3 Models and Data

The following explanation of Prof. George Box succinctly summarises the link between data and models; see https://williamghunter.net/george-box-articles/statistics_as_a_catalyst_to_learning

> The iterative inductive-deductive process between model and data is not esoteric but is part of our every day experience. For example, suppose I park my car every morning in my own particular parking place. On a particular day after leave my place of work, I might go through a series of inductive-deductive problem solving cycles like this: Model: Today is like every day. Deduction: My car will be in my parking place. Data: It isn't! Induction: Someone must have taken it. Model: My car has been stolen. Deduction: My car will not be in the parking lot. Data: No. It is over there! Induction: Someone took it and brought it back. Model: A thief took it and brought it back. Deduction: My car will be broken into. Data: No. It's unharmed and it's locked! Induction: Someone who had a key took it. Model: My wife used my car. Deduction: She has probably left a note. Data: Yes. Here it is!

4 Some dichotomies of this course

- **Theory versus Applications** Data and applications are the cornerstones of this course but occasionally we must introduce some theory (some of which you may have come across already) as this will dictate or add understanding to the analyses.
- **Teaching versus Learning** As the emphasis is on doing analysis rather than mastering the mathematical theory behind the statistical tools, you should spend time playing with the statistical software and trying out different ways to analyse, summarise and display data. The teaching sessions or online forums will usually revolve around software output and hopefully they will take the form of discussions.
- **Acquisition of knowledge versus Assessment** This is difficult. Ideally, we would play with data and have fun but assessment must enter the picture. Hopefully the assessment will not dominate the course. Much of the paper has remained unchanged in content for a few years so you should be clear on the general directions for studying.
- **Past Practice versus Future Needs** Many of the approaches we use have been in circulation for many years although the wide availability of computers have made the analyses more accessible. What statistics will you be using in five or ten years time when you are out in the workplace? Obviously, we do not know, but the "power of computer" will be one of the significant factors of the practice of statistics.

Bibliography

[1] John W Tukey. *Exploratory data analysis*. Vol. 2. Reading, Mass., 1977.