Introduction to Statistical Modelling

Semester 2, 2022

Dr Gentry White

Tarang Janawalkar

Contents

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 Elements of Statistical Modelling

1.1.1 Data

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

1.1.2 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

1.1.3 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.

1.1.4 Probability

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

1.2 Experimental Units and Populations

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.

1.2.1 Sample vs. Population

Definition 1.2 (Population)**.** We might have questions about a very large collection of things called a **population**.
A dataset collected from a population is called a census.

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.

1.3 Types of Data

1.3.1 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**.

1.3.2 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.

1.3.3 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.

1.3.4 Qualitative Data

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

1.4 Summarising and Describing Data

1.4.1 Tables

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

1.5 Bar Charts

Graphical depictions of the data can also be useful but are limited in the number of variables displayed in one picture.
Bar charts are most useful for categorical data where 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.
When the categories are **ordered** from left to right in descending order counts, the plot is called a **Pareto plot**.

1.6 Line Charts

Line charts illustrate a *trend* of change based on **two** quantitative variables. Typically line charts display trends over time (or other ordinal variables).
Often trends over time need to be aggregated by plotting the average or median per year to avoid a "busy" plot which can sometimes 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.

1.7 Histograms

Histograms are a special kind of bar chart that give a visual description of 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.
The binning process is performed by the computer, however in most cases we override the automatic settings and select either the number of bins, or the width of each bin.

1.8 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*.

1.9 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.

1.9.1 Centrality

Histograms are a graphical representation of the distribution or density of observations. Centrality is the degree to which an observation is central to the distribution. Additionally, the data can be multi-modal if there are multiple "peaks" or "centres" in the distribution.
Altering the number of bins or bin width may reveal the centrality of the observations.

1.9.2 Skew

Another characteristic of histograms is the degree to which the distribution is skewed. Skew is the deviation from symmetry about the centre of the data. Skew is either "right" skew where the tail of the density or histogram is heavier to the right, or "left" skew if otherwise.
This can be observed by looking at how much the left/right tails are stretched in comparison to one another, i.e., the tail to the right of a right skewed chart stretches further on the x-axis than on the left.

1.9.3 Trends

Trends refer to changes in a line chart and are often described as a constant (first-derivative) pattern of increasing or decreasing values.

2 Numerical Summaries of Data

Although graphical summaries are useful for developing a general understanding data, they are limited to subjective interpretations. To form a precise understanding of the data, we need to use numerical summaries. Here we must make a distinction between sample and population summaries as measurements may vary between samples, whereas population summaries are generally constant.

2.1 Measures of Centrality

2.1.1 Mean

Given a set of n observations x_1, x_2, ..., x_n, the **arithmetic mean** or **average** is defined as

$$\frac{1}{n}\sum_{i=1}^{n} x_i \equiv \frac{1}{n}\left(x_1 + x_2 + \cdots + x_n\right)$$

If the data is taken from a sample, the sample mean is denoted $\overline{x}$.

2.1.2 Median

A drawback to the mean is that it can be misleading when the data is skewed. The **median** is the middle value of a set of n observations when arranged from largest to smallest.
If n is odd:

$$\text{median} = x^{(\frac{n+1}{2})}$$

or the $(n+1)/2$th value of the sorted list. If n is even, the median is the :

$$\text{median} = \frac{x^{(\frac{n}{2})} + x^{(\frac{n}{2}+1)}}{2}$$

2.1.3 Mode

Given discrete data, the mode is defined as the most common value in a set of observations.

2.1.4 Population Mean

The mean of a finite population is computed in the same way as the mean of a sample, but the population mean is denoted by μ.

2.2 Measures of Dispersion

Dispersion refers to how much variation there is in a set of observations.

2.2.1 Range

Given a set of observations that are ordered such that

$$x^{(1)} \leq x^{(2)} \leq \cdots \leq x^{(n)}$$

the range is defined as

$$x^{(n)} - x^{(1)}.$$

2.2.2 Variance

The variance is the average of the squared deviations from the mean.

- Given the observations x_1, x_2, ..., x_N, from a population of size N with mean μ, the **population variance** is defined as

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2.$$

- Given the observations x_1, x_2, ..., x_n, from a sample of size n with mean $\overline{x}$, the **sample variance** is defined as

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

The **population variance** is given by

2.2.3 Standard Deviation

The standard deviation is the square root of the variance. This is conceptually easier to understand as it has the same units are the data.

- The **population standard deviation** is defined as
$$\sigma = \sqrt{\sigma^2}.$$
- The **sample standard deviation** is defined as
$$s = \sqrt{s^2}.$$

Theorem 2.2.1 (Chebyshev's Theorem)**.** *Given a set of n observations, at least*
$$1 - \frac{1}{k^2}$$
of them are within k standard deviations of the mean, where $k \geq 1$*. Formally,*
$$\frac{\#\left\{x | \overline{x} - ks < x < \overline{x} + ks\right\}}{n} \geq 1 - \frac{1}{k^2}$$

Theorem 2.2.2 (Empirical Rule)**.** *If a histogram of the data is approximately unimodal and symmetric, then,*

- *68% of the data falls within* ***one*** *standard deviation of the mean*
- *95% of the data falls within* ***two*** *standard deviations of the mean*
- *99% of the data falls within* ***three*** *standard deviations of the mean*

Often the standard deviation cannot be computed directly, but can be approximated using the Empirical rule. Here we assume that
$$\text{range} \approx 4s$$
so that
$$s = \frac{\text{range}}{4}.$$

2.3 Skew

The **skew** describes the asymmetry of the distribution. For a finite population of size N, the **population skew** is defined as
$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \mu}{\sigma}\right)^3$$
For a sample of size n, the **sample skew** is defined as
$$\frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^3}{\left(\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}\right)^3}$$

- When the skew is **positive**, the data is **right-skewed** and the "tail" of the distribution is **longer on the right**
- When the skew is **negative**, the data is **left-skewed** and the "tail" of the distribution is **longer on the left**

2.4 Measures of Rank

It is often useful to know the rank or *relative standing* of a value in a set of observations. This is natural for ordinal data whose ordering has implicit meaning, but it can also be useful for nominal data as a means of measuring dispersion.

2.4.1 Z-Score

The Z-score is a unitless quantity and can be used to make comparisons of relative rank between members of a population.

$$Z = \frac{x - \mu}{\sigma} \quad \text{or} \quad \frac{x - \overline{x}}{s}$$

2.4.2 Quantiles

In addition to Z-scores, quantiles can be used to make comparisons of relative ranking between populations, as well as construct intervals bounding a given proportion of the observations. For a set of n observations, x_q is the q-th quantile, if $q\%$ of the observations are less than x_q.

2.5 Inter-Quartile Range

The inter-quartile range (IQR) is the difference between the 75th and 25th quantiles, or the range covered by the middle 50% of data. It is a robust measure of the dispersion of the data, as it is not affected by extreme values unlike the range or variance.

2.6 Boxplots

2.7 Five Number Summary

The five number summary is set of measurements that indicates the

- minimum value
- 25% quartile
- median
- 75% quartile
- maximum value

A boxplot is a graphical display of the five number summary. It is a plot of the values of the data mapped to the y-axis.
Using the `ggplot2` package, the function `geom_boxplot()` draws a box encompassing the IQR with a horizontal line indicating the median. Vertical lines extend 1.5 times the IQR above and below the box. The points not within the ends of the vertical lines are also plotted to indicate outliers.

2.7.1 Outliers

Outliers are extreme observations that fall outside some interval defined either by quantiles (above 95% or below 5% quantiles) or in terms of the Empirical rule (outside two standard deviations from the mean). They should be investigated to determine if they are errors or naturally occurring extreme values.

3 Bivariate Data

Data in two dimensions is often used to describe relationships between two variables.

3.1 Bivariate Categorical Data

Bivariate categorical data is a dataset with two qualitative or categorical variables that have a relationship we want to summarise. This can be done using contingency tables or side-by-side (or stacked) bar charts.

3.1.1 Contingency Tables

Contingency tables or crosstabs are tabular representations of the frequency of occurrence of pairs of values. The categories for each variable are assigned to an axis of the table so that each cell represents the frequency of occurrence of a pair of categories, one from each variable.

3.1.2 Bar Plots

Often the data is presented more effectively as a stacked bar chart or side-by-side bar chart. Here the counts for each pair of categories are plotted on the same axis, and stacked on top of one another to display relative proportion, or side-by-side if too busy.

3.2 Bivariate Quantitative Data

Bivariate quantitative data is a dataset with one qualitative variable and one quantitative variable. This can be represented as a table, or through various charts, by comparing charts side-by-side for each category.

3.3 Scatter Plots

When both variables are quantitative, the data can be represented as a scatter plot with each variable assigned to an axis and the points plotted on the axes.

3.3.1 Covariance and Correlation Coefficients

For such data, the covariance is the measure of the linear correlation between the variables. For variables x and y,

$$s_{xy} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{n-1}.$$

Note that when $x = y$, the formula simplifies to the sample variance of x. The covariance has the following characterstics:

- $s_{xy} > 0$: As x increases, y also increases.
- $s_{xy} < 0$: As x increases, y decreases.
- $s_{xy} \approx 0$: No relationship between x and y.

Although the covariance is a useful tool to measure relationships, it is only generalisable in terms of its sign. Thus, if we want to compare across data sets, we need to use the correlation coefficient.

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

The **correlation coefficient** is a measure of the strength of the relationship between the variables. It is a scale-free and unitless measure bounded between -1 and 1 and has the same characteristics as the covariance.
Note that a correlation coefficient of 0 indicates **no linear relationship** between the variables, and not necessarily indicative of **no relationship**.

3.4 Regression and Least Squares

In addition to the numerical summaries above, a regression or least squares line of best fit provides both a graphical and numerical summary of the relationship between the variables. A linear relationship between two variables x and y is defined as

$$y = a + bx.$$

The least squares best fit determines the coefficients a and b that minimise the sum of the squares of the residuals (errors) between y and the line $\hat{y} = a + bx$. Mathematically,

$$\min_{a,\,b} \sum_{i=1}^{n} \left(y_i - \hat{y}\left(x_i\right)\right)^2.$$

The coefficients can be summarised by the formula

$$b = r\frac{s_y}{s_x} = \frac{s_{xy}}{s_x^2}$$
$$a = \overline{y} - b\overline{x}.$$