STAT2203 STAT7203

Probability models and data analysis

Lecture 1.2

Describing Data

Data as a spreadsheet

Data is often stored in a table or spreadsheet. In this course, we will store data in CSV (*Comma Separated Values*) format.

A statistical convention is to denote variables as columns and the individual items (or units) as rows.

ID	group	age	height
1	1	10	100
2	1	12	120
3	2	9	90
4	2	8	87

Types of variables

- **Quantitative** variables (take on numerical values):
  - **Continuous** - represent measurements that take values in a continuous range, e.g., one's height
  - **Discrete** - have only a finite/countable number of numeric possibilities, e.g., number of eggs one eats for breakfast
- **Categorical/qualitative/factor** variables:
  - **Nominal** - represent groups without order, e.g., "Yes/No", "Red/Blue/Green"
  - **Ordinal** - represent groups with order, e.g., "Monday/.../Sunday", "Likert scale"

Categories are often referred to as **levels**.

Summarising categorical data

A categorical variable can be summarised by a **table of counts or proportions**.

Two-way tables are used to cross-tabulate between two categorical variables.

Visualising categorical data

A **barplot/barchart** is a graphical representation of the table of counts.

For each level of the variable, a rectangle is drawn whose height corresponds to the number of observations with that level.

Summarising quantitative data

We often extract information from the quantitative data by summarizing it into a few numbers or **summary statistics**. This typically involves finding the **location (or centre)** of the data, the **spread** of variability in the data (how far the values extend from the centre), and the **shape/skewness** of the variability, e.g., are values spread symmetrically on either side of the centre? If not, then is the data right skewed, or is it left skewed?

Summary statistics (mean)

The simplest, and arguably most useful, information about the data is its center.

Suppose we have data $x_1, \ldots, x_n$. A simple measure of the center of the data is the **sample mean**. This is simply the average of the data values:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \ .$$

Sample mean is very sensitive to outliers.

Summary statistics (Median)

A more robust measure of the center of the data is the **sample median**. This is simply the value M such that

- 50% of samples are smaller than or equal to M, and
- 50% are greater than or equal to M.

Sample mean is more robust to outliers.

Summary statistics (variance)

The next useful information from the data is a measure of its spread, i.e., the degree of variability among the samples.

The simplest measure of the spread of the data is the **sample variance** defined as

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

The **sample standard deviation** $s = \sqrt{s^2}$ is preferred as a measure of spread because it has the same units as the original data.

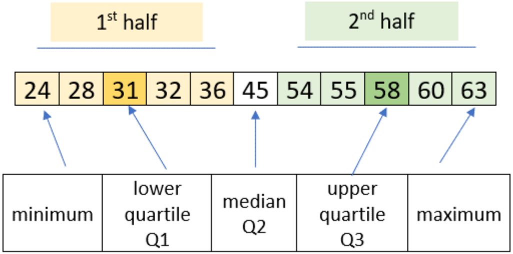
Summary statistics (Quantiles)

We can define more sophisticated summary statistics using **quantiles**.

Roughly speaking, for any $0 < p < 1$, the p-**quantile** (or **100**p-**percentile**) of the data $x_1, \ldots, x_n$ is a value y that is greater than or equal to a fraction p (or $100p\%$) of the data and is smaller than or equal to a fraction $1 - p$ (or $100(1 - p)\%$) of the data.

Note: There are nine different "formulas" for computing y that commonly appear in statistical software packages.

Five-number Summary



Fine-number Summary

- 0.5-quantile is the **median (M)**
- 0.25-quantile is the **first quartile (Q1)**
- 0.75-quantile is the **third quartile (Q3)**
- The **interquartile range** $IQR = Q3 - Q1$ is another measure of the spread of the data