MATH2801: Theory of Statistics

2025, Term 2

These lecture notes were written by Matt Wand then modified and edited by David Warton, Diana Combe, Zdravko Botev, Libo Lee, Jakub Stoklosa and others.

A brief introduction

Statistics is “learning from data” – the science of designing studies and analysing their results.

Statistics uses a lot of mathematics, and some like statistics because of its challenging mathematics! **It’s a powerful toolkit for uncovering hidden truths and making sense of a complex world.**

It also involves many other skills arising from the application of statistical thinking in practice, and some like statistics because of its usefulness in a range of interesting applications.

Statistics is pervasive – everyone has data they need to analyse, and we now live in an age of data! **This means the power to understand and predict is more accessible and crucial than ever before.**

Approximately 90% of the world's data has been created in the last few years. This highlights the exponential growth of data generation in our modern digital age!

Statistics plays a major role in psychology, social sciences, biology, chemistry, engineering, physics, economics/finance, medical sciences, ecology/zoology, computer sciences, aviation, education, astronomy, agriculture, sports, political studies, robotics, etc. **From understanding the human mind to exploring the cosmos, statistics is a key to unlocking discoveries across all these fields.**

Let's have a look at a few real-world examples.

Example 1: The NSW election

Example

Shortly before the last New South Wales (NSW) state election, a poll was held asking a selection of NSW voters people who they would vote for.

Out of 365 respondents, 54.5% said they would vote the Labor party ahead of the Liberal/National coalition.

Some important questions where statistics can help are:

- How accurate is this estimate of the proportion of people voting for Labor (ahead of Liberal/National coalition)?
- Is this sample, of just 365 NSW voters, sufficient to predict that Labor party would win the election?

Example 2: Guinea pigs

Example

Does smoking while pregnant affect the cognitive development of the foetus?

Johns *et al.* (1993) conducted a study to look at this question using guinea pigs as a model. They planned to inject nicotine tartate in a saline solution into some pregnant guinea pigs, inject no nicotine into others, and compare the cognitive development of offspring by getting them to complete a simple maze where they look for food.

Some important questions where statistics can help are:

- How many guinea pigs should they include in the experiment?
- How should the "no nicotine" treatment have been applied?
- How should the data be analysed to assess the effect of nicotine on cognitive development?

Statistics in action

Here are some questions where statistical thinking is essential:

- Does a new medical treatment work better than the existing one?
- How are climate change patterns affecting where species can live?
- What are your odds of winning the lottery? Is it a worthwhile investment?
- Is there evidence of gender bias in promotion processes?
- Does intercessory prayer impact patient recovery rates (Benson *et al.*, 2005)?
- Is Sydney experiencing a long-term decline in rainfall?
- What are the ecological consequences of a new invasive pest?
- Can we predict future sales to effectively trial new products?

Unlock the world with statistics

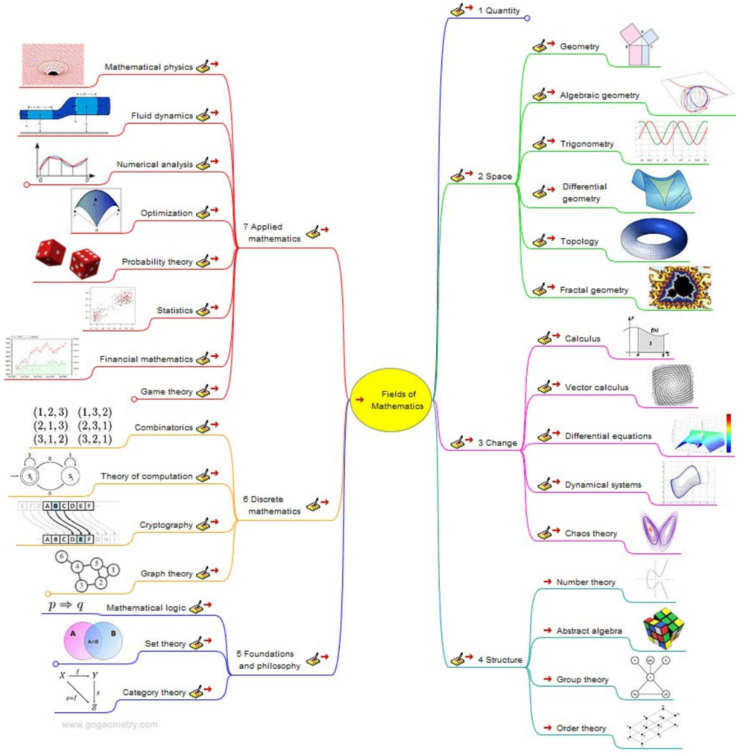
Ever wonder what's really going on behind the news headlines, advertisements, or political polls? Statistical reasoning provides a powerful lens to critically evaluate information and gain a deeper understanding of the world around you.

In fact, some argue that statistical literacy is now crucial for navigating modern life.

Beyond everyday understanding, statistics is of fundamental importance across a vast range of disciplines. Consequently, statistical skills are highly valued, and there's a significant shortage of statisticians in the job market – great news if you decide to major in statistics!

Furthermore, the recent explosion in artificial intelligence (AI) and machine learning has placed statistics at center stage. The very algorithms powering AI are built upon core statistical principles, making this field an exciting frontier for those with statistical expertise.

You've seen how statistics helps us understand the world. Now, you might be wondering why we delve into the mathematics. Think of it this way: understanding the underlying theory gives you the power to go beyond simply applying methods – you'll know why they work, how to adapt them, and when to be cautious.



We'll start by building a strong foundation with key statistical concepts and terminology. These fundamental ideas will reappear throughout this course and are central to the entire field of statistics. So, let's begin our journey!

Zooming in: Samples vs. the whole picture (census)

Definition

A **population** is the entire group of individuals or items we want to learn about.

Definition

A **sample** is a smaller, manageable subset of the population from which we collect data.

A **census** is when we attempt to gather data from the entire population.

Why do most studies focus on samples instead of a full census? Primarily for practical reasons: it's often more cost-effective and logistically feasible to study a subset.

Moreover, by concentrating our efforts on a smaller group, we can often gather more in-depth and higher-quality data compared to a broad but shallow census. It's often a trade-off between breadth and depth!

Samples in action

Example

Consider the guinea pigs study of the effects of smoking during pregnancy on offspring.

The study used a **sample** of guinea pigs – the alternative would be to enroll every guinea pig on the planet in the study! That's just not going to happen. . .

Example

The Australian Bureau of Statistics (ABS) coordinate a census of all Australians every five years. This is designed to find out demographic information such as population size, age of Australians, education, *etc*. However, for more frequent data like the unemployment rate, the ABS uses surveys of a **sample** of people.

So, if we only look at a sample, how do we draw conclusions about the *entire* population? That's a key question we'll explore!

Description vs. inference

Definition

Descriptive statistics: Tools for **summarizing** what our data shows.

Definition

Inferential statistics: Using sample data to make **generalizations** (inferences, predictions, decisions) about the wider population.

You've likely encountered descriptive statistics before – calculating averages, creating graphs, etc. These help us see patterns within our sample.

While describing our sample is crucial, often our real goal is to go further: to use the sample to **infer** something about the **entire population**. That's where the power of inferential statistics comes in!

Example

Consider the NSW poll, which included 365 registered NSW voters.

When we say 54.5% of respondents would vote Labor party, we are reporting a **descriptive statistic**.

When we use the data to answer the question "How much evidence does this study provide that the Labor party will win the next election?" we are making an **inference** about the population of all 5.5 million (!) NSW voters, based on a sample of just 365.

Now, inference is where things get a bit more brain-bending, in a good way! It involves the mathematical and conceptual challenge of using samples to understand entire populations.

We'll tackle this later in the course with some essential techniques, starting with simpler situations.

Sampling introduces variation

A fundamental idea in statistical inference is that sampling induces variation – different samples will give different data, depending on which subjects end up getting included in the sample.

Example

Consider again the NSW election example, recall that 365 NSW voters were sampled, and of these, 54.5% of them would vote for the Labor party ahead of the Liberal/National coalition.

If a different 365 NSW voters were sampled, would we expect to get exactly 54.5% voting for the Labor party again?

The challenge of inference: Accounting for uncertainty

When we use a sample to understand a larger population, we face a key challenge: sample variation. Different samples will naturally give slightly different results. We need to account for this inherent variability when making inferences.

If we choose our sample randomly (which is often the goal!), then the data we collect is also random. This is where the power of **probability theory** comes in – it provides the framework for understanding and quantifying this randomness in our data.

Because probability is fundamental to making sound statistical inferences, we will be building key probabilistic concepts throughout this course.

A chapter covering the essential basics of probability is available on Moodle. Please review this material independently.

Focusing our inquiry: The research question

Building on our understanding of sampling and the need for probability to handle sample variation, the very first step in any statistical endeavor is crystal clear: What is the central question we're trying to answer?

The research question dictates how we design our study and ultimately how we analyze the data. Understanding the primary purpose of a study is paramount – everything else flows from it.

The question shapes the data collection!

Example

Remember the NSW election example? If our research question is "Who is likely to win the next election?", we immediately know we need to start by gathering opinions from a **representative sample** of NSW voters.

In obtaining a representative sample of NSW voters we need to make sure we **don't** include people who are not eligible to vote in the election, for example:

- People under the age of 18.
- People who have not registered to vote.
- People not registered to vote in NSW.

We also need to sample in a manner that gives all NSW voters the opportunity of being in the sample, to make sure all types of voter are represented.

One way of collecting such a sample is to use "random digit dialing" – dialing random (landline) phone numbers in houses in NSW.

Do you think this sample is truly representative of the NSW voter population?

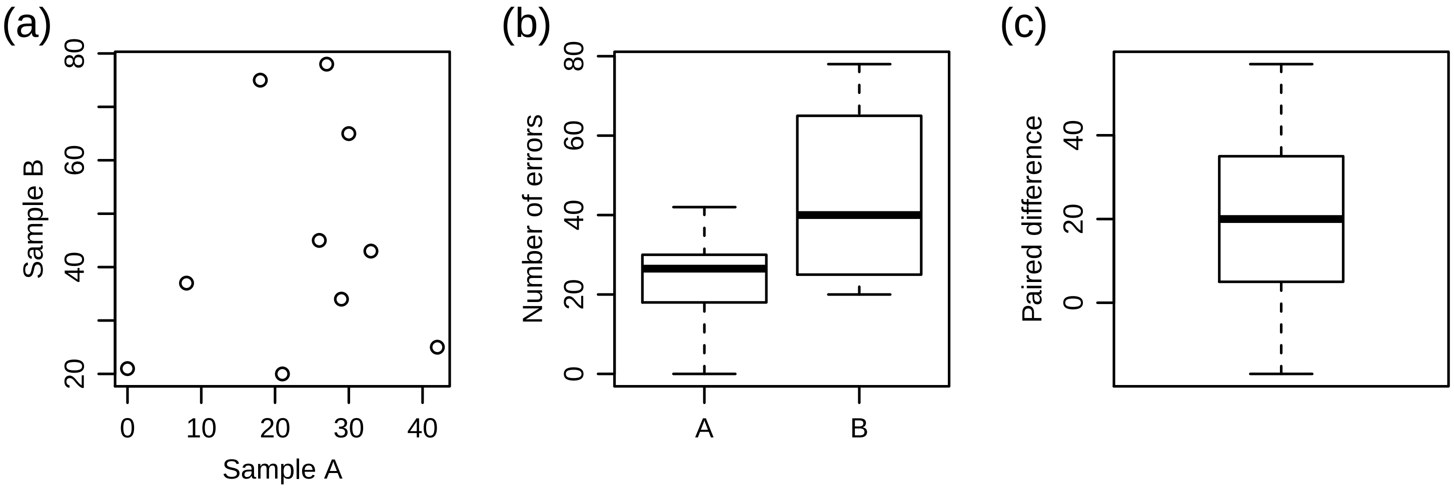
How we analyse data depends on the question!

Example

Consider the following data:

Sample A:	8	30	29	27	26	33	0	42	21	18
Sample B:	37	65	34	78	45	43	21	25	20	75

Here are 3 graphs that would all be appropriate for their own research questions:



These data are actually from the guinea pig experiment, where Sample A is the number of errors in the maze made by control guinea pigs (no nicotine treatment) and Sample B is the number of errors by treatment guinea pigs (with nicotine treatment).

Which plot is most suitable for visualising the effect of nicotine on cognitive development of guinea pigs? Why have you chosen this plot?

The main lesson here is that whenever collecting or analysing data, or answering questions on how to do it, we need to keep in mind the primary purpose of the study!

Furthermore, as we will see later in this chapter, visualizing our data through plots is an important tool used in statistics[1].

[1] For deeper insights into data visualization, see W. S. Cleveland's *Elements of graphing data*, Hobart Press.

Statistics packages/software

Graphs (and indeed most statistical procedures) are most easily implemented using a computer, and a **statistics package** developed for data analysis.

In MATH2801, we will use the freely-available **R/RStudio** software: https://www.r-project.org/

As we will see, `RStudio` will be used for most graphs in the lecture notes.

Some other common programs used for statistics are:

- Modern Languages: Python, Julia.
- Older software (less frequently used): SAS, SPSS (PASW), Minitab, S (S-PLUS).
- Spreadsheets: Excel.

Chapter 1: Descriptive statistics

Before diving into the theoretical concepts, we'll start with a quick review of summary statistics. This is an excellent way to get a feel for the data and refresh some fundamental ideas you've likely encountered before.

The challenge: Given a dataset $\{x_1, x_2, \ldots, x_n\}$ for some sample size n, how can we effectively summarise its key features, both visually and numerically?

In this section, we'll briefly revisit some essential tools for this purpose. Much of this will be revision from high school or other courses, so we'll move at a brisk pace.

In practice, we very rarely calculate numerical and graphical summaries by hand, but it's crucial to understand how they are generated, how to create them using `R`/`RStudio`, and, most importantly, how to interpret what they tell us about the data!

The data analysis roadmap: Two key questions

When faced with data, where do we even begin? Here are the crucial first two questions to tackle:

1. **What's the burning question?** (The research question).
2. **What kind of data do we have?** (Properties of the primary variables).

Think of it this way: our initial descriptive summaries should directly shed light on the research question we've identified.

Once we know our question, the next vital step is to understand the nature of our variables. Specifically, is each variable **categorical** or **quantitative**? Let's explore this distinction on the next slide.

Every piece of information we collect can generally be classified into one of two fundamental types:

Categorical variables: These sort responses into distinct groups or categories. Think of things like gender (male/female/other) or political preference (Party A/Party B/Independent).

Quantitative variables: These yield numerical responses, typically measured on a scale. Examples include height (in cm), temperature (in degrees Celsius), or counts, like the number of emails you received today.

So, if our data $\{x_1, x_2, \ldots, x_n\}$ comes from measuring a **quantitative variable**, each x_i is a real number ($x_i \in \mathbb{R}$).

If it comes from a **categorical variable**, each x_i belongs to one of a finite set of categories or "levels" ($x_i \in \{C_1, C_2, \ldots, C_K\}$).

Example

Consider our NSW election poll. The raw data might look like $\{\text{Labor}, \text{Labor}, \text{Liberal}, \ldots\}$. We can then represent these categories numerically, for example, as $\{1, 1, 0, ...\}$.

Examples

Let's practice in the following scenarios:

1. Would you rather use Windows/PC or a Mac?
2. Are the number of errors made in a maze by offspring of pregnant guinea pigs affected by a nicotine treatment (vs. no treatment)?
3. Is a Titanic passenger's gender related to their survival (yes/no)?
4. How does brain mass change in dinosaurs as their body mass increases?

What are the variables of interest in these questions? Are each of these variables categorical or quantitative?

1.

2.

3.

4.

Summary of descriptive methods

Useful descriptive methods for when we wish to summarise one variable, or the association between two variables, depend on whether these variables are categorical or quantitative.

	Does the research question involve:				
	One variable		**Two variables**		
Data type:	*Categorical*	*Quantitative*	*Both categorical*	*One of each*	*Both quantitative*
Numerics:	Table of frequencies	Mean/sd Median/quantiles	Two-way table	Mean/sd per group	Correlation
Graphs:	Bar chart	Dotplot Boxplot Histogram *etc.*	Clustered bar chart	Clustered dotplot Clustered boxplot Clustered histogram *etc.*	Scatterplot

We will work through each of the methods mentioned in the above table.

Example

Consider again the research questions of the previous example.

Using the table from the previous slide, what method(s) could we use to construct a graph to answer each research question?

1.

2.

3.

4.

Categorical data

We will simultaneously treat the problems of summarising **one** categorical variable and studying the association between **two** categorical variables, because similar methods are used for these problems.

Numerical summaries of categorical data:

The main tool for summarising categorical data is a table of frequencies (or percentages).

Definition

A **table of frequencies** consists of the counts of how many subjects fall into each level of a categorical variable.

A **two-way table** (of frequencies) counts how many subjects fall into each combination of levels from a pair of categorical variables.

Example

Suppose we have a sample consisting of whether a student would prefer to use Windows/PC or a Mac. We can summarise the data as follows:

Preference	Windows/PC	Mac
Frequency	237	128

Example

Consider the question of whether there is an association between gender and whether or not a passenger on the Titanic survived. We can summarise the results from passenger records as follows:

		Survival outcome	
		Survived	Died
Gender	Male	142	709
	Female	308	154

If studying the association between two categorical variables, a two-way table cross-classifies subjects according to how many fall in each combination of categories across the two variables.

Whenever one of the variables of interest has only two possible outcomes, then a list (or table) of percentages is a useful alternative way to summarise the data.

In the Titanic example (previous slide), an alternative summary is to use the percentage survival conditional on gender.

We see that a much higher percentage of females survived compared to males: their survival rate was

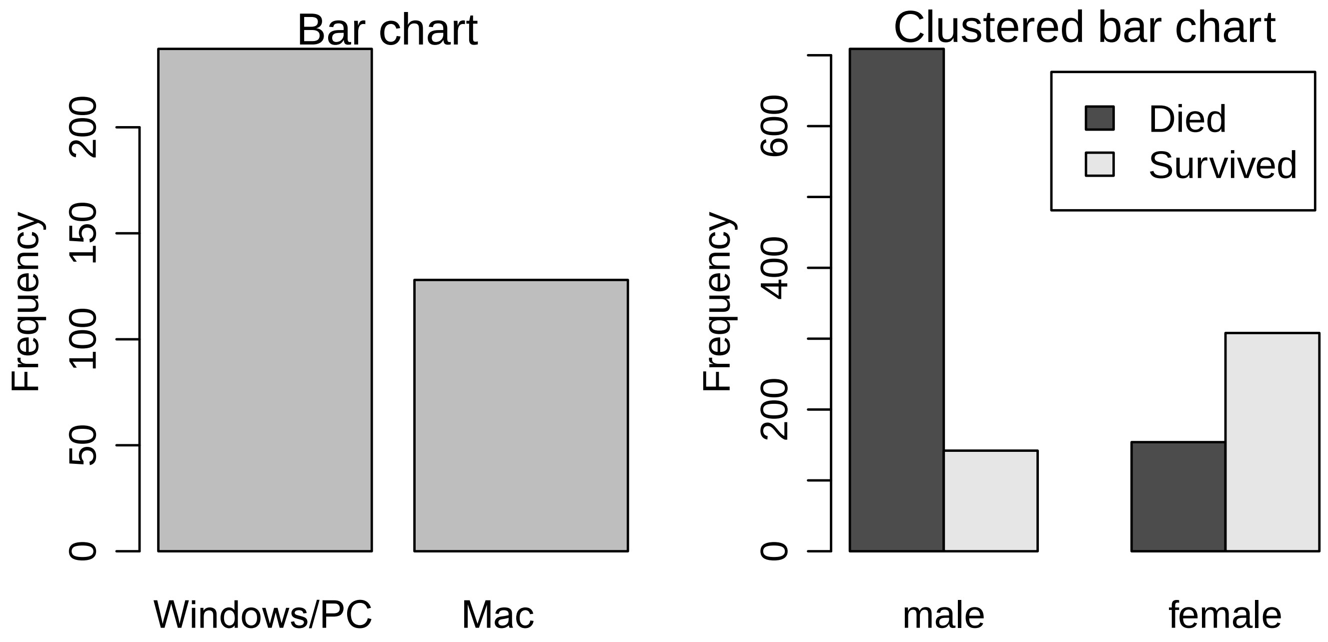
$$100 \times \{308/(308+154)\} \approx 67\% \text{ vs. } 100 \times \{142/(142+709)\} \approx 17\%!$$

If we are interested in an association between more than two categorical variables, it's possible to extend the above ideas, *e.g.* construct a three-way table. . .

Graphical summaries of categorical data

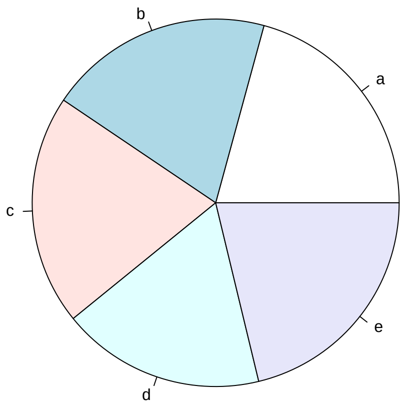
A **bar chart** is a graph of a table of frequencies.

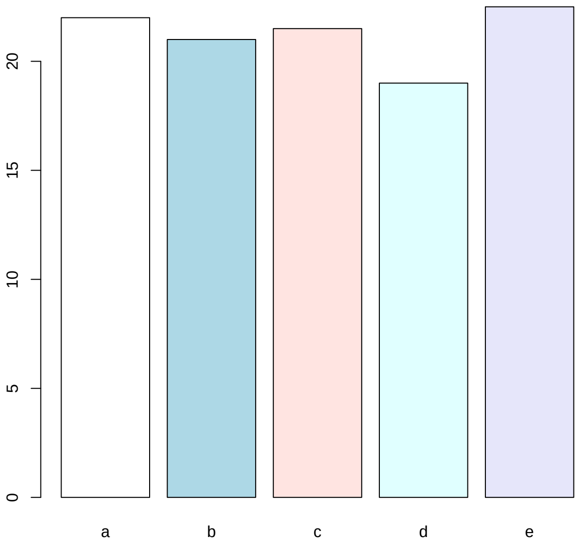
A **clustered bar chart** graphs a two-way table, spacing the “bars” out as clusters to indicate the two-variable structure:



Pie charts are often used to graph categorical variables, however these are **not** generally recommended.

It has been shown that readers of pie charts find it more difficult to understand the information that is contained in them, *e.g.* comparing the relative size of frequencies across categories.





(For details, see the Wikipedia entry on pie charts and references therein http://en.wikipedia.org/wiki/Pie_chart)

Quantitative data

When summarising a quantitative variable, we are usually interested in three things:

- **Location** or "centre" or "central location" – a value around which most of the data lie.
- **Spread** – how variable the values are around their centre.
- **Shape** – other information about a variable apart from location and spread. Skewness is an important example, which may or may not be a result of suspected outliers/unusual observations.

Numerical summaries of quantitative data

Commonly used numerical summaries (or **statistics** or **measures**) of a quantitative variable are the sample mean, variance and standard deviation:

Definition

The **sample mean**

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

is a natural measure of location of a quantitative variable.

The **sample variance**

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

is a common measure of spread.

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

Note that $(x_1 - \bar{x}) + (x_2 - \bar{x}) + \cdots + (x_n - \bar{x}) = 0$.

The variance is a useful quantity for theoretical purposes, as we will see in the coming chapters. We divide by $n - 1$ as this gives an unbiased estimator of the population variance (sometimes this is called Bessel's correction). We will learn what an unbiased estimator is later in the course.

The standard deviation however is of more practical interest because it is on the same scale as the original variable and hence is more readily interpreted.

The sample mean and variance are very widely used and we will derive a range of useful results about these estimators in this course. There are many others!

Can you think of any other well-known statistics for the central location and/or spread?

Summarizing with position: Median and quantiles

Definition

First, imagine we line up all our n data points from smallest to largest: $\{x_{(1)}, x_{(2)}, \dots, x_{(n)}\}$. Here, $x_{(1)}$ is the smallest observation, $x_{(2)}$ is the second-smallest observation, ..., $x_{(n)}$ is the largest observation in the dataset.

Now, let's talk about the **median**: it's the middle value of our ordered data.

$$\tilde{x}_{0.5} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd (the single middle value)} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n+2}{2}\right)}\right) & \text{if } n \text{ is even (the average of the two middle values)} \end{cases}$$

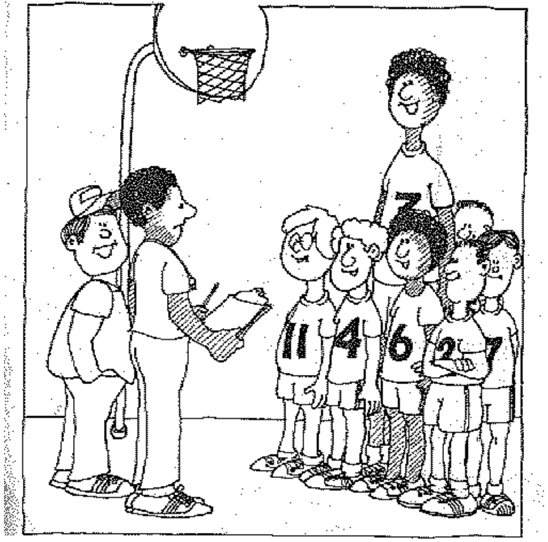
More generally, **quantiles** tell us about values at different positions in our sorted data.

The pth sample **quantile**, $\tilde{x}_p$, is roughly the value below which a fraction p of the data falls. One way to estimate it is to find the k-th ordered value, $\tilde{x}_p = x_{(k)}$ where $p \approx \frac{k-0.5}{n}$.

For other values of p we can estimate by drawing a line between the points (also called linear interpolation).

The median is sometimes suggested as a measure of location, instead of $\bar{x}$, because it is much less sensitive to unusual observations (outliers). However, it is much less widely used in practice.

1. Measuring center or average



" SHOULD WE SCARE THE OPPOSITION BY ANNOUNCING OUR MEAN HEIGHT OR LULL THEM BY ANNOUNCING OUR MEDIAN HEIGHT ? "

moore

There are a number of alternative (but very similar) ways of defining sample quantiles.

A different method again is used as the default approach on the statistics package `R/RStudio`.

Example

The following (ordered) dataset is the number of mistakes made when ten subjects are each asked to do a repetitive task 500 times.

2 4 5 7 8 10 14 17 27 35

Calculate the 5th and 15th sample quantiles of the data, then using these calculate the 10th sample quantile.

There are ten observations ($n = 10$) in the dataset. For the 5th sample quantile we have $p = 0.05$, so

$$p = \frac{k - 0.5}{n} \Rightarrow 0.05 = \frac{k - 0.5}{10} \Rightarrow k = 1 \Rightarrow \tilde{x}_{0.05} = x_{(1)} = 2.$$

Similarly, we can show that the 15th sample quantile is 4 since $\tilde{x}_{0.15} = x_{(2)} = 4$.

For the 10th sample quantile, we get $k = 1.5$ which is in the middle of $x_{(1)}$ and $x_{(2)}$. To estimate this value we can take the average (that is, we interpolate),

$$\tilde{x}_{0.1} = \frac{1}{2}\left(x_{(1)} + x_{(2)}\right) = \frac{1}{2}(2 + 4) = 3.$$

Apart from $\tilde{x}_{0.5}$, the two important quantiles are the **first and third quartiles**, $\tilde{x}_{0.25}$ and $\tilde{x}_{0.75}$ respectively.

Definition

These terms are used to define the **interquartile range**

$$IQR = \tilde{x}_{0.75} - \tilde{x}_{0.25}$$

which is sometimes suggested as an alternative measure of spread to the sample standard deviation, because it is much less sensitive to unusual observations (also called "outliers").

It is however rarely used in practice.

Definition

Another useful numerical summary of sample data is the *five number summary*, which consists of the median and quartiles of the sample, together with the minimum and maximum observed values, often written as follows:

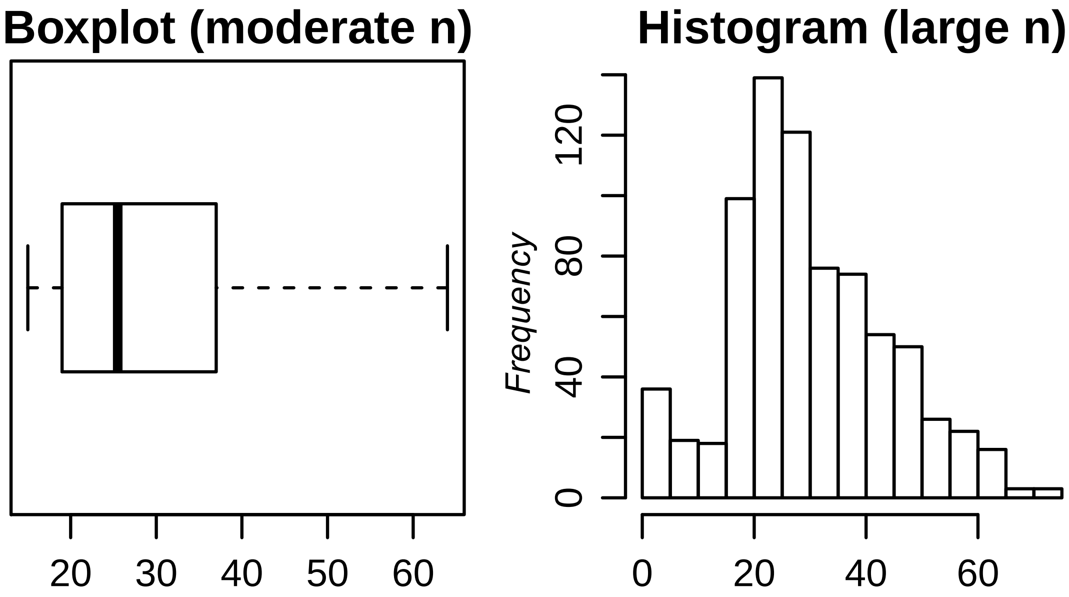
$$\left\{x_{(1)},\ \tilde{x}_{0.25},\ \tilde{x}_{0.5},\ \tilde{x}_{0.75},\ x_{(n)}\right\}$$

This ordered selection of numbers can tell us a lot of useful information about a variable at a glance.

Graphical summaries of quantitative data

There are many ways to summarise a variable, and a key thing to consider when choosing a graphical method is the sample size (n).

Two common plots are:

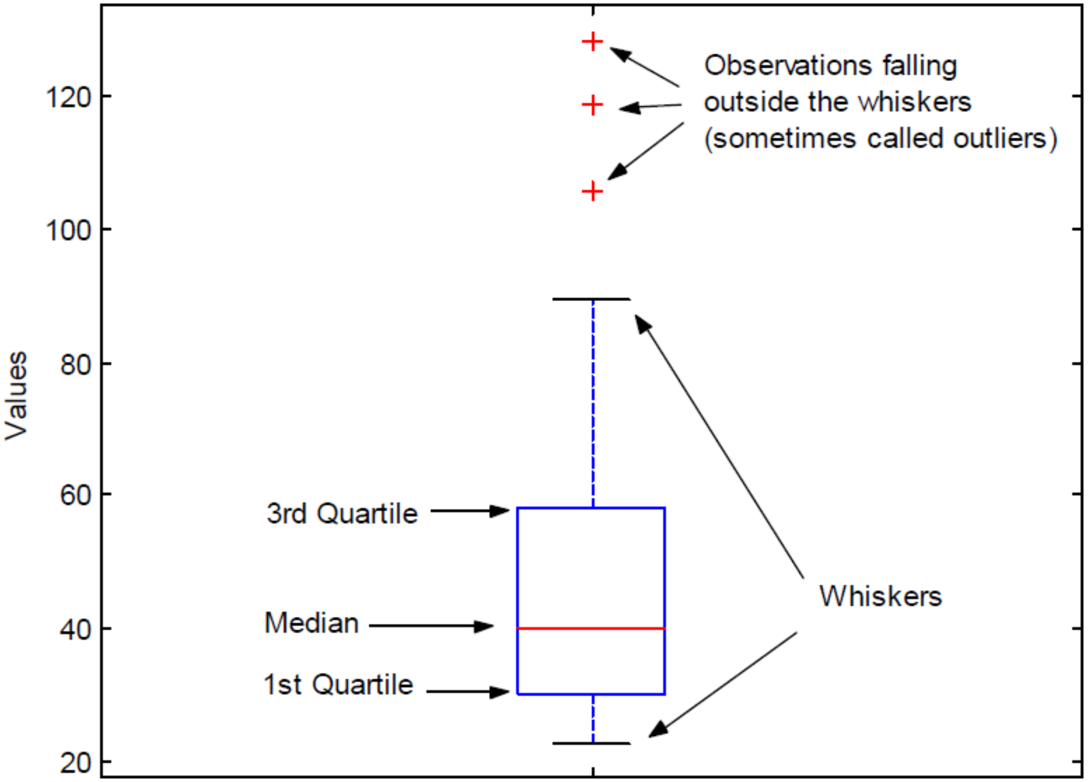


Boxplot (moderate n)

Histogram (large n)

Boxplot

A **boxplot** concisely describes location, spread and shape via the median, quartiles and extremes.


Values
120
100
80
60
40
20
Observations falling outside the whiskers (sometimes called outliers)
3rd Quartile
Median
1st Quartile
Whiskers

- The line in the middle of the box is the median, the measure of centre.

- The box is bounded by the upper and lower quartiles, so box width is a measure of spread (the interquartile range, IQR).

- The whiskers extend until the most extreme value within one and a half interquartile ranges ($1.5 \times IQR$) of the nearest quartile.

- Any value farther than $1.5 \times IQR$ from its nearest quartile is classified as an extreme value (or "outlier"), and labelled as a dot or open circle. Specifically, an observation is deemed an outlier if is:

$$> (\tilde{x}_{0.75} + 1.5 \times IQR) \quad \text{or} \quad < (\tilde{x}_{0.25} - 1.5 \times IQR).$$

Boxplots are most useful for moderate-sized samples (*e.g.* $10 < n < 50$).

Definition

A **histogram** is a plot of the frequencies or relative frequencies of values within different intervals or *bins* that cover the range of all observed values in the sample.

This involves breaking the data up into smaller subsamples, and as such it will only find meaningful structure if the sample is large enough (*e.g.* $n > 30$) for the subsamples to contain non-trivial counts.

An issue in histogram construction is choice of number of bins.

A useful rough rule-of-thumb is to use

$$\text{number of bins} = \sqrt{n}.$$

Kernel density estimator*

A histogram is a step-wise rather than smooth function. A quantitative variable that is continuous (*i.e.* a variable that can take any value within some interval) might be better summarised by a smooth function.

Definition

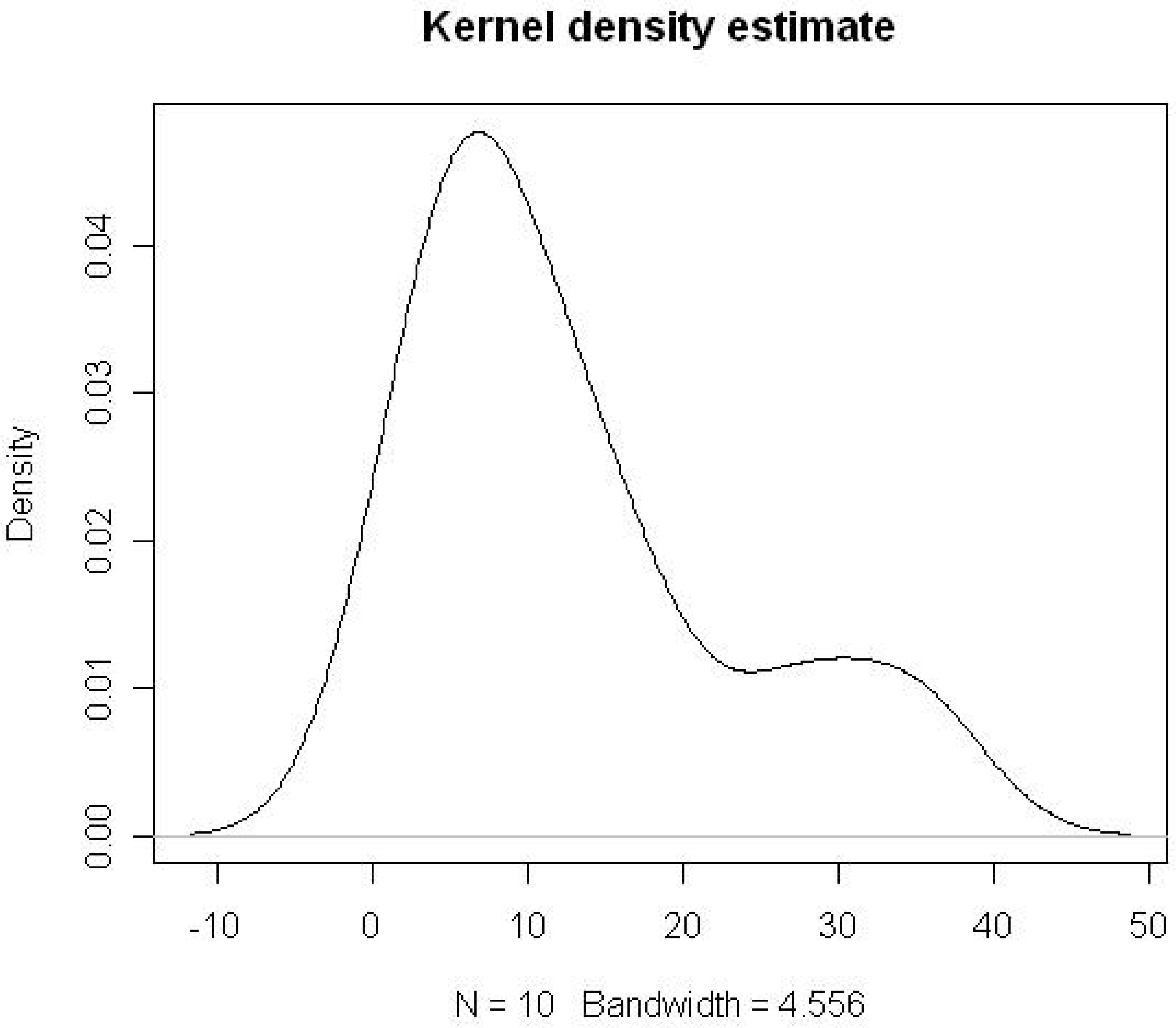
An alternative estimator that often has better properties for continuous variables is a **kernel density estimator**:

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} w_h(x - x_i)$$

for some choice of weighting function $w_h(x)$ which includes a "bandwidth parameter" h.

Usually, $w(x)$ is chosen to be the normal density (defined in Chapter 3) with mean 0 and standard deviation h. A lot of research has studied the issue of how to choose a bandwidth h, and most statistics packages are now able to automatically choose an estimate of h that usually performs well.

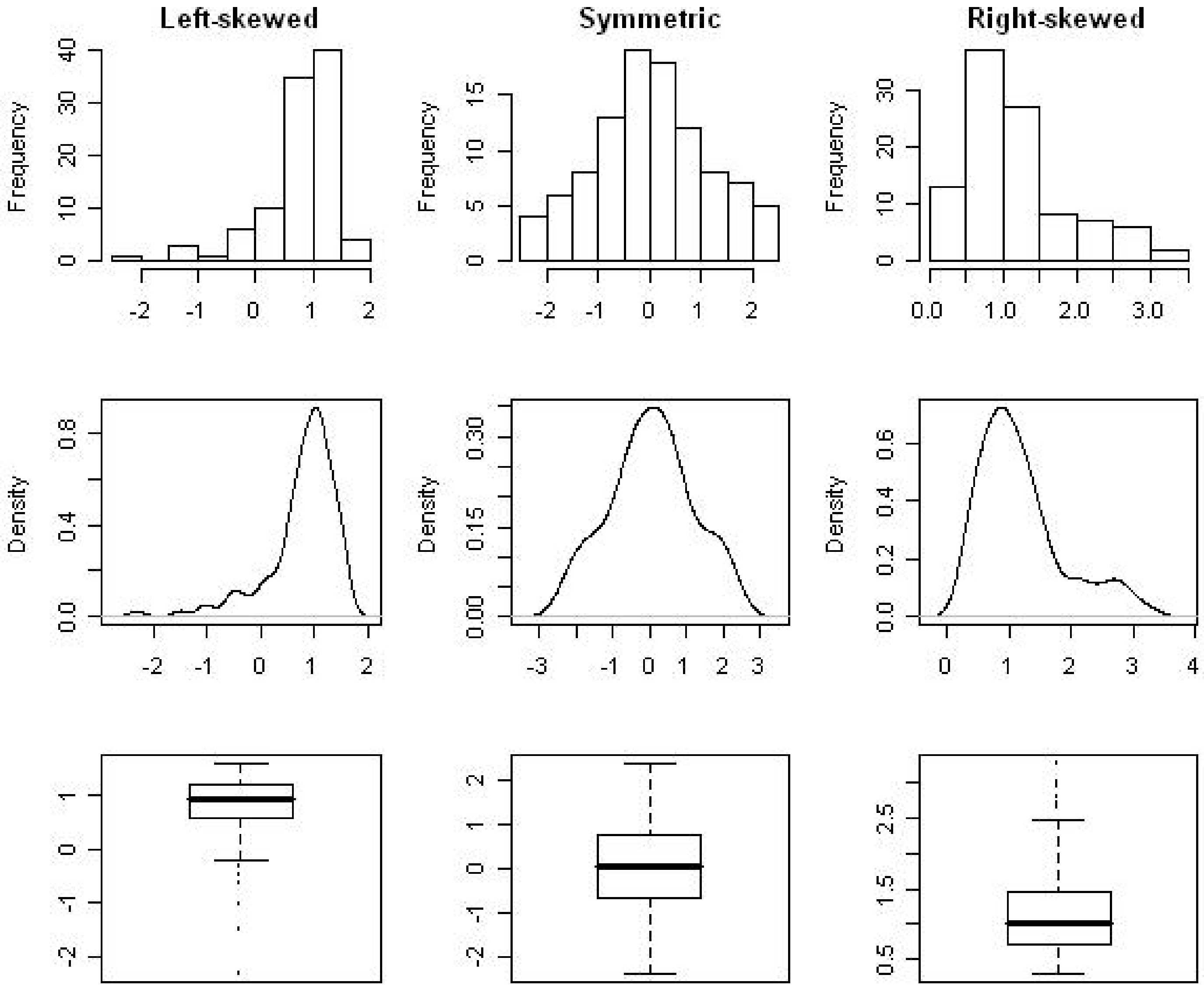
The larger h is, the larger the bandwidth that is used *i.e.* the larger the range of observed values x_i that influence estimation of $\hat{f}_h(x)$ at any given point x.


Kernel density estimate
Density
0.00
0.01
0.02
0.03
0.04
-10
0
10
20
30
40
50
N = 10   Bandwidth = 4.556

Shape of a distribution

Something we can see from a graph that is hard to see from numerical summaries is the **shape** of a distribution. Shape properties, broadly, are characteristics of the distribution apart from location and spread.

An example of an important shape property is **skew** – if the data tend to be asymmetric about its centre, it is skewed. We say data are "left-skewed" if the left tail is longer than the right, conversely, data are right-skewed if the right-tail is longer.


Left-skewed
Symmetric
Right-skewed
Frequency
0 10 20 30 40
-2 -1 0 1 2
0 5 10 15
-2 -1 0 1 2
0 10 20 30
0.0 1.0 2.0 3.0
Density
0.0 0.4 0.8
-2 -1 0 1 2
0.00 0.15 0.30
-3 -1 0 1 2 3
0.0 0.2 0.4 0.6
0 1 2 3 4
-2 -1 0 1
-2 -1 0 1 2
0.5 1.5 2.5

Coefficient of skewness and outliers

There are some numerical measures of shape, *e.g.* the coefficient of skewness κ_1:

$$\hat{\kappa}_1 = \frac{1}{(n-1)s^3} \sum_{i=1}^{n} (x_i - \bar{x})^3$$

but they are rarely used – perhaps because of extreme sensitivity to outliers, and perhaps because shape properties can be easily visualised as above.

Definition

Another important thing to look for in graphs is **outliers** – unusual observations that might carry large weight in analysis.

Such values need to be investigated – are they errors, are they "special cases" that offer interesting insights, how dependent are results on these outliers.

Summarising associations between variables

We have already considered the situation of summarising the association between categorical variables, which leaves two possibilities to consider...

Associations between quantitative variables:

Consider a pair of samples from two quantitative variables

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}.$$

Often, we would like to understand how the x and y variables are related. Analysis of two quantitative variables is commonly referred to as (linear) regression[2].

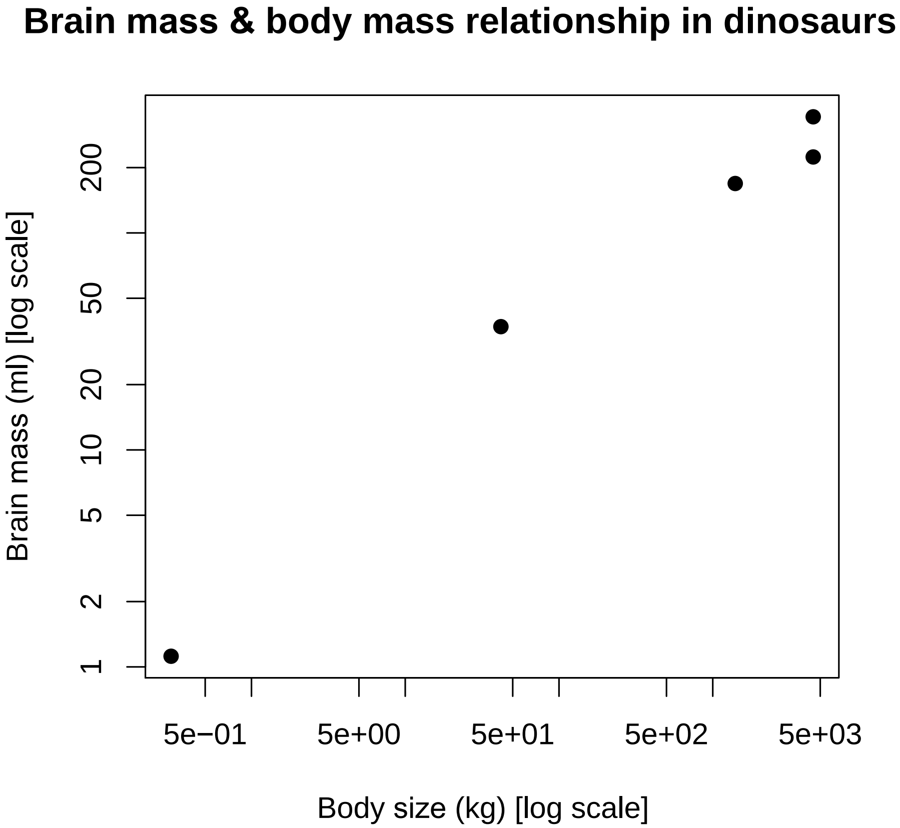
Definition

An effective graphical display of the relationship between two quantitative variables is a **scatterplot** – a plot of the y_i against the x_i.

[2] We won't spend too much time on regression in MATH2801; we will only cover a few concepts. More on regression and linear models will be covered in MATH2831/2931 in T3.

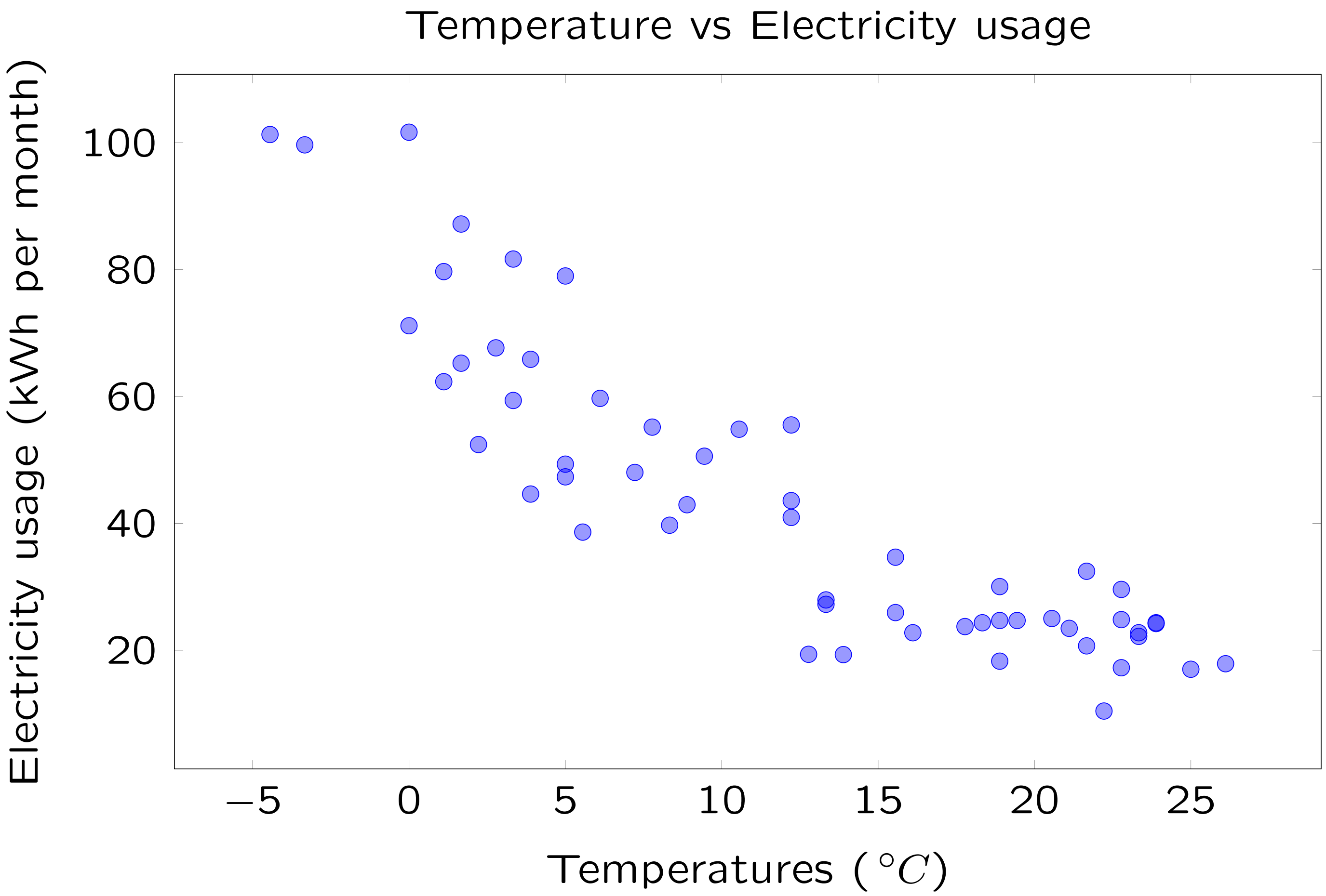
Example

How did brain mass change as a function of body size in dinosaurs?



Example

How does annual electricity usage change as temperature changes?



Definition

An effective numerical summary of the *linear* relationship between two quantitative variables is the **correlation coefficient** (r):

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$$

where $\bar{x}$ and s_x are the sample mean and standard deviation of x, similarly for y.

Here, r measures the strength and direction of the association between x and y:

Result

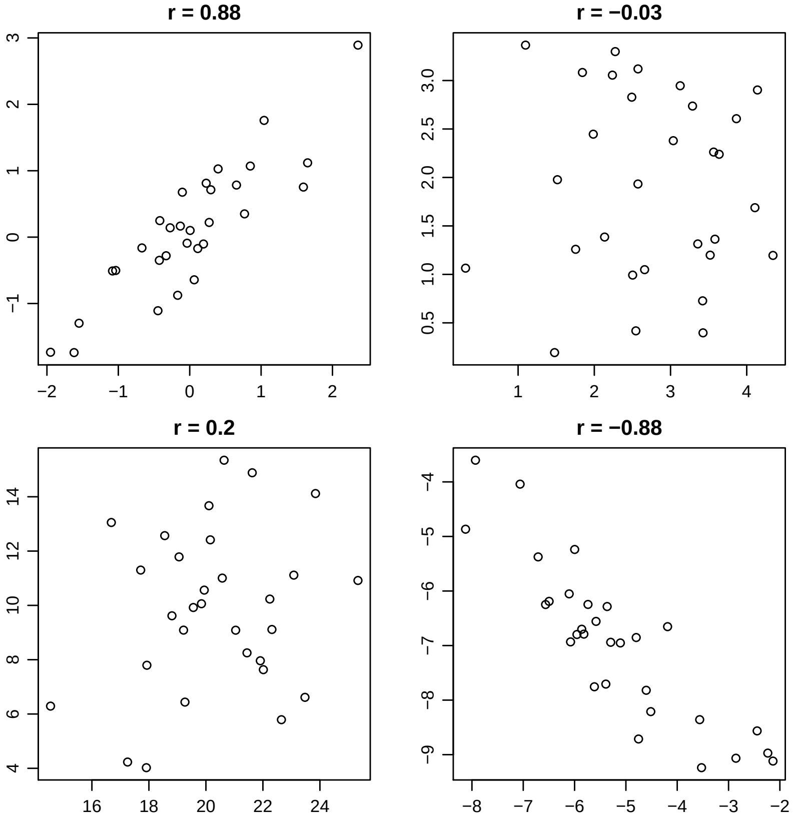
1. $|r| \leq 1$ (or $-1 \leq r \leq 1$).
2. $r = -1$ if and only if $y_i = a + bx_i$ for each i, for some constants a, b such that $b < 0$.
3. $r = 1$ if and only if $y_i = a + bx_i$ for each i, for some constants a, b such that $b > 0$.

*Can you prove these results? (*Hint:* consider using the square of $\frac{x_i-\bar{x}}{s_x} + \frac{y_i-\bar{y}}{s_y}$).

These results imply that r measures the strength and direction of associations between x and y:

- Strength of (linear) association – values closer to 1 or -1 suggest that the relationship is closer to a straight line.

- Direction of association – values less than zero suggest a decreasing relationship, values greater than zero suggest an increasing relationship.



Associations between categorical and quantitative variables

When studying whether categorical and quantitative variable are associated, an effective strategy is to summarise the quantitative variable(s) separately for each level of the categorical variable(s).

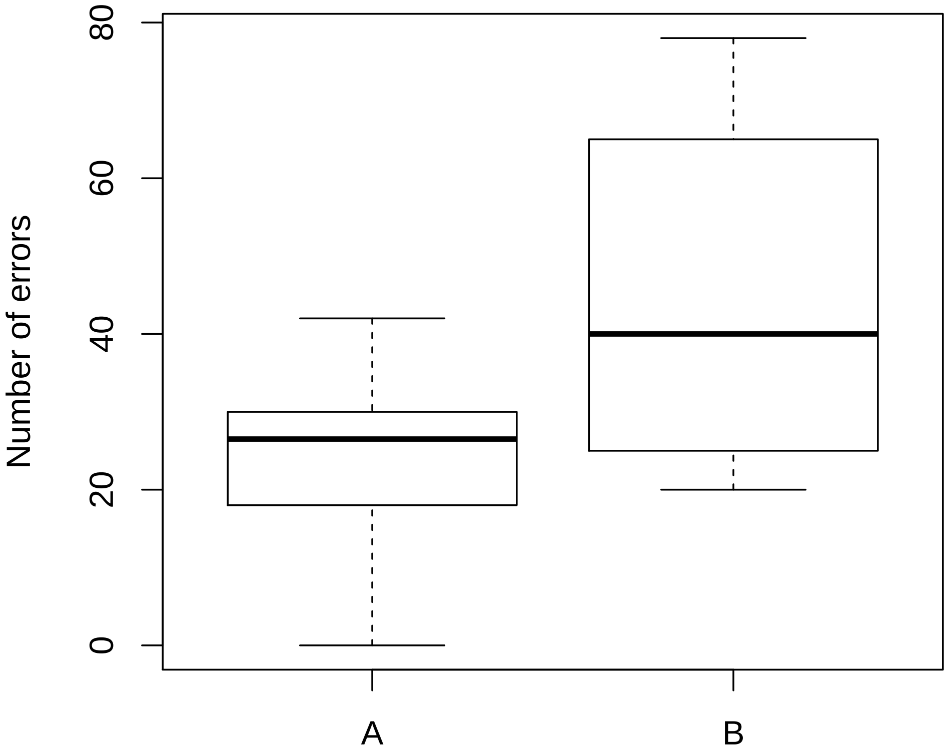
Example

Recall the guinea pig experiment – we want to explore whether there is an association between a nicotine treatment (categorical) and number of errors made by offspring (quantitative).

To summarise number of errors, we might typically use mean/sd and a boxplot.

Instead of looking at the association between number of errors and nicotine treatment, we calculate mean/sd of number of errors for each of the two levels of treatment (nicotine and no nicotine), and construct a boxplot for each level of treatment:

	$\bar{x}$	s
Sample A	23.4	12.3
Sample B	44.3	21.5



In the above example, the boxplots are presented on a common axis – sometimes this is referred to as **comparative boxplots** or “side-by-side boxplots”.

An advantage of boxplots over histograms is that they can be quite narrow and hence readily compared across many samples by stacking them side-by-side.

Some interesting extensions are reviewed in the article “40 years of boxplots” by Hadley Wickham and Lisa Stryjewski at Rice University.

This idea can be naturally extended to when there are more than two variables.

Transforming data

Transforming data is typically done for one of two reasons – to change the scale data were measured on (linear transformation), or to improve data properties (non-linear transformation).

We will treat each of these in turn.

Definition

A **linear transformation** of a sample from a quantitative variable, from $\{x_1, x_2, \dots, x_n\}$ to $\{y_1, y_2, \dots, y_n\}$, satisfies:

$$y_i = a + bx_i \quad \text{for each } i \text{ and } b \neq 0.$$

Linear transformations do not affect the shape of a distribution – only its location and spread.

Effects of linear transformation on statistics

Result

Consider a linear transformation $y_i = a + bx_i$, $i = 1, \dots, n$, and its effects on some statistic to be calculated from the x_i (m_x) and the y_i (m_y).

If

$$m_y = a + bm_x,$$

then we say that m is a measure of location – *e.g.* the mean: $\bar{y} = a + b\bar{x}$ or the median: $\tilde{y}_{0.5} = a + b\tilde{x}_{0.5}$.

If m_x is a measure of spread in the same units as x,

$$m_y = |b|m_x$$

– *e.g.* the standard deviation: $s_y = |b|s_x$.

If m_x is a measure of shape, then:

$$m_y = \begin{cases} m_x & \text{if } b > 0 \\ -m_x & \text{if } b < 0. \end{cases}$$

– *e.g.* the skewness coefficient: $\kappa_y = \kappa_x$ for $b > 0$.

These results are necessary by definition – to be a measure of location of x, m_x has to “move with the data” under changes of scale.

For m_x to be a measure of spread, it needs to be invariant under translation but it needs to vary with resizing.

A measure of shape on the other hand should be invariant under any change of scale.

Let’s have a look at some examples.

Examples

Show the following sequence of results: under linear transformation:

1. The sample mean $\bar{x} = \frac{1}{n}\Sigma_{i=1}^{n} x_i$ behaves as a measure of location.
2. The standard deviation $s_x = \sqrt{\frac{1}{n-1}\Sigma_{i=1}^{n}(x_i - \bar{x})^2}$ behaves as a measure of spread.
3. The correlation coefficient

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

behaves as a measure of shape (consider linear transformations of x_i and y_i).

Example

Dinosaur body mass (x) was measured (well, in this case it was estimated!) in kilograms.

If we transform the body mass data into grams instead (denoted y), how will the following values calculated from y relate to their counterparts calculated from x?

1. $\bar{y}$, mean body mass in grams.
2. s_y, standard deviation of body mass in grams.
3. r_y, the correlation between body mass (in grams) and brain mass.

What about when the above three statistics are calculated on log-transformed data, rather than the raw data?

A particularly important example of a linear transformation is “standardisation” of data to z-scores, as below:

Definition

The z-**score**, or **standardised score** of a quantitative variable is defined as

$$z = \frac{x - \bar{x}}{s_x}$$

The z-score is a measure of unusualness – it measures how many standard deviations above/below the mean a value is (extreme values being unusual ones, far from zero).

We will attach probabilities to precisely how unusual a given z-score is in the coming chapters.

Examples

Sydney's daily maximum temperature in March has a mean of about 25 degrees Celsius, and a standard deviation of 2.2. Hence the following z-scores:

A March maximum temperature of 20 degrees in Sydney: $z = -2.3$, since

$$z = \frac{x - \bar{x}}{s_x} = \frac{23 - 25}{2.2} = -2.3$$

A maximum temperature of 35 degrees in Sydney: $z = 4.5$.

Some other unusually large z-scores:
Sachin Tendulkar's cricket batting average: $z = 1.5$, and
Don Bradman's cricket batting average: $z = 5.5$.

Your winnings if you win the jackpot in the Powerball lotto: $z = 7367$!!!

Non-linear transformations

If we have a quantitative variable that is strongly skewed, then the patterns we see (in scatterplots or elsewhere) can be dominated by a few outlying values.

In such cases transforming data can be a good idea – applying a non-linear transformation to a dataset will change its shape, often changing it for the better!

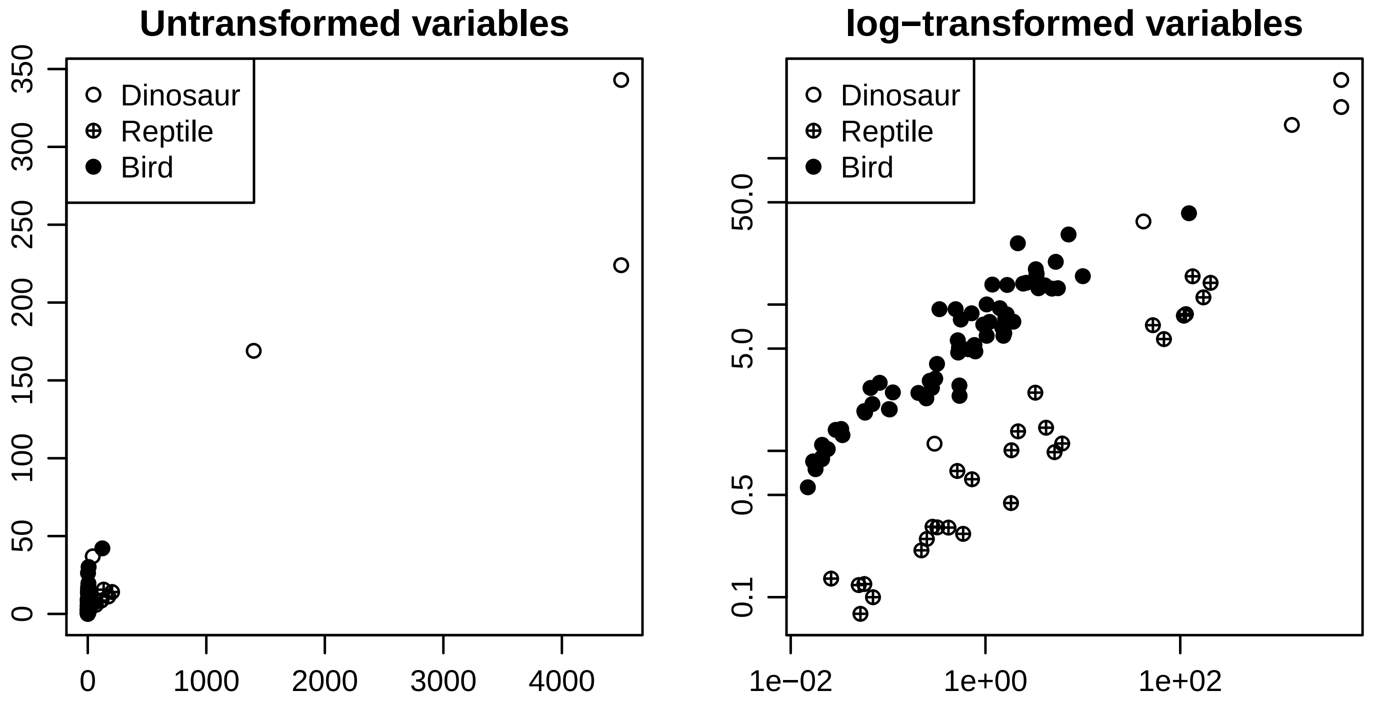
The most common transformation is a **log-transformation**.

You can use any base (base 2 and 10 are most common for interpretability) and it doesn't really matter. However, in this course we interpret $\log()$ as the natural logarithm $\log_e()$.

Example

Consider brain-mass – body-mass data for dinosaurs, reptiles and birds.

Compare scatterplots of log-transformed data and untransformed data below:



In the untransformed plot, little can be seen except for three outlying values (*Tyrannosaurus*, *Carhcarodontosaurus* and *Allosaurus*). On transformation, a lot of interesting structure becomes apparent.

One reason why the log-transformation often works so well in revealing structure is the special property: $\log(ab) = \log(a) + \log(b)$.

This can be understood as taking multiplicative processes and making them additive – that is, a variable that "grows" in a multiplicative way (*e.g.* virus transmission, account balance, size, profit, population size, etc.) can be understood as growing in an additive way once log-transformed.

This is useful because in graphs (and in most analyses) additive patterns are the easiest to perceive.

Result

Let $y = h(x)$ be some non-linear transformation of real-numbered values x. In most cases, we have

$$\bar{y} \neq h(\bar{x}).$$

This point should be kept in mind when analysing transformed data – the **mean** of transformed data is a different quantity to the mean of the originally observed variable, and they do not even have a one-to-one correspondence in most cases!