

Introduction to Statistical Modelling

Semester 2, 2022

Dr Gentry White

Tarang Janawalkar

This work is licensed under a Creative Commons
“Attribution-NonCommercial-ShareAlike 4.0 International” license.



Contents

Contents	1
1 Introduction	2
1.1 Data	2
1.1.1 Collecting information	2
1.1.2 Randomness	2
1.1.3 Probability	2
1.1.4 Sample vs. Population	2
1.1.5 Univariate, Bivariate, and Multivariate	3
1.1.6 Experimental vs. Observational Data	3
1.1.7 Quantitative Data	3
1.1.8 Qualitative Data	3
1.2 Summarising and Describing Data	3
1.2.1 Tables	3
1.3 Bar Charts	4
1.4 Line Charts	4
1.5 Histograms	4
1.6 Plots, Graphs, and Charts	4
1.7 Interpreting Graphical Descriptions	4

1 Introduction

Statistics is a field of mathematics that deals with data. It includes the study of summarising data, constructing probabilistic models, estimating parameters, and making statistical inferences. Statistical modelling includes asking questions, obtaining data and determining a mathematical model.

1.1 Data

Data is a collection of numbers that describes some characteristic that can be ranked, counted, or measured.

1.1.1 Collecting information

Statistical modelling relies upon reliably sourced data. When collecting data, we must consider

- what questions are we trying to answer,
- what information is needed to answer these questions,
- what is the best source for that information

1.1.2 Randomness

We must be aware that everything is different and that randomness introduces uncertainty in data. Random events are events whose exact outcome cannot be predicted. We can assume that all variation in the world is observed due to randomness.

1.1.3 Probability

Probability is a mathematical construct for dealing with randomness and uncertainty.

Definition 1.1 (Experimental unit). An **experimental unit** is an individual that generates information for the data collection process. Careful consideration of what constitutes an experimental unit must be made to ensure that it aligns with the questions of interest.

1.1.4 Sample vs. Population

Definition 1.2 (Population). We might have questions about a very large collection of things called a **population**.

As it is not feasible to collect data from an entire population, we must use a sample of the population.

Definition 1.3 (Sample). A **sample** is a subset of a population that is representative of the population, in some cases a random sample is sufficient.

Definition 1.4 (Random sample). A **random sample** is one where the sample members are selected from the population by chance.

1.1.5 Univariate, Bivariate, and Multivariate

Data can be described in terms of dimension, that is, how many measurements were collected from each experimental unit. By collecting multiple measurements from each experimental unit, we can ask questions about the relationship between the measurements.

- When a single measurement is collected, the resulting dataset is **univariate**.
- If two measurements are collected, the dataset is **bivariate**.
- If more than two measurements are collected, the dataset is **multivariate**.

1.1.6 Experimental vs. Observational Data

Data sets that have been collected without any specific analyses or modelling in mind are called **observational data**. By contrast, when a collection procedure is specifically designed to obtain data with a specific intent, i.e., a laboratory test, the data is called **experimental data**.

Observational data may contain biases that limit its usefulness and bias any modelling or analysis results.

1.1.7 Quantitative Data

Quantitative data is data that is expressed numerically. This data can be classified as *discrete*, *continuous*, or *ordinal*.

- Count data is classified as discrete, i.e., integer values or finite sets of real values.
- Continuous data is a measurement on a continuum or a measure that can be subdivided infinitely, i.e., time and lengths.
- Ordinal data is data where the order or ranking of values (discrete or continuous) is important.

When data is not ordinal, it is called **nominal** data.

1.1.8 Qualitative Data

Qualitative (categorical) data is data where the variable of interest is membership to a group or category.

1.2 Summarising and Describing Data

1.2.1 Tables

Tables are the most immediate way of summarising a data set; typically, we might organise data in a table with one row for each subject and a column for each measurement.

1.3 Bar Charts

Graphical depictions of the data can also be useful but are limited in the number of variables displayed in one picture. The bar chart is one of the simplest and easiest to understand. It is most useful for categorical data; categories are listed on the x-axis of the plot, and bars for each category are drawn with their heights corresponding to the counts for that category. The simple bar chart can be very useful if the categories are ordered from left to right in descending order counts. This plot is sometimes called a Pareto Plot after the Italian economist Vilfredo Pareto.

1.4 Line Charts

Line charts illustrate a trend of change based on two quantitative variables. Line Charts are especially useful for displaying trends over time (or other ordinal variables). It is important to be clear on what is being plotted. Often trends over time need to be aggregated, e.g., plotting the average or median per year. Otherwise, the plots can look too “busy” and be difficult to read. While the resulting chart can explain overall trends, they can obscure how much variability or “noise” is in the data and may be misleading if the overall trend is obscured by variability.

1.5 Histograms

Histograms give a visual description of our data by “binning” or grouping data into data ranges, then plotting bars with heights equal to the count of the bins’ contents or the relative proportion of the bins’ contents. Histograms give us a picture of the shape of the data and help identify patterns in the distribution of values.

1.6 Plots, Graphs, and Charts

- A **chart** is a visual display of data, i.e., a table, a graph, or a diagram
- A **graph** is a diagram showing the relationship between variables, each measured along orthogonal axes.
- A **plot** is used as a synonym for graph but is less precise in its definition; it also sometimes refers specifically to a graph *produced by a computer*.

1.7 Interpreting Graphical Descriptions

Graphical descriptions of data should ensure that all information about the data is expressed.

- The x and y axes should be clear in what they are measuring, including any units.
- Consider how the graph or chart was made. What choices were made and how might different options change how the graph is perceived.
- Does the graph contain any outliers that merit investigation to determine if they are accurate measurements, or if they result from either measurement or recording error.
- For Pareto charts and histograms; the y -axis should measure proportion or density rather than frequency to make comparisons easier.

1.7.1 Centrality**1.7.2 Skew****1.7.3 Trends**