

F20SA / F21SA Statistical Modelling and Analysis

Contents

1	Introduction	1–1
1.1	What is Statistics and Probability?	1–1
1.2	Example of uses	1–1
1.2.1	Examples of the Application of Probability	1–1
1.2.2	Examples of Statistical Analysis	1–2
1.3	Sampling	1–2
1.3.1	Choosing a Sample	1–2
1.4	Types of Data	1–3
2	Data Summary	2–1
2.1	First Steps with Data	2–1
2.2	Displaying Data	2–1
2.2.1	Example: Histogram	2–2
2.3	Measures of Centre	2–3
2.3.1	Mean	2–3
2.3.2	Median	2–5
2.3.3	Robustness of Measures of Centre and Skewness	2–5
2.4	Measures of Spread	2–7
2.4.1	Standard Deviation	2–7
2.4.2	Interquartile Range	2–9
2.4.3	Boxplot	2–10
2.5	Further properties of Mean and Standard Deviation	2–11
3	Introduction to Probability	3–1
3.1	Terminology	3–1
3.2	Definition of Probabilities	3–2
3.2.1	Relation to relative frequencies	3–2
3.2.2	Basic rules for Probabilities	3–2
3.2.3	Equally likely outcomes	3–3
3.2.4	Venn Diagram and Addition Law	3–3
3.2.5	Complementary Events	3–4
3.2.6	Example of probability calculations	3–4
3.3	Conditional Probability	3–7
3.3.1	Partitioning of an event	3–8
3.3.2	Bayes Rules	3–9
3.3.3	Independence	3–9
3.3.4	Tree Diagrams	3–10

3.4	Summary	3–11
3.5	Further Probability Examples	3–12
4	Random Variables	4–1
4.1	Motivating Example	4–1
4.2	Definitions	4–2
4.2.1	Discrete Random Variables and Probability Mass Functions	4–2
4.2.2	Cumulative Distribution Function for Discrete Random Variables	4–3
4.2.3	Continuous Random Variables and Probability Density Function	4–4
4.2.4	Cumulative Distribution Function for Continuous Random Variables	4–4
4.3	Mean and Variance	4–5
4.3.1	Definition of Expectation	4–5
4.3.2	Sample mean and Expectation	4–6
4.3.3	Variance	4–7
4.3.4	Change of Origin and scale	4–8
4.4	Median and Quartiles	4–8
4.5	Examples	4–10
5	Special Distributions	5–1
5.1	Expectation and Variance revisited	5–2
5.2	Specific Discrete Random Variables	5–2
5.2.1	Bernoulli Random Variables $Bernoulli(p)$	5–2
5.2.2	Uniform Random Variables $U(k)$	5–3
5.2.3	Binomial Random Variables, $Bin(n, p)$	5–3
5.2.4	Poisson Random Variables $Po(\lambda)$	5–6
5.2.5	Geometric Random Variables, $Geo(p)$	5–7
5.3	Specific Continuous Random Variables	5–7
5.3.1	Uniform Random Variables, $U(a, b)$	5–7
5.3.2	Exponential Random Variables $Exp(\lambda)$	5–8
5.3.3	Gamma Random Variables $Gamma(\alpha, \beta)$	5–9
5.3.4	Beta Random Variables $Beta(\alpha, \beta)$	5–9
5.3.5	Normal Random Variables, $N(\mu, \sigma^2)$	5–10
5.4	Examples of Random Variable Calculations.	5–12
5.5	Sampling Distributions and the Central limit Theorem	5–16
5.5.1	Sampling Distributions	5–16
5.5.2	Properties of the sample sum and sample mean	5–16
5.5.3	Central Limit Theorem (CLT)	5–18
5.5.4	Distribution of sample variance	5–19
5.6	Sampling from a normal variable	5–19
5.6.1	χ^2 , t and F distributions	5–20
5.6.2	Sampling distributions- mean and variance	5–21
5.7	Two samples	5–21
5.7.1	Difference between sample means, two independent samples	5–21
5.7.2	Ratio of two sample variances, independent normal samples	5–22

5.8	Examples: Sampling distributions and CLT	5–23
6	Model Fitting 1: Parameter Estimation	6–1
6.1	Motivation	6–1
6.2	Introduction	6–1
6.3	Properties of estimators	6–2
6.3.1	Bias of estimators	6–2
6.3.2	Consistency	6–3
6.3.3	Efficiency	6–3
6.3.4	Sufficiency	6–4
6.4	Methods of constructing estimators	6–4
6.4.1	Method of moments estimators (MME)	6–4
6.4.2	Method of least squares estimators (LSE)	6–4
6.4.3	Method of maximum likelihood	6–5
6.5	Examples	6–10
7	Model Fitting 2: Confidence Intervals	7–1
7.1	Introduction	7–1
7.2	Confidence Intervals for Population Mean	7–2
7.2.1	σ^2 known	7–3
7.2.2	σ^2 unknown	7–3
7.2.3	One sided confidence intervals	7–4
7.3	CIs for population variance	7–4
7.4	CIs for population proportion	7–4
7.5	CIs for a Poisson mean	7–5
7.6	CIs based on general MLEs	7–5
7.7	Examples	7–6
8	Decision Making: Hypothesis Testing	8–1
8.1	Introduction	8–1
8.2	Definitions	8–2
8.3	Standard test statistics	8–2
8.3.1	Testing a population mean	8–2
8.3.2	Testing a population variance	8–3
8.3.3	Testing a population proportion	8–3
8.3.4	Testing a Poisson mean	8–3
8.3.5	Examples	8–3
8.4	Significance and P-values	8–5
8.4.1	Examples	8–5
8.5	Final Comments	8–6
9	Regression	9–1
9.1	Motivation	9–1
9.2	Linear Regression	9–2
9.2.1	Fitting the Model	9–3

9.2.2	Residuals	9–5
9.2.3	Model checking	9–7
9.2.4	Terminology and Warning	9–11
9.2.5	Standard Errors, Confidence Intervals and Hypothesis testing	9–11
9.3	Correlation	9–14
9.3.1	Linear Regression Examples	9–16
9.4	Multiple Regression	9–23
9.4.1	ANOVA	9–23
9.4.2	Multiple Linear Regression Case Study	9–24
9.4.3	Model Selection	9–26
10	Bayesian inference	10–1
10.1	Bayes' Theorem, priors, likelihoods and posteriors	10–1
10.1.1	A simple example	10–1
10.1.2	Estimating parameters using Bayesian inference	10–2
10.1.3	Example of Bayesian inference.	10–3
10.2	Reporting conclusions from a Bayesian analysis	10–6
10.2.1	Deriving Bayesian estimators	10–6
10.2.2	Some examples	10–8
10.2.3	Reporting an interval	10–9
10.2.4	More on selecting priors	10–11
10.3	Predictive distributions	10–13
11	Non Parametric Methods	11–1
11.1	Introduction	11–1
11.1.1	Example data set	11–1
11.2	Permutation and Randomization Test	11–2
11.2.1	Setup	11–2
11.2.2	Permutation Method	11–2
11.2.3	Randomization Method	11–4
11.3	Bootstrap Methods	11–8
11.3.1	Bootstrap Confidence Interval	11–8
11.3.2	Smoothed Bootstrap	11–10
12	Principal component analysis and factor analysis	12–1
12.1	Principal component analysis	12–1
12.2	Factor analysis	12–3

F20SA / F21SA Statistical Modelling and Analysis

Chapter 1: Introduction

Contents

1.1	What is Statistics and Probability?	1-1
1.2	Example of uses	1-1
1.2.1	Examples of the Application of Probability	1-1
1.2.2	Examples of Statistical Analysis	1-2
1.3	Sampling	1-2
1.3.1	Choosing a Sample	1-2
1.4	Types of Data	1-3

1.1 What is Statistics and Probability?

Statistics is the development of mathematical tools to study data and make inferences and decisions. This includes design of experiments to minimise cost but provide the necessary data, and how to update beliefs given observations.

In comparison probability is the mathematical study of uncertainty, chance and risk. It was originally developed for the study of gambling games but has found applications in many other areas from finance, to energy systems and health diagnosis.

This course only touches on the basic ideas. It will provide you an introduction to the tools and ideas that can be used when considering data and modelling risk and chance.

1.2 Example of uses

Statistics and probability have found applications in many areas, which can be seen in the fact that in newspapers there are regularly stories making use of statistics.

1.2.1 Examples of the Application of Probability

- Given a positive test result for a rare disease, how likely are you to have the disease?
- Given a suspect has the same rare type of shoe as discovered at a crime scene, what does this tell us about their guilt?
- How likely is a given poker hand?
- What is the optimal strategy for gambling, for example the Monty Hall problem?
- How much redundancy is needed for a given performance guarantee, e.g. in electrical generators?
- How likely is a reservoir to be empty in summer?

1.2.2 Examples of Statistical Analysis

- What do unemployment statistics tell us?
- Does the observed behaviour imply wrongdoing, e.g.
 - Insider trading
 - Harold Shipman
- Is there evidence of global warming?
- Are hospitals improving?
- What is the effect of mobile phones on cancer rates?
- Is the quality of concrete consistent across batches?
- What can we learn about flood intensity and duration?
- Can we predict the rate of road rutting?

1.3 Sampling

When carrying out scientific studies, it is important to consider the **population** of interest. This could be all the salmon alive at a fish farm or all of the houses on sale in Edinburgh. Typically, measurements are taken on a **sample** from this population. This can be for reasons of cost (time or money) or because it is physically impossible to measure everything.

For some topics, choices need to be made about the population of interest, these choices affect the conclusions that can be drawn.

Example 1.1. Testing water quality from a borehole. Water can be collected and tested, but there are choices of the amount to collect and the frequency of collection.

Chemists (and doctors) typically use the word ‘sample’ for a single collection; statisticians use the word for a set of collections. **Sample Size** means the number of **Units** in the collection.

1.3.1 Choosing a Sample

What does a sample tell us about the population from which it was taken? If our sample is chosen in such a way that each member of the population is **equally likely** to be selected, this course will show how to answer this. Any sampling scheme that does not do this is likely to give misleading answers and care must be taken.

If the whole of the population can be listed, it is easy to take a random sample:

- put names in a hat, mix and draw some out;
- use tables of random numbers;
- use suitable computer program.

If you are using tables, it is important to vary the starting point. Most computer packages do this automatically. Some packages allow you to set the seed of the pseudo-random generator, so that a repeatable sequence is obtained.

In practice, it is often necessary or desirable to use several stages in drawing the sample.

Example 1.2. For farmed salmon, one might take a random sample of fish farms and then take a sample of the fish from each of the chosen farms.

If it is not possible to list the whole population, randomisation should be introduced in other ways (for example quadrat sampling of vegetation).

The sampling scheme will affect the way in which the data should be analysed and the usefulness of the results.

1.4 Types of Data

Different types of information need to be dealt with in different ways. The following classification scheme is quite general.

- **Categorical Variables** are qualitative.
E.g. the name of the degree programme
 - If there are only 2 values possible, it is called a **Binary Variable**.
E.g. Coin toss: Heads, Tails.
 - **Ordinal Variables** have categories that are in a definite order.
E.g. Poor, Satisfactory, Good.
E.g. Olympic medals: Gold, Silver, Bronze.
- **Quantitative Variables** take values on a scale with natural units.
 - **Discrete Variables** which only contain certain values, often integers.
E.g. Counts.
 - **Continuous Variables** which, in principle, can be measured to arbitrary precision.
E.g. Times or Distances.

Example 1.3. Further examples of data types:

Data	Data Type
Weight of Girder	Quantitative - Continuous
Rock type	Categorical
Number of Floors of Buildings	Quantitative - Discrete
Patient Health	Categorical - Ordinal
Flood risk for Property	Categorical - Binary
Height of Bridge	Quantitative - Continuous

F20SA / F21SA Statistical Modelling and Analysis

Chapter 2: Data Summary

Contents

2.1	First Steps with Data	2-1
2.2	Displaying Data	2-1
2.2.1	Example: Histogram	2-2
2.3	Measures of Centre	2-3
2.3.1	Mean	2-3
2.3.2	Median	2-5
2.3.3	Robustness of Measures of Centre and Skewness	2-5
2.4	Measures of Spread	2-7
2.4.1	Standard Deviation	2-7
2.4.2	Interquartile Range	2-9
2.4.3	Boxplot	2-10
2.5	Further properties of Mean and Standard Deviation	2-11

2.1 First Steps with Data

In many situations you will be provided with raw data, often in the form of a table or spreadsheet, and asked to analyse the data. For example, a survey was conducted of 100 UK-domiciled university undergraduates. The height (in metres, to the nearest 0.01m) and the sex (M/F) of each undergraduate were recorded. The raw data can be seen in Table 2.1. The issue is how to start to make sense of such data as the table alone is not enlightening and does not provide us with any insight to the experiment.

The first step is to summarise the data, this process is called Exploratory Data Analysis (EDA). Depending on the type of data we have the possible methods which can be utilised will vary, but there are generally two classes of methods:

- Plotting the data,
- Calculating summary statistics.

2.2 Displaying Data

In many cases it is often useful to display the data first. This serves to provide us with an initial feel for the data and often a way of presenting our results.

Examples of types of plots include:

- Dotplot
- Bar Chart

1.72(M)	1.53(F)	1.73(M)	1.64(M)	1.76(M)	1.53(F)	1.66(F)
1.67(M)	1.78(M)	1.74(M)	1.55(M)	1.70(F)	1.73(M)	1.62(M)
1.54(F)	1.61(F)	1.85(M)	1.75(M)	1.66(M)	1.65(F)	1.87(M)
1.48(F)	1.71(M)	1.77(M)	1.80(F)	1.58(M)	1.94(M)	1.72(M)
1.69(F)	1.73(M)	1.69(M)	1.80(M)	1.60(F)	1.75(M)	1.56(F)
1.78(M)	1.83(M)	1.80(M)	1.62(F)	1.84(M)	1.59(F)	1.97(M)
1.81(M)	1.60(F)	1.51(M)	1.83(M)	1.89(M)	1.73(F)	1.56(F)
1.76(M)	1.63(M)	1.61(F)	1.91(M)	1.75(M)	1.69(F)	1.88(M)
1.70(M)	1.59(F)	1.50(F)	1.66(M)	1.64(F)	1.80(M)	1.62(F)
1.67(F)	1.63(M)	1.71(M)	1.60(F)	1.75(M)	1.58(M)	1.74(M)
1.77(M)	1.81(M)	1.65(M)	1.62(F)	1.73(M)	1.60(M)	1.57(F)
1.75(M)	1.82(M)	1.76(F)	1.61(F)	1.69(M)	1.55(F)	1.85(M)
1.76(M)	1.54(M)	1.78(M)	1.43(F)	1.60(F)	1.66(M)	1.92(M)
1.66(F)	1.74(M)	1.62(F)	1.81(M)	1.79(M)	1.67(M)	1.51(F)
1.86(M)	1.64(F)					

Table 2.1: Heights and sex of a sample of 100 UK undergraduates

- Histogram
- Stem-and-leaf Plot,
- Boxplot
- Scatter plot

2.2.1 Example: Histogram

Histograms are the most common form of plot when dealing with continuous data. The data values are grouped into ‘Bins’ and we count the number of values within each bin. It is important to note that the bins need not be of the same width and in many circumstances the data is not suited to equal bin widths.

To construct a histogram:

1. Choose bin width so that there are between 6 and 20 bins. Use more bins for larger sample sizes. The first bin should contain the minimum value and the last should contain the maximum value.
2. Construct the frequency table.
3. Draw histogram. The **Area** of each rectangle should be proportional to the number in the bin.

In the case that the bins have equal width, the height of each rectangle is proportional to the frequency within each bin.

Example 2.1. In Figure 2.1 we have a histogram of the student heights given in Table 2.1. We can note from this that there are two peaks within the data, one around 1.6m and 1.75m. We say this data is bimodal.

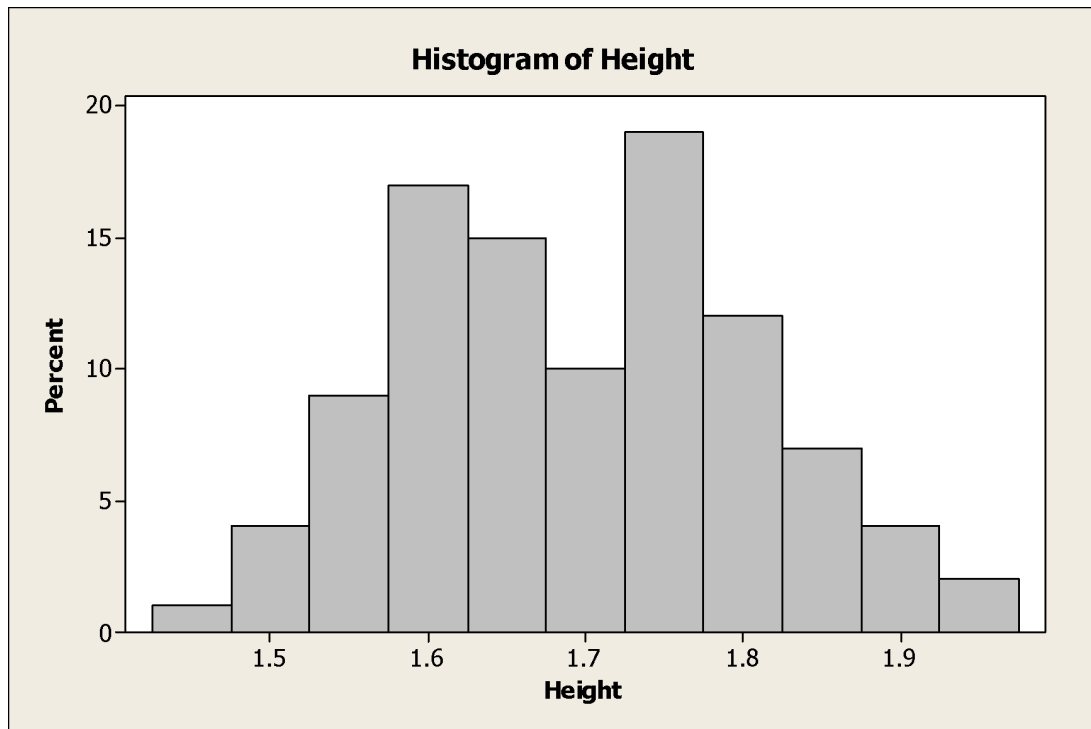


Figure 2.1: Histogram of Student heights, equal bin sizes

2.3 Measures of Centre

Graphical methods are useful as a first step, but to draw conclusions we need numerical methods. There are many summary statistics which can be calculated but the most valuable are those which provide a measure of the centre of the data. There are several ways to measure the centre of data but we consider the two most common measures:

- The sample mean,
- The median.

2.3.1 Mean

Definition 2.1. For a sample with **sample size** n (number of values in sample) and data values: x_1, x_2, \dots, x_n we define the **sample mean** (\bar{x}) by

$$\begin{aligned}
 \bar{x} &= \frac{\text{sum of data values}}{\text{sample size}} \\
 &= \frac{x_1 + x_2 + \dots + x_n}{n} \\
 &= \frac{1}{n} \sum_{i=1}^n x_i
 \end{aligned}$$

Example 2.2.

Data: 1, 2, 3 Total = 6 Mean = 2.0

Data: 4, 6 Total = 10 Mean = 5.0

Data: 1, 2, 3, 4, 6 Total = 16 Mean = 3.2

Note that the overall mean is **not** the average of the two separate means.

Example 2.3. Fourteen electrical components were tested to destruction and the failure times in hours were recorded. The raw data can be seen in Table 2.2.

55	283	76	197	5	45	102
28	37	4	139	10	82	75

Table 2.2: Failure times of 14 electrical components in hours.

For this data we have:

$$\begin{aligned}
 n &= 14 \\
 \sum x_i &= 55 + 283 + \cdots + 75 \\
 \bar{x} &= \frac{1138}{14} \\
 &= 81.3
 \end{aligned}$$

Mean failure time = 81.3 hours

Sample Means for combined samples

To calculate an overall mean from group means, we need to take into account the different sample sizes.

- Calculating the individual totals for each group.
- Add these together to obtain the overall total.
- Use this to calculate the overall sample mean.

Example 2.4. Mean age of 50 males = 22.6 years

Mean age of 30 females = 19.4 years

Mean age of combined group of 80 people?

$$\begin{aligned}
 \text{Male total} &= 22.6 \times 50 = 1130 \\
 \text{Female total} &= 19.4 \times 30 = 582 \\
 \text{Mean age} &= \frac{1130 + 582}{50 + 30} = 21.4 \text{ years}
 \end{aligned}$$

Example 2.5. Mean salary of 50 people = £18500

Mean salary of 30 males = £18100

Mean salary of the females?

$$\begin{aligned}
\text{Overall total} &= 18500 \times 50 = 925000 \\
\text{Male total} &= 18100 \times 30 = 543000 \\
\text{Female total} &= 925000 - 543000 = 382000 \\
\text{Female mean} &= \frac{382000}{20} = 19100
\end{aligned}$$

2.3.2 Median

The **median** is the **middle value** when the data have been **sorted**. It is used as an alternative measure of the centre of the data when the data set contains extreme values which would “distort” the mean. If we consider the following two samples

$$\{1, 2, 3, 4, 5\}$$

and

$$\{1, 2, 3, 4, 90\}.$$

In the first case the sample mean is 3 but in the second the sample mean is 20 which is not representative of most of the values. In comparison in both cases the median is 3.

Calculating the median

The data must be sorted before the median can be calculated.

If we have n data points then:

- if n is **odd**, it is the $\left(\frac{n+1}{2}\right)$ sorted value.
- if n is **even**, it is the average of the $\left(\frac{n}{2}\right)$ and the $\left(\frac{n}{2} + 1\right)$ sorted values.

Example 2.6. We find the median for the failure times of the electrical components given in table 2.2. The first step is to sort the data so we can find the middle value.

Here the sorted data is

$$4, 5, 10, 28, 37, 45, 55, 75, 76, 82, 102, 139, 197, 283$$

Since $n = 14$ we require the average of the 7th and 8th sorted values.

So the median = $\frac{55+75}{2} = 65$ hours.

2.3.3 Robustness of Measures of Centre and Skewness

If the data is roughly symmetric the mean is the best measure of centre. In comparison the median is a more **robust** measure of the ‘typical’ value (not so affected by extreme values). If the data has a larger tail on one side than the other we say the data is skewed. If the larger tail is on the right we say it is positively skewed, see Figure 2.2, and in this case the mean will be larger than the median. Where as if the larger tail is on the left we say it is negatively skewed, see Figure 2.3, and the mean will be less than the median.

There are alternative robust methods for measuring the centre of data including the trimmed mean (the sample mean after removing the largest and smallest 5% of data) and mode for discrete data (value with the highest frequency).

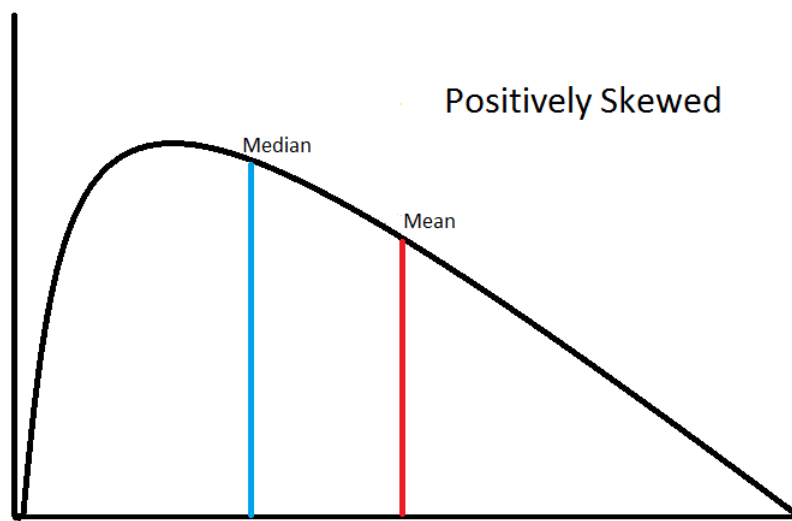


Figure 2.2: A positively skewed distribution

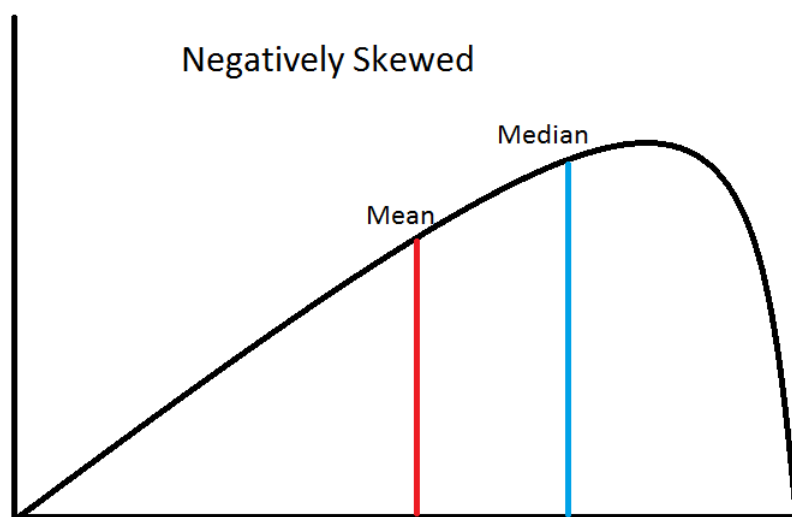


Figure 2.3: A negatively skewed distribution

2.4 Measures of Spread

It is important to compute a measure of **spread** as well as a measure of centre. In other words, a measure of whether the data values are spread out or are bunched together. The simplest measure is the range:

$$\text{Range} = \text{Maximum} - \text{Minimum}$$

but this tends to increase with sample size and hence can be not very informative. Instead we consider two other measures of spread:

- Sample standard deviation,
- Inter-quartile range.

2.4.1 Standard Deviation

The first measure of spread we consider is the sample standard deviation. In order to compute it, we make use of the sample mean.

Definition 2.2. 1. Find the **squared** distance between x_i and \bar{x}

$$d_i^2 = (x_i - \bar{x})^2$$

2. Add these up and divide the total by $n - 1$ (instead of n).
3. “Undo the squaring”

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

Example 2.7.

x_i	d_i	d_i^2
1	-2	4
2	-1	1
3	0	0
4	1	1
5	2	4

$$\bar{x} = 3$$

$$\sum d_i^2 = 10$$

$$n - 1 = 4$$

$$s = \sqrt{\frac{10}{4}} = 1.58$$

Remark 2.1. • Alternative formula (easier to compute).

$$s = \sqrt{\frac{1}{n - 1} \left(\sum x_i^2 - \frac{(\sum x_i)^2}{n} \right)}$$

- On the rare occasions that the **population** mean is known exactly, the divisor n is used instead of $n - 1$. This is called the population standard deviation.

- The standard deviation is in the same units as the data. E.g. if x_i is measured in cm, then s is measured in cm.
- The larger the value of s , the more the data is spread out about the mean.
- The square of the standard deviation is called the variance.
- The Greek letter σ (sigma) is often used to denote standard deviation and σ^2 for the variance.

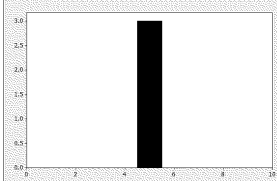
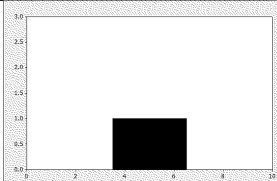
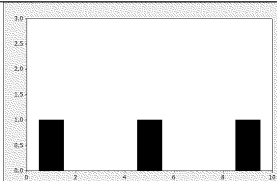
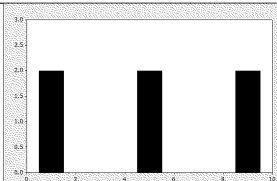
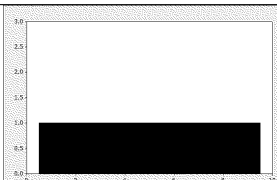
Example 2.8. We find the standard deviations for the failure times of the electrical components given in Table 2.2. The sample size is $n = 14$.

The sum of the observations is $\sum x_i = 1138$.

The sum of squares is $\sum x_i^2 = 55^2 + 283^2 + \cdots + 75^2 = 174092$. Therefore the sample standard deviation is

$$\begin{aligned}
 s &= \sqrt{\frac{1}{13} \left(174092 - \frac{1138^2}{14} \right)} \\
 &= \sqrt{\frac{81588.86}{13}} \\
 &= 79.2 \text{ hours}
 \end{aligned}$$

Example 2.9. In each case, the mean is 5.0

Sample data: 5, 5, 5 Standard Deviation = 0.0	
Sample data: 4, 5, 6 Standard Deviation = 1.0	
Sample data: 1, 5, 9 Standard Deviation = 4.0	
Sample data: 1, 1, 5, 5, 9, 9 Standard Deviation = 3.58	
Sample data: 1, 2, 3, 4, 5, 6, 7, 8, 9 Standard Deviation = 2.74	

An earlier example showed that when groups are combined, the overall mean is obtained by adding totals rather than averaging means. To obtain an overall standard deviation, it is necessary to add up the sums of squared values.

Example 2.10.

Data: 1, 2, 3 Total SS = 14 S.D. = 1.00

Data: 4, 6 Total SS = 52 S.D. = 1.41

Data: 1, 2, 3,
4, 6 Total SS = 66 S.D. = 1.92

The overall s.d. = $\sqrt{(66 - 16^2/5)/4} = 1.92$

Note that the abbreviations SS and S.D. are often used for the ‘Sums of Squares’ and for ‘Standard Deviation’ respectively.

2.4.2 Interquartile Range

If the data is skewed, the inter-quartile range may be a better measure of spread than the standard deviation.

- The **lower quartile** $Q1$ is value such that a quarter of the sample takes values **less** than $Q1$.
- The **upper quartile** $Q3$ is value such that a quarter of sample takes values **greater** than $Q3$.
- The **inter-quartile range** (IQR) is defined to be $IQR = Q3 - Q1$.

You may also encounter Deciles and Percentiles; these divide data into tenths and hundredths. With n data points we arrange them in ascending order and then

1. $Q1$ is the $\left(\frac{n+1}{4}\right)$ st observation.
2. $Q3$ is the $\left(\frac{3(n+1)}{4}\right)$ st observation or the $\left(\frac{n+1}{4}\right)$ st observation when counting down from the largest value.
3. $IQR = Q3 - Q1$.

Example 2.11. For 2, 4, 6, 7, 8, 9. $n = 6$ so

1. $Q1$ is the $\left(\frac{7}{4}\right) = 1.75$ st observation, $Q1 = 3.5$.
2. $Q3$ is the $\left(\frac{21}{4}\right) = 5.25$ st observation, $Q3 = 8.25$.
3. $IQR = Q3 - Q1 = 8.25 - 3.5 = 4.75$.

Example 2.12. We find the inter-quartile range for the failure times of the electrical components given in Table 2.2. The sorted data is

4	5	10	28	37	45	55
75	76	82	102	139	197	283

$n = 14$

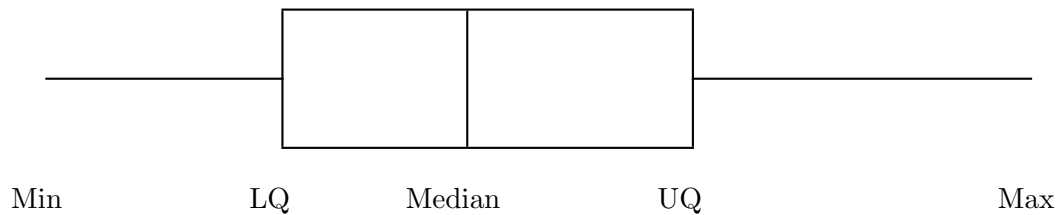
- $Q1$ is the $\frac{n+1}{4} = \frac{14+1}{4} = 3.75$ th observation.
- 3^{rd} observation = 10 and 4^{th} = 28
- $Q1 = 10 + (0.75)(28-10) = 10 + 13.5 = 23.5$ hours

- $Q3$ is the $\frac{3(n+1)}{4} = \frac{3(14+1)}{4} = 11.25$ th observation.
- 11^{th} observation=102 and 12^{th} observation=139
- $Q3 = 102 + (0.25)(139-102) = 102+9.25 = 111.25$ hours

The inter-quartile range $IQR = 111.25 - 23.5 = 87.75$ hours.

2.4.3 Boxplot

The quartiles can be used to create a display of the data called a **box-and-whisker plot** or **box plot**. The “box” is formed from the quartiles and the “whiskers” connect the box to the maximum and the minimum.



If the data is skewed, the median will not be near the middle of the box, and one whisker will be much longer than the other. The values used in drawing a boxplot are called a **five number summary**.

Example 2.13. The five number summary for the failure data in Table 2.2 is $\{4, 23.5, 65, 111.25, 283\}$ and the boxplot for the data can be seen in Figure 2.4.

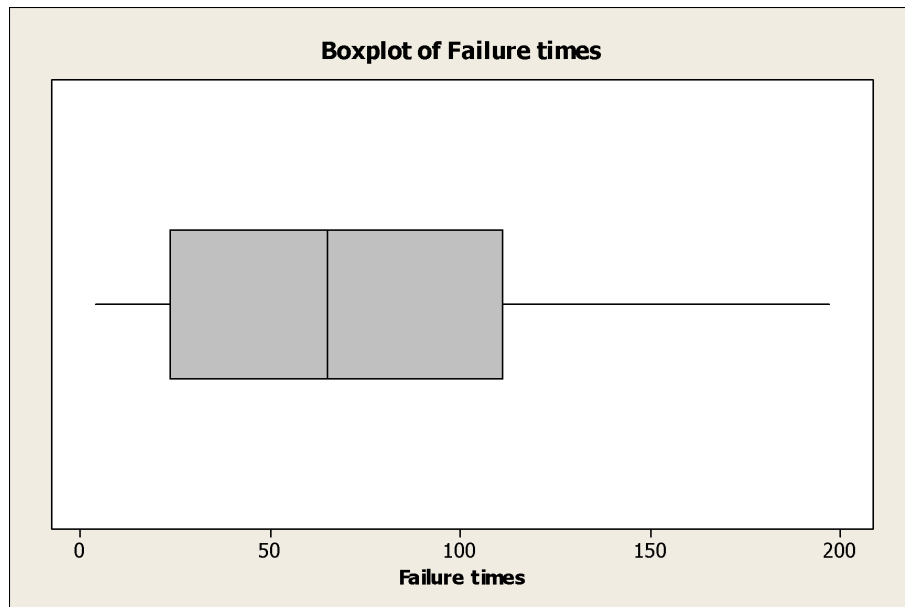


Figure 2.4: Boxplot for failure times

2.5 Further properties of Mean and Standard Deviation

If data is roughly symmetric about the mean, then:

- Approximately $\frac{2}{3}$ of the data will be within 1 s.d. of the mean
- Approximately 95% of the data will be within 2 s.d. of mean
- Usually all the data will be within 3 s.d. of the mean.
- The Inter-Quartile Range will be approximately 1.35 standard deviations.

For an explanation of this see the normal random variable in chapter 4.

We can use this idea to decide whether an observation is extreme given the other data points.

- An individual data point is considered to be extreme if it is several standard deviations away from the mean.
- The **standard score** (z-score, standardised value) for x is:

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

- This measures how many standard deviations x is above or below the mean.

Example 2.14. For data with $\bar{x} = 81.3$ and $s = 79.2$ we find the standard scores for 55 and 283.

$$x_i = 55 \quad z = \frac{55 - 81.3}{79.2} = -0.33 \quad \text{small } z \text{ (typical)}$$

$$x_i = 283 \quad z = \frac{283 - 81.3}{79.2} = 2.55 \quad \text{large } z \text{ (extreme)}$$

This suggests that 283 might be unusual.

Change of scales effect on Mean and Standard Deviation

- Addition of a constant ($y_i = x_i + c$):
 - Mean is increased by the same constant ($\bar{y} = \bar{x} + c$).
 - Standard Deviation is unchanged ($s_y = s_x$).
- Multiplication by a constant ($y_i = c \times x_i$):
 - Mean is multiplied by the constant ($\bar{y} = c \times \bar{x}$).
 - Standard Deviation is multiplied by the constant ($s_y = c \times s_x$).

Example 2.15. Temperature conversion from °C to °F.

Need to multiply by 1.8 and add 32.

Celsius Mean = 15°C and S.D. = 5.5°C

Fahrenheit Mean = $15 \times 1.8 + 32 = 59^\circ\text{F}$

Fahrenheit S.D. = $5.5 \times 1.8 = 9.9^\circ\text{F}$

F20SA / F21SA Statistical Modelling and Analysis

Chapter 3: Introduction to Probability

Contents

3.1	Terminology	3-1
3.2	Definition of Probabilities	3-2
3.2.1	Relation to relative frequencies	3-2
3.2.2	Basic rules for Probabilities	3-2
3.2.3	Equally likely outcomes	3-3
3.2.4	Venn Diagram and Addition Law	3-3
3.2.5	Complementary Events	3-4
3.2.6	Example of probability calculations	3-4
3.3	Conditional Probability	3-7
3.3.1	Partitioning of an event	3-8
3.3.2	Bayes Rules	3-9
3.3.3	Independence	3-9
3.3.4	Tree Diagrams	3-10
3.4	Summary	3-11
3.5	Further Probability Examples	3-12

3.1 Terminology

In this and the next chapter we are interested in **random experiments**, i.e., experiments for which the outcome is uncertain but is one of a known and describable set of possible **outcomes**. The **sample space** S is the set of all possible outcomes for an experiment. An **event** is a collection of possible outcomes from an experiment.

Example 3.1.

Experiment	Event
The roll of a dice	The result is more than 3
Weather in Edinburgh tomorrow	It will be snowing
Wind turbine state	The turbine is working
State of Road after 1 year	The ruts are greater than 1cm in depth

The union of two events A and B , denoted $A \cup B$, is the event which occurs if and only if at least one of events A or B occurs. This should be thought as A OR B .

The intersection of two events A and B , denoted $A \cap B$, is the event which occurs if and only if both A and B occur. This should be thought as A AND B .

The **complement** of an event A , denoted A' , occurs if and only if the event A does not occur.

The empty set, \emptyset , contains none of the possible outcomes of the experiment and so corresponds to the impossible event.

Events A and B are mutually exclusive if and only if $A \cap B = \emptyset$ (that is, the events cannot occur simultaneously).

Example 3.2.

Let us consider the experiment of rolling a die. Let:

- A be the event of an even roll ($\{2, 4, 6\}$),
- B be the event that the die roll is more than 3 ($\{4, 5, 6\}$),
- C be the event that the die roll is in $\{1, 5\}$.

Then we have:

- $A \cap B$ the die roll is $\{4, 6\}$, (A AND B)
- $A \cup B$ the die roll is $\{2, 4, 5, 6\}$, (A OR B)
- A' the die roll is odd ($\{1, 3, 5\}$), (NOT A)
- $A \cap C = \emptyset$ so A and C are mutually exclusive events.

3.2 Definition of Probabilities

Probability is a numerical scale to describe how likely events are to occur. Given an event E we assign a probability $\mathbb{P}(E)$ to the event such that $0 \leq \mathbb{P}(E) \leq 1$.

- $\mathbb{P}(E) = 0$ means ‘(almost) never happens’.
- $\mathbb{P}(E) = 1$ means ‘(almost) always happens’.

3.2.1 Relation to relative frequencies

We observe that relative frequency tends to settle down as the number of trials increases: formally we define the probability of the event occurring as the limit of the relative frequency, so

$$\begin{array}{c} \text{relative frequency} \rightarrow \text{probability} , \\ \text{as number of trials} \rightarrow \infty. \end{array}$$

The graph in Figure 3.1 shows the relative frequency of the occurrence of an event with probability 0.4 after 1, 2, \dots , 200 trials.

3.2.2 Basic rules for Probabilities

Using the definition of probability as the limit of relative frequency we formulate some initial rules that probabilities obey.

1. For any event $\mathbb{P}(E) \geq 0$ (no event can occur a negative number of times).
2. For S , the sample space, $\mathbb{P}(S) = 1$, (something has to happen).
3. For E_1, E_2 such that $E_1 \cap E_2 = \emptyset$ (they can not occur together),

$$\mathbb{P}(E_1 \cup E_2) = \mathbb{P}(E_1) + \mathbb{P}(E_2).$$

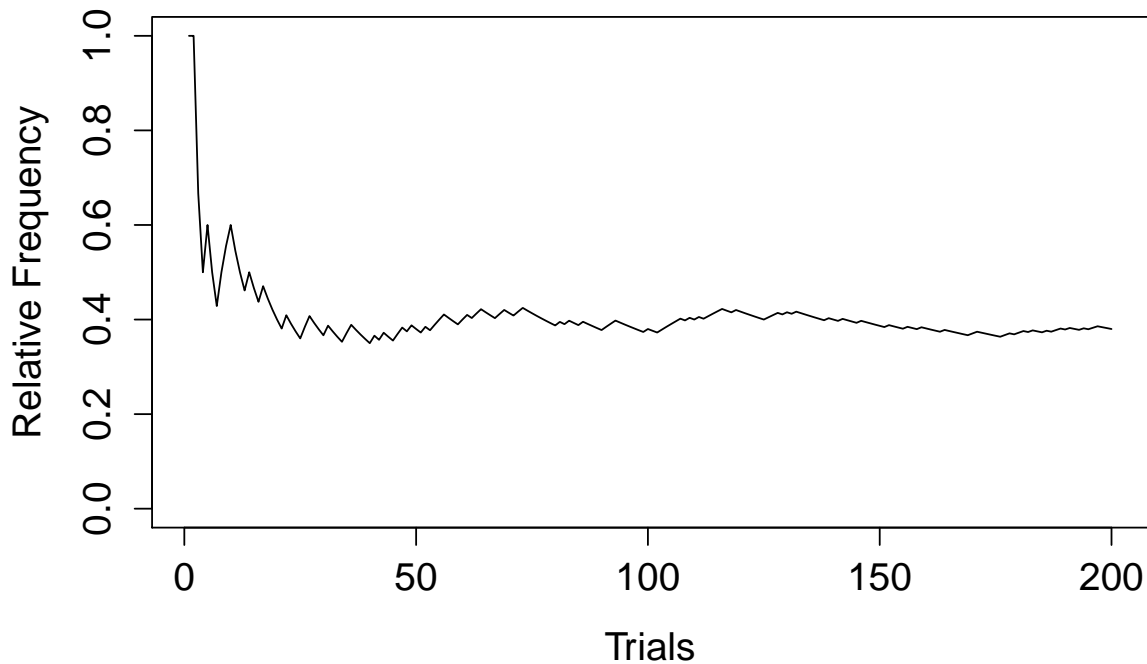


Figure 3.1: Relative frequency of a probability 0.4 event over 1, 2, ..., 200 trials

3.2.3 Equally likely outcomes

With a finite number, n , of equally likely outcomes, each outcome has probability $\frac{1}{n}$. Hence for an event E which contains r favourable outcomes we have

$$\mathbb{P}(E) = \frac{\text{number of favourable outcomes}}{\text{number of possible outcomes}} = \frac{r}{n}.$$

The key to evaluating probabilities in this case is to define a sample space in a convenient way and then to count the numbers of outcomes corresponding to various events.

Example 3.3.

Consider a throw of two fair six-sided dice, one red and the other blue. Let A be the event that the score = 7 or 8. Let

$$S = \{(i, j) : i = 1, 2, 3, 4, 5, 6; j = 1, 2, 3, 4, 5, 6\}$$

where i and j are the scores on the red and blue die respectively. S consists of 36 elements (equally-likely outcomes).

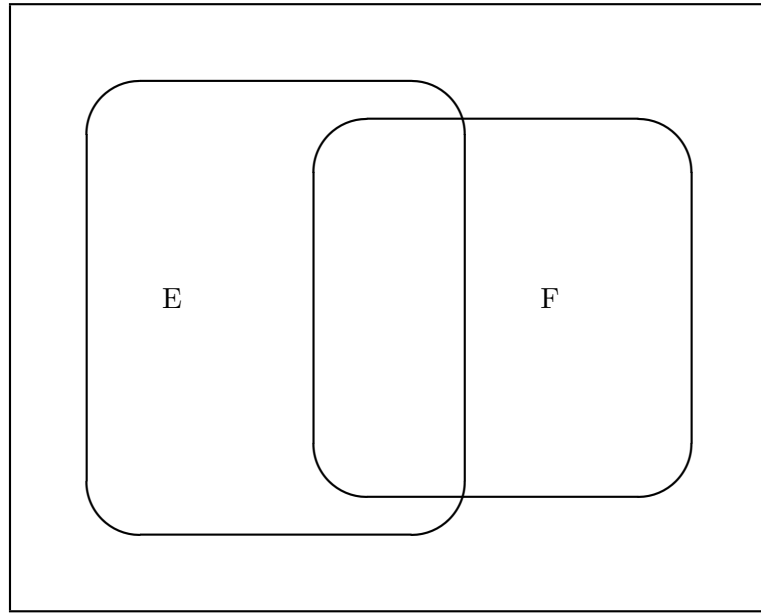
$$A = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1), (2, 6), (3, 5), (4, 4), (5, 3), (6, 2)\}.$$

A consists of 11 elements, so

$$\mathbb{P}(A) = 11/36 = 0.3056.$$

3.2.4 Venn Diagram and Addition Law

Often it is useful to have a graphical representation of events and how they relate to each other. For this a Venn diagram can be useful. An example of a Venn diagram can be found in Figure

Figure 3.2: A Venn diagram with two events E and F

3.2

The outer rectangle represents all the possible outcomes. The regions labelled E and F represent the outcomes in these events. The overlap region of E and F represents the event $E \cap F$ (“ E And F ”). The larger outline that contains both letters represents the event $E \cup F$ (“ E Or F ”). In the Venn diagram, it is possible to draw all of the events so that the **area** corresponds to the probability of an event. This implies that the area enclosed by the outer rectangle is one unit. It is clear from the diagram that:

$$\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F).$$

The subtraction of the ‘AND’ term is because adding the two probabilities means the overlap region has been counted twice.

This is called the “Addition Law” or the “Or Law”. This generalizes the basic rule for adding mutually exclusive events ($\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F)$ if E and F are mutually exclusive).

3.2.5 Complementary Events

Remember that if A is an event, then the event A' is said to be the complementary event. Since $S = A \cup A'$ and $A \cap A' = \emptyset$ we have:

$$\mathbb{P}(A) + \mathbb{P}(A') = \mathbb{P}(S) = 1.$$

The use of a complementary event can make the calculation of a probability easier.

3.2.6 Example of probability calculations

Example 3.4.

Athletes data.

Distance	USA	GB	Kenya	
Sprinter	7	5	0	12
400m	2	2	0	4
Middle	2	4	7	13
Long	1	1	5	7
	12	12	12	36

If an athlete is picked at random (equally likely):

$$\mathbb{P}(400\text{m}) = \frac{4}{36} = \frac{1}{9}$$

$$\mathbb{P}(\text{GB}) = \frac{12}{36} = \frac{1}{3}$$

$$\begin{aligned}\mathbb{P}(\text{USA or GB}) &= \mathbb{P}(\text{USA}) + \mathbb{P}(\text{GB}) \text{ because mutually exclusive} \\ &= \frac{12}{36} + \frac{12}{36} = \frac{2}{3}\end{aligned}$$

$$\mathbb{P}(\text{USA or Sprinter})$$

$$= \mathbb{P}(\text{USA}) + \mathbb{P}(\text{Sprinter}) - \mathbb{P}(\text{USA and Sprinter})$$

$$= \frac{12}{36} + \frac{12}{36} - \frac{7}{36}$$

$$= \frac{17}{36}$$

Note that the events “Sprinter” and “Kenyan” are mutually exclusive, because there are no athletes who belong to both events.

Example 3.5.

Roulette Wheel There are 37 numbers on the wheel: 0 – 36.

18 are red (9 odd and 9 even)

18 are black (9 odd and 9 even)

1 is green (zero)

We can display this in a table:

	Red	Black	Green	
Odd	9	9	0	18
Even	9	9	0	18
Zero	0	0	1	1
	18	18	1	37

Consider a spin of the wheel and assume it is fair, so that all numbers are equally likely

Let $A = \text{'Red'}$ and $B = \text{'Odd'}$.

$$\mathbb{P}(\text{Red}) = \mathbb{P}(A) = \frac{\text{Number of reds}}{\text{Number of outcomes}} = \frac{18}{37}$$

$$\mathbb{P}(\text{Odd}) = \mathbb{P}(B) = \frac{\text{Number of odds}}{\text{Number of outcomes}} = \frac{18}{37}$$

$$\begin{aligned}\mathbb{P}(\text{Red AND Odd}) &= \frac{\text{Number Red AND Odd}}{\text{Number of outcomes}} \\ &= \frac{9}{37}\end{aligned}$$

For ‘OR’ questions, it is usually best to use the Addition Law:

$$\begin{aligned}
\mathbb{P}(\text{Red OR Odd}) &= \mathbb{P}(\text{Red}) + \mathbb{P}(\text{Odd}) \\
&\quad - \mathbb{P}(\text{Red AND Odd}) \\
&= \frac{18}{37} + \frac{18}{37} - \frac{9}{37} \\
&= \frac{27}{37}
\end{aligned}$$

Example 3.6.

Two dice are thrown. It helps to suppose that one dice is **Red** while the other is **Blue**. Possible outcomes are:

		Blue					
		1	2	3	4	5	6
Red	1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
	2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
	3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
	4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
	5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
	6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

There are 36 possible outcomes – all equally likely.

$$\mathbb{P}(\text{Total} = 3) = \frac{2}{36}$$

$$\mathbb{P}(\text{Total} = 7) = \frac{6}{36}$$

$$\mathbb{P}(\text{Both odd}) = \frac{9}{36}$$

$$\mathbb{P}(\text{Red} = \text{Blue}) = \frac{6}{36}$$

$$\mathbb{P}(\text{Red} > \text{Blue}) = \frac{15}{36}$$

Example 3.7 (Two Dice continued). E be the event ‘Sum is 4’

F be the event ‘Dice show same even number’.

Find the probability that at least one of these events occurs.

$$\mathbb{P}(E) = \mathbb{P}(1, 3) + \mathbb{P}(2, 2) + \mathbb{P}(3, 1) = \frac{3}{36}$$

$$\mathbb{P}(F) = \mathbb{P}(2, 2) + \mathbb{P}(4, 4) + \mathbb{P}(6, 6) = \frac{3}{36}$$

These are not mutually exclusive because

$$\mathbb{P}(E \cap F) = \mathbb{P}(2, 2) = \frac{1}{36}$$

Using the addition rule of probability:

$$\begin{aligned}
\mathbb{P}(E \cup F) &= \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F) \\
&= \frac{3}{36} + \frac{3}{36} - \frac{1}{36} \\
&= \frac{5}{36}
\end{aligned}$$

Let G be the event ‘At least one 6’

$$\begin{aligned}\mathbb{P}(G) &= \mathbb{P}(\text{Red } 6) + \mathbb{P}(\text{Blue } 6) - \mathbb{P}(\text{Both } 6) \\ &= \frac{1}{6} + \frac{1}{6} - \frac{1}{36} \\ &= \frac{11}{36}\end{aligned}$$

3.3 Conditional Probability

We introduce the concept of the probability that an event occurs, **conditional on** another specified event occurring (or, in other language, given that another specified event occurs).

Example 3.8.

Consider the event that in a throw of a fair six-sided die we score 6, conditional on scoring more than 2. The event ‘scoring more than 2’ corresponds to the 4 equally-likely outcomes $\{3, 4, 5, 6\}$ and of these only 1 outcome corresponds to ‘score of 6’, so the probability required is $1/4$. Imposing the condition has effectively reduced/restricted the sample space from $\{1, 2, 3, 4, 5, 6\}$ to $\{3, 4, 5, 6\}$. Note that the conditional probability can be expressed as the ratio of two ‘unconditional’ probabilities of events defined in terms of the original sample space of size 6 by $\frac{1}{4} = \frac{1/6}{4/6}$.

Example 3.9.

Again, consider a throw of two fair six-sided dice, one red and the other blue. Let A be the event ‘score = 7 or 8’ and let B be the event ‘score = 8, 9 or 10’. Let

$$S = \{(i, j) : i = 1, 2, 3, 4, 5, 6; j = 1, 2, 3, 4, 5, 6\}$$

where i and j are the scores on the red and blue die respectively.

$$\begin{aligned}A &= \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1), (2, 6), (3, 5), (4, 4), (5, 3), (6, 2)\} & 11 \text{ elements} \\ B &= \{(2, 6), (3, 5), (4, 4), (5, 3), (6, 2), (3, 6), (4, 5), (5, 4), (6, 3), (4, 6), (5, 5), (6, 4)\} & 12 \text{ elements}\end{aligned}$$

So $\mathbb{P}(A) = 11/36$ and $\mathbb{P}(B) = 12/36$.

The event $A \cap B$ (A AND B) is the event ‘score of 8’ and $\mathbb{P}(A \cap B) = 5/36$.

Consider the event A conditional on B , that is ‘a score of 7 or 8 given that the score is 8, 9, or 10’. The outcomes in B favourable to A are (2,6), (3,5), (4,4), (5,3), (6,2) so the probability of event A conditional on B is $5/12$. This probability is $(5/36)/(12/36) = \mathbb{P}(A \cap B)/\mathbb{P}(B)$.

Using these examples we have a general definition of conditional probability.

Definition 3.1. The probability of event A conditional on event B is denoted $\mathbb{P}(A|B)$ and is defined as

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \text{ for } \mathbb{P}(B) \neq 0.$$

The **multiplication rule** for probabilities follows, namely $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B|A) = \mathbb{P}(B)\mathbb{P}(A|B)$.

Example 3.10.

Suppose we draw two balls at random, one after the other and without replacement, from a bag containing 6 red and 4 blue balls. Define the events A = first ball drawn is red, and let B = second ball drawn is blue. We are then interested in the event:

$$\mathbb{P}(\text{1}^{st} \text{ ball drawn is red and } \text{2}^{nd} \text{ ball drawn is blue}) = \mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B|A).$$

We then have $\mathbb{P}(A) = 6/10$ and $\mathbb{P}(B|A) = 4/9$ so

$$\mathbb{P}(A \cap B) = \frac{6}{10} \frac{4}{9} = \frac{4}{15}.$$

Chain rule: The multiplication rule introduced above can be extended to any finite number of events. More precisely, for any integer $n \geq 2$ and any events A_1, \dots, A_n we have

$$\mathbb{P}\left(\bigcap_{j=1}^n A_j\right) = \mathbb{P}(A_1) \prod_{k=2}^n \mathbb{P}\left(A_k \mid \bigcap_{j=1}^{k-1} A_j\right).$$

For example, when $n = 4$, we have

$$\mathbb{P}(A_4 \cap A_3 \cap A_2 \cap A_1) = \mathbb{P}(A_4|A_3 \cap A_2 \cap A_1) \cdot \mathbb{P}(A_3|A_2 \cap A_1) \cdot \mathbb{P}(A_2|A_1) \cdot \mathbb{P}(A_1).$$

We will now apply the chain rule in an example.

Example 3.11.

Ten bats out of a colony of 50 bats have been ringed. Five bats are caught at random from this colony.

Assume that each bat is equally likely to be caught. Then the probability that the first four bats caught are unmarked and the last one caught has a ring:

$$\begin{aligned} &= \frac{40}{50} \times \frac{39}{49} \times \frac{38}{48} \times \frac{37}{47} \times \frac{10}{46} \\ &= 0.08627 \end{aligned}$$

In a similar way, the probability could be calculated that only the first bat of five already had a ring. You should check that the probability comes out to be the same!

This is true for any other position in the sequence, so the probability that only one bat from five trapped at this colony already has a ring is:

$$5 \times 0.08627 = 0.431$$

Example 3.12.

Four playing cards, two red and two black. Two cards are chosen at random. What is the probability that they are the same colour?

Answer: Suppose that one card has been chosen. There are 3 possibilities for the second card. Only one of these is the same colour as the first card.

This can also be considered as the number of ways of choosing 2 items out of 4.

Example 3.13.

3 cards. One card is blank on both sides, one has **X** on both sides and the third has a side of each type. A card is selected at random and one side shown. If this is blank, what is the probability that the other side is blank?

Answer: There are 3 sides that are blank out of the 6 possible.

So $\mathbb{P}(\text{Blank side chosen}) = \frac{3}{6} = \frac{1}{2}$.

However, of these 3 sides, 2 have blank sides on the reverse.

So $\mathbb{P}(\text{Blank on reverse} \mid \text{Blank side chosen}) = \frac{2}{3}$.

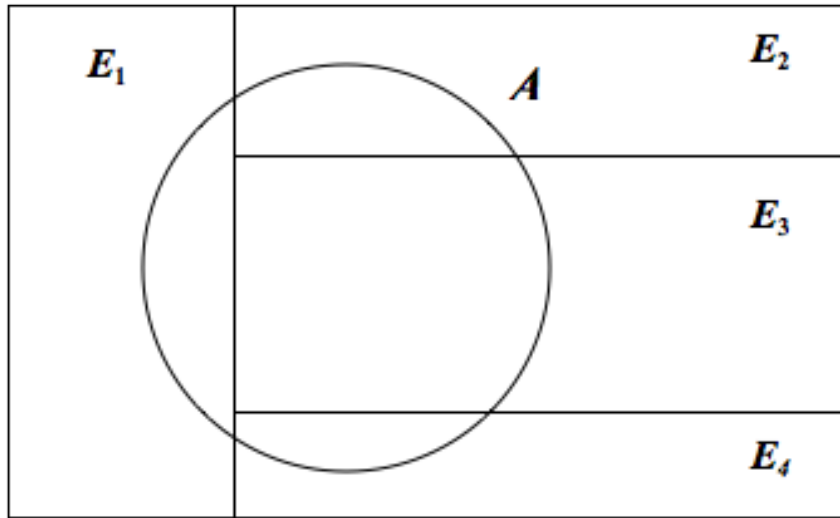
3.3.1 Partitioning of an event

Often it is useful to consider partitioning an event into smaller events.

Let $\{E_1, E_2, \dots, E_k\}$ be a partition of S and let A be an event. Then $A = A \cap S = \cup(A \cap E_i)$ so

$$\mathbb{P}(A) = \sum_{i=1}^k \mathbb{P}(A \cap E_i) = \sum_{i=1}^k \mathbb{P}(E_i) \mathbb{P}(A|E_i).$$

The event A has been partitioned into events $A \cap E_i$ for $i = 1, 2, \dots, k$. For example with $k = 4$:



3.3.2 Bayes Rules

If we apply the definition of conditional probability twice we find

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}.$$

This is called Bayes Rule or Bayes Theorem.

When using the partition of events as described in the previous section we have

$$\mathbb{P}(E_i|A) = \frac{\mathbb{P}(E_i)\mathbb{P}(A|E_i)}{\sum_{j=1}^k \mathbb{P}(E_j)\mathbb{P}(A|E_j)}.$$

Example 3.14.

Suppose a population is made up of 60% men and 40% women. The percentages of men and women in the population who have an iPhone are 30% and 40% respectively. A person is selected at random from the population and is found to have an iPhone. What is the probability that the selected person is male?

$$\mathbb{P}(\text{Selected is Male}) = \mathbb{P}(\text{Male} | \text{iPhone}) = \frac{\mathbb{P}(\text{iPhone} | \text{Male}) \mathbb{P}(\text{Male})}{\mathbb{P}(\text{iPhone})}.$$

So

$$\begin{aligned} \mathbb{P}(\text{Selected is Male}) &= \frac{\mathbb{P}(\text{iPhone} | \text{Male}) \mathbb{P}(\text{Male})}{\mathbb{P}(\text{iPhone} | \text{Male}) \mathbb{P}(\text{Male}) + \mathbb{P}(\text{iPhone} | \text{Female}) \mathbb{P}(\text{Female})} \\ &= \frac{0.6 \times 0.3}{0.6 \times 0.3 + 0.4 \times 0.4} = 0.5294. \end{aligned}$$

3.3.3 Independence

Events A and B are independent if and only if $\mathbb{P}(A \cap B) = P(A)P(B)$. This is equivalent to $\mathbb{P}(A|B) = \mathbb{P}(A)$ and $\mathbb{P}(B|A) = \mathbb{P}(B)$. So events A and B are independent if and only if the occurrence of one does not affect the probability of occurrence of the other.

Note that events A_1, A_2, \dots, A_k are (mutually) independent if and only if the probability of the intersection of any $2, 3, \dots, k$ of the events equals the product of their respective probabilities.

So, for three events A , B , and C to be independent, we require

- $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$
- $\mathbb{P}(A \cap C) = \mathbb{P}(A)\mathbb{P}(C)$
- $\mathbb{P}(B \cap C) = \mathbb{P}(B)\mathbb{P}(C)$
- $\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$

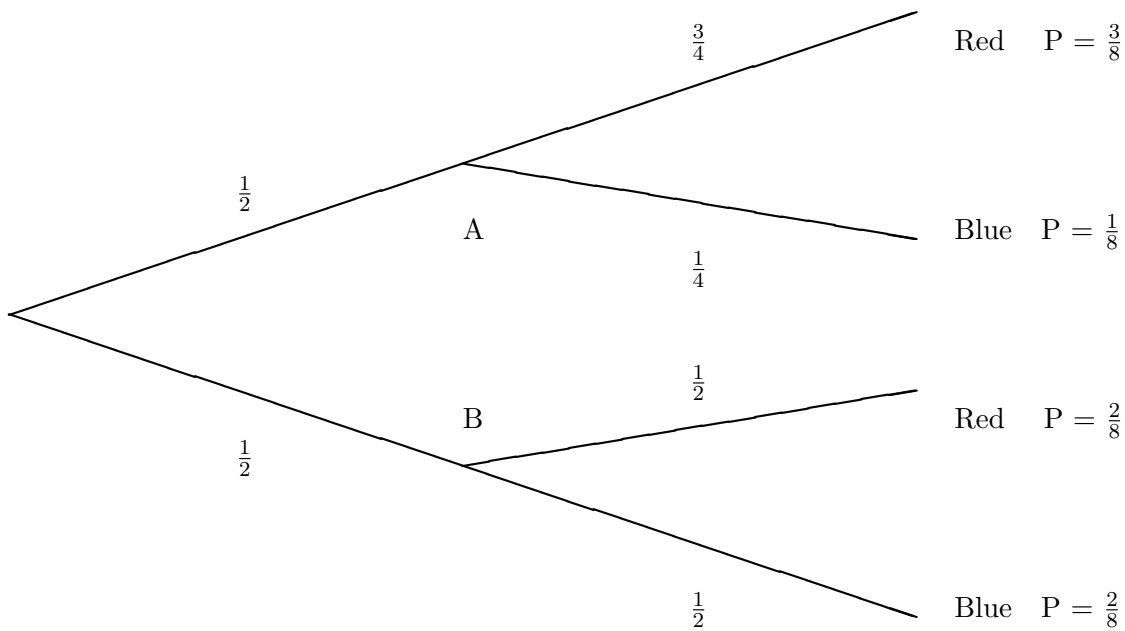
A **trial** is a single repetition of a random experiment. T_1 and T_2 are independent trials if and only if all events defined on the outcome of T_1 are independent of all events defined on the outcome of T_2 .

3.3.4 Tree Diagrams

Tree diagrams can be useful way to graphically represent small problems with nested probabilities.

Example 3.15.

Suppose there are 2 identical opaque bags A and B. Bag A contains 3 red counters and 1 blue counter. Bag B contains 2 red counters and 2 blue counters. One of the bags is chosen at random and a counter removed. If the selected counter is red, what is the probability that the chosen bag was bag A?



$$\mathbb{P}(\text{Bag A} \mid \text{Red counter}) = \frac{\frac{3}{8}}{\frac{3}{8} + \frac{2}{8}} = \frac{3}{5} = 0.6$$

3.4 Summary

Suppose E, F and C are events.

1. $0 \leq \mathbb{P}(E) \leq 1$

2. If outcomes are equally likely, then: $\mathbb{P}(E) = \frac{\text{Favourable outcomes}}{\text{Total outcomes}}$

3. $\mathbb{P}(E) = 1 - \mathbb{P}(E')$

4. $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F)$

5. $\mathbb{P}(E \mid C) = \frac{\mathbb{P}(E \cap C)}{\mathbb{P}(C)}$

$$\mathbb{P}(E \cap F) = \mathbb{P}(E) \times \mathbb{P}(F \mid E)$$

6. If $\mathbb{P}(E \mid C) = \mathbb{P}(E)$

then, E and C are independent events.

7. If $\mathbb{P}(E \text{ and } F) = \mathbb{P}(E) \times \mathbb{P}(F)$

then, E and F are independent events.

3.5 Further Probability Examples

Example 3.16 (Birthday Paradox).

How many people do you need in a room until the chance of a shared birthday is higher than getting a head on a flip of a fair coin?

Example 3.17 (Using an unfair coin). You are given an unfair coin which comes up heads with probability 0.7 and tails with probability 0.3. Can you use this coin to produce a fair result (i.e. the probability of heads being 0.5)?

Example 3.18. 5 cards are selected from a deck of 52 cards, 26 red and 26 black. Calculate the probability that we select at least one black card given we select the cards

- with replacement?
- without replacement?

Example 3.19. One half percent of the population has a particular disease. A test is developed for the disease. The test gives a false positive 3% of the time and a false negative 2% of the time.

1. What is the probability that Joe (a random person) tests positive?
2. Joe just got the bad news that the test came back positive; what is the probability that Joe has the disease?

Example 3.20 (Cystic Fibrosis). Cystic Fibrosis is the most serious of the human genetic diseases in the UK. 5% of population are carriers.

There is a diagnostic test which is positive with probability 0.85 for carriers and is always negative for non-carriers. If someone tests negative, what is the probability of being a carrier?

Example 3.21 (The Prosecutor's Fallacy). A large factory has 1000 male employees. An assault on a female worker takes place. DNA profiling of a suspect gives rise to a rare trait, X , such that

$$\mathbb{P}(X \text{ occurs in general population}) = \frac{1}{10000}$$

which is equivalent to

$$\mathbb{P}(\text{Match} \mid \text{Innocent}) = \frac{1}{10000}$$

A worker tests positive for the trait is this enough to convict him?

Example 3.22. The foundation of a wall can fail either by excessive settlement or from bearing capacity. Let event A be a failure caused by excessive settlement and B be a failure in bearing capacity with associated probabilities, $\mathbb{P}(A) = a$ and $\mathbb{P}(B) = b$. The probability of a failure in bearing capacity given that the foundation displays excessive settlement is $\mathbb{P}(B|A) = \beta_a$.

1. What is the probability of the foundation failing?
2. What is the probability that there is excessive settlement but no failure in bearing capacity?
3. What is the probability of excessive settlement given that the wall fails in bearing capacity?
4. For $a = 0.005$, $b = 0.002$ and $\beta_a = 0.2$, evaluate these probabilities.

Example 3.23. To get from Turin (Italy) to Grenoble (France) one of two routes can be used. The two options are either a direct route or via Chambéry (France). During heavy snow there is a chance that various parts of the route will be closed. We define the events of road closures as

- A the event the road Turin to Grenoble is open.
- B the event the road Turin to Chambéry is open.
- C the event the road Chambéry to Grenoble is open.

From past records we know that $\mathbb{P}(A) = 0.6$, $\mathbb{P}(B) = 0.7$, $\mathbb{P}(C) = 0.4$, $\mathbb{P}(C|B) = 0.5$, $\mathbb{P}(A|B \cap C) = 0.4$.

1. What is the probability that a traveller will be able to get from Turin to Grenoble?
2. What is the probability that a traveller will be able to get from Turin to Grenoble via Chambéry?

Example 3.24. Before accepting a new stretch of road the government tests the thickness of a 30cm stretch using an ultrasonic instrument to check for compliance. The whole length is accepted if the 30cm stretch passes this test. From past experience it is known that 85% of all constructed sections meet specifications but the ultrasound testing is only 75% reliable. So there is a 25% chance of erroneous conclusion based on the ultrasonic equipment.

1. What is the probability that poorly constructed section is accepted on the basis of the test?
2. What is the probability of a well constructed section is rejected based on the ultrasound test?

Example 3.25. We are interested in the life span of a reservoir. The lifetime of a reservoir can come to an end either by a flood which exceeds the spillway capacity or because excessive sedimentation makes the reservoir useless. We want to calculate the probability the reservoir will come to the end of its useful life in each year after its construction.

- The probability of a large flood in a year independent of previous years and of sedimentation is q .
 - Given the reservoir has not silted up prior to year i the probability it silts up in year i is $p_i = 1 - e^{-\beta i}$ with $\beta > 0$.
1. What is the probability the reservoir will survive n years?
 2. What is the probability the reservoir will come to an end in year n ?

F20SA / F21SA Statistical Modelling and Analysis

Chapter 4: Random Variables

Contents

4.1	Motivating Example	4–1
4.2	Definitions	4–2
4.2.1	Discrete Random Variables and Probability Mass Functions	4–2
4.2.2	Cumulative Distribution Function for Discrete Random Variables	4–3
4.2.3	Continuous Random Variables and Probability Density Function	4–4
4.2.4	Cumulative Distribution Function for Continuous Random Variables . .	4–4
4.3	Mean and Variance	4–5
4.3.1	Definition of Expectation	4–5
4.3.2	Sample mean and Expectation	4–6
4.3.3	Variance	4–7
4.3.4	Change of Origin and scale	4–8
4.4	Median and Quartiles	4–8
4.5	Examples	4–10

4.1 Motivating Example

Probability theory can be used to build models of the way in which data can arise.

Example 4.1. In a large population of oysters, 10% of the oysters contain a pearl. A random sample of oysters is opened until the first pearl is found. Let R be the oyster in which the first pearl is found.

R can take the values 1, 2, 3,

$$\mathbb{P}(R \text{ is } 1) = 0.1$$

$$\mathbb{P}(R \text{ is } 2) = 0.9 \times 0.1 = 0.09$$

[failure then success]

$$\mathbb{P}(R \text{ is } 3) = 0.9^2 \times 0.1 = 0.081$$

[2 failures then success]

$$\mathbb{P}(R \text{ is } 4) = 0.9^3 \times 0.1 = 0.0729$$

[3 failures then success]

and in general

$$\mathbb{P}(R \text{ is } r) = 0.9^{r-1} \times 0.1$$

This is an example of a discrete random variable.

4.2 Definitions

So far we have considered the probability of general events, for example the probability that it will rain tomorrow but in many circumstances we will be interested in a numerical variable whose value is uncertain, unpredictable or nondeterministic. Examples include the strength of concrete, the number of millimetres of rain in the next month, the number of days since the last component failure. We call them random variables, they assume a value which depends on the occurrence or outcome of a random experiment.

Formally a random variable is a map from the sample space S to a numerical value. We use the following conventions when dealing with random variables:

- Capital letters will be used for random variables: X, Y, \dots
- Lower case letters will be used for the possible values of random variables: x, y, \dots

We will be interested in two classes of random variables:

1. **Discrete Random Variables**, the random variable can only take a countable number of values (e.g. only integer values).
2. **Continuous Random Variables** (e.g. taking any real value).

In this chapter and the next we will provide a brief introduction to random variables and look at some common examples often used in modelling applications. This will only be a brief introduction to a much richer theory.

4.2.1 Discrete Random Variables and Probability Mass Functions

We start by considering discrete random variables. They can only take a countable number of values, for example taking only integer values $\{1, 2, 3, \dots\}$. This is the most common form. Examples of such random variables are the number of flaws in a concrete beam, the number of closed roads during a flood.

To describe these random variables we make use of a probability mass function which is often abbreviated to pmf.

Definition 4.1. The probability mass function of a discrete random variable X gives the probabilities of the values taken by X .

$$p_X(x) = \mathbb{P}(X = x)$$

Using our definition of probability from the previous chapter we know

- $0 \leq p_X(x) \leq 1$ for all possible x ,
- $p_X(x) = 0$ for all values of x which are unattainable,
- $\sum p_X(x) = 1$ where the sum is over all possible values of X ,
- $\mathbb{P}(X \in A) = \sum_{x \in A} p_X(x)$ for a set A of values.

Example 4.2. For the previous example of the number of oysters required to be opened until we found a pearl we had

$$\mathbb{P}(R \text{ is } r) = 0.9^{r-1} \times 0.1,$$

so the probability mass function for this random variable is

$$p_R(r) = 0.9^{r-1} \times 0.1.$$

Non-Examinable

We need to check that $\sum p_R(r) = 1$. We have

$$\sum_{r=1}^{\infty} p_R(r) = \sum_{r=1}^{\infty} 0.9^{r-1} \times 0.1 = 0.1 \times \sum_{r=0}^{\infty} 0.9^r = \frac{0.1}{1-0.9} = 1.$$

Example 4.3. Two dice are rolled and we are interested in two random variables, X the sum of the results and Y the maximum result. By counting the number of favourable outcomes in each case we obtain the probability mass functions for the two random variables. Firstly for X :

$$\begin{aligned} p_X(2) &= 1/36, & p_X(8) &= 5/36, \\ p_X(3) &= 2/36, & p_X(9) &= 4/36, \\ p_X(4) &= 3/36, & p_X(10) &= 3/36, \\ p_X(5) &= 4/36, & p_X(11) &= 2/36, \\ p_X(6) &= 5/36, & p_X(12) &= 1/36, \\ p_X(7) &= 6/36, \end{aligned}$$

and for Y we have:

$$\begin{aligned} p_Y(1) &= 1/36, \\ p_Y(2) &= 3/36, \\ p_Y(3) &= 5/36, \\ p_Y(4) &= 7/36, \\ p_Y(5) &= 9/36, \\ p_Y(6) &= 11/36. \end{aligned}$$

4.2.2 Cumulative Distribution Function for Discrete Random Variables

Rather than describing a discrete random variable by its pmf we may instead consider the cumulative distribution function which we abbreviate as cdf and denote by $F_X(x)$.

Definition 4.2. For a random variable X the cumulative distribution function is the probability that the random variable does not exceed value x . Therefore this is a non-decreasing function that is bounded below by 0 and above by 1. We denote

$$F_X(x) = \mathbb{P}(X \leq x)$$

and we have

$$0 \leq F_X(x) \leq 1.$$

For a discrete random variable this means that the distribution function is the sum of the probabilities of all possible values of X that are less or equal to x ,

$$F_X(x) = \sum_{y \leq x} \mathbb{P}(X = y) = \sum_{y \leq x} p_X(y).$$

Example 4.4. For the random variable R (number of oysters until first pearl found) we want to find the cdf,

$$F_R(r) = \mathbb{P}(R \leq r) = 1 - \mathbb{P}(R > r).$$

Note that $\mathbb{P}(R > r)$ is the probability that we did not find a pearl in first r oysters and hence

$$\mathbb{P}(R > r) = 0.9^r.$$

This gives

$$F_R(r) = \mathbb{P}(R \leq r) = 1 - 0.9^r.$$

Example 4.5. For the roll of two dice we consider the random variables X the sum of the results and Y the maximum result. The associated cdf for X is

$$\begin{aligned} F_X(2) &= 1/36, & F_X(8) &= 26/36, \\ F_X(3) &= 3/36, & F_X(9) &= 30/36, \\ F_X(4) &= 6/36, & F_X(10) &= 33/36, \\ F_X(5) &= 10/36, & F_X(11) &= 35/36, \\ F_X(6) &= 15/36, & F_X(12) &= 36/36, \\ F_X(7) &= 21/36, \end{aligned}$$

and for Y the cdf is

$$\begin{aligned} F_Y(1) &= 1/36, \\ F_Y(2) &= 4/36, \\ F_Y(3) &= 9/36, \\ F_Y(4) &= 16/36, \\ F_Y(5) &= 25/36, \\ F_Y(6) &= 36/36. \end{aligned}$$

4.2.3 Continuous Random Variables and Probability Density Function

We now move to consider continuous random variables. We need a different approach to specifying continuous random variables since the probability that it takes a single specific value is 0 so we are unable to make use a probability mass functions. Instead we use a probability density function which is abbreviated as pdf. The probability density function represents the intensity of probability or probability rate.

A density function has to obey the following properties

1. The pdf has to be non-negative,

$$f_X(x) \geq 0.$$

2. The integral over all possible values of X has to be one,

$$\int_{-\infty}^{\infty} f_X(x) dx = 1.$$

For a given pair of values x_1 and x_2 with $x_1 < x_2$ the area under the curve between the two values represents the probability that the random variable X will take a value within the interval $[x_1, x_2]$,

$$\mathbb{P}(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f_X(x) dx \leq 1.$$

4.2.4 Cumulative Distribution Function for Continuous Random Variables

The definition we gave for the cumulative distribution function F_X for discrete random variables is still valid when considering continuous random variables, i.e., $F_X(x)$ is the probability that the random variable X does not exceed value x . Hence we have

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^x f_X(x) dx,$$

from which it follows

$$\frac{dF_X(x)}{dx} = f_X(x).$$

Example 4.6. Consider the strength of wood samples in N/mm^2 , which is modelled by a random variable X that has probability density function:

$$f_X(x) = \begin{cases} \frac{x}{1400} & \text{for } 0 \leq x < 40 \\ \frac{70-x}{1050} & \text{for } 40 \leq x \leq 70 \\ 0 & \text{for } x < 0 \text{ and } 70 > x \end{cases}$$

So the cdf for X is

$$F_X(x) = \int_{-\infty}^x f_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ \frac{x^2}{2800} & \text{for } 0 \leq x < 40 \\ \frac{140x-x^2}{2100} - \frac{4}{3} & \text{for } 40 \leq x \leq 70 \\ 1 & \text{for } x > 70 \end{cases}.$$

If we were interested in finding the probability that the strength of an individual wood sample lies between 20 N/mm^2 and 50 N/mm^2 , we could either integrate the probability density function or take the difference between two values of the distribution function.

Method A:

$$\begin{aligned} \mathbb{P}(20 < X < 50) &= \int_{20}^{50} f_X(x) dx = \int_{20}^{40} \frac{x}{1400} dx + \int_{40}^{50} \frac{70-x}{1050} dx \\ &= \frac{1200}{2800} + \frac{500}{2100} = \frac{2}{3} \end{aligned}$$

Method B:

$$\mathbb{P}(20 < X < 50) = F_X(50) - F_X(20) = \frac{140 \times 50 - 50^2}{2100} - \frac{4}{3} - \frac{20^2}{2800} = \frac{2}{3}.$$

Example 4.7. The intensity of earthquakes is modelled by a random variable X which takes non-negative values ($X \geq 0$) and has probability density function:

$$f_X(x) = 0.2e^{-0.2x}, \text{ for } x \geq 0.$$

Hence the cdf for X is

$$F_X(x) = \int_0^x 0.2e^{-0.2y} dy = 1 - e^{-0.2x}.$$

We consider the distribution of the earthquake intensity given that the earthquake is larger than a given threshold t . Hence we are interested in

$$\mathbb{P}(X > s+t | X > t) = \frac{\mathbb{P}(X > s+t)}{\mathbb{P}(X > t)} = \frac{1 - F_X(s+t)}{1 - F_X(t)} = e^{-0.2(t+s)} / e^{-0.2t} = e^{-0.2s}.$$

4.3 Mean and Variance

4.3.1 Definition of Expectation

The expectation (or expected value) of the random variable X , denoted $\mathbb{E}[X]$, is given by

- $\mathbb{E}[X] = \sum xp_X(x)$ for X a discrete random variable,
- $\mathbb{E}[X] = \int xf_X(x)dx$ for X a continuous a random variable.

This is the mean of the random variable X and is often denoted by μ .

The expectation of $h(X)$, a function of the random variable X , denoted $\mathbb{E}[h(X)]$, is given by

- $\mathbb{E}[h(X)] = \sum h(x)p_X(x)$ for X a discrete random variable,
- $\mathbb{E}[h(X)] = \int h(x)f_X(x)dx$ for X a continuous a random variable.

Example 4.8. Consider the number of oysters R we need to open before we find a pearl. From a previous example we know that the probability generating function for R is

$$p_R(r) = 0.9^{r-1} \times 0.1.$$

Hence the mean is

$$\mathbb{E}(R) = 10.$$

Example 4.9. For the roll of two dice we consider the random variables: X the sum of the results and Y the maximum result. We find the mean of X is

$$\mathbb{E}(X) = \sum_{x=2}^{12} xp_X(x) = 7$$

and

$$\mathbb{E}(Y) = \sum_{y=1}^6 yp_Y(y) = 1 \times \frac{1}{36} + 2 \times \frac{3}{36} + \dots + 6 \times \frac{11}{36} = \frac{161}{36}.$$

Example 4.10. The strength of wood samples in N/mm^2 is described by the random variable X which has probability density function:

$$f_X(x) = \begin{cases} \frac{x}{1400} & \text{for } 0 \leq x < 40 \\ \frac{70-x}{1050} & \text{for } 40 \leq x \leq 70 \\ 0 & \text{for } x < 0 \text{ and } 70 > x \end{cases}$$

Non-Examinable

The mean strength is given by

$$\mathbb{E}(X) = \int_0^{70} xf_X(x)dx = \int_0^{40} x \frac{x}{1400} dx + \int_{40}^{70} x \frac{70-x}{1050} dx = \left[\frac{x^3}{1400 \times 3} \right]_0^{40} + \left[\frac{105x^2 - x^3}{1050 \times 3} \right]_{40}^{70} = 36\frac{2}{3}.$$

Example 4.11. The intensity of earthquakes is modelled by a random variable X which takes non-negative values ($X \geq 0$) and has probability density function:

$$f_X(x) = 0.2e^{-0.2x}.$$

The mean of the earthquake intensity

$$\mathbb{E}(X) = \int_0^{\infty} x0.2e^{-0.2x}dx = 5.$$

4.3.2 Sample mean and Expectation

For independent identically distributed (i.i.d) samples of a random variable X the sample mean and the mean of X , $\mathbb{E}[X]$, will be close for large sample sizes (here large is normally greater than 30).

Theorem 4.1. *Let X be a random variable with mean μ and X_1, X_2, X_3, \dots be independent identical distributed samples with the same distribution as X . Then the probability that the sample mean, \bar{X} , is more than a given distance from the mean of the random variable, $\mathbb{E}[X]$, tends to 0 as the number of samples tends to infinity, i.e., for any $\varepsilon > 0$ we have*

$$\mathbb{P} \left(\left| \frac{1}{n} \sum_{i=1}^n X_i - \mu \right| > \varepsilon \right) \rightarrow 0.$$

This theorem is related to the ideas that we saw in the previous chapter regarding the relationship between relative frequency and probability but now extended to the expectation of a random variable and sample mean.

4.3.3 Variance

As for summary statistics for real data, we are interested in the mean value but also a measure of variability. The same is true when we consider random variables, here we define the variance of a random variable which we denote $Var(X)$.

Definition 4.3. The variance of a random variable X is

$$Var(X) = \mathbb{E}((X - \mathbb{E}(X))^2) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2.$$

The standard deviation is the square root of the variance, $SD(X) = \sqrt{Var(X)}$.

Example 4.12. Consider the number of oyster we need to open before we find a pearl, R , from previously we have the probability generating function for this is

$$p_R(r) = 0.9^{r-1} \times 0.1$$

and

$$\mathbb{E}(R) = 10.$$

We find $\mathbb{E}(R^2)$

$$\mathbb{E}(R^2) = 190.$$

So

$$Var(X) = 190 - 10^2 = 90.$$

Example 4.13. For the roll of two dice we consider the random variables X the sum of the results and Y the maximum result. We start by finding the variance for X ,

$$\mathbb{E}(X^2) = 54\frac{5}{6}, \quad Var(X) = 54\frac{5}{6} - 49 = 5\frac{5}{6}.$$

Now we do the same for Y

$$\mathbb{E}(Y^2) = 21\frac{35}{36}, \quad Var(Y) = 1.97.$$

Example 4.14. The strength of wood samples in N/mm^2 are described by the random variable X which has probability density function:

$$f_X(x) = \begin{cases} \frac{x}{1400} & \text{for } 0 \leq x < 40 \\ \frac{70-x}{1050} & \text{for } 40 \leq x \leq 70 \\ 0 & \text{for } x < 0 \text{ and } 70 > x \end{cases}$$

So we want to find $\mathbb{E}(X^2)$ so as to be able to find the variance of X ,

Non-Examinable

$$\begin{aligned}\mathbb{E}(X^2) &= \int_0^{70} x^2 f_X(x) dx = \int_0^{40} x^2 \frac{x}{1400} dx + \int_{40}^{70} x^2 \frac{70-x}{1050} dx \\ &= \left[\frac{x^4}{1400 \times 4} \right]_0^{40} + \left[\frac{70x^3}{1050 \times 3} - \frac{x^4}{1050 \times 4} \right]_{40}^{70} = \frac{40^4}{1400 \times 4} + \frac{70 \times 70^3 - 70 \times 40^3}{1050 \times 3} + \frac{40^4 - 70^4}{1050 \times 4} \\ &= 1550\end{aligned}$$

So the variance of X is

$$\text{Var}(X) = 1550 - \left(36\frac{2}{3} \right)^2 = 205.56.$$

Example 4.15. The intensity of earthquakes are modelled by a random variable X which takes non-negative values ($0 \leq x$) and has probability density function:

$$f_X(x) = 0.2e^{-0.2x}.$$

We want to find the variance of X so we start by finding $\mathbb{E}(X^2)$,

$$\mathbb{E}(X^2) = \int_0^{\infty} x^2 0.2e^{-0.2x} dx = 50$$

So the variance is given by

$$\text{Var}(X) = 50 - 5^2 = 25.$$

4.3.4 Change of Origin and scale

As when considering the sample mean and sample standard it is often useful to consider the effect of linear transformations on the mean and variance of a random variable.

Consider the linear transformation $Y = a + bX$. We then have

$$\begin{aligned}\mathbb{E}[Y] &= \mathbb{E}[a + bX] = a + b\mathbb{E}[X] & \text{i.e. } \mu_Y &= a + b\mu_X \\ \text{Var}[Y] &= b^2 \text{Var}[X] & \text{i.e. } \sigma_Y^2 &= b^2 \sigma_X^2\end{aligned}$$

Example 4.16. The temperature of the water in a lake on summers day in degrees C is modelled by the random variable X with mean $\mathbb{E}(X) = 31$ and variance $\text{Var}(X) = 5$. If instead we wanted to model this in degrees F we know the conversion is carried out by multiply by 1.8 and add 32. Let Y be the random variable which describes the temperature in degrees F, we therefor have

$$\mathbb{E}(Y) = 1.8 \times 31 + 32 = 87.8$$

and variance

$$\text{Var}(Y) = 1.8^2 \times 5 = 16.2.$$

4.4 Median and Quartiles

When we considered measures of centre and spread for data we also considered the alternative measures median and interquartile range. We now look at the definition of median and interquartile range for random variables.

Previously for the median it was the value such that half the data lies above and half below, so when we move to random variables we are interested in the value such that the probability that X is less than the value is exactly a half. The quartiles are defined in a similar fashion. In the case of continuous random variables we make use of the cdf.

Definition 4.4. For a random variable X the median \tilde{x} is the value such that

$$F_X(\tilde{x}) = \mathbb{P}(X \leq \tilde{x}) = 1/2.$$

Similarly the lower quartile, x_1 , is the value such that

$$F_X(x_1) = \mathbb{P}(X \leq x_1) = 1/4,$$

and similar for the upper quartile, x_3 , we have

$$F_X(x_3) = \mathbb{P}(X \leq x_3) = 3/4.$$

Example 4.17. The intensity of earthquakes are modelled by a random variable X which takes non-negative values ($0 \leq x$) and has cdf:

$$F_X(x) = \int_0^x 0.2e^{-0.2y} dy = 1 - e^{-0.2x}.$$

So we have

$$\tilde{x} = -5\ln(0.5) = 3.47$$

$$x_1 = -5\ln(0.75) = 1.44$$

$$x_3 = -5\ln(0.25) = 6.93.$$

For a discrete random variable we consider the following example.

Example 4.18. For the random variable R (number of oyster until first pearl found) we have

$$F_R(r) = \mathbb{P}(R \leq r) = 1 - 0.9^r.$$

From the cdf we can find

$$\begin{array}{ll} F_R(6) = 0.469 & F_R(7) = 0.521 \\ F_R(2) = 0.19 & F_R(3) = 0.271 \\ F_R(13) = 0.746 & F_R(14) = 0.771 \end{array}$$

From this we can see that the $\tilde{x} = 7$, $x_1 = 3$ and $x_3 = 14$.

4.5 Examples

Example 4.19. Let X be the number of sixes which turn up in 4 throws of a fair six-sided die.

- What is the probability mass function?
- What is the mean and variance of X ?

Example 4.20. Consider the discrete random variable X with probability mass function

x	10	100	1000
$f_X(x)$	0.2	0.5	0.3

- What is the mean of X ?
- What is $\mathbb{E}(2X + 3)$?
- What is $\mathbb{E}(\log_{10}(X))$?

Example 4.21. Consider the continuous random variable X with pdf $f_X(x) = 3/x^4$ for $x > 1$.

- Does this obey the specifications of a pdf?
- What is the probability of the the random variable being between 1 and 2 in value?
- What is the cdf for X ?

Example 4.22. Suppose X has a distribution with constant density, that is $f_X(x) = 1$, $0 < x < 1$. Find the distributions of

1. $Y = X^2$
2. $Y = -\log(X)$.

F20SA / F21SA Statistical Modelling and Analysis

Chapter 5: Special Distributions

Contents

5.1	Expectation and Variance revisited	5-2
5.2	Specific Discrete Random Variables	5-2
5.2.1	Bernoulli Random Variables $Bernoulli(p)$	5-2
5.2.2	Uniform Random Variables $U(k)$	5-3
5.2.3	Binomial Random Variables, $Bin(n, p)$	5-3
5.2.4	Poisson Random Variables $Po(\lambda)$	5-6
5.2.5	Geometric Random Variables, $Geo(p)$	5-7
5.3	Specific Continuous Random Variables	5-7
5.3.1	Uniform Random Variables, $U(a, b)$	5-7
5.3.2	Exponential Random Variables $Exp(\lambda)$	5-8
5.3.3	Gamma Random Variables $Gamma(\alpha, \beta)$	5-9
5.3.4	Beta Random Variables $Beta(\alpha, \beta)$	5-9
5.3.5	Normal Random Variables, $N(\mu, \sigma^2)$	5-10
5.4	Examples of Random Variable Calculations.	5-12
5.5	Sampling Distributions and the Central limit Theorem	5-16
5.5.1	Sampling Distributions	5-16
5.5.2	Properties of the sample sum and sample mean	5-16
5.5.3	Central Limit Theorem (CLT)	5-18
5.5.4	Distribution of sample variance	5-19
5.6	Sampling from a normal variable	5-19
5.6.1	χ^2 , t and F distributions	5-20
5.6.2	Sampling distributions- mean and variance	5-21
5.7	Two samples	5-21
5.7.1	Difference between sample means, two independent samples	5-21
5.7.2	Ratio of two sample variances, independent normal samples	5-22
5.8	Examples: Sampling distributions and CLT	5-23

We study a series of distributions that have wide applicability - we give examples of situations in which the distributions are relevant and summarise some of their properties. Each of these is a parametric family of distributions, they have a set of parameters which can be varied so as to control their form. In later chapters we will explore methods for fitting these distributions through the selection of the parameters using data.

5.1 Expectation and Variance revisited

It is often useful to know how the expectation and variance behave when we add two random variables together.

Let X and Y be two random variables then

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y).$$

Furthermore we can extend the idea of independence to random variables. Previously when considering events we saw that events were independent if the outcome of one did not effect the likelihood of the other event. Similarly we say two random variables are independent if the value of one does not effect distribution of the other, i.e.

$$\mathbb{P}(X \leq x, Y \leq y) = \mathbb{P}(X \leq x)\mathbb{P}(Y \leq y).$$

We then have if X and Y are independent then

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y).$$

since

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y).$$

5.2 Specific Discrete Random Variables

We will start by considering 5 different families of discrete random variables:

- Bernoulli Random Variables
- Uniform Random Variables
- Binomial Random Variables
- Poisson Random Variables
- Geometric Random Variables

5.2.1 Bernoulli Random Variables *Bernoulli*(p)

The simplest discrete random variable which takes values $\{0, 1\}$ and has a single parameter $0 \leq p \leq 1$. X is the outcome of a single random trial with success probability p . X takes value 1 if the trial was successful and takes value 0 otherwise. So the probability mass function is

$$p_X(x) = \begin{cases} 1 - p & \text{if } x = 0 \\ p & \text{if } x = 1 \end{cases}.$$

$$\mu = \mathbb{E}(X) = p,$$

$$\mathbb{E}(X^2) = p,$$

$$\sigma^2 = \text{Var}(X) = p - p^2 = p(1 - p).$$

5.2.2 Uniform Random Variables $U(k)$

Uniform on $\{1, 2, \dots, k\}$: parameter k is a positive integer; X is the outcome in the situation in which all outcomes $1, 2, \dots, k$ are equally likely, with probability $1/k$.

$$p_X(x) = 1/k \quad x = 1, 2, \dots, k,$$

$$\mu = \mathbb{E}(X) = \sum_{x=1}^k \frac{1}{k} x = \frac{1}{k} \frac{1}{2} k(k+1) = \frac{k+1}{2},$$

$$\mathbb{E}(X^2) = \sum_{x=1}^k \frac{1}{k} x^2 = \frac{1}{k} \frac{1}{6} k(k+1)(2k+1) = \frac{(k+1)(2k+1)}{6},$$

$$\sigma^2 = \text{Var}(X) = \frac{1}{12}(k^2 - 1).$$

Example 5.1. Let X be the number showing face up when a fair six-sided die is thrown once. Then $X \sim U(6)$ with

$$p_X(x) = \frac{1}{6}$$

$$\mu = \mathbb{E}(X) = \frac{7}{2},$$

$$\sigma^2 = \text{Var}(X) = \frac{35}{12}.$$

5.2.3 Binomial Random Variables, $\text{Bin}(n, p)$

Permutations and The Binomial Coefficient

How many ways are there to arrange n different objects in order?

Example: 3 letters A, B, C

6 possibilities: ABC, ACB, BAC, BCA, CAB, CBA

There are 3 choices for the first position, 2 for the next and 1 for the last. So there are:

$3! = 3 \times 2 \times 1 = 6$ ways (called: 3 factorial)

Definition 5.1. n factorial is:

$n! = n \times (n-1) \times \dots \times 2 \times 1$ for all $n \geq 1$

$0! = 1$

More generally:

How many ways are there to arrange r objects in order, if they are chosen from n different objects?

There are n choices for first position, $(n-1)$ for the next position, down to $(n-r+1)$ for the last.

Thus there are $\frac{n!}{(n-r)!}$ ways.

Example: Can arrange 3 items chosen from 5 in 60 different ways.

Definition 5.2. The Binomial Coefficient is

$$\binom{n}{r} = \frac{n!}{r! \times (n-r)!}$$

and is the number of ways of choosing r items out of n .

The New Cambridge Statistical Tables tabulate values of this for $n \leq 30$.

Note: Some books use a different notation for the binomial coefficient. For example: C_r^n or $C(n, r)$.

One way to think of the Binomial coefficient is that it is the number of ways of permuting the original n objects divided by the number of ways of permuting both the r chosen objects and the $(n - r)$ objects that were not chosen.

Example: The number of ways of choosing two items out of five is:

$$\binom{5}{2} = \frac{5!}{2! \times 3!} = \frac{5 \times 4 \times 3 \times 2 \times 1}{(2 \times 1) \times (3 \times 2 \times 1)} = 10$$

Binomial Random Variable

$Bin(n, p)$: parameters n a positive integer and p , $0 < p < 1$; X is the number of successes in a sequence of n Bernoulli trials (i.e. n independent, identical trials) with $\mathbb{P}(\text{success}) = p$; Alternative notations $X \sim \text{Binomial}(n, p)$ or $X \sim Bi(n, p)$ or $X \sim B(n, p)$.

Let $\mathbb{P}(\text{failure}) = q = 1 - p$. The event $X = x$ occurs when x trials result in successes and $n - x$ trials result in failures; any such sequence has probability $p^x \times q^{n-x}$ and there are $\binom{n}{x}$ such sequences, so

$$\begin{aligned} p_X(x) &= \binom{n}{x} p^x \times q^{n-x}, x = 0, 1, \dots, n \\ \mu &= \mathbb{E}(X) = \sum_{x=1}^n \binom{n}{x} p^x \times q^{n-x} x = np, \\ \mathbb{E}(X^2) &= \sum_{x=1}^n \binom{n}{x} p^x \times q^{n-x} x^2 = n(n-1)p^2 + np \\ \sigma^2 &= \text{Var}(X) = np(1-p) = npq. \end{aligned}$$

The $Bin(n, p)$ distribution is positively skewed for $p < 0.5$, negatively skewed for $p > 0.5$, and symmetric for $p = 0.5$. The skewness increases as p tends to 0 or 1.

For us to use the binomial distribution as a model we require:

1. A series of trials with only two possible outcomes, success or failure.
2. The probability of success, p , is constant throughout the trials.
3. The number of trials n is fixed.
4. The outcomes of the trials are independent.
5. The random variable X is the total number of successes, with the order in which the events occur being irrelevant.

Example 5.2. A road is flooded with probability $p = 0.1$ during a year and not more than one flood occurs in a single year. What is the probability that there will be exactly 3 floods in a 5 year period? What is the probability that there is at least one flood in a 5 year period? What is the expected number of floods in a 5 year period?

Let X be the number of floods in a five year period. We assume that the occurrence of a flood in one year is independent of the other years, so $X \sim \text{Bin}(5, 0.1)$ and we want to find $\mathbb{P}(X = 3)$ and $\mathbb{P}(X > 0)$.

$$\mathbb{P}(X = 3) = \binom{5}{3} (0.1)^3 \times (0.9)^2 = 0.0081.$$

$$\mathbb{P}(X > 0) = 1 - \mathbb{P}(X = 0) = 1 - \binom{5}{0} (0.1)^0 \times (0.9)^5 = 0.41.$$

$$\mathbb{E}(X) = 5 \times 0.1 = 0.5.$$

Using tables

In many cases carrying out the calculations to find the probabilities for a binomial random variable can take a substantial amount of time. For example if we are interested in $X \sim \text{Bin}(20, 0.3)$ and want to find $\mathbb{P}(X \leq 15)$. Instead in New Cambridge Statistical Tables (NCST) the cumulative distribution for the binomial distribution has been tabulated for $n = \{2, 3, \dots, 20\}$ and $0 \leq p \leq 0.5$ by 0.01 increments. We can use these to calculate the probabilities of interest.

Example 5.3. The number of floods of a river during a 15 year period are modelled by $X \sim \text{Bin}(15, 0.3)$. We want to find:

- $\mathbb{P}(\text{exactly 7 floods})$
- $\mathbb{P}(\text{less than 4 floods})$
- $\mathbb{P}(\text{between 6 and 10 floods (inclusive)})$

We are interested in $n = 15$ which is tabulated on pages 14 and 15.

- $\mathbb{P}(\text{exactly 7 floods}) = \mathbb{P}(X = 7) = \mathbb{P}(X \leq 7) - \mathbb{P}(X \leq 6) = 0.9500 - 0.8689 = 0.0811$
- $\mathbb{P}(\text{less than 4 floods}) = \mathbb{P}(X < 4) = \mathbb{P}(X \leq 3) = 0.2969$
- $\mathbb{P}(\text{between 6 and 10 floods (inclusive)}) = \mathbb{P}(6 \leq X \leq 10) = \mathbb{P}(X \leq 10) - \mathbb{P}(X \leq 5) = 0.9993 - 0.7216 = 0.2777$

For $p > 0.5$ this situation is not immediately covered in the tables but we can think of the number of failures rather than the number of successes. Let X be the number of successes from n trials and Y be the number of failures, so we have

$$\mathbb{P}(X = x) = \mathbb{P}(Y = n - x),$$

as for x successes we need $n - x$ failures. As well if $X \sim \text{Bin}(n, p)$ we have $Y \sim \text{Bin}(n, 1 - p)$.

Example 5.4. The number of floods of a river during a 15 year period are modelled by $X \sim \text{Bin}(15, 0.7)$. We want to find:

- $\mathbb{P}(\text{exactly 7 floods})$
- $\mathbb{P}(\text{less than 4 floods})$
- $\mathbb{P}(\text{between 6 and 10 floods (inclusive)})$

We are interested in $n = 15$ which is tabulated on pages 14 and 15. We have $p > 0.5$ therefore to use the tables we will need to consider the number of years without floods $Y \sim \text{Bin}(15, 0.3)$.

- $\mathbb{P}(\text{exactly 7 floods}) = \mathbb{P}(X = 7) = \mathbb{P}(Y = 15 - 7)$

$$\mathbb{P}(Y = 8) = \mathbb{P}(Y \leq 8) - \mathbb{P}(Y \leq 7) = 0.9848 - 0.9500 = 0.0348$$

- $\mathbb{P}(\text{less than 4 floods}) = \mathbb{P}(X \leq 4) = \mathbb{P}(Y \geq 11)$

$$\mathbb{P}(Y \geq 11) = 1 - \mathbb{P}(Y \leq 10) = 1 - 0.9993 = 0.0007$$

- $\mathbb{P}(\text{between 6 and 10 floods (inclusive)}) = \mathbb{P}(6 \leq X \leq 10) = \mathbb{P}(5 \leq Y \leq 9)$

$$\mathbb{P}(5 \leq Y \leq 9) = \mathbb{P}(Y \leq 9) - \mathbb{P}(Y \leq 4) = 0.9963 - 0.5155 = 0.4808$$

5.2.4 Poisson Random Variables $Po(\lambda)$

X is the number of events which occur in a unit of time or space in the situation in which events occur ‘at random’ one after another through time or space with rate λ (the situation is more formally described as being a ‘Poisson process’ with intensity λ); Alternative notations $X \sim \text{Poisson}(\lambda)$ or $X \sim \text{Poi}(\lambda)$ or $X \sim P(\lambda)$.

$$p_X(x) = \frac{e^{-\lambda} \lambda^x}{x!} \quad x \geq 0$$

$$\mu = \mathbb{E}(X) = \sum_{x=0}^{\infty} \frac{e^{-\lambda} \lambda^x}{x!} x = \lambda,$$

$$\mathbb{E}(X^2) = \sum_{x=0}^{\infty} \frac{e^{-\lambda} \lambda^x}{x!} x^2 = \lambda^2 + \lambda$$

$$\sigma^2 = \text{Var}(X) = \lambda.$$

The $Po(\lambda)$ distribution is positively skewed, less strongly as λ increases.

Example 5.5. Industrial accidents in the plants of a large multinational company occur at random through time, one after the other and at a rate of 1 per week. Let X be the number of such accidents in a 4-week period. $\mathbb{E}(X) = 1 \times 4 = 4$. We model the number of accidents using $X \sim Po(4)$. For example we maybe interested in

$$\mathbb{P}(\text{At least one accident}) = \mathbb{P}(X > 0) = 1 - \mathbb{P}(X = 0) = 1 - e^{-4} = 1 - 0.0183 = 0.9817$$

or

$$\mathbb{P}(\text{Exactly 3 accidents}) = \mathbb{P}(X = 3) = \frac{e^{-4} 4^3}{3!} = 0.1954.$$

Using tables

As for the binomial distribution the cumulative distribution function it tabulated for the Poisson distribution in NCST (pp 25–32). In the tables μ is used instead of λ .

Example 5.6. Let $X \sim \text{Poi}(4.5)$, calculate

- $\mathbb{P}(X \leq 8)$
- $\mathbb{P}(X \geq 3)$
- $\mathbb{P}(2 \leq X \leq 7)$

Since we are interested in $X \sim \text{Poi}(4.5)$ we need to look at the table on pages 26 and 27.

- $\mathbb{P}(X \leq 8) = 0.9597$
- $\mathbb{P}(X \geq 3) = 1 - \mathbb{P}(X \leq 2) = 1 - 0.1736 = 0.8264$
- $\mathbb{P}(2 \leq X \leq 7) = \mathbb{P}(X \leq 7) - \mathbb{P}(X \leq 1) = 0.9134 - 0.0611 = 0.8523$

5.2.5 Geometric Random Variables, $\text{Geo}(p)$

X is the number of trials until the first success occurs in a sequence of Bernoulli trials with $\mathbb{P}(\text{success}) = p$. So we require $0 \leq p \leq 1$. It is a discrete ‘waiting time distribution’ in the sense ‘how long do we have to wait to get a success?’ We will set $q = 1 - p$ as before. The event $X = x$ occurs when the first $x - 1$ trials result in failures and the next trial results in success; the probability of this sequence of outcomes occurring gives the probability mass function:

$$p_X(x) = p(1 - p)^{x-1} \quad x > 0$$

$$\mu = \mathbb{E}(X) = \frac{1}{p},$$

$$\sigma^2 = \text{Var}(X) = \frac{(1 - p)}{p^2},$$

$$F(x) = \mathbb{P}(X \leq x) = 1 - (1 - p)^x.$$

Example 5.7. A company produces treatment plants and delivers them to rural communities. From previous experience they know that 8% of the plants are faulty. If they deliver a faulty plant they send a replacement plant. What is the probability they have to deliver more than 2 plants for a single order?

The number of plants that have to be delivered until a non faulty unit is delivered is X and has a geometric distribution with $p = 0.92$. We want

$$\mathbb{P}(X > 2) = 1 - \mathbb{P}(X = 1) - \mathbb{P}(X = 2) = 1 - 0.92 - 0.08 \times 0.92 = 0.0064.$$

Example 5.8. Alice plays a game, she rolls a die until she rolls either a five or a six. She receives a pound for each time she has had to roll the die. How much should she pay to play the game?

The number of rolls X has a geometric distribution with $p = 1/3$. So we have $\mathbb{E}(X) = 1/p = 3$, this is the expected number of rolls until a five or six is rolled. So Alice’s expected return is 3 pounds and so this is the amount she should pay to play the game.

5.3 Specific Continuous Random Variables

We now move on to consider continuous random variables and the standard families of distributions which are used both in modelling and to help in making statistical decisions.

5.3.1 Uniform Random Variables, $U(a, b)$

X is the position of a point chosen ‘at random’ in the interval (a, b) ; all outcomes in the interval are ‘equally likely’ (i.e. events defined by subintervals within (a, b) of the same length have the same probability). The probability density function is constant (graph is flat).

$$f_x(x) = \frac{1}{b - a}, \quad a < x < b$$

$$F_X(x) = \mathbb{P}(X \leq x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & x > b \end{cases}$$

$$\mu = \mathbb{E}(X) = \int_a^b x \frac{1}{b-a} dx = \frac{1}{b-a} \frac{1}{2} (b^2 - a^2) = \frac{a+b}{2}.$$

$$\mathbb{E}(X^2) = \int_a^b x^2 \frac{1}{b-a} dx = \frac{(b^3 - a^3)}{3(b-a)}$$

$$\sigma^2 = \text{Var}(X) = \frac{1}{12} (b-a)^2.$$

Example 5.9. A unit length stick is broken into two parts with the location of the break being uniformly distributed along its length. We want to find out the probability that the longest section after the break is longer than 0.8 in length. We are interested in

$$\mathbb{P}(\text{longest section longer than } 0.8) = \mathbb{P}(X < 0.2 \text{ or } X > 0.8).$$

The two events $\{X < 0.2\}$ and $\{X > 0.8\}$ are mutually exclusive so

$$\mathbb{P}(X < 0.2 \text{ or } X > 0.8) = \mathbb{P}(X < 0.2) + \mathbb{P}(X > 0.8) = 0.2 + 0.2 = 0.4.$$

5.3.2 Exponential Random Variables $\text{Exp}(\lambda)$

X is the waiting time between consecutive events in the situation in which events occur ‘at random’ one after another through time with rate λ .

The probability density function is

$$f_X(x) = \lambda e^{-\lambda x}, \quad x > 0$$

and the cumulative distribution function is

$$F_X(x) = \int_0^x \lambda e^{-\lambda u} du = 1 - e^{-\lambda x}.$$

So

$$\mu = \mathbb{E}(X) = \int_0^\infty x \lambda e^{-\lambda x} dx = \frac{1}{\lambda}.$$

$$\mathbb{E}(X^2) = \int_0^\infty x^2 \lambda e^{-\lambda x} dx = \frac{2}{\lambda^2}$$

$$\sigma^2 = \text{Var}(X) = \frac{1}{\lambda^2}.$$

The $\text{Exp}(\lambda)$ distribution is positively skewed.

Example 5.10. The length of times between buses are modelled by an exponential random variable with mean 20.

1. What is the probability that the time between buses is longer than 20 minutes?
2. A passenger sees a bus leave from outside his window and goes to make a cup of tea which takes 5 minutes and then goes to the stop and is willing to only wait 15 minutes. What is the probability he will catch the next bus?
3. Given a passenger waits 25 minutes for a bus, what is the probability density function for the extra time they have to wait for the next bus?

Let X be the length of time until the next bus arrives so we have $X \sim \text{Exp}(1/20)$.

1.

$$\mathbb{P}(X > 20) = 1 - F_X(20) = e^{-\frac{20}{20}} = 0.368$$

2.

$$\mathbb{P}(5 \leq X \leq 20) = F_X(20) - F_X(5) = e^{-\frac{5}{20}} - e^{-\frac{20}{20}} = 0.411$$

3. We start by finding the cumulative distribution function, let Y be the remaining time the passenger has to wait.

$$\mathbb{P}(Y \leq y) = 1 - \mathbb{P}(Y > y) = 1 - \mathbb{P}(X > y + 25 | X > 25)$$

So we have

$$\mathbb{P}(X > y + 25 | X > 25) = \frac{\mathbb{P}(X > y + 25)}{\mathbb{P}(X > 25)} = e^{-\frac{(y+25)}{20}} e^{\frac{25}{20}} = e^{-\frac{y}{20}}.$$

Therefore the cumulative distribution function is

$$F_Y(y) = 1 - e^{-\frac{y}{20}},$$

and so probability density function

$$p_Y(y) = \frac{1}{20} e^{-\frac{y}{20}}.$$

So $Y \sim \text{Exp}(\frac{1}{20})$.

5.3.3 Gamma Random Variables $\text{Gamma}(\alpha, \beta)$

A gamma random variables X has probability density function

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \quad x \geq 0,$$

where the parameters $\alpha > 0$ and $\beta > 0$ control the expectation and the variance of X , and where Γ is the so called gamma function. The expectation and variance of X are given by

$$\mathbb{E}(X) = \frac{\alpha}{\beta},$$

and

$$\text{Var}(X) = \frac{\alpha}{\beta^2}.$$

This distribution is widely used to model positive quantities, particularly in the context of Bayesian inference. Notice that it coincides with the exponential distribution when $\alpha = 1$.

5.3.4 Beta Random Variables $\text{Beta}(\alpha, \beta)$

A beta random variables X takes values in the interval $[0, 1]$ with probability density function

$$f_X(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} \quad 0 \leq x \leq 1,$$

where the parameters $\alpha > 0$ and $\beta > 0$ control the expectation and the variance of X , and where $B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha + \beta)$ is the so called Beta function. The expectation and variance of X are given by

$$\mathbb{E}(X) = \frac{\alpha}{\alpha + \beta},$$

and

$$\text{Var}(X) = \frac{\alpha\beta}{(\alpha + \beta + 1)(\alpha + \beta)^2}.$$

This distribution is widely used to model probabilities, particularly in the context of Bayesian inference. Notice that this distribution coincides with the uniform distribution when $\alpha = \beta = 1$.

5.3.5 Normal Random Variables, $N(\mu, \sigma^2)$

Also called the Gaussian distribution and is important in statistical theory and practice. It is also a good empirical model for some kinds of physical data and provides an approximation to other distributions. As well it models distributions of certain sample statistics, in particular the sample mean and sample proportion and so is the basis of much of statistical methodology. We confirm that the parameters do indeed represent the mean and standard deviation of the distribution (as suggested by the choice of symbols).

The probability density function is

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \quad \text{for } -\infty < x < \infty.$$

It is important to note that the probability density function is symmetric around the μ .

So we have the mean of a $N(\mu, \sigma^2)$ is μ and the variance is σ^2 .

Linear transformation and the standard normal

Let $X \sim N(\mu, \sigma^2)$, then the linear transformation of X , $Y = a + bX$ has distribution $Y \sim N(a + b\mu, b^2\sigma^2)$.

The $N(0, 1)$ distribution is called the standard normal distribution.

Let $X \sim N(\mu, \sigma^2)$ and let $Z = \frac{X-\mu}{\sigma}$. This transformation is called standardising and $Z \sim N(0, 1)$ is a standard normal random variable.

For $Z \sim N(0, 1)$ the associated pdf is

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}.$$

We cannot find an explicit expression for $F_X(x)$ where $X \sim N(\mu, \sigma^2)$. To find probabilities associated with the distribution of X , we standardise the variable and find the probability using published tables of $P(Z \leq z)$ for $Z \sim N(0, 1)$ and $z > 0$ (pages 34 and 35 in NCST). In some cases we will need to make use of the fact that the normal distribution is symmetric, so for $z < 0$ we have $P(Z \leq z) = P(Z \geq -z)$.

Example 5.11. The diameter of bolts measured in mm are modelled by X which has a normal distribution with mean 20 and variance 4.

1. What is the probability that the bolts diameter is smaller than 22mm?
2. What is the probability that the bolts diameter is larger than 19mm?
3. For a bolt to fit in the associated nut the diameter has to lie between 18mm and 21mm, what proportion of bolts meet this specification?

We have $X \sim N(20, 4)$.

- 1.

$$\mathbb{P}(X < 22) = \mathbb{P}(Z < (22 - 20)/\sqrt{4}) = \mathbb{P}(Z < 1) = 0.8413$$

2.

$$\mathbb{P}(X > 19) = \mathbb{P}(Z > (19 - 20)/\sqrt{4}) = \mathbb{P}(Z > -1/2) = 1 - \mathbb{P}(Z < -1/2) = 0.6915$$

3.

$$\mathbb{P}(18 < X < 21) = \mathbb{P}(-1 < Z < 1/2) = \mathbb{P}(Z < 1/2) - \mathbb{P}(Z < -1) = 0.5328.$$

Similarly, if we want to find the x such that $\mathbb{P}(X > x) = \alpha$ with $X \sim N(\mu, \sigma^2)$ we start by finding the equivalent value for the standard normal Z and then using the reverse transformation to find the equivalent value for X .

Example 5.12. The diameter of bolts measured in mm are modelled by X which has a normal distribution with mean 20 and variance 4.

1. What is the diameter such that only 1% of the population are larger?
2. What is the diameter such that only 5% of the population are smaller?

We use the tables to find the appropriate values for the standard normal and reverse the standardisation transformation.

1. We have $\mathbb{P}(Z > 2.326) = 0.01$ so we now reverse the transformation to get $x = 2.326 * 2 + 20 = 24.652$,

$$\mathbb{P}(X > 24.652) = 0.01.$$

2. We have $\mathbb{P}(Z > 1.645) = 0.05$ giving $\mathbb{P}(Z < -1.645) = 0.05$ so the reverse transformation gives $x = 20 - 2 * 1.645$,

$$\mathbb{P}(X < 16.71) = 0.05.$$

5.4 Examples of Random Variable Calculations.

Example 5.13. Consider a multiple choice exam with 9 questions for each of which 4 answers are given. Let X be the number of correct answers obtained by a candidate who simply guesses (i.e. who selects the answer to each question at random from the 4 answers provided).

- What is the probability they get three correct?
- What is the probability they get at most 3 wrong?

Example 5.14. Let X be the number of times a pair of fair dice have to be thrown until a double six comes up.

- What is the mean and the variance of X ?
- What is the probability it will take more than 10 throws to get a double 6?

Example 5.15. A drug causes serious side effects in approximately 0.1% of users. Consider a group of 2000 users of the drug. Let X be the number of people in the group who suffer serious side effects. The appropriate model for distribution of X is $X \sim \text{bin}(2000, 0.001)$.

- What is the probability that either no-one or one person has serious side effects?
- What is approximately the probability that more than 4 people have serious side effects?

Example 5.16. A computer simulation program selects six numbers independently and at random from the interval $(0,5)$. Find the probabilities that, of the numbers selected:

- three are less than 2.5 and three are greater than 2.5 ,
- four are less than 2 and two are greater than 2.

Example 5.17. Suppose the length of times between failures of generators at a dam have an exponential distribution. The average time, as measured by the median, is 624 minutes.

- What is the average times as measured by the mean?
- What percentage of times are greater than 1000 minutes ?

Example 5.18. Suppose that the sizes of tree trunks can be modelled by a normal distribution with mean $\mu = 600$ and standard deviation $\sigma = 90$. The size of a particular trunk is known to be greater than 510. Find the probabilities that the size of this tree trunk is

- greater than 600,
- between 510 and 630.

Example 5.19. A proposed bridge across a stream is supported at the two ends and a central pier. The design allows for relative settlement of the foundations but these need to be kept within limits. We assume that the settlements are assumed to be normally distributed and by comparison to similar sites the means are ascertained to be 3cm on each end and 5cm at the central pier and the standard deviations to be 1cm at each end and 1.5cm for the central pier. We assume the settlements are independent.

- What is the probability that the maximum settlement is in excess of 7.5cm?
- Specify the maximum relative settlement of the center pier for which the engineer should design on the basis that this will be exceeded with probability of 0.0001. Assume we can ignore the settlement at the ends.

5.5 Sampling Distributions and the Central limit Theorem

5.5.1 Sampling Distributions

Let $X_i, i = 1, 2, \dots, n$ be independent identically distributed (i.i.d) random variables each with the same distribution as a random variable X with mean μ and variance σ^2 . Then $\mathbf{X} = (X_1, X_2, \dots, X_n)$ is a **random sample of (or from) the population variable X** .

A function of a random sample which does not involve unknown parameters is called a **statistic**. A statistic is a random variable. The value assumed by a statistic for any particular sample can be calculated from the sample data, and the values of the statistic vary from sample to sample. The distribution of a statistic is called a **sampling distribution**- its properties depend on those of the population variable X and on the sample size. Important examples of statistics are the **sample mean** and the **sample variance**. Other examples are the **sample median**, the **sample maximum**, and the **sample range**.

Reminder:

Sample Mean: $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$.

Sample Variance: $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2 / n \right)$.

5.5.2 Properties of the sample sum and sample mean

$$\mathbb{E}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathbb{E}(X_i) = n\mu.$$

$$Var\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n Var(X_i) = n\sigma^2 \text{ by independence.}$$

So for the sample mean we have

$$\mathbb{E}(\bar{X}) = \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} n\mu = \mu.$$

$$Var(\bar{X}) = Var\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} Var\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}.$$

We note that the expected value of the sample mean is the population mean, and the variance of the sample mean is the population variance divided by the sample size.

$SD(\bar{X}) = \sigma/\sqrt{n}$ is called the **standard error** of the sample mean, is denoted $s.e.(\bar{X})$ and is a measure of the precision of the sample mean as an estimate of the population mean - the smaller $s.e.(\bar{X})$ is, the better. As n increases, $s.e.(\bar{X})$ decreases, so larger samples provide more precise estimates of population parameters.

Example 5.20. Consider $X \sim Exp(1)$, hence we have the mean $\mu = 1$ and the variance $\sigma^2 = 1$. In Figure 5.1 the first plot shows 200 observations from the population; the second shows the means of 200 samples of size 2 from the population; the third shows the means of 200 samples of size 30 from the population; the fourth shows the means of 200 samples of size 1000 from the population.

The summary statistics for the four data sets are:

Population data	mean = 1.098	s.d.=0.919	min 0.006	max 4.103
Means of samples of size 2	mean = 0.957	s.d.=0.681	min 0.0424	max 3.737
Means of samples of size 30	mean = 1.008	s.d.=0.187	min 0.655	max 1.628
Means of samples of size 100	mean = 0.997	s.d.=0.098	min 0.727	max 1.259

The four data sets are centred on the same value (the population mean 1) and the spread of the means decreases as the sample size on which the means are based increases. Further, the last two data sets all look as though they could come from normal distributions.

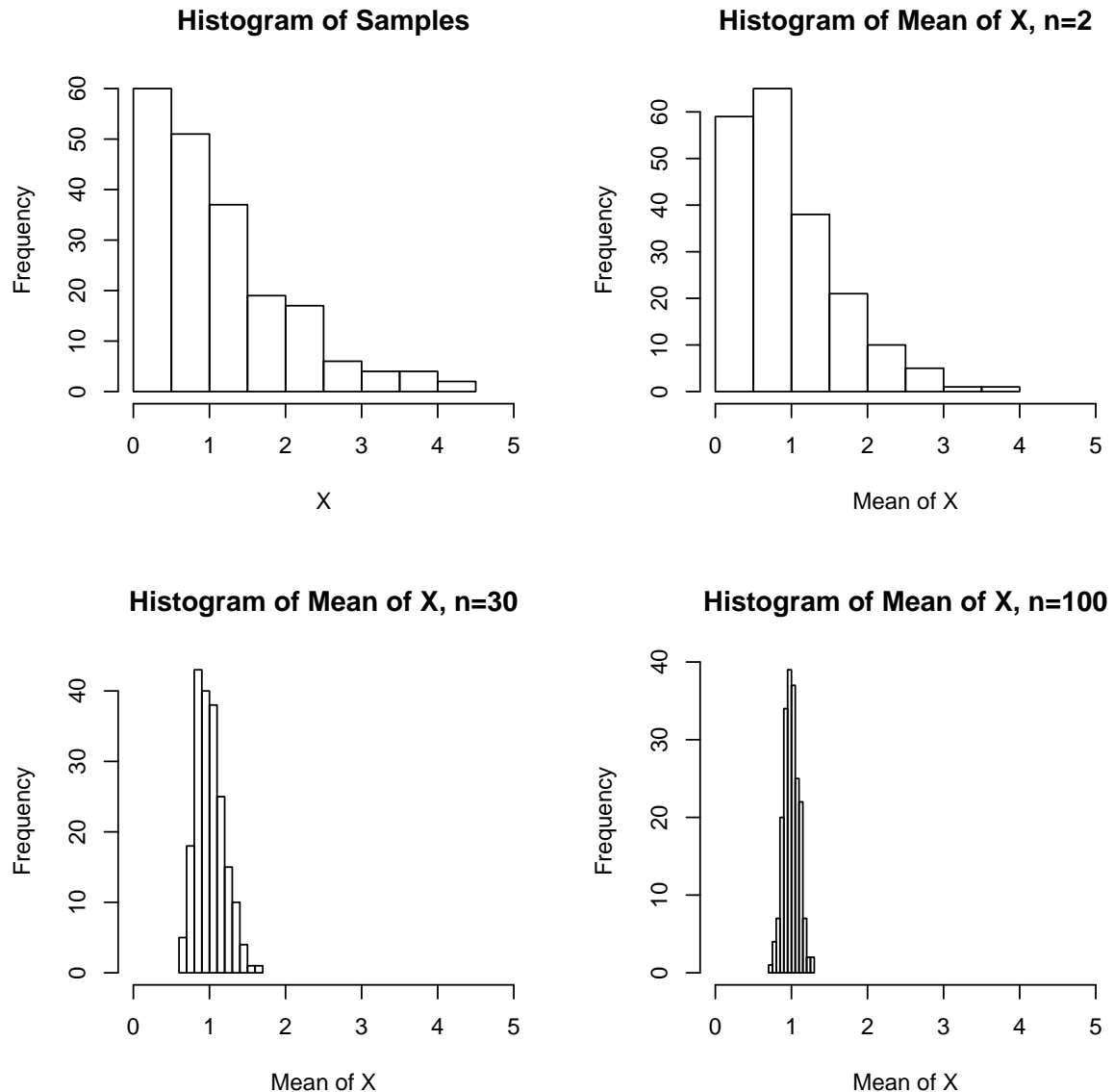


Figure 5.1: Histograms of sample means

Note: be careful not to confuse the sum of n i.i.d. random variables with n times any one of them.

For example: $n = 2$

Let X_1 and X_2 be i.i.d. with mean μ and variance σ^2 .

$\mathbb{E}(2X_1) = 2\mu$ and $\mathbb{E}(X_1 + X_2) = 2\mu$ but $\text{Var}(2X_1) = 4\sigma^2$ while $\text{Var}(X_1 + X_2) = 2\sigma^2$.

The random variables $2X_1$ and $(X_1 + X_2)$ have the same mean but have different variances.

5.5.3 Central Limit Theorem (CLT)

The CLT is a cornerstone of statistical theory and practice - it is hard to overstate its importance. Informally, it states that sample means become ‘more and more normally distributed’ as the size of the sample on which the means are based increases, almost regardless of the form of the distribution of the population from which the sample is drawn.

Theorem 5.1. *Let $\mathbf{X} = (X_1, X_2, \dots, X_n)$ be a random sample of a population variable X with mean $\mu < \infty$ and variance $\sigma^2 < \infty$ and let \bar{X} be the sample mean. Then*

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \rightarrow N(0, 1) \text{ as } n \rightarrow \infty.$$

(that is, the limiting distribution of the standardised sample mean, as the sample size tends to infinity, is the standard normal distribution).

From this we have the asymptotic (‘large sample’) distribution of \bar{X} :

$$\text{for large } n, \bar{X} \sim N(\mu, \sigma^2/n) \text{ approximately}$$

Equivalently, we have the asymptotic (‘large sample’) distribution of the sample sum:

$$\text{for large } n, \sum X_i \sim N(n\mu, n\sigma^2) \text{ approximately}$$

This result is true regardless of the nature of the population variable X (provided only it has finite mean and variance). We saw this in the previous example before when examining the sample means with sample sizes 30 or 100.

Warning! The CLT is sometimes misunderstood. It concerns the distribution of sample means/sums, not of individual values from a population. A large sample from a skewed population will be a skewed sample.

The statement ‘we have a lot of data, so it will be approximately normally distributed’ is nonsense.

The statement ‘we have a large sample, so the mean (or sum) of the sample is approximately normally distributed’ is correct.

The CLT in practice - important special cases

1. The normal approximation to the binomial distribution

$X \sim \text{Bin}(n, p)$ is the sum of n i.i.d. Bernoulli random variables with success probability p . Hence the CLT applies:

$$\text{for large } n, X \sim N(np, np(1-p)) \text{ approximately}$$

2. The normal approximation to the distribution of a sample proportion

Let P be the proportion of successes in a series of n i.i.d. trials, i.e., $P = X/n$ where $X \sim \text{Bin}(n, p)$. The random variable P has mean p and variance

$$\frac{p(1-p)}{n}.$$

The CLT applies:

$$\text{for large } n, P \sim N\left(p, \frac{p(1-p)}{n}\right) \text{ approximately}$$

3. The normal approximation to the Poisson distribution

Let $X \sim Po(\lambda)$, then for a positive integer λ , the random variable X is the sum of λ i.i.d. random variables each distributed $Po(1)$. So the CLT applies:

for large λ , $X \sim N(\lambda, \lambda)$ approximately

Continuity Correction

When we use a normal approximation to calculate a probability associated with a discrete random variable we are using a continuous distribution to approximate a discrete one. We are effectively superimposing a continuous probability density function curve over a set of probability mass function rectangles and matching the areas - this is best done if we think of the rectangle for, say $X = 16$, as covering the interval from $X = 15.5$ to $X = 16.5$ and we use this interval when using the approximating distribution - this improves the approximation.

So, for example, for $X \sim Bin(200, 0.4)$, we approximate the probability $\mathbb{P}(X \geq 84)$ by calculating $\mathbb{P}(Y > 83.5)$, and $\mathbb{P}(73 \leq X \leq 85)$ by calculating $\mathbb{P}(72.5 \leq Y \leq 85.5)$, where $Y \sim N(80, 48)$.

For $X \sim Po(120)$, we approximate $\mathbb{P}(X \geq 110)$ by calculating $\mathbb{P}(Y \geq 109.5)$ where $Y \sim N(120, 120)$.

5.5.4 Distribution of sample variance

Let $X_i, i = 1, 2, \dots, n$ be independent identically distributed (i.i.d) random variables each with the same distribution as a random variable X with mean μ and variance σ^2 .

The sample variance is defined to be

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

We now want to find the mean of S^2 . To this end, observe that we have

$$\sum (X_i - \mu)^2 = \sum (X_i - \bar{X} + \bar{X} - \mu)^2 = \sum (X_i - \bar{X})^2 + \sum (\bar{X} - \mu)^2 + 2(\bar{X} - \mu) \sum (X_i - \bar{X}).$$

Note that $\sum (X_i - \bar{X}) = 0$ and hence

$$\sum (X_i - \mu)^2 = \sum (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2.$$

Taking expectations on both sides, we obtain

$$\begin{aligned} nVar(X) &= \mathbb{E}((n-1)S^2) + nVar(\bar{X}) \\ n\sigma^2 &= (n-1)\mathbb{E}(S^2) + n(\sigma^2/n) \end{aligned}$$

This implies

$$\mathbb{E}(S^2) = \sigma^2.$$

We note that the expected value of the sample variance is the population variance.

5.6 Sampling from a normal variable

In the special case that the samples are independent identically distributed from a normal distribution we are able to obtain exact results for the sampling distributions of the sample mean and sample variance.

A useful property of the normal distribution is that the sum of two independent normal random variables has a normal distribution. Let $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ and X and Y be independent, then $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

5.6.1 χ^2 , t and F distributions

Before we can state the results about sampling distributions we need to define three further family of distributions.

Chi-squared, χ_n^2

The parameter n is a positive integer. A χ_n^2 random variable is the sum of the squares of n independent $N(0, 1)$ random variables. Namely, let Z_i be i.i.d. $N(0, 1)$. Then

$$X = \sum_{i=1}^n Z_i^2$$

is such that $X \sim \chi_n^2$. The cdf is given in NCST Table 7 pp37-39; percentage points (quantiles) are given in Table 8 pp40-41.

Example 5.21. Let $X \sim \chi_3^2$ and $Y \sim \chi_9^2$, then

$$\mathbb{P}(X < 10) = 0.9814 \quad \mathbb{P}(Y < 10) = 0.6495,$$

$$\mathbb{P}(X < 7.815) = 0.95 \quad \mathbb{P}(Y < 21.67) = 0.99.$$

Student's t distribution

The t -distribution is symmetric but has higher variation than $N(0, 1)$. It has a single parameter, which is referred to as the number of ‘degrees of freedom’.

Let $U \sim N(0, 1)$ and $V \sim \chi_n^2$ such that U and V are independent. Then

$$X = \frac{U}{\sqrt{V/n}}$$

is such that $X \sim t_n$ (X has t distribution with n degrees of freedom). The t distribution is similar to $N(0, 1)$ but with ‘fatter/heavier tails’. A random variable $X \sim t_n$ has mean 0 and variance $n/(n-2)$ for $n > 2$.

Tables are available in NCST: cdf pp42 - 44, percentage points p45 (note in each column the percentage point tends to the corresponding $N(0, 1)$ point as the number of degrees of freedom increases - the bottom row corresponds to $N(0, 1)$).

F distribution

The F distributions has two integer parameters n, m .

Let $U \sim \chi_n^2$ and $V \sim \chi_m^2$ such that U and V are independent. Then

$$X = \frac{U/n}{V/m}$$

is such that $X \sim F_{n,m}$. It is useful to note that if $X \sim F_{n,m}$ then $1/X \sim F_{m,n}$. This is used when looking up values in the cumulative distribution tables. Tables of upper percentage points are available in NCST pp50 - 55.

5.6.2 Sampling distributions- mean and variance

Let $\mathbf{X} = (X_1, X_2, \dots, X_n)$ be a random sample of a variable $X \sim N(\mu, \sigma^2)$. Also let \bar{X} be the sample mean and S^2 the sample variance. Then

1. \bar{X} and S^2 are independent random variables.

2. We have

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$$

3. We have

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

4. When σ^2 is unknown and instead we want to estimate it by S^2 , we can use

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

5.7 Two samples

In some situations we want to study the difference, if any, between the means of two populations - we do this via the difference between the two sample means. In other cases, we may want to study the difference, if any, between the variances of two populations - we do this via the ratio of the sample variances.

Let $\mathbf{X} = (X_1, X_2, \dots, X_n)$ be a random sample of size n of a population variable X with mean μ_X and variance σ_X^2 , and let $\mathbf{Y} = (Y_1, Y_2, \dots, Y_m)$ be a random sample of size m of a population variable Y with mean μ_Y and variance σ_Y^2 , where X and Y independent. Let the sample means be denoted by \bar{X} , \bar{Y} and the sample variances by S_X^2 , S_Y^2 .

5.7.1 Difference between sample means, two independent samples

$\bar{X} - \bar{Y}$ has mean $\mu_X - \mu_Y$ and variance $\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}$.

For large n and m :

$$\bar{X} - \bar{Y} \sim N\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right) \text{ approximately (exact for normally distributed samples).}$$

In the case of sampling from normal distributions with $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ we have

$$\bar{X} - \bar{Y} \sim N\left(\mu_X - \mu_Y, \sigma^2 \left(\frac{1}{n} + \frac{1}{m}\right)\right)$$

so

$$\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\sigma^2 \left(\frac{1}{n} + \frac{1}{m}\right)}} \sim N(0, 1).$$

Also

$$\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{S_p^2 \left(\frac{1}{n} + \frac{1}{m}\right)}} \sim t_{n+m-2},$$

where S_p^2 is a pooled estimate of the common variance σ^2 given by

$$S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}.$$

5.7.2 Ratio of two sample variances, independent normal samples

Recall $\frac{(n-1)S_X^2}{\sigma_X^2} \sim \chi_{n-1}^2$ and $\frac{(m-1)S_Y^2}{\sigma_Y^2} \sim \chi_{m-1}^2$, independent.

We scale the sample variances by the corresponding population variances and construct the ratio to give

$$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F_{n-1, m-1}.$$

5.8 Examples: Sampling distributions and CLT

Example 5.22. A random sample of 9 observations is taken from a $N(35, 16)$ distribution. Find the probability that the sample mean exceeds 36.

Example 5.23. A random sample of 200 observations is taken from a random variable with mean 35 and variance 16. Find the (approximate) probability that the sample mean assumes a value between 34.6 and 35.3.

Example 5.24. In sampling from a $N(\mu, 25)$ distribution, find how many observations are required to ensure that the sample mean differs from the population mean by at most 1 with probability at least 0.9, i.e., to ensure $\mathbb{P}(|\bar{X} - \mu| < 1) \geq 0.9$.

Example 5.25. Consider a random sample of 50 lifetimes of light bulbs. Suppose that the lifetimes are distributed exponentially with mean 1000 hrs. Let X_i be a lifetime.

1. Find the 95% probability level (approximately) for the sample mean claim amount (i.e. the interval such that with a 95% chance the mean lies within it).
2. Find the approximate probability that the total of all 50 lifetimes exceeds 60,000hrs.

Example 5.26. Find the (approximate) probability of getting

- at least 108 heads when a fair coin is tossed 200 times; and
- at most 138 heads when a fair coin is tossed 300 times.

Example 5.27. A fair six-sided die is thrown repeatedly - until the total accumulated score is higher than 500. By considering the distribution of the score on 140 throws, find the (approximate) probability that more than 140 throws are required.

Example 5.28. Accidents at a building happen on a building site at a rate of 5 per month. Find the probability that at most 75 accidents happen on this site during a given year assuming they are modelled by a Poisson random variable.

Example 5.29. Political opinion polls in the UK typically have sample sizes between 1000 and 1500. The media usually quote that the results (percentages in favour of particular parties) have ‘margin of error $\pm 3\%$ ’. Show that, with a poll size of 1034 or more, there is a probability of at least 0.95 that the percentage in favour of a party will be ‘out’ (i.e. will differ from the true underlying percentage) by at most 3%.

Example 5.30. A company wishes to estimate the proportion, p , of logs that have a certain property. The proportion is to be estimated by finding the proportion, P , of such policies with the property in a random sample.

1. How large a sample should be selected to be 90% confident that the error of estimation does not exceed 0.015?
2. How large a sample should be selected to be 90% confident that the error of estimation does not exceed 0.015, given that the true proportion p is known not to exceed 0.3?

F20SA / F21SA Statistical Modelling and Analysis

Chapter 6: Model Fitting 1: Parameter Estimation

Contents

6.1	Motivation	6-1
6.2	Introduction	6-1
6.3	Properties of estimators	6-2
6.3.1	Bias of estimators	6-2
6.3.2	Consistency	6-3
6.3.3	Efficiency	6-3
6.3.4	Sufficiency	6-4
6.4	Methods of constructing estimators	6-4
6.4.1	Method of moments estimators (MME)	6-4
6.4.2	Method of least squares estimators (LSE)	6-4
6.4.3	Method of maximum likelihood	6-5
6.5	Examples	6-10

6.1 Motivation

We now move back to looking at processing data and drawing conclusions. In Chapter 2 we saw an example of real data in the form of modules of rupture for a series of wood samples, see Table 6.1. We might believe this data is described by a continuous random variable with probability density function:

$$f_X(x) = \begin{cases} \frac{1}{w} \left(1 - \frac{|x-c|}{w}\right) & \text{if } |x-c| < w, \\ 0 & \text{otherwise,} \end{cases}$$

where w and c are unknown parameters. The issue is how to use the data to estimate w and c , and how do we compare different methods for selecting w and c ?

6.2 Introduction

The aim of statistical inference is to extract relevant, useful information from a set of data $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$ and to use this information in an efficient way to inform us about the population from which the data has arisen.

We perform inference about a population in the presence of a model, which is a mathematical representation of the random process which is assumed to have generated the data - the model involves a probability distribution for a population variable X . This distribution contains one (or more) unknown parameter(s) θ , whose value(s) we want to estimate.

In the classical (frequentist) approach we think of θ as having a fixed (but unknown) value, and we regard the data \mathbf{x} as coming from a repeatable sampling procedure.

We have a **random sample** $\mathbf{X} = (X_1, X_2, \dots, X_n)$ available - each observation X_i is a sampled value of the random variable X with pdf $f(x; \theta)$.

48.78	32.02	45.54	32.40	48.37	50.98	35.58	40.53	29.11	65.35
41.64	39.34	34.12	33.06	29.93	40.71	28.97	47.25	65.61	45.19
39.77	46.33	45.92	33.47	36.38	34.63	34.56	32.68	37.78	70.22
35.89	46.99	36.47	35.67	46.86	24.84	28.69	43.26	43.33	41.75
54.04	22.67	28.98	28.46	36.00	28.83	38.64	47.61	53.63	37.51
35.43	39.62	40.85	23.16	23.19	42.31	24.25	28.13	41.85	31.60
22.75	44.78	56.60	44.51	36.88	39.33	44.54	32.48	33.19	37.65
44.78	26.63	28.76	42.47	44.30	39.93	40.85	36.81	39.15	28.00
43.99	43.48	47.42	48.39	44.59	39.60	39.97	35.88	54.71	46.01
47.74	30.05	33.61	38.05	44.00	38.16	37.69	33.92	43.64	43.48
25.39	30.33	44.36	35.03	40.39	43.33	41.78	57.99	56.80	40.27
38.00	39.21	35.30	31.33	41.72	69.07	33.14	49.57	43.07	39.05
25.98	51.39	33.18	27.31	29.90	51.90	55.23	40.20	43.12	32.76
36.84	50.91	36.85	53.99	35.17	33.71	36.53	49.59	30.02	45.97
34.49	49.65	17.98	43.41	34.44	46.50	22.74	32.03	38.81	23.14
38.71	47.83	27.90	28.71	27.93	36.92	34.40	39.20	24.09	53.00
30.53	44.07	44.36	58.34						

Table 6.1: Modulus of rupture data from 50 mm \times 150 mm Swedish redwood and whitewood timber in Newtons per square millimetre.

A **statistic** is a function of \mathbf{X} which does not involve θ . It is a random variable with its own probability distribution, called its **sampling distribution** - the properties of which depend on those of X .

An **estimator** of θ is a statistic whose value is used as an estimate of θ .

We denote the estimator $\hat{\theta}(\mathbf{X})$ or just $\hat{\theta}$ or $\tilde{\theta}$ or θ^* , and the estimate (its value for a particular observed sample) $\hat{\theta}(\mathbf{x})$, or again just $\hat{\theta}$.

Example 6.1. Suppose we have a random sample $\mathbf{X} = (X_1, X_2, \dots, X_n)$ of $X \sim \text{Poisson}(\theta)$. Consider the following four estimators of θ :

1. $\hat{\theta} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}$,
2. $\hat{\theta} = X_1$,
3. $\hat{\theta} = \sum_{i=1}^n w_i X_i$ where w_i are known constants,
4. $\hat{\theta} = \frac{1}{n} \sum_{i=1}^n X_i^2$.

Which do you think is the ‘best’ one to use? Why? What makes it a good estimator?

We consider desirable properties of $\hat{\theta}$, recognising that it has its own sampling distribution with its own properties.

The SD of an estimator $\hat{\theta}$ is referred to as its **standard error**: we write $se(\hat{\theta})$ (or $ese(\hat{\theta})$ for the **estimated standard error**).

6.3 Properties of estimators

6.3.1 Bias of estimators

An estimator $\hat{\theta}$ is **unbiased** for θ if $\mathbb{E}(\hat{\theta}) = \theta$

The **bias** of $\hat{\theta}$ is $B(\hat{\theta}) = \mathbb{E}(\hat{\theta} - \theta) = \mathbb{E}(\hat{\theta}) - \theta$.

$\hat{\theta}$ is **asymptotically unbiased** for θ if $\mathbb{E}(\hat{\theta}) \rightarrow \theta$ as $n \rightarrow \infty$, i.e., if $B(\hat{\theta}) \rightarrow 0$ as $n \rightarrow \infty$.

Example 6.2. Sampling from X with mean μ and variance σ^2 (recall that S is the sample standard deviation):

- $\mathbb{E}(\bar{X}) = \mu$ so \bar{X} is unbiased for μ .
- $\mathbb{E}(S^2) = \sigma^2$ so S^2 is unbiased for σ^2 .
- $\mathbb{E}(S) \neq \sigma$ so S is biased for σ .

The **mean square error** of an estimator $\hat{\theta}$ is $MSE(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2]$.

Note the difference from the expression for the variance of the estimator: $V(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2]$. The MSE and the variance of an estimator are the same in the case of an unbiased estimator.

It is easy to show $MSE(\hat{\theta}) = V(\hat{\theta}) + (B(\hat{\theta}))^2$.

So, if $V(\hat{\theta})$ and $B(\hat{\theta}) \rightarrow 0$ as $n \rightarrow \infty$, then $MSE(\hat{\theta}) \rightarrow 0$ as $n \rightarrow \infty$.

6.3.2 Consistency

An estimator $\hat{\theta}$ is **consistent** (some authors say ‘consistent in probability’) for θ if, for any $\varepsilon > 0$,

$$\mathbb{P}(|\hat{\theta} - \theta| \geq \varepsilon) \rightarrow 0 \text{ as } n \rightarrow \infty.$$

It can be shown that

$$\mathbb{P}(|\hat{\theta} - \theta| \geq \varepsilon) \leq \frac{1}{\varepsilon^2} MSE(\hat{\theta}).$$

Hence $V(\hat{\theta}) \rightarrow 0$ and $B(\hat{\theta}) \rightarrow 0$ as $n \rightarrow \infty$ implies $MSE(\hat{\theta}) \rightarrow 0$ and so in this case $\hat{\theta}$ is consistent for θ .

Example 6.3. \bar{X} is unbiased for μ with variance $\sigma^2/n \rightarrow 0$ as $n \rightarrow \infty$. So \bar{X} is consistent for μ .

Example 6.4. Sampling from $X \sim N(\mu, \sigma^2)$. We have

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2 \text{ so } V(S^2) = \frac{2\sigma^4}{n-1}.$$

We know that S^2 is an unbiased estimator for σ^2 and we see that the variance tends to 0 as n increases. This gives us that S^2 is consistent for σ^2 .

6.3.3 Efficiency

An estimator $\hat{\theta}_1$ is **more efficient** than $\hat{\theta}_2$ if $MSE[\hat{\theta}_1] < MSE[\hat{\theta}_2]$.

Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be unbiased estimators of θ with $V[\hat{\theta}_1] \leq V[\hat{\theta}_2]$. Then the **relative efficiency** of the estimators is $V[\hat{\theta}_1]/V[\hat{\theta}_2]$.

Example 6.5. For $X \sim Po(\theta)$ we see that both \bar{X} and X_1 are unbiased estimators for θ . The variance of \bar{X} is θ/n but the variance of X_1 is θ . So \bar{X} is more efficient than X_1 and the relative efficiency is $(\theta/n)/\theta = 1/n$.

For a given estimation problem it can be shown that there is a lower bound on the efficiency of possible unbiased estimators. An unbiased estimator which is the most efficient for a given estimation problem is called a MVUE (minimum variance unbiased estimator). We will see later examples of estimators which are guaranteed to be MVUE in the limit as the sample size tends to infinity.

6.3.4 Sufficiency

An estimator $\hat{\theta}$ is **sufficient** for θ if the conditional distribution of \mathbf{X} given $\hat{\theta}$ does not depend on θ .

A sufficient statistic contains all the information in the data relevant to estimating θ . There is a ‘factorisation criterion’ for establishing sufficiency and the concept is important theoretically.

Example 6.6. \bar{X} is sufficient for μ .

Example 6.7. Coin tossing: $\mathbb{P}(\text{head}) = \theta$.

Data: $X_i, i = 1, 2, \dots, n$ where $X_i = 1$ (head) or 0 (tail).

$\sum X_i$ (total number of heads) is sufficient for $n\theta$. The only other information we may have available (order of heads and tails) tells us nothing further of relevance to estimating θ .

6.4 Methods of constructing estimators

Having considered how to compare estimators we now explore three methods for constructing estimators given a model. These are

1. Method of moments estimators (MME)
2. Method of least squares estimators (LSE)
3. Maximum likelihood estimators (MLE)

In the case of the maximum likelihood estimators we will also consider the behaviour of the estimates for large sample sizes.

6.4.1 Method of moments estimators (MME)

We simply equate the population mean and variance to the corresponding sample means and variance (in as convenient a manner as possible) and solve for parameter estimates. The method is often convenient - but does not always produce efficient estimators.

Example 6.8. Consider $X \sim U(0, \theta)$, then we have

$\mathbf{E}(X) = \theta/2$ and we set $\hat{\theta}/2 = \bar{X}$. Therefore the MME of θ is $\hat{\theta} = 2\bar{X}$.

Example 6.9. Consider $X \sim N(0, \sigma^2)$ and suppose we want to estimate σ . We have $\mathbf{E}(X) = 0$ which is of no use in this case, but $V(X) = \sigma^2$ so we set $\hat{\sigma}^2 = S^2$, where S^2 is the sample variance. Hence our estimate of σ is $\hat{\sigma} = \sqrt{S^2}$.

6.4.2 Method of least squares estimators (LSE)

We choose parameter values which minimise the sum of the squares of the deviations of the observations from their means as given by the model. Hence for a sample $\mathbf{X} = (X_1, \dots, X_n)$ we are looking to select the estimator to minimise

$$\sum (X_i - \mathbf{E}(X))^2.$$

Example 6.10. X has an unspecified distribution with mean μ . We can write $X = \mu + \varepsilon$ where ε is a variable with mean 0; ε represents the deviation of X from its mean and is an ‘error’ or ‘noise’ variable.

Sum of squares of errors is $S = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (X_i - \mu)^2$. We want to look for a value of μ that minimizes S . Hence we differentiate by μ to obtain

$$\frac{dS}{d\mu} = -2 \sum_{i=1}^n (X_i - \mu) = 0.$$

So the LSE is $\hat{\mu} = \bar{X}$.

Example 6.11. Suppose a variable Y is proportional to x (a constant) but there is a normal measurement error (with constant variance) present.

Model: $Y = \beta x + \varepsilon$ where $\varepsilon \sim N(0, \sigma^2)$; note that $Y \sim N(\beta x, \sigma^2)$.

We observe Y_1, Y_2, \dots, Y_n independently at x_1, x_2, \dots, x_n respectively. This means $Y_i = \beta x_i + \varepsilon_i$, so $Y_i \sim N(\beta x_i, \sigma^2)$. Hence the sum of squares is $S = \sum_{i=1}^n (Y_i - \beta x_i)^2$. Differentiating with respect to β we get

$$\frac{dS}{d\beta} = -2 \sum_{i=1}^n x_i (Y_i - \beta x_i),$$

so we have an estimator $\hat{\beta}$ of β given as

$$\hat{\beta} = \frac{\sum x_i Y_i}{\sum x_i^2} = \sum c_i Y_i \text{ where } c_i = \frac{x_i}{\sum x_i^2}.$$

We want to know how $\hat{\beta}$ is distributed:

$$\mathbb{E}(\hat{\beta}) = \sum c_i \mathbb{E}(Y_i) = \sum c_i \beta x_i = \beta \sum c_i x_i = \beta.$$

$$V(\hat{\beta}) = \sum c_i^2 V(Y_i) = \sigma^2 \sum c_i^2 = \frac{\sigma^2}{\sum x_i^2}.$$

Since $\hat{\beta}$ is a linear combination of independent normal random variables it is a normal random variable, $\hat{\beta} \sim N(\beta, \sigma^2 / \sum x_i^2)$.

We will return to this in later chapters.

6.4.3 Method of maximum likelihood

Example 6.12. You are given a coin by a friend and know that the probability of a head is either $\frac{1}{2}$ (fair coin) or $\frac{1}{4}$ (biased coin). You flip the coin 10 times and observe that the number of heads is 4. Using this information do you now believe the coin is fair or not?

We can calculate the probability of observing 4 heads out of 10 flips in both cases. Let X be the number of heads observed.

Fair coin: $X \sim \text{Bin}(10, 0.5)$

$$\mathbb{P}(X = 4) = \binom{10}{4} (0.5)^4 (1 - 0.5)^6 = 0.205.$$

Biased coin: $X \sim \text{Bin}(10, 0.25)$

$$\mathbb{P}(X = 4) = \binom{10}{4} (0.25)^4 (1 - 0.25)^6 = 0.146.$$

Since there was a higher probability of observing the result with a fair coin, we would believe the coin is fair.

Likelihood and Score functions and Fisher's Information

The **likelihood function** for a parameter θ , given a sample \mathbf{x} from independent observations, is

$$L(\theta, \mathbf{x}) = f_X(x_1; \theta) f_X(x_2; \theta) \dots f_X(x_n; \theta),$$

where $f_X(\cdot; \theta)$ is the pmf for a discrete distribution or the pdf for a continuous distribution. It is regarded as a function of the unknown parameters θ .

It represents a relative measure of how strongly the data support different possible values of the parameter θ , and is usually written simply as $L(\theta)$.

Log-likelihood function:

$$l(\theta) = l(\theta; \mathbf{x}) = \ln[L(\theta, \mathbf{x})]$$

Score function:

$$U(\theta) = U(\theta; \mathbf{x}) = \frac{\partial l(\theta)}{\partial \theta}.$$

We also denote $l_i(\theta) := \ln(f_X(x_i; \theta))$. Since $l(\theta)$ is a log of a product (and hence the sum of the individual logs) it is clear that $l(\theta) = \sum l_i(\theta)$ and $U(\theta) = \sum U_i(\theta)$ where l_i and U_i are the corresponding functions for the i^{th} observation alone.

Fisher's Information about θ contained in the sample is

$$I(\theta) = \mathbb{E} \left[\left(\frac{\partial l}{\partial \theta} \right)^2 \right].$$

Furthermore, for independent identically distributed samples we see that $I(\theta)$ for a sample of size n is just $n \times I(\theta)$ for a sample of size 1 (i.e. a single observation). Also we have an important and useful alternative form

$$I(\theta) = -\mathbb{E} \left[\frac{\partial^2 l}{\partial \theta^2} \right].$$

Maximum likelihood estimate (MLE)

If we know the value of θ , then $L(\theta; x)$ gives the probability of (value of the pdf associated with) observing that particular sample. But we don't know θ , it's the x_i 's that are known.

The **maximum likelihood estimator** (MLE) of θ is the value $\hat{\theta}$ that maximises $L(\theta)$. In many cases it is easier to maximise $l(\theta)$ which provides the same value for $\hat{\theta}$.

Example 6.13. $X \sim \text{Bin}(10, \theta)$. Suppose we observe 7 successes in 10 trials.

$$L(\theta) = \binom{10}{7} \theta^7 (1 - \theta)^3, \quad l(\theta) = k + 7 \ln(\theta) + 3 \ln(1 - \theta)$$

where $k = \ln \binom{10}{7}$. Differentiating with respect to θ ,

$$\frac{dl}{d\theta} = \frac{7}{\theta} - \frac{3}{1 - \theta}.$$

Solving $\frac{dl}{d\theta} = 0$, we obtain the MLE for θ which is $\hat{\theta} = 0.7$.

In general for $X \sim \text{Bin}(n, \theta)$, with known n and unknown θ , we have

$$L(\theta, x) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}, \quad l(\theta) = k + x \ln(\theta) + (n - x) \ln(1 - \theta).$$

Differentiating with respect to θ ,

$$\frac{dl}{d\theta} = \frac{x}{\theta} - \frac{n-x}{1-\theta}.$$

Solving $\frac{dl}{d\theta} = 0$, we obtain the MLE for θ which is $\hat{\theta} = x/n$.

The MLE of a binomial probability (a population proportion) is the sample proportion.

Example 6.14. We are given a random sample $\mathbf{X} = (X_1, \dots, X_n)$ of size n from $X \sim Po(\lambda)$.

$$L(\lambda, \mathbf{X}) = \frac{e^{-n\lambda} \lambda^{\sum X_i}}{\prod X_i!}, \quad l(\lambda) = -n\lambda + \left(\sum X_i\right) \ln(\lambda) + k.$$

Differentiating with respect to λ ,

$$\frac{dl}{d\lambda} = -n + \frac{\sum X_i}{\lambda} = 0,$$

so the MLE is $\hat{\lambda} = \bar{X}$.

The MLE of a Poisson mean is the sample mean.

Example 6.15. We are given a random sample of size n from $X \sim N(\mu, 1)$.

$$L(\mu, \mathbf{X}) = \frac{1}{(2\pi)^{n/2}} e^{-\sum (X_i - \mu)^2 / 2}, \quad l(\mu) = -\sum (X_i - \mu)^2 / 2 + k.$$

Differentiating with respect to μ ,

$$\frac{dl}{d\mu} = \sum (X_i - \mu) = 0,$$

so the MLE is $\hat{\mu} = \bar{X}$.

The MLE of a normal distribution mean is the sample mean.

Notes:

1. MLEs are invariant under transformations of the parameter, i.e., if $\hat{\theta}$ is the MLE of θ then $g(\hat{\theta})$ is the MLE of $g(\theta)$.
e.g. if $\hat{\theta} = 0.7$ then the MLE of $\nu = \theta^2$ is $\hat{\nu} = 0.7^2 = 0.49$.
2. It is not always possible to solve the equation(s) for the MLE(s) explicitly - we may need a numerical solution.
3. Likelihood methods are not restricted to situations in which we have full data from a random sample from a single distribution. The method can be used whenever we can specify the probability distribution for each observation of the observed data.

Example 6.16. A case of incomplete information: random sample, size 10, of $X \sim Exp(\lambda)$. It turns out that 7 observations are less than 5 and 3 observations exceed 5. Calculate the MLE of λ .

Let $\theta = \mathbb{P}(X > 5) = \exp(-5\lambda)$.

We think about ‘proportions’ of observations larger than 5. The MLE of θ is the proportion of observations greater than 5 in the sample (here 3 observations out of 10).

Hence the MLE of $\theta = \exp(-5\lambda)$ is 0.3. So the MLE of λ is given by solving $\exp(-5\lambda) = 0.3$.

We obtain the MLE of λ as 0.241.

Classical large sample properties of MLEs

MLEs have very attractive properties - at least for large samples.

Firstly, it can be shown that the MLE $\hat{\theta}$ is consistent for θ .

The ‘maximum likelihood theorem’ states that in most cases the limit as $n \rightarrow \infty$

$$(\hat{\theta} - \theta)\sqrt{I(\theta)} \sim N(0, 1),$$

where I is the Fisher information. This does not apply if the domain of the random variable depends on the parameter, for example if $X \sim U(0, \theta)$.

So, in large samples, an MLE $\hat{\theta}$ is approximately:

- unbiased
- efficient (asymptotically a minimum variance unbiased estimator)
- normally distributed

So asymptotically

$$\hat{\theta} \sim N\left(\theta, \frac{1}{I(\theta)}\right).$$

So asymptotically,

$$se(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}}$$

and we estimate the standard error by

$$ese(\hat{\theta}) = \frac{1}{\sqrt{I(\hat{\theta})}}$$

Example 6.17. $X \sim Po(\lambda)$

$\hat{\lambda} = \bar{X}$, $V(\hat{\lambda}) = V(\bar{X}) = V(X)/n = \lambda/n$ and the $ese(\hat{\lambda}) = \sqrt{\hat{\lambda}/n}$.

Example 6.18. Random sample, size n (large), of $X \sim Exp(\lambda)$.

$$L(\lambda) = \lambda^n \exp(-\lambda \sum X_i), \quad l(\lambda) = n \ln(\lambda) - \lambda \sum X_i.$$

$$\frac{dl}{d\lambda} = \frac{n}{\lambda} - \sum X_i = 0.$$

So $\hat{\lambda} = 1/\bar{X}$, which is a biased estimate but is asymptotically unbiased. It can be shown that $\mathbb{E}(\hat{\lambda}) = n\lambda/(n-1)$.

We also have

$$-\frac{d^2l}{d\lambda^2} = \frac{n}{\lambda^2} \text{ so } I(\lambda) = \frac{n}{\lambda^2}.$$

So asymptotically we have

$$\hat{\lambda} \sim N\left(\lambda, \frac{\lambda^2}{n}\right),$$

and the $ese(\hat{\lambda}) = \hat{\lambda}/\sqrt{n}$.

MLEs in multi-parameter models

We now have an r -vector of parameters $\theta = (\theta_1, \theta_2, \dots, \theta_r)$ and we use partial differentiation to maximise L (if appropriate).

Example 6.19. Random sample, size n , of $X \sim N(\mu, \sigma^2)$. Find the MLE of $\theta = (\mu, \varphi)$ where $\varphi = \sigma^2$. Then $l(\mu, \varphi) = (-n/2) \ln(2\pi\varphi) - \sum ((X_i - \mu)^2 / 2\varphi)$,

$$\frac{\partial l}{\partial \mu} = 2 \sum ((X_i - \mu) / 2\varphi) = \sum ((X_i - \mu) / \varphi) = 0$$

$$\frac{\partial l}{\partial \varphi} = -n/(2\varphi) + \sum ((X_i - \mu)^2 / 2\varphi^2) = 0.$$

Therefore the MLE estimates are

$$\hat{\mu} = \sum X_i / n = \bar{X} \quad \hat{\varphi} = \frac{1}{n} \sum (X_i - \hat{\mu})^2 = \frac{1}{n} \sum (X_i - \bar{X})^2.$$

Note that $\hat{\varphi}$ is biased but asymptotically unbiased.

It is possible to obtain large sample results as for a single parameter by extending the definition of the Fisher information function to a matrix version.

6.5 Examples

Example 6.20.

Random sample, size 5, of $X \sim U(0, \theta)$. Data $x = (0.46, 1.14, 0.83, 0.21, 0.59)$.

- (a) Find the MME and the MLE of θ .
- (b) Compare the expectations and the MSEs of the two estimators.

Example 6.21.

Random sample, size n , of a random variable X with probability density function

$$f(x) = 2\theta x \exp(-\theta x^2), \quad x > 0,$$

where $\theta > 0$.

- (a) Find the MME of θ .
- (b) Find the MLE of θ , the score function $U(\theta)$, and Fishers information function $I(\theta)$. State the limiting distribution of the MLE.

Example 6.22. A crude model for the number of boys, X , in families with three children is:

X	0	1	2	3
$\mathbb{P}(X = x)$	θ	$3\theta/2$	$3\theta/2$	$1 - 4\theta$

where θ is a parameter to be estimated. In a random sample of n such families there were n_i with i boys, $i = 0, 1, 2, 3$.

- (a) Find the method of moments estimate of θ .
- (b) Find the maximum likelihood estimator of θ and its asymptotic distribution.

Example 6.23. Random variables X , Y , and Z are independently distributed as normal random variables with unit variances and means a , $a + b$, and $a + 2b$ respectively, where a and b are unknown constants.

- (a) Find the MLEs of a and b and show that they are unbiased.
- (b) If α , β , and γ are constants subject to the restriction $\alpha + \beta + \gamma = 0$, find a further restriction on these constants which ensures that the estimator $\alpha X + \beta Y + \gamma Z$ is unbiased for b . Hence find the estimator of b with minimum variance in the class of unbiased estimators of the form $\alpha X + \beta Y + \gamma Z$, and verify that this MV estimator is in fact the MLE of b obtained in part (a).

Example 6.24.

Random sample, size n , from $X \sim N(\mu, 1)$. All we know is that k of the n observations are positive. Find an expression for the MLE of μ and evaluate it in the case $n = 20, k = 14$.

Example 6.25. Suppose Y_i , $i = 1, 2, \dots, n$ are independent random variables with $Y_i \sim Po(\lambda x_i)$, and that we have one observation of each random variable, so the data are (y_i, x_i) , $i = 1, 2, \dots, n$. Find the MLE for λ .

Example 6.26. The lifetime T of a bulb of a certain type is to be modelled as a $Exp(\theta)$ random variable. A random sample of n such bulbs are put on test and are observed for a period t_0 . The times to failure of bulbs which fail are recorded. It is observed that by time t_0 , m of the n bulbs fail, with lifetimes t_1, t_2, \dots, t_m respectively. The remaining $n - m$ bulbs are still working at the end of the observation period. Find a MLE for θ .

F20SA / F21SA Statistical Modelling and Analysis

Chapter 7: Model Fitting 2: Confidence Intervals

Contents

7.1	Introduction	7-1
7.2	Confidence Intervals for Population Mean	7-2
7.2.1	σ^2 known	7-3
7.2.2	σ^2 unknown	7-3
7.2.3	One sided confidence intervals	7-4
7.3	CI's for population variance	7-4
7.4	CI's for population proportion	7-4
7.5	CI's for a Poisson mean	7-5
7.6	CI's based on general MLEs	7-5
7.7	Examples	7-6

7.1 Introduction

Instead of quoting a single value of an estimator (giving a 'point estimate') and its standard error, we can construct a random interval which contains the parameter θ with given probability. An interval $I = I(\mathbf{X})$ such that

$$\mathbb{P}(\text{interval } I \text{ contains } \theta) \geq 1 - \alpha$$

is a $100(1 - \alpha)\%$ **confidence interval** for θ .

We write this as

$$\mathbb{P}(\theta \in I) \geq 1 - \alpha,$$

for convenience, but it is important to understand that it is the interval I which is 'random' (the end points of the interval are random variables and their values are calculated from the data, hence we write $I = I(\mathbf{X})$). The parameter θ is an unknown constant - values inside the calculated interval are 'reasonable' as potential values for the parameter - the data is consistent/compatible with θ having such a value. Values outside the interval are 'unreasonable' - the data is incompatible with θ having such a value.

Definition 7.1. A $100(1 - \alpha)\%$ confidence interval (CI) for θ is $(L_1, L_2) = (L_1(\mathbf{X}), L_2(\mathbf{X}))$, where

$$\mathbb{P}(L_1 \leq \theta \leq L_2) \geq 1 - \alpha.$$

L_1 and L_2 are the lower and upper confidence limits, respectively.

With $\alpha = 0.05$ we have a 95% confidence interval.

A particular CI, once calculated from a sample, either does contain the true value of θ or it does not. In repeated sampling, about 95% of '95% CIs' will contain θ . In Figure 7.1 we see confidence intervals for 100 samples plotted which are calculated for the mean, the true mean

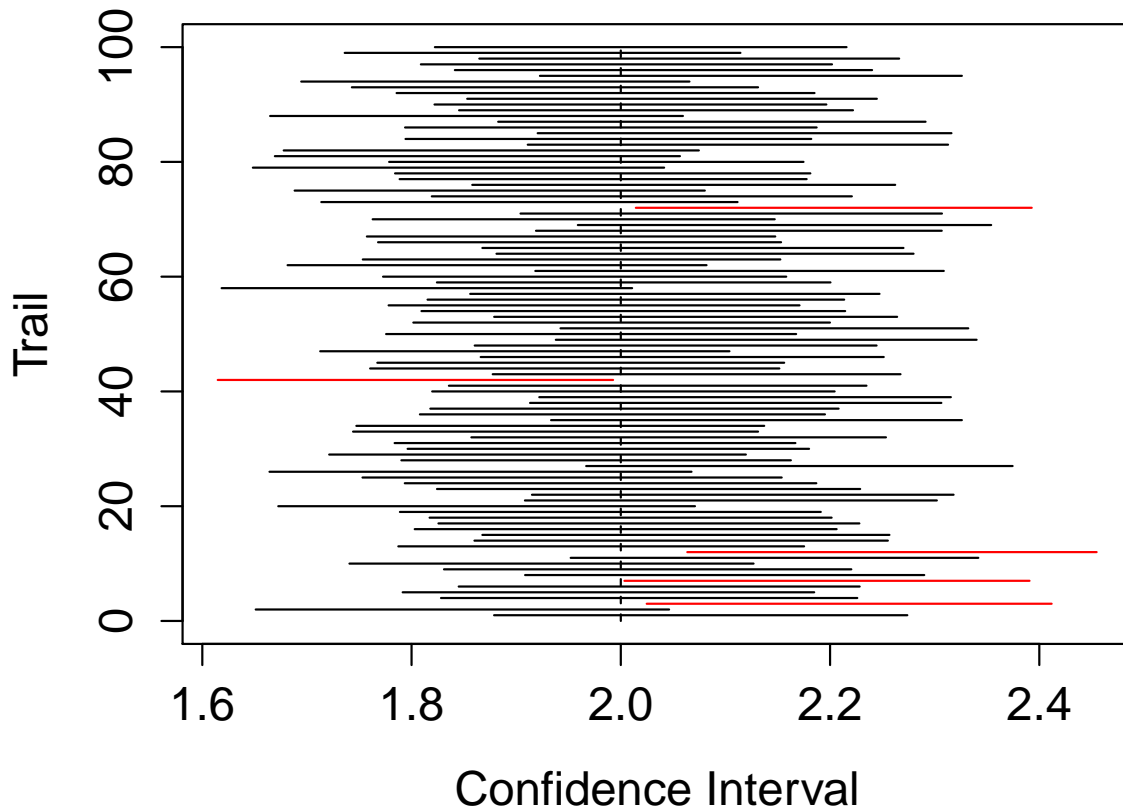


Figure 7.1: 95 % Confidence Intervals for Sample Mean.

is marked and those intervals which do not contain it are marked. In this case we can see that 5 out of the 100 confidence intervals do not contain the true mean.

A CI for a parameter θ is constructed using a **pivotal quantity**, which is a quantity whose distribution is completely known and which is monotonic in θ .

We will consider a series of standard cases, in general we will produce the 95% confidence interval but this can easily be adapted to other confidence intervals.

7.2 Confidence Intervals for Population Mean

Suppose $X \sim N(\mu, \sigma^2)$ and we want to construct a CI for μ . From the previous discussion we know that

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right), \quad \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1), \quad \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1},$$

where \bar{X} is the sample mean and S the sample standard deviation. We will use these as the basis for our pivotal quantities.

7.2.1 σ^2 known

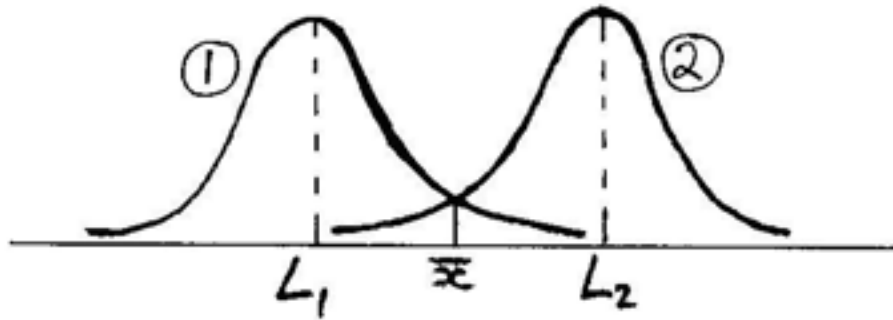
This situation is unrealistic and generally not appropriate but serves as a good introduction to the concepts. We use $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ as the pivotal quantity since this has a $N(0, 1)$ distribution. This gives

$$\mathbb{P}\left(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.95,$$

so

$$\mathbb{P}\left(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}\right) = 0.95.$$

Therefore a 95% confidence interval for μ is given by $\bar{X} \pm \left(1.96 \frac{\sigma}{\sqrt{n}}\right)$ i.e. $\bar{X} \pm (1.96 \times s.e(\bar{X}))$.



(1) is the distribution of \bar{X} which gives the lower limit L_1 , and (2) is the distribution of \bar{X} which gives the upper limit L_2 - given the data, L_1 is the smallest, and L_2 the largest value for μ which is compatible with the data we have observed.

Let $z_{\alpha/2}$ be the upper $100(\alpha/2)\%$ percentage point of the $N(0, 1)$ distribution, i.e., let $z_{\alpha/2}$ be the value such that $\mathbb{P}(Z > z_{\alpha/2}) = \alpha/2$ where $Z \sim N(0, 1)$. Then a $100(1 - \alpha)\%$ for μ is derived in the same way from

$$\mathbb{P}\left(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha,$$

so

$$\mathbb{P}\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

Therefore the confidence interval is $\bar{X} \pm (z_{\alpha/2} \times s.e(\bar{X}))$.

7.2.2 σ^2 unknown

We can use $\frac{\bar{X} - \mu}{S/\sqrt{n}}$ as the pivotal quantity as we know it has a t_{n-1} distribution. Let $t_{\alpha/2}$ be the upper $100(\alpha/2)\%$ percentage point of the t_{n-1} distribution, so it is the value such that $\mathbb{P}(t_{n-1} > t_{\alpha/2}) = \alpha/2$. Then

$$\mathbb{P}\left(-t_{\alpha/2} < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{\alpha/2}\right) = 1 - \alpha,$$

so

$$\mathbb{P}\left(\bar{X} - t_{\alpha/2} \frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2} \frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

therefore a $100(1 - \alpha)\%$ CI for μ is given by $\bar{X} \pm \left(t_{\alpha/2} \frac{S}{\sqrt{n}}\right)$ i.e. $\bar{X} \pm (t_{\alpha/2} \times e.s.e(\bar{X}))$.

In the case where X is not assumed to have a normal distribution, we can appeal to the large

sample normality of \bar{X} and use: large sample $100(\alpha/2)\%$ CI for μ is $\bar{X} \pm (z_{\alpha/2} \times e.s.e(\bar{X}))$.

7.2.3 One sided confidence intervals

We can construct one-sided intervals if desired: for example in the case of known variance we have

$$\mathbb{P}\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.645\right) = 0.95,$$

so

$$\mathbb{P}\left(\bar{X} - 1.645\frac{\sigma}{\sqrt{n}} < \mu\right) = 0.95.$$

giving $\left(\bar{X} - 1.645\frac{\sigma}{\sqrt{n}}, \infty\right)$ as a lower 95% confidence interval for μ .

This approach is appropriate if we are concerned with establishing the smallest plausible value for μ .

7.3 CIs for population variance

Suppose $X \sim N(\mu, \sigma^2)$ and we want to construct a CI for σ^2 . We know $\frac{(n-1)S^2}{\sigma^2}$ has a χ_{n-1}^2 distribution so we use this as our pivotal quantity.

Let a and b be the lower and upper $100(\alpha/2)\%$ percentage points of χ_{n-1}^2 , i.e., let a and b be such that for $X \sim \chi_{n-1}^2$ we have

$$\mathbb{P}(X < a) = \mathbb{P}(X > b) = \alpha/2.$$

Then we have

$$\mathbb{P}\left(a < \frac{(n-1)S^2}{\sigma^2} < b\right) = 1 - \alpha$$

so

$$\mathbb{P}\left(\frac{(n-1)S^2}{b} < \sigma^2 < \frac{(n-1)S^2}{a}\right) = 1 - \alpha.$$

This gives us a $100(1 - \alpha)\%$ CI for σ^2 given by

$$\left(\frac{(n-1)S^2}{b}, \frac{(n-1)S^2}{a}\right).$$

This procedure is robust to lack of normality in the distribution of the population variable, and performs well in non-normal situations, at least for large samples.

7.4 CIs for population proportion

Let X be the number of successes in n Bernoulli trials with $\mathbb{P}(\text{success}) = \theta$, and let $P = X/n$ be the sample proportion of successes (P is the MLE for θ).

We have $X \sim \text{Bin}(n, \theta)$ and so by the CLT we know that for large n , P is approximately $N(\theta, \theta(1 - \theta)/n)$ distributed.

Therefore for large samples we have a $100(1 - \alpha)\%$ CI for θ given by

$$P \pm (z_{\alpha/2} \times ese(P))$$

where

$$ese(P) = \sqrt{\frac{P(1-P)}{n}}.$$

7.5 CIs for a Poisson mean

Suppose $X \sim Po(\lambda)$ and we want to construct a CI for λ . The MLE for λ is \bar{X} . For large n by the CLT we know that \bar{X} is approximately $N(\lambda, \lambda/n)$ distributed.

Hence for large samples we have a $100(1 - \alpha)\%$ CI for λ , given by

$$\bar{X} \pm (z_{\alpha/2} \times ese(\bar{X})),$$

with

$$ese(\bar{X}) = \sqrt{\frac{\bar{X}}{n}}.$$

7.6 CIs based on general MLEs

For large sample MLEs we have a very useful, simple result based on the normal distribution. In the previous chapter we saw that in many cases an MLE, $\hat{\theta}$, is approximately $N(\theta, 1/I(\theta))$ distributed, where I is the Fisher information function.

Hence a $100(1 - \alpha)\%$ confidence interval for θ is

$$\hat{\theta} \pm (z_{\alpha/2} \times se(\hat{\theta})),$$

with $se(\hat{\theta}) = \sqrt{1/I(\hat{\theta})}$.

In many cases we will need to use the $ese(\hat{\theta}) = \sqrt{1/I(\hat{\theta})}$ rather than $se(\hat{\theta})$. For example we have,

$$95\% \text{ CI for } \hat{\theta} : \hat{\theta} \pm (1.96 \times ese(\hat{\theta})),$$

$$99\% \text{ CI for } \hat{\theta} : \hat{\theta} \pm (2.576 \times ese(\hat{\theta})),$$

Note: All these two sided confidence intervals have equivalent one sided versions which should be used if the problem so requires.

7.7 Examples

Example 7.1. A random sample of 20 observations from a $N(\mu, \sigma^2)$ distribution gives the following data summaries: $\sum x_i = 4035$, $\sum x_i^2 = 857115$.

- (a) Find a 95% CI for μ .
- (b) Find a 95% CI for σ^2 .

Suppose we now discover that the largest original observation (which was 359, and much bigger than all the other 19 observations) was of dubious accuracy. Recalculate the confidence intervals for the 19 samples.

Example 7.2. Calculate 95% CIs for θ , the proportion of a population with a certain property, given that

- (a) a random sample of 200 includes 70 with the property;
- (b) a random sample of 1200 includes 420 with the property.

Note: the sample proportion in (b) is the same as in (a), but the sample size is much bigger

Example 7.3. In a given piece of road, potholes appear and the number of potholes within a given month is modelled by a Poisson random variable with mean λ . The numbers of holes in different months are independent and identically distributed. The total number of holes to have arisen in the past 36 months is 442. Find a CI for λ .

Example 7.4. Consider a random sample, size n , of $X \sim \text{Exp}(\lambda)$ (so with mean $\mu = 1/\lambda$). Construct an approximate 95% CI for λ (for large n) based on large sample ML theory. For $n = 50$ and $\sum x_i = 316.2$ calculate the CI.

F20SA / F21SA Statistical Modelling and Analysis

Chapter 8: Decision Making: Hypothesis Testing

Contents

8.1	Introduction	8-1
8.2	Definitions	8-2
8.3	Standard test statistics	8-2
8.3.1	Testing a population mean	8-2
8.3.2	Testing a population variance	8-3
8.3.3	Testing a population proportion	8-3
8.3.4	Testing a Poisson mean	8-3
8.3.5	Examples	8-3
8.4	Significance and P-values	8-5
8.4.1	Examples	8-5
8.5	Final Comments	8-6

8.1 Introduction

As before, we have a random sample x of size n of a population random variable X with pdf/pmf $f(x; \theta)$. The distribution we assign to X is our model for the process which has generated our data, for example $X \sim N(\mu, 1)$ or $X \sim Po(\lambda)$.

A **hypothesis** H is a statement about the distribution of X - in particular, in this chapter, it is a statement about the unknown value of a parameter θ . A **simple hypothesis** is a statement which completely specifies the distribution: for example if $X \sim N(\mu, 1)$ then ' $H : \mu = 5$ ' is a simple hypothesis. If H is not simple, it is **composite** for example ' $H : \mu > 5$ '.

A **test** of H is a rule which partitions the sample space into two subsets:

- **critical region:** data in this subset are not consistent with H and we reject H ,
- **acceptance region:** data in this subset are consistent with H and we do not reject H .

The **null hypothesis** H_0 represents the current theory (the 'status quo'). Examples include: $H_0 : \theta = 2$, $H_0 : \theta = \theta_0$, $H_0 : P < 0.4$, $H_0 : P > P_0$, $H_0 : \mu = 5$, $H_0 : \mu > 5$, $H_0 : \mu < \mu_0$. $H_0 : \mu_1 - \mu_2 = 0$ (this is the 'no difference' or 'no treatment effect' hypothesis) $H_0 : \sigma_1^2 = \sigma_2^2$ (this is the 'equal variances' or 'homoscedasticity' hypothesis)

The null hypothesis H_0 is contrasted with an **alternative hypothesis** H_1 and our test is written, for example, as follows:

- $H_0 : \theta = \theta_0$ v $H_1 : \theta = \theta_1$ a test with simple null and alternative hypotheses
- $H_0 : \theta = \theta_0$ v $H_1 : \theta > \theta_0$ a one-sided test with simple null and composite alternative hypotheses

- $H_0 : \theta \leq \theta_0$ v $H_1 : \theta > \theta_0$ a one-sided test with composite null and alternative hypotheses
- $H_0 : \theta = \theta_0$ v $H_1 : \theta \neq \theta_0$ a two-sided test with simple null and composite alternative hypotheses

The fundamental questions we are asking are:

- does our data provide strong enough evidence to justify our rejecting the null hypothesis?
- how strong is our evidence against the null hypothesis?

The decision is based on the value of an appropriate function of the data called the **test statistic** (examples include the sample mean \bar{x} , sample proportion P , sample variance S^2 , maximum value in the sample), whose distribution is completely known ‘under H_0 ’, that is, when H_0 is true.

8.2 Definitions

There are two types of testing errors we are exposed to when making our decision:

- type I error: not accepting H_0 when it is true
- type II error: accept H_0 when it is false

The probabilities of making these errors are conventionally denoted α and β :

- $\alpha = \mathbb{P}(\text{commit a type I error}) = \mathbb{P}(\text{reject } H_0 | H_0 \text{ true})$
- $\beta = \mathbb{P}(\text{commit a type II error}) = \mathbb{P}(\text{accept } H_0 | H_0 \text{ false})$

$1 - \beta = \mathbb{P}(\text{reject } H_0 | H_0 \text{ false})$ is called the power of the test - it is the probability of making a correct decision to reject the null hypothesis. It measures the effectiveness of the test at detecting departures from the null hypothesis. We also call α the *size* or *level* of the test.

We want both α and β to be small, but, for a fixed sample size, it is not possible to lower both probabilities of error simultaneously - we can of course lower the probabilities by increasing the sample size.

In general we will select α which is called the **level of significance** of the test (popular choices are $\alpha = 0.05$, giving a ‘5% test’ and $\alpha = 0.01$, giving a ‘1% test’) and use this to define the critical region.

8.3 Standard test statistics

8.3.1 Testing a population mean

Suppose $X \sim N(\mu, \sigma^2)$, and we have a random sample of size n . We are interested in testing $H_0 : \mu = \mu_0$.

σ^2 **known**

Test statistic is

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

which has a $N(0, 1)$ distribution under H_0 .

σ^2 **unknown**

Test statistic is

$$\frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

which has a t_{n-1} distribution under H_0 . This is often referred to as the ‘t-test’.

Large sample non-normal data

For large samples from ‘any’ distribution: Test statistic is

$$\frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

which is $N(0, 1)$ (approximately) under H_0 .

8.3.2 Testing a population variance

Suppose $X \sim N(\mu, \sigma^2)$, and we have a random sample size n . We are interested in testing $H_0 : \sigma = \sigma_0$.

Test statistic is

$$\frac{(n-1)S^2}{\sigma_0^2}$$

which has a χ_{n-1}^2 distribution under H_0 .

8.3.3 Testing a population proportion

Let X be the number of successes in n Bernoulli trials with $\mathbb{P}(\text{success}) = \theta$, we want to test $H_0 : \theta = \theta_0$.

Test statistic is X , which is $\text{Bin}(n, \theta_0)$ under H_0 .

For large n the test statistics is

$$\frac{\bar{X} - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}}$$

which is $N(0, 1)$ (approximately) under H_0 .

8.3.4 Testing a Poisson mean

Suppose $X \sim \text{Po}(\lambda)$, and we have a random sample size n , and we want to test $H_0 : \lambda = \lambda_0$.

The test statistic is $\sum X_i$ which is $\text{Po}(n\lambda_0)$ under H_0 .

For large n we have $\sum X_i$ is approximately $N(n\lambda_0, n\lambda_0)$ or \bar{X} is approximately $N(\lambda_0, \lambda/n)$ under H_0 .

8.3.5 Examples

Example 8.1. Random sample of $X \sim N(\mu, \sigma^2)$. We want to test $H_0 : \mu = 10.5$ versus $H_1 : \mu < 10.5$ at the 5% level. We have data from a random sample of size 10:

$$\sum x_i = 92.1 \quad \sum x_i^2 = 877.47.$$

We reject H_0 for small values of \bar{X} . The test statistic is

$$\frac{\bar{X} - \mu_0}{S/\sqrt{n}},$$

which has a t_9 distribution under H_0 . The data gives

$$\bar{x} = 9.21, \quad s^2 = \frac{1}{9} \left(877.47 - \frac{92.1^2}{10} \right) = 3.2477.$$

Lower 5% point for t_9 is -1.833, so we reject H_0 for

$$\frac{\bar{X} - 10.5}{S/\sqrt{10}} < -1.833.$$

This defines the critical region for the test.

For our sample $\bar{x} = 9.21$, $s^2 = 3.2477$; the test statistic has value -2.26 so we do not accept H_0 in favour of H_1 .

An alternative, and simpler, approach is to calculate the observed value of the test statistic for the sample in hand and compare it with the tabulated percentage point (or go further - see P-values later). Here, our observed $t = (9.21 - 10.5)/0.32477^{1/2} = -2.264$, which is lower than the relevant percentage point (-1.833) - our observed value is low enough to be 'in the tail' of the reference distribution - and we reject H_0 .

Example 8.2. A coin is tossed 20 times and lands 'heads' 5 times and 'tails' 15 times. Investigate whether the coin is fair or biased in favour of tails (i.e., do we have strong enough evidence to conclude that the coin is biased in favour of tails?)

Let X be the number of heads. Then $X \sim \text{Bin}(20, \theta)$ where $\mathbb{P}(\text{head}) = \theta$.

We will test $H_0 : \theta = 0.5$ versus $H_1 : \theta < 0.5$ at 5%. We reject H_0 for 'small' values of X .

From NCST (p22),

$$\mathbb{P}(X \leq 5 | \theta = 0.5) = 0.0207$$

which is less than 0.05. Our observation ' $x = 5$ ' is in the lower tail of the reference binomial distribution so we reject H_0 . We conclude that the coin is biased in favour of tails.

Suppose the coin was tossed 20 times and landed 'heads' 8 times.

$$\mathbb{P}(X \leq 8 | \theta = 0.5) = 0.2517.$$

This is far too high to provide evidence against H_0 , which can stand.

But suppose now that the coin was tossed 100 times and landed 'heads' 40 times (same proportion of 'heads', but on many more tosses). Now $X \sim \text{Bin}(100, \theta)$ and we can use the test statistic

$$\frac{\bar{X} - n\theta_0}{\sqrt{n\theta_0(1 - \theta_0)}}$$

which is approximately $N(0,1)$ distribution under H_0 .

Our observed statistic is $(40 - 50)/5 = -2$ which is less than the lower 5% point of the $N(0,1)$ distribution (-1.645) - our observed value is in the tail of the reference distribution - this time we have sufficiently strong evidence against H_0 to justify our rejecting it. We reject H_0 and conclude that the coin is biased in favour of tails.

8.4 Significance and P-values

A typical conclusion of a significance test is simply ‘reject H_0 at the 5% level of significance’ or just ‘reject H_0 at 5%’. This is not as informative as we can be. It is more informative to quantify the strength of the evidence the data provide against H_0 .

We do this by calculating the probability value (P-value) of our observed test statistic. The P-value is the observed significance level of the test statistic - it is the probability, assuming H_0 is true, of observing a value of the test statistic as extreme (that is, as inconsistent with H_0) as the value we have actually observed.

The P-value is the probability of the smallest critical region which includes the observed test statistic.

Given the data we have, the P-value is the lowest level at which we can reject H_0 .

The smaller is the P-value, the stronger is our evidence against H_0 .

The use of P-values is very widespread in published statistical work and is strongly recommended.

P-value	Suitable language for your conclusions (in most applications)
> 0.05	insufficient evidence against H_0 to justify rejecting it evidence not strong enough to justify rejecting H_0 H_0 can stand
< 0.05	we have some evidence against H_0 we can reject H_0 at the 5% level of testing
< 0.01	we have strong evidence against H_0 we can reject H_0 at the 1% level of testing we can reject H_0 at levels of testing down to 1%
< 0.001	we have overwhelming evidence against H_0 we can reject H_0 at the 0.1% level of testing we can reject H_0 at levels of testing down to 0.1%

Table 8.1: Suitable conclusions from P-values

8.4.1 Examples

Example 8.3. Consider $X \sim N(\mu, 1)$ with the null hypothesis, H_0 , $\mu_0 = 10$, and the alternative hypothesis, H_1 , $\mu_1 = 10.5$, and $n = 25$.

Suppose we observe $\bar{x} = 10.41$. This value is in the critical region (which is ‘ $\bar{x} > 10.329$ ’) and has P-value given by

$$\mathbb{P}(\bar{X} \geq 10.41 | \mu = 10) = \mathbb{P}(Z \geq 2.05) = 0.020$$

or 2.0%. So we have strong enough evidence to justify rejecting H_0 , at levels of testing down to 2%.

Suppose however we observe $\bar{x} = 10.27$. This value is not in the critical region and has P-value given by

$$\mathbb{P}(\bar{X} \geq 10.27 | \mu = 10) = \mathbb{P}(Z \geq 1.35) = 0.089$$

or 8.9%. The P-value is higher and the evidence is not strong enough to justify rejecting H_0 .

Example 8.4. In the first example, the observed test statistic is -2.264 and the P-value of this statistic is $\mathbb{P}(t_9 < -2.264) = 0.025$ (from NCST). So we have strong enough evidence to justify rejecting H_0 , at levels of testing down to 2.5%.

Example 8.5. In the second example (with 100 tosses), under H_0 , $X \sim N(50, 25)$ (approximately), and the P-value of our observation ‘40 heads’ is calculated as

$$\mathbb{P}(X \leq 40|H_0) = \mathbb{P}\left(Z < \frac{40.5 - 50}{5}\right) = \mathbb{P}(Z < -1.9) = 0.029.$$

We have strong enough evidence to justify rejecting H_0 , at levels of testing down to about 3%. [Note the use of the ‘continuity correction’ when using the normal distribution (which is continuous) to calculate an approximation to a probability for the binomial distribution (which is discrete).]

8.5 Final Comments

1. We may be able to reject H_0 at a specified level simply by using so much data that our test statistic has a small enough standard error to enable us to detect a departure from H_0 . This departure may, however, be of little or no physical significance.
2. A failure to reject H_0 does not imply that H_0 is true. It indicates that we have failed to reject it - our data do not provide sufficiently strong evidence against it. H_0 represents a theory which lives on to fight another day.
3. Good practice in testing
State:
 - the model
 - the hypotheses
 - the test statistic
 - the distribution of the test statistic under H_0
 - the critical region
 - the observed value of the test statistic
 - the P-value (at least approximately) of the test statistic
 - your conclusion as regards the hypotheses
 - your conclusion in words which relate to the physical situation concerned

F20SA / F21SA Statistical Modelling and Analysis

Chapter 9: Regression

Contents

9.1	Motivation	9–1
9.2	Linear Regression	9–2
9.2.1	Fitting the Model	9–3
9.2.2	Residuals	9–5
9.2.3	Model checking	9–7
9.2.4	Terminology and Warning	9–11
9.2.5	Standard Errors, Confidence Intervals and Hypothesis testing	9–11
9.3	Correlation	9–14
9.3.1	Linear Regression Examples	9–16
9.4	Multiple Regression	9–23
9.4.1	ANOVA	9–23
9.4.2	Multiple Linear Regression Case Study	9–24
9.4.3	Model Selection	9–26

9.1 Motivation

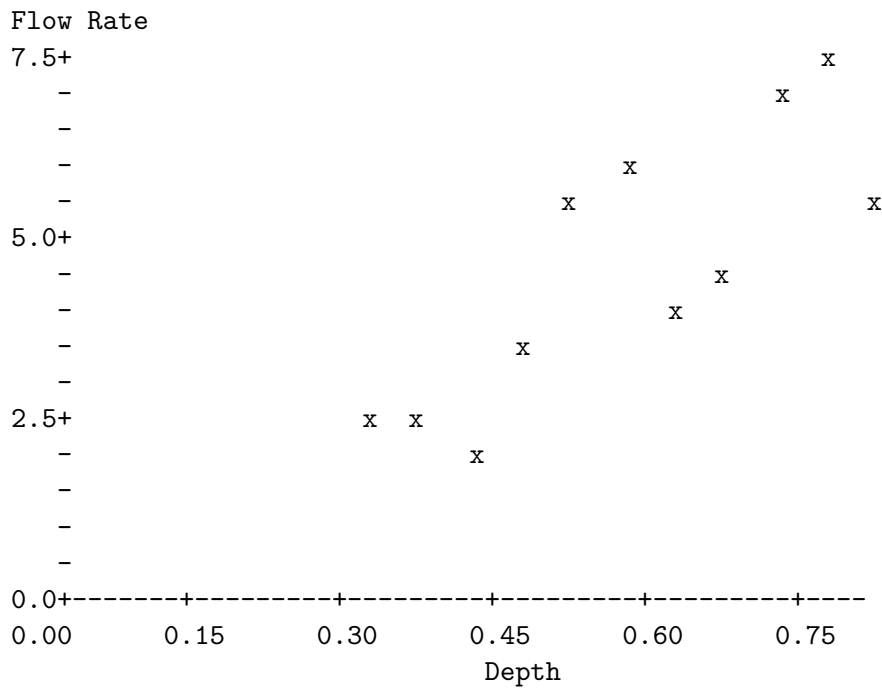
We are often interested in the relationship between two or more variables. This can arise from surveys in which several variables are measured on each unit or from experiments in which some variables are modified and other variables observed.

Data sets arising from these two types of situation cannot be distinguished in general, but **the interpretation is different**.

Example 9.1. A study for an environmental impact assessment measured the flow rate against the depth at a site on a stream.

Depth (m)	0.30	0.35	0.40	0.45
Flow Rate (m/s)	2.3	2.4	2.0	3.5
Depth (m)	0.50	0.55	0.60	0.65
Flow Rate (m/s)	5.7	6.1	4.2	4.7
Depth (m)	0.70	0.75	0.80	
Flow Rate (m/s)	6.9	7.4	5.7	

Note that the choice of which depths to use was made in advance; that is why they are spaced at fixed intervals.



The plot suggests a straight line relationship.

It is often useful to be able to predict the values of one variable from another variable. To do this, we need to formulate a model.

9.2 Linear Regression

A possible model is

$$\text{Flow} = \alpha + \beta \times \text{Depth}.$$

However, the model should allow for the variability that is present in the data. This can be achieved by stipulating

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

where $\varepsilon_i \sim N(0, \sigma^2)$ and independent.

This is called a linear regression model.

Notes:

- The model includes two parts: the functional part and the part which models the variability of the function.
- Many of the laws of physics and chemistry started out as empirical observations of this sort. The observed variability was often just “measurement error”.
- In other sciences, such as biology, there is often variability inherent in the material which is much greater than any “errors”.

Since a linear transformation of the normal random variable has a normal distribution we have

$$Y_i \sim N(\alpha + \beta x_i, \sigma^2)$$

and each observation is independent of the other observations.

9.2.1 Fitting the Model

We saw three methods for fitting parameters in Chapter 6; for this model we use the **Method of Least Squares** (LSE). Assuming the errors are normally distributed with constant variance then the Least Square Error and the Maximum Likelihood Estimators agree. This method is optimal for predicting the 'y' variable from the 'x' variable(s) if the model for the variability is as specified above.

Reminder: to fit the model, we minimise squared deviations in the 'y' direction. i.e. find $\hat{\alpha}$, $\hat{\beta}$ to minimise:

$$\sum_i (y_i - \alpha - \beta x_i)^2.$$

Notes:

- Predicting 'x' from 'y' gives different answers.
- If the 'x' values were chosen, different methods are needed if we wish to predict 'x' from 'y'. This is needed in assay systems.

We calculate:

$$\begin{aligned} S_{xx} &= \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n} \\ S_{xy} &= \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} \\ S_{yy} &= \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{n} \end{aligned}$$

$$\begin{aligned} \text{Note: } S_{yy} &= (n-1) \times \text{Variance}(y) \\ &= \sum (y_i - \bar{y})^2 \end{aligned}$$

Then:

$$\begin{aligned} \hat{\beta} &= \frac{S_{xy}}{S_{xx}} \\ \hat{\alpha} &= \frac{1}{n} \left(\sum y_i - \hat{\beta} \sum x_i \right) \\ &= \bar{y} - \hat{\beta} \bar{x} \end{aligned}$$

Thus the line goes through the point (\bar{x}, \bar{y}) .

The fitted line $y = \hat{\alpha} + \hat{\beta}x$ is said to be the **regression of Y on X**. In the motivating example, there was only one y value for each x value, but the method is also applicable when there are many y values (flow rate at different points with same depth).

If the model assumptions are correct, the fitted line $y = \hat{\alpha} + \hat{\beta}x$ gives the best estimate of y for any given value of x . This value is called the **fitted value at x** .

The model can also make predictions of y for values of x where no measurements were taken.

Example 9.2. A study for an environmental impact assessment measured the flow rate against the depth at a site on a stream. We want to fit the model

$$\text{Flow} = \alpha + \beta \times \text{Depth} + \varepsilon_i$$

where $\{\varepsilon_i\}$ are independent identically distributed $N(0, \sigma^2)$.

Depth (m)	0.30	0.35	0.40	0.45
Flow Rate (m/s)	2.3	2.4	2.0	3.5
Depth (m)	0.50	0.55	0.60	0.65
Flow Rate (m/s)	5.7	6.1	4.2	4.7
Depth (m)	0.70	0.75	0.80	
Flow Rate (m/s)	6.9	7.4	5.7	

We use the data to estimate the parameters α and β :

$$\begin{aligned}
 n &= 11 \\
 \sum x_i &= 6.05 \Rightarrow \bar{x} = 0.55 \\
 \sum y_i &= 50.9 \Rightarrow \bar{y} = 4.627 \\
 \sum x_i^2 &= 3.6025 \Rightarrow S_{xx} = 0.275 \\
 \sum x_i y_i &= 30.625 \Rightarrow S_{xy} = 2.63 \\
 \sum y_i^2 &= 271.59 \Rightarrow S_{yy} = 36.0618 \\
 \hat{\beta} &= \frac{S_{xy}}{S_{xx}} = \frac{2.63}{0.275} = 9.5636 \\
 \hat{\alpha} &= -0.6327
 \end{aligned}$$

The fitted model can be used to predict the flow rate for any specified depth: i.e. $\hat{y} = \hat{\alpha} + \hat{\beta}x$. However, the value of $\hat{\alpha}$ suggests that it could be dangerous to extrapolate from this model!

9.2.2 Residuals

Predictions and estimates of parameters are not very useful without some idea of their precision. This implies that we need to estimate the amount of variability about the fitted regression line. This is done by calculating the differences between the observed values and the fitted values. These are called the **residuals**. The sum of squared residuals is used to estimate σ^2 .

For instance, from the previous example we have:

Depth	Flow Rate	Fitted	Residual	Residual ²
0.30	2.3	2.24	0.06	0.0036
0.35	2.4	2.71	-0.31	0.0961
0.40	2.0	3.19	-1.19	1.4161
0.45	3.5	3.67	-0.17	0.0289
0.50	5.7	4.15	1.55	2.4025
0.55	6.1	4.63	1.47	2.1609
0.60	4.2	5.11	-0.91	0.8281
0.65	4.7	5.58	-0.88	0.7744
0.70	6.9	6.06	0.84	0.7056
0.75	7.4	6.54	0.86	0.7396
0.80	5.7	7.02	-1.32	1.7424
			0.00	10.8982

The total at the bottom right of the table is called the **Residual Sums of Squares**. It is sometimes denoted by SS_{RES} . We have

$$SS_{RES} = \sum (y_i - \hat{y}_i)^2,$$

where $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$. We can show that

$$\sum (y_i - \hat{y}_i)^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

and hence, for our data, the residual sums of squares can be found more easily (and more accurately) from

$$S_{yy} - \frac{S_{xy}^2}{S_{xx}} = 10.9094.$$

We started with $n = 11$ observations and have estimated 2 parameters (α, β) , so have 9 **degrees of freedom** left. We define the **Residual Mean Square** as

$$s^2 = \frac{1}{n-2} \sum (y_i - \hat{y}_i)^2,$$

which, for our data, gives

$$s^2 = \frac{10.9094}{9} = 1.2122.$$

Residual Mean Square is an estimate of σ^2 . Its square root s is called the **Residual Standard Deviation**. For our data we have $s = \sqrt{1.2122} = 1.101$

The quantity

$$\frac{S_{xy}^2}{S_{xx}}$$

is called the **Fitted Sums of Squares**. It is sometimes denoted by SS_{FIT}

Note that: $SS_{yy} = SS_{FIT} + SS_{RES}$

Note also: $S_{xx} \geq 0$ and $S_{yy} \geq 0$

$|S_{xy}| \leq$ at least one of S_{xx}, S_{yy}

9.2.3 Model checking

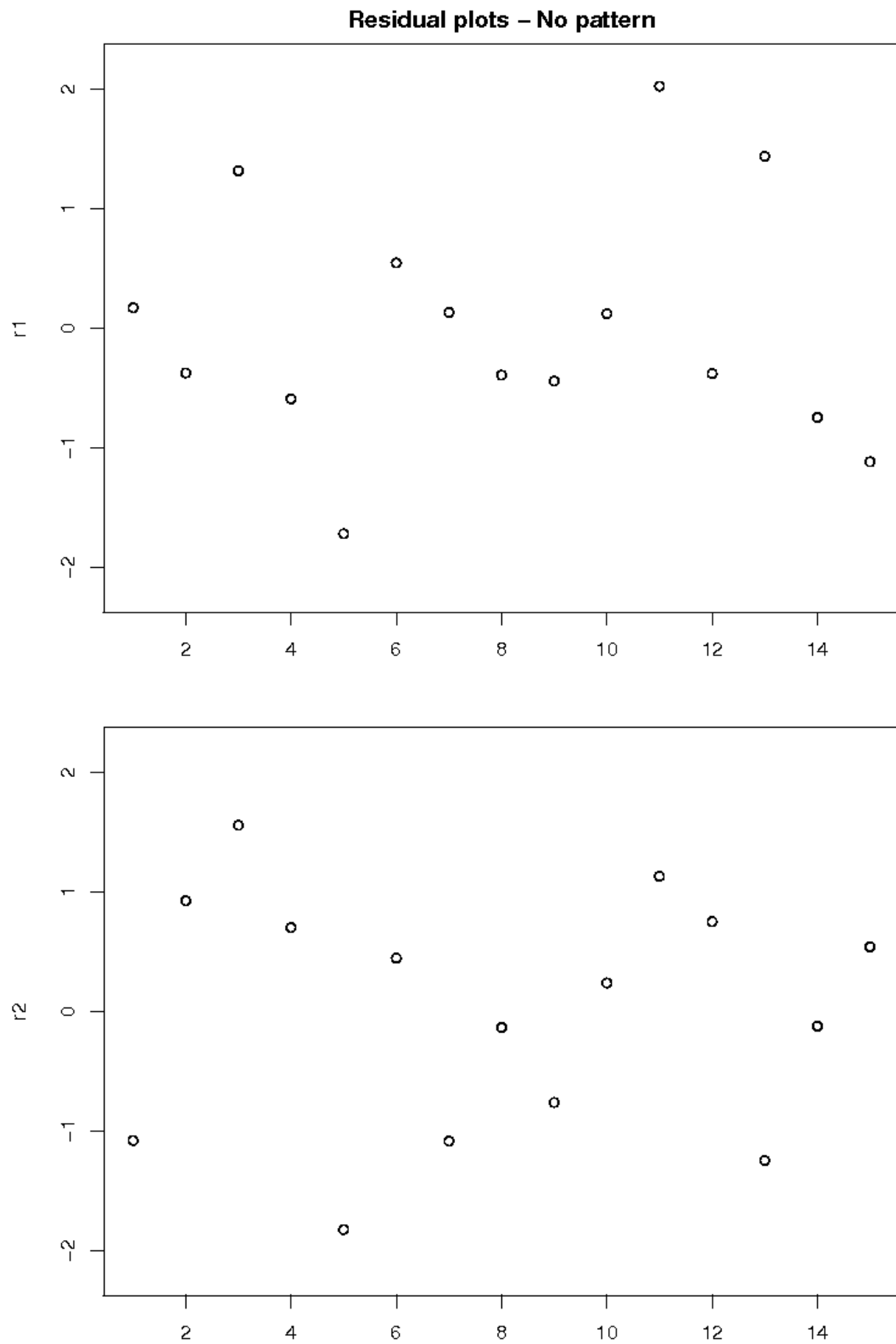
The above regression model has made various assumptions:

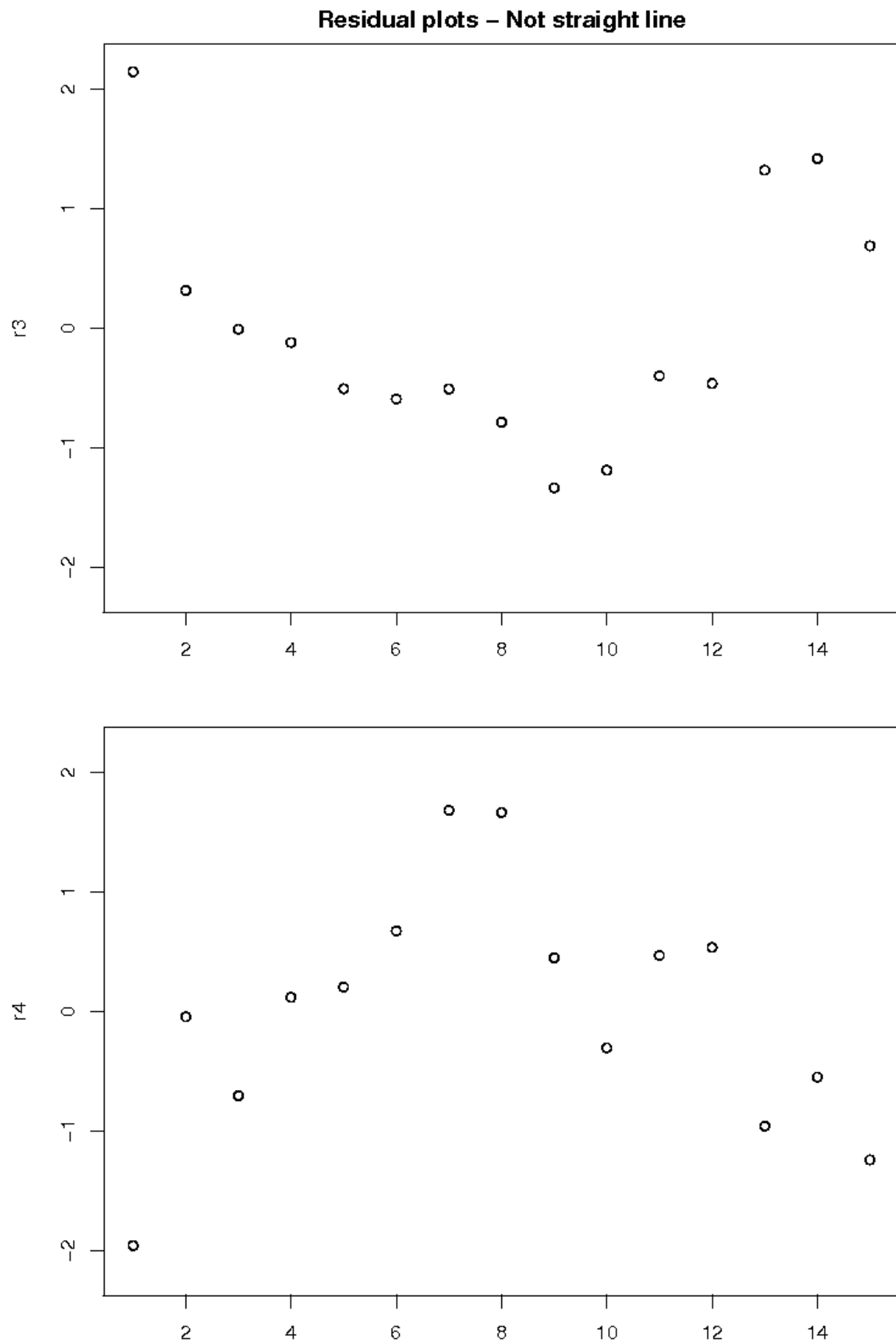
- The form of the relationship:
 $\text{Flow} = \alpha + \beta \times \text{Depth}$
- Variability about the relationship:
 - has constant standard deviation;
 - comprises independent observations;
 - follows a Normal distribution.

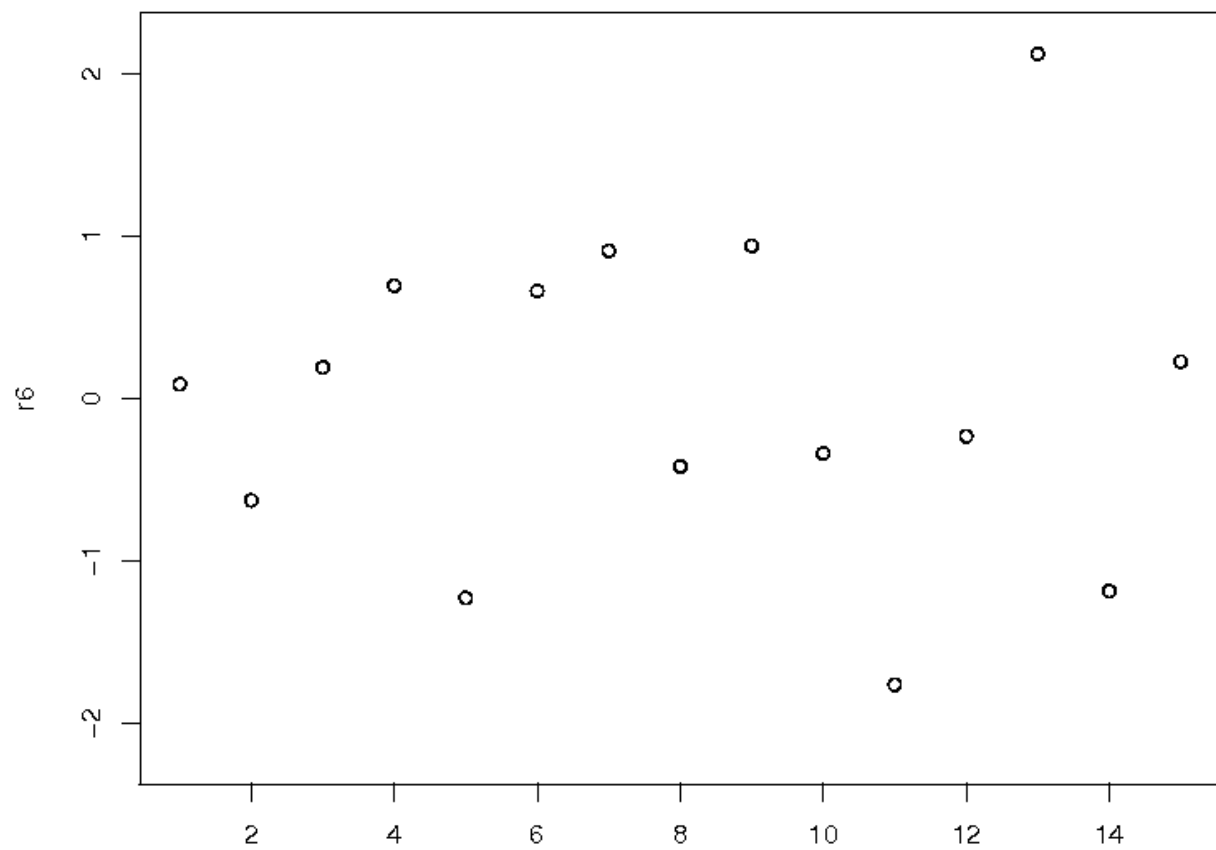
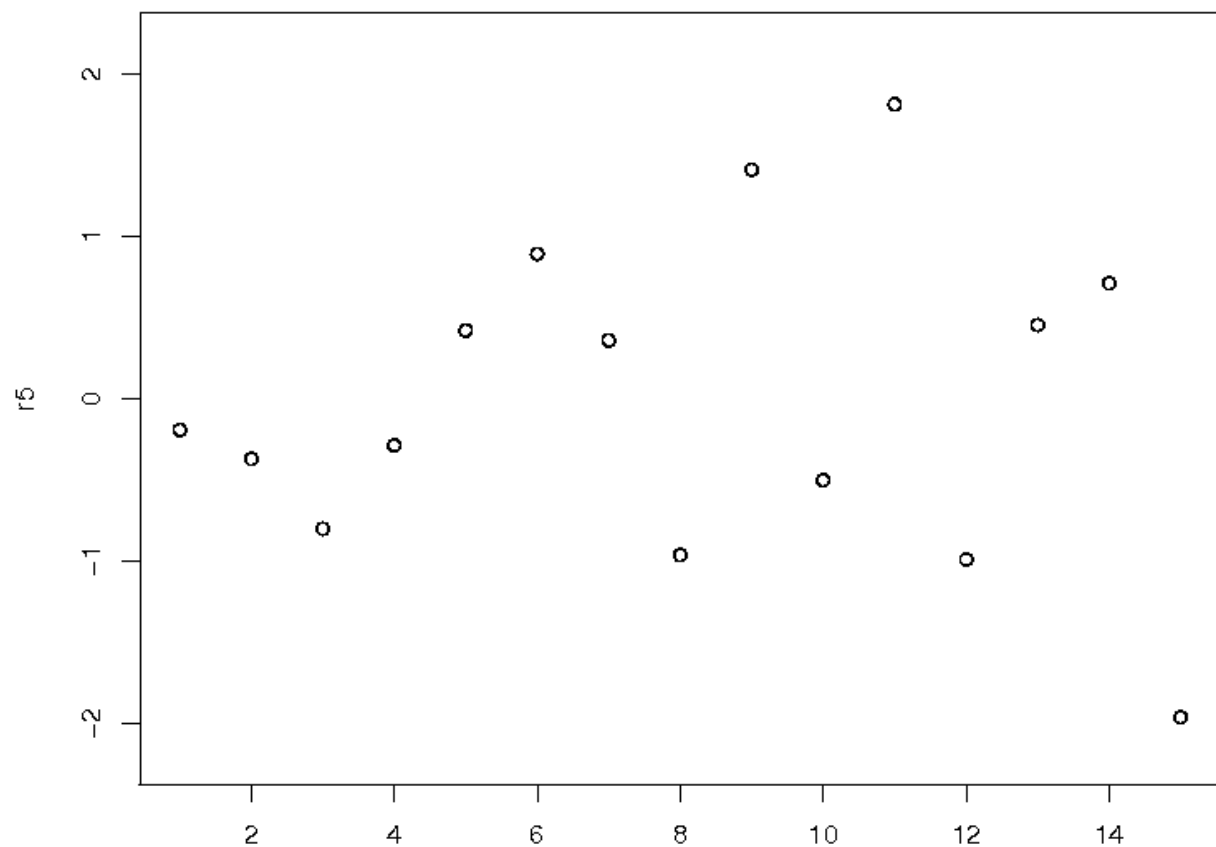
The model can be useful even if some of the assumptions fail, but it is important to be aware of the limitations. Thus if the form is wrong, the model can still be used for interpolation. However, it becomes dangerous to extrapolate. Failure of the first variability assumption means that the estimates are no longer optimal. Other failures mainly affect the tests and confidence intervals that will be covered later.

Plots of the residuals are often a good way of checking for incorrect assumptions. The most useful plots are usually of the residuals against the ‘x’ variable(s) or against the fitted values. These can often detect either an inappropriate function or variability not constant. It is sometimes possible to detect a failure of the assumption of independent observations, but this is much harder. The next three pages provide examples of the residuals plotted against the ‘x’ variable. The first page shows two examples where the residuals meet the conditions, the second page shows residuals with an unaccounted for trend and finally the third page shows examples with increasing variability.

Note: Many of the regression calculations depend on intermediate values. For example, $\hat{\alpha}$ depends on $\hat{\beta}$; fitted values depend on both of these. It is important that these intermediate values are not rounded off too early. Thus the slope estimate might be reported as 9.56, but a more accurate value such as 9.563636 should be used in the calculation of fitted values.





Residual plots – Variability increasing

9.2.4 Terminology and Warning

The example of linear regression we have examined has assumed that the depths are fixed and the flow rate was measured. It is important that the fixed value is used as the X value and the measured value used as the Y value.

The estimated slope of the regression line is:

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}}$$

This is not symmetrical in X and Y , so regressing depth on flow rate will give different answers. Whether these estimates are sensible or useful depends on the practical context.

If X is fixed and Y is then measured, regressing X on Y is **always** misleading. This can easily be seen by considering an extreme example:

Suppose several Y measurements are taken at a single value of X , then there is no information about how Y changes as X changes. In this case $S_{xx} = S_{xy} = 0$ and $S_{yy} > 0$. The regression of Y on X gives an indeterminate slope ($\frac{0}{0}$) which is a reasonable conclusion. However, the regression of X on Y gives a slope of zero, implying no change!

Regression methods use one or more variables to predict another variable. This course uses the terms **predictor variables** for those being used to make a prediction and **response variable** for the one that is being predicted.

The methods are used in many different application areas and these areas may use other names.

The **response** variable can also be called:

dependent variable, criterion variable, measured variable, endogenous variable or regressand.

Predictor variables can also be called:

independent variables, input variables, covariates, explanatory variables, exogenous variables or regressors.

It is recommended that the terms dependent and independent should be avoided when talking about variables in regression models. This is because the same terms are used in probability with different meanings.

9.2.5 Standard Errors, Confidence Intervals and Hypothesis testing

As described previously, we use the residual mean square as the estimate of σ^2 which is given by

$$s^2 = \frac{1}{n-2} \left(S_{yy} - \frac{S_{xy}^2}{S_{xx}} \right).$$

It can be shown that

$$\frac{(n-2)s^2}{\sigma^2} \sim \chi_{n-2}^2,$$

where n is the number of samples. In addition s^2 , $\hat{\alpha}$ and $\hat{\beta}$ are independent. Furthermore, since both $\hat{\alpha}$ and $\hat{\beta}$ are linear combinations of $\{y_i\}$, they are normally distributed.

Inference on the slope, β

It can be shown that

$$\hat{\beta} \sim N \left(\beta, \frac{\sigma^2}{S_{xx}} \right)$$

Since

$$\frac{(n-2)s^2}{\sigma^2} \sim \chi_{n-2}^2,$$

we find that

$$\frac{\hat{\beta} - \beta}{ese(\hat{\beta})} \sim t_{n-2},$$

where t_{n-2} is the Student t-distribution with $n-2$ degrees of freedom and

$$ese(\hat{\beta}) = \sqrt{\frac{s^2}{S_{xx}}}.$$

From this we can obtain confidence intervals for β , for example a two sided $100(1-\alpha)\%$ confidence interval is given by

$$\hat{\beta} \pm t_{\alpha/2} \times ese(\hat{\beta}), \quad ese(\hat{\beta}) = \sqrt{\frac{s^2}{S_{xx}}},$$

where $t_{\alpha/2}$ is the upper $\alpha/2$ percentage point of t_{n-2} . In addition we can use this as the test statistic for a hypothesis test on the value of β . The null hypothesis $H_0: \beta = 0$ is the no linear relationship hypothesis. If this hypothesis is not rejected one should discard the x information and base inferences on the y_i 's alone.

Inferences on a mean response

At $X = x_0$, the mean response is $\mathbb{E}[Y|x_0] = \mu_0 = \alpha + \beta x_0$, which is estimated by $\hat{\mu}_0 = \hat{\alpha} + \hat{\beta}x_0$. This is normally distributed and unbiased with

$$V[\hat{\mu}_0] = \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right] \sigma^2.$$

From this and the distribution of the sample variance we obtain

$$\frac{\hat{\mu}_0 - \mu_0}{ese(\hat{\mu}_0)} \sim t_{n-2}$$

where

$$ese(\hat{\mu}_0) = \sqrt{\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right] s^2}.$$

From this we can obtain confidence intervals for μ_0 , for example a two sided $100(1 - \alpha)\%$ confidence interval is given by

$$\hat{\alpha} + \hat{\beta}x_0 \pm t_{\alpha/2} \times \sqrt{\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right] s^2}$$

where $t_{\alpha/2}$ is the upper $\alpha/2$ percentage point of t_{n-2} . In addition we can use this as the test statistic for a hypothesis test on the value of μ_0 .

A special case of importance occurs if we consider $x_0 = 0$ then $\mu_0 = \alpha$. So we have

$$\frac{\hat{\alpha} - \alpha}{ese(\hat{\alpha})} \sim t_{n-2}$$

where

$$ese(\hat{\alpha}) = \sqrt{\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right] s^2}.$$

This can be used to produce a confidence interval for α and also test the null hypothesis $H_0 : \alpha = 0$ (the linear regression should pass through the origin).

Inferences on an individual response

Suppose we want to estimate/predict the value of an individual response (not the expected/mean response) at $X = x_0$. The estimate is the same as above, that is $\hat{y}_0 = \hat{\alpha} + \hat{\beta}x_0$, but we are in a more uncertain situation now, since we are estimating a particular response at x_0 , rather than the mean response at x_0 . The standard error of estimation is greater. We must now include the variation about the mean response, and the variance of our prediction is now

$$\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right] \sigma^2$$

from which we get the estimated standard errors of the prediction.

9.3 Correlation

Regression can also be used if a **random sample** is taken from a population and two variables are measured on each unit of the sample. In this case, one variable can often be used to predict the values of the other variable. The variable being predicted is taken as Y in the formulae above.

Example: Suppose that we take a random sample of students and measure heights and weights. Then either measurement could be used to predict the other:

$$\text{Weight} = \hat{\alpha}_H + \hat{\beta}_H \text{Height}$$

$$\text{Height} = \hat{\alpha}_W + \hat{\beta}_W \text{Weight}$$

We may not wish to predict one variable from the others. We may be just interested in whether they go up or down together. Correlation measures this.

There are several different ways of calculating a correlation coefficient. The best known of these is Pearson's r .

$$r = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}$$

This is related to the regression slopes that could be calculated when predicting Y from X and when predicting X from Y .

$$\hat{\beta}_X \hat{\beta}_Y = \frac{S_{XY}^2}{S_{XX}S_{YY}} = r^2$$

Notes:

1. $-1 \leq r \leq 1$
2. $r = +1 \Rightarrow$ points lie exactly on a line with a positive slope.
 $r = -1 \Rightarrow$ points lie exactly on a line with a negative slope.
3. $r = 0 \Rightarrow$ No association: Possible independence.
4. It is possible to test if a sample correlation coefficient could be zero, or another specified value, but this is rarely useful in practise.

If two variables have a non-zero correlation, it does not mean that either causes the other. For example, in Holland (Netherlands) there is a strong positive correlation between the number of babies born each year and the number of breeding white storks – both have declined.

If the x values have been selected, it is still possible to calculate r , but it is **meaningless**. This is because the wider the spacing of the x values, the closer r^2 gets to 1, unless x and y are independent of each other.

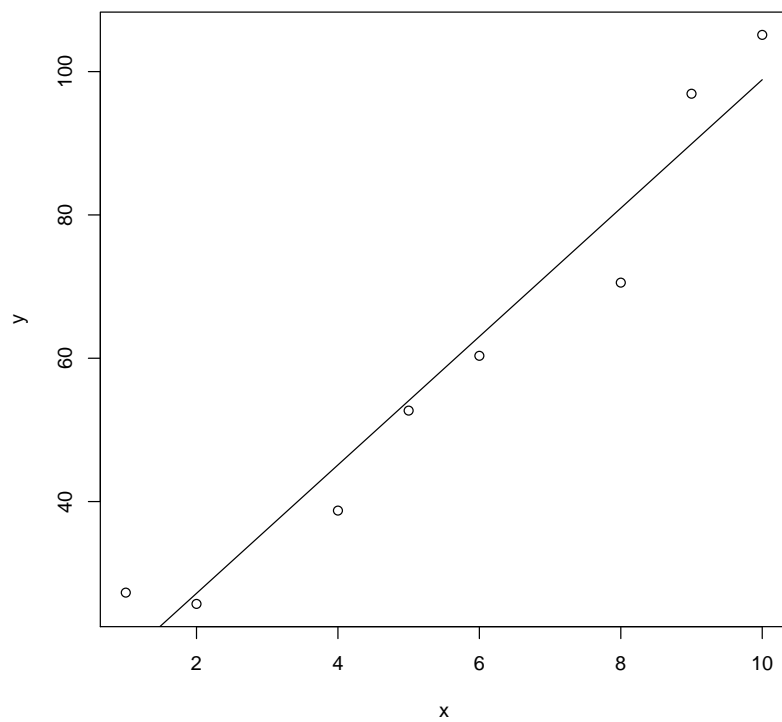
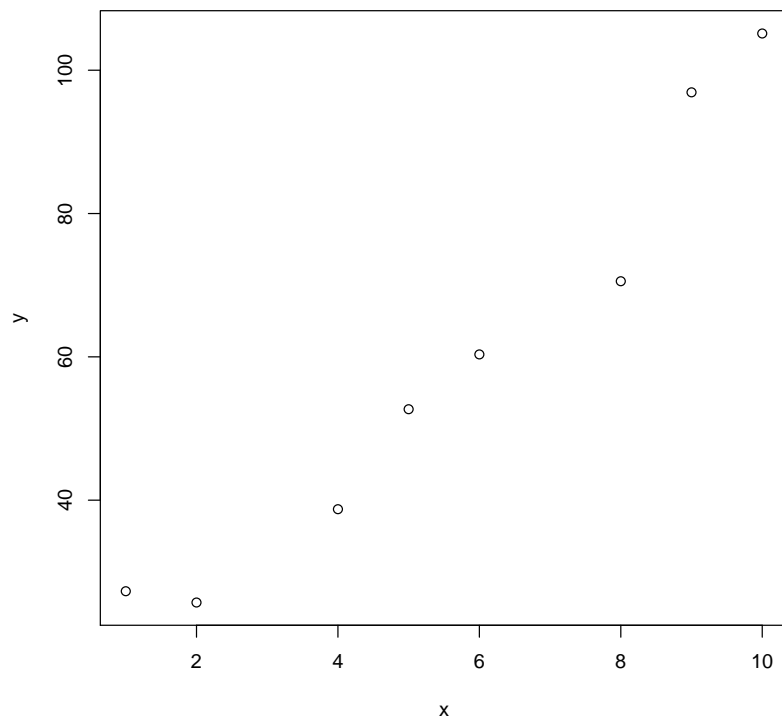
9.3.1 Linear Regression Examples

Example 9.3. A steel manufacturing company is interested in the relationship between the load bearing capacity of steel girders and the steel's thickness. They take a sample of 8 girders of various steel thicknesses, test their maximum load and obtain the following data.

Steel Thickness, T (cm)	1	8	10	5	6	2	9	4
Maximum Load, L (kg)	27.3	70.5	105.1	52.7	60.3	25.7	96.9	38.7

$$\sum T^2 = 327, \sum L^2 = 34722.72, \sum TL = 3345.9.$$

- Find the median and interquartile range for the Maximum Loads.
- Produce a scatter plot of Steel Thickness versus Maximum Load.
- Calculate S_{TT} , S_{LL} , S_{TL} , the corrected sums of squares and cross products for Maximum Loads and Steel Thickness.
- State the assumptions used in the standard linear regression model.
- Calculate estimates of the slope (β) and intercept (α) of the regression line $L = \alpha + \beta T$.
- Add the calculated regression line to the scatter plot produced previously.
- Find a 95% confidence interval for the expected Maximum Load for girders with Steel Thickness of 7cm.

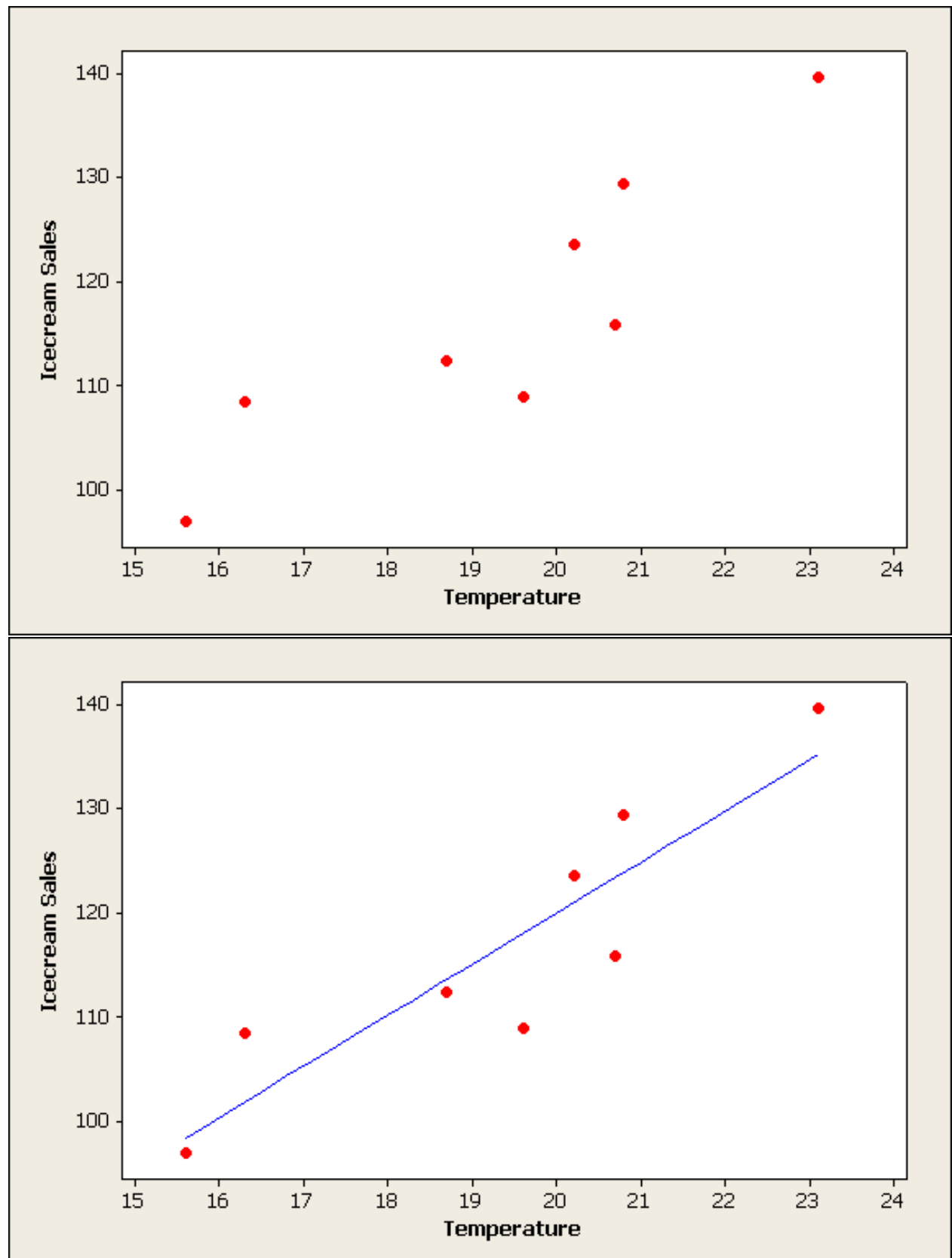


Example 9.4. An ice cream manufacturer is interested in the relationship between temperature at 9am and the daily sales figure that day. It collects data from 8 random days over the summer.

Temperature, x , (°C)	23.1	20.8	19.6	20.7	15.6
Ice cream sales, y , (£)	139.6	129.4	109.0	115.8	97.0
Temperature, x , (°C)	20.2	16.3	18.7		
Ice cream sales, y , (£)	123.6	108.5	112.4		

$$\sum x^2 = 3045.68, \sum y^2 = 110615.1, \sum xy = 18330.1.$$

- Produce a scatter plot of Temperature versus Ice cream sales.
- Calculate S_{xx} , S_{xy} and S_{yy} , the corrected sums of squares and cross products for Temperature and Ice cream sales.
- Calculate the Pearson's correlation coefficient and comment on the result.
- The ice cream manufacturer is interested in predicting sales from the temperature at 9am. Carry out a linear regression of Ice cream sales on Temperature.
- Add the regression line to your previously plotted graph.
- State the assumptions used in the standard linear regression model.
- Find the predicted mean volume of Ice cream sales and associated confidence interval given the Temperature is 22 °C.
- Would the confidence interval for the predicted mean volume of sales be larger or smaller for a Temperature of 25 °C rather than 22 °C?



9.4 Multiple Regression

Suppose that we have a response and that there are 2 possible predictors. Then we could consider the model:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \quad \text{where } \varepsilon \sim N(0, \sigma^2)$$

As before, we can use the method of least squares to fit this model. Solving the resulting equations is straightforward but requires matrix algebra and does not give ‘simple’ formulae. So we use the computer!

A model with two predictors is easy to understand – it corresponds to finding a plane surface in 3 dimensions. However, deciding whether this model is an improvement on using one of the single predictor models is not always straightforward.

In most cases, the two predictors x_1 and x_2 will not be independent (i.e. $r \neq 0$), which can make interpretation difficult.

1. Compare the two models:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 \quad \text{and} \quad y = \alpha + \beta_1 x_1$$

In general, the two slope estimates $\hat{\beta}_1$ will be different. The two slope estimates can even differ in sign!

2. If the correlation of x_1 and x_2 is high (positive or negative), then the model with both present may suggest that neither is a useful predictor, even if both models containing just one of the predictors are useful.

A common example of the use of two predictors is where $x_2 = x_1^2$.

Note that these are only independent if $\bar{x}_1 = 0$

9.4.1 ANOVA

Most computer programs that carry out regression calculations give a summary of some of these calculations called an ANOVA table. It is not very useful in simple linear regression, but is useful in more complicated situations, such as the analysis of experiments and surveys.

For the stream data, the table will look something like:

Source	df	ss	ms	F	P
Depth	1	25.1524	25.1524	20.75	0.0014
Resid.	9	10.9095	1.2122		
Total	10	36.0618			

- The Total row gives the variability (S_{YY}) in the variable being analysed (Flow Rate).
- The Residual row gives the variability about the fitted model. This is sometimes labelled ‘Error’.
- There is usually a row for each term in the model giving the increase in variability if that term is dropped. This can be used to help decide whether it is possible to simplify the model. The column labelled ‘df’ gives the degrees of freedom.
- The column labelled ‘ss’ gives the corrected sums of squares.
- The column labelled ‘ms’ gives the mean square. This is just the the sums of squares divided by the corresponding degrees of freedom.

- The column labelled ‘F’ gives the ratio of the mean square to the residual mean square. This is a test statistic – values larger than one suggest that that term in the model make a useful contribution to the model.
- The column labelled ‘P’ is the significance probability corresponding to the F test statistic. This column may be omitted.
- The residual mean square is the estimate of s^2 that is used in the calculation of all standard errors.

9.4.2 Multiple Linear Regression Case Study

Suppose that we have a response and that there are multiple possible predictors. Then we could consider the model:

$$y = \alpha + \sum_i \beta_i x_i + \varepsilon \quad \text{where } \varepsilon \sim N(0, \sigma^2)$$

Here, each x_i represent a different predictor.

Computer programs can use the method of least squares to fit this model. In practice we usually need to strike a balance between two competing objectives:

- We wish to omit variables that are unimportant because this leads to better predictions;
- We do not wish to leave out important variable because this leads to worse predictions.

This choice is often not straightforward. The following example illustrates some of the problems that can arise.

Data

Data were collected on cereal yields (litres/hectare), spring rainfall (mm) and spring temperature (°C) for 12 years from an arid country in the Middle East. The aim was to investigate the effect of rainfall on yield, to see whether it would be worth introducing an irrigation scheme.

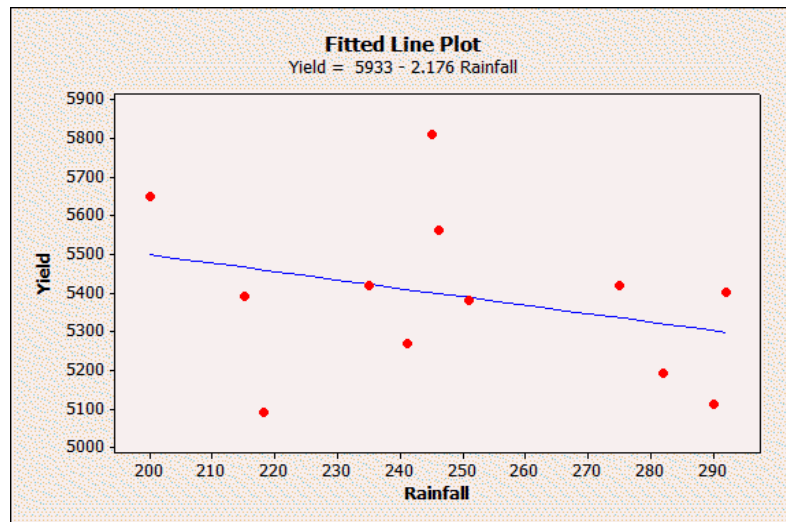
Yield	Rainfall	Temperature
5390	215	11.7
5190	282	9.0
5090	218	10.8
5380	251	10.1
5420	235	10.4
5110	290	8.8
5810	245	12.1
5400	292	8.4
5560	246	10.9
5270	241	10.3
5420	275	9.2
5650	200	12.4

Analysis

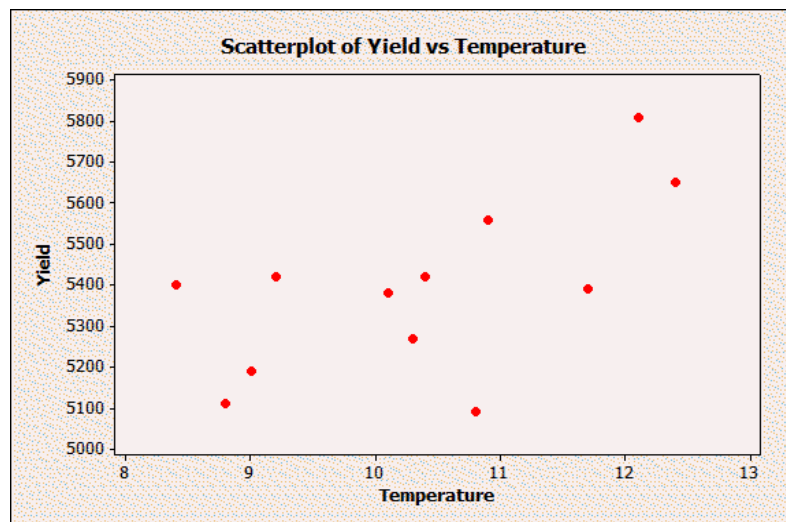
We start by fitting yield against rainfall as this is the problem of interest. The fitted line is: $\text{Yield} = 5933 - 2.176 \text{ Rainfall}$

This has a negative slope, which was not expected!

A plot of Yield against Rainfall suggests that the assumptions of the simple regression model are reasonable.



If the only available data was on rainfall and yield, the researchers would have had to decide either that irrigation was not useful or that more data was needed. The standard error of the slope estimate is 2.110, so the true value could be positive even though our estimate is negative. Temperature information is available and a plot shows a rather stronger relationship with yield.



If both terms are included in the model, then the prediction equation is:
Yield = 531 + 8.295 Rainfall + 270 Temperature

Source	df	ss	ms	F	P
Rain	1	140029	140029	7.395	0.024
Temp.	1	279432	279432	14.756	0.004
Resid.	9	170427	18936		
Total	11	497692			

The ANOVA table shows that the significance probability for temperature is small, so this variable appears to be a useful predictor of yield; the value of the coefficient is positive, which implies that higher temperatures lead to larger yields.

The significance probability for rainfall is 0.024, so there is evidence that rainfall does have an effect on yield, with higher rainfall giving greater yields.

The table summarises the correlations between the three variables.

yield	1.000		
rainfall	-0.310	1.000	
temperature	0.613	-0.8941	1.000
	yield	rainfall	temperature

The correlation between rainfall and temperature is large and negative. This is not really surprising because if it is raining, the sun is not shining. In that region at that time of year, sunny implies higher temperatures.

Conclusions

Higher temperatures and increased rainfall both had a positive effect on cereal yield. Temperature appears to have a greater effect on yield than rainfall.

9.4.3 Model Selection

Finally, there is no need to restrict ourselves to just two predictors; the same methods can be used with many predictors, but the choice of a suitable model becomes harder. The usual aim is to try to find the simplest model that will give good predictions. We need to balance two conflicting objectives; we do not want to omit anything important because this will bias the answers, and we do not wish to include anything unimportant because this will give poorer predictions.

If there are k possible predictors, there will be 2^k linear regression models that could be fitted. This number becomes large rather quickly. Thus computer methods are essential.

For example, if $k = 9$, there are 512 different models available. If the only models considered are those involving a few different predictors, the numbers can still be large – there are $\binom{9}{3} = 84$ models using 3 predictors.

If we only consider models that involve the same number of predictors, then the obvious choice is the model that gives us the best predictions. In other words, choose the model with the smallest residual mean square. Equivalently, we use:

$$R^2 = 1 - \frac{\text{Residual sums of squares}}{\text{Total sums of squares}}$$

R^2 is called the coefficient of determination. For linear regression, it is the square of the correlation between the observed values and the fitted values.

An ANOVA table and its F tests **may** allow us to compare the models with different numbers of predictors. However, this is **only valid if the models being compared are nested**. When there are many possible predictors, the best model involving k of these often does not contain the best model involving $k - 1$ of them as a special case.

A number of criteria have been suggested to allow the choice of model to be automated. Some of them are described below. There is no general agreement as to which of these is best.

In the methods below, n is the number of units and k is the number of parameters in the model. A popular method is to use ‘Adjusted R^2 ’. When an extra term is added to a model, the value of R^2 always increases. The adjustment tries to allow for this.

$$\text{Adjusted } R^2 = 1 - \frac{\text{Residual mean square}}{\text{Total mean square}}$$

If this is plotted against k , the curve typically flattens out. The method chooses k to be the smallest number on the flatter section.

Note: Adjusted R^2 does **not** have the same interpretation as R^2 . In particular, it can take negative values!

Another criterion is Mallows' C_p . This compares the residual sums of squares from a model with k parameters with the residual mean square s^2 obtained by using **all** available predictors.

$$C_p = \frac{\text{Residual sums of squares}}{s^2} - n + 2k$$

A common way to choose the value of k is to take that with the smallest value of Mallows' Statistic C_p , as long as $C_p < k$.

Alternatives exist, for example the AIC (Akaike Information Criterion).

F20SA / F21SA Statistical Modelling and Analysis

Chapter 10: Bayesian inference

Contents

10.1 Bayes' Theorem, priors, likelihoods and posteriors	10–1
10.1.1 A simple example	10–1
10.1.2 Estimating parameters using Bayesian inference	10–2
10.1.3 Example of Bayesian inference.	10–3
10.2 Reporting conclusions from a Bayesian analysis	10–6
10.2.1 Deriving Bayesian estimators	10–6
10.2.2 Some examples	10–8
10.2.3 Reporting an interval	10–9
10.2.4 More on selecting priors	10–11
10.3 Predictive distributions	10–13

10.1 Bayes' Theorem, priors, likelihoods and posteriors

Bayesian inference is now a major area in statistics and is used in many areas including biology, medicine, engineering, physics and epidemiology. At one time nearly all statisticians were classical (frequentist) and Bayesian methods were little used. However, thanks to the advent of powerful computing resources, Bayesian methods can be applied to many real-world problems that could not easily be tackled using classical methods.

10.1.1 A simple example

The key idea in Bayesian statistics is that all quantities about which there is uncertainty should be represented as random variables and characterised using distributions. Consider the following toy example:

A particular coin of the realm is known to exist in two forms. The majority of coins (type 1) have a “H” and a “T” but a small proportion (1%) have been minted with 2 “H” (type 2). It is known that the coins, when tossed, will land one side or the other with equal probability.

I take a coin from my pocket and toss it 3 times, obtaining 3 Hs. How can you analyse the result of this experiment to infer the particular type of coin I have tossed?

Let p denote the probability of getting a ‘H’ with my coin. Before tossing the coin you know that

$$p = 0.5 \text{ OR } p = 1.0$$

depending on whether the coin is type 1 or type 2. If you assume that the coin is randomly drawn from the population of coins then

$$\begin{aligned}\mathbb{P}(p = 0.5) &= 0.99 \\ \mathbb{P}(p = 1.0) &= 0.01.\end{aligned}$$

For the observed outcome from the experiment you know that

$$\begin{aligned}\mathbb{P}(HHH|p = 0.5) &= 0.125 \\ \mathbb{P}(HHH|p = 1.0) &= 1.0.\end{aligned}$$

You can now calculate the following conditional probabilities using Bayes' Theorem:

$$\begin{aligned}\mathbb{P}(p = 0.5|HHH) &= \frac{(0.99 \times 0.125)}{(0.99 \times 0.125 + 0.01 \times 1.0)} \\ &= 0.925(3dp)\end{aligned}$$

Also

$$\mathbb{P}(p = 1.0|HHH) = 0.075(3dp)$$

This calculation is an illustration of Bayesian reasoning and the main steps in achieving it are as follows:

- (A) You began with an initial (prior) belief about the value of p .
- (B) You carried out an experiment.
- (C) You computed the conditional distribution of p given the observed outcome 'HHH'. (This is the 'posterior' distribution of p .)

In Bayesian inference, we express our conclusions from an experiment in the form of a probability distribution for the quantity(ies) or parameters of interest. Before the experiment your prior belief was: $\mathbb{P}(\text{Type 1}) = 0.99, \mathbb{P}(\text{Type 2}) = 0.01$. Denote this probability function as $\pi(p)$, where $p \in \{0.5, 1\}$. We refer to this as the *prior distribution of p* .

After observing data $\mathbf{x} = 'HHH'$ your posterior belief becomes: $\Pr(\text{Type 1}) = 0.925; \Pr(\text{Type 2}) = 0.075$. We denote this as $\pi(p|\mathbf{x})$ and we call it the *posterior distribution of p* .

How did you go from prior to posterior? Note that the above calculation of the posterior could be written as

$$\pi(p|\mathbf{x}) = \frac{\pi(p)\mathbb{P}(\mathbf{x}|p)}{\sum_{q \in \{0.5, 1\}} \pi(q)\mathbb{P}(\mathbf{x}|q)}$$

Note also that $\mathbb{P}(\mathbf{x}|p)$ is just our likelihood function $L(p; \mathbf{x})$ so that you obtain your 'posterior' distribution $\pi(p|\mathbf{x})$ by multiplying the prior $\pi(p)$ by $L(p; \mathbf{x})$ and 'normalising' to turn it into a probability mass function. This, in essence, is how Bayesian inference works and all the calculations we subsequently do are based on this idea.

To a classical statistician, there is nothing contentious in the above example since the value of p could be considered to be generated through a repeatable experiment (random sampling of a coin from the whole population). On the other hand Bayesian inference requires that probability distributions need to be assigned to any unknown parameter of interest even when its value has not been assigned through some repeatable sampling process. This is seen by some statisticians as being controversial as it introduces subjective belief into the process of statistical inference.

10.1.2 Estimating parameters using Bayesian inference

We give a general description of the role of Bayes' theorem in Bayesian inference when estimating parameters in a statistical model.

Bayesian Inference: Suppose that we carry out an experiment where the outcome \mathbf{x} has probability function $f(\mathbf{x}; \theta)$ where θ is an unknown parameter. Suppose my prior belief about θ is

described by the probability function $\pi(\theta)$. Then, on observing \mathbf{x} my posterior belief about θ is given using Bayes' theorem as:

$$\pi(\theta|\mathbf{x}) = \frac{\pi(\theta)L(\theta;\mathbf{x})}{\int \pi(\theta')L(\theta';\mathbf{x})d\theta'}, \quad (**)$$

where $L(\theta;\mathbf{x}) = f(\mathbf{x};\theta) = \pi(\mathbf{x}|\theta)$ is the likelihood.

We have written the above rule for the case where θ is a continuously-valued parameter so that $\pi(\theta)$ is a density. If $\pi(\theta)$ is a discrete distribution as in the above example with the coin, then the integral on the denominator just becomes a sum. Conceptually, there is no real difference between the situations.

Some things to note:

- In Bayesian statistic we tend to use $\pi()$ to denote a probability density or mass function (be careful not to get confused with the number π !)
- Although the definition of $\pi(\theta|\mathbf{x})$ involves an integral over the whole parameter space on the denominator often we don't need to be able to compute this integral (see later).
- To calculate the posterior distribution you first need to identify a prior $\pi(\theta)$. This is the main source of controversy in Bayesian statistics (see later).
- The only way in which the observations enter Bayesian inferences is through the likelihood function. Therefore Bayesian methods respect the likelihood principle.

10.1.3 Example of Bayesian inference.

Example 10.2.1. You want to estimate the proportion of the population who intend to vote Labour in a forthcoming council election in your town. To do this you select a random sample of $n = 100$ names from the voting register and contact them to ask their voting intentions. At the last election 40% of the people in the town voted Labour. In your survey you find that $x = 25$ people in your sample say they will vote Labour. Carry out a Bayesian analysis using this information.

Clearly the sampling distribution for x is $\text{Bin}(n, p)$ where p is the unknown proportion of Labour voters. We can therefore identify our likelihood as:

$$L(p; x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

We need to specify our prior distribution for p . This should be done before the data are collected. Since p is a probability, we need to find a distribution whose range is the interval $(0, 1)$. Now we have prior information from last election, but the picture may have changed since then - so how could we use this information in selecting a prior?

Maybe we could select $\pi(p)$ to have mean 0.4 but some variance around that value? This would then allow for the possibility that the picture had changed. What about using a Beta distribution of some sort?

Let's look at $\text{Beta}(4, 6)$ by plotting its density.

```
> x = rpoints(1000)
> plot(x, dbeta(x, 4, 6), type = "l")
```

This distribution has mean 0.4 but also considerable variation around this value, representing uncertainty in true proportion - so let's go with this choice of prior. Now we consider the observation $x = 25$ and derive the posterior using the formula (**).

We can see without calculating the normalising constant (integral on the denominator) that this must be a Beta(29, 81) distribution!

Exercise. Plot of density of posterior Beta(29, 81). Now compare this density with the prior Beta(4, 6).

Note how these plots illustrate the way in which our belief changes in response to the observation $x = 25$.

Further comments

- We didn't actually need to compute the normalising constant because the dependence on p in the numerator $\pi(p)L(p; r)$ distinguished the posterior as a Beta distribution!

For this reason Bayesians often specify the posterior distribution only up to a constant of proportionality.

Instead of (**) they simply write

$$\pi(\theta|\mathbf{x}) \propto \pi(\theta)L(\theta; \mathbf{x})$$

and then hope to identify the form of the posterior distribution from the right-hand-side.

- In Example 10.2.1 we selected our prior distribution using a combination of historical data and subjective judgement. How can we select priors in cases where we don't have any information? (see later)

Example 10.2.2. (Bayesian inference for $\text{Exp}(\lambda)$): In a certain population lifetimes are known to be exponentially distributed with parameter λ . We collect a random sample of n lifetimes $\mathbf{x} = (x_1, \dots, x_n)$. Carry out inference on λ using Bayesian methods.

The main steps are as follows:

- **Step 1:** We need to specify a prior distribution for λ representing belief about λ before any observations are taken. Given λ is positive, which family of distributions might be sensible to consider? What about Gamma(α, β)?

Set $\pi(\lambda) \sim \text{Gamma}(\alpha, \beta)$ where α, β are chosen to reflect your prior belief.

- **Step 2:** Note that the likelihood is:

$$L(\lambda; \mathbf{x}) = \lambda^n \exp\{-\lambda \sum_{i=1}^n x_i\}$$

- **Step 3:** Now derive the posterior

$$\pi(\lambda|\mathbf{x}) \propto \pi(\lambda) \times L(\lambda; \mathbf{x}).$$

Derivation:

- Can we recognise the form of posterior $\pi(\lambda|\mathbf{x})$? There is only one density on the positive real line that is proportional to $\lambda^{\alpha+n-1} \exp(-\lambda(\beta + \sum x_i))$.

That is the $\text{Gamma}(\alpha + n, \beta + \sum x_i)$ density, so this must be our posterior density $\pi(\lambda|\mathbf{x})$!

- The above derivation shows that for any choice of Gamma prior we will always end up with a Gamma posterior in example 10.2.2.

Example 10.2.3. Suppose now we have a random sample of size n , $\mathbf{x} = (x_1, \dots, x_n)$, from a $\text{Poisson}(\lambda)$ distribution where our prior belief regarding λ is represented as a $\text{Gamma}(\alpha, \beta)$ distribution. *What is the posterior density $\pi(\lambda|\mathbf{x})$?*

Derivation:

- Likelihood is given by:

$$L(\lambda; \mathbf{x}) =$$

- Therefore the posterior is

$$\pi(\lambda|\mathbf{x}) \propto \pi(\lambda) \times L(\lambda; \mathbf{x})$$

- By inspection we can recognise $\pi(\lambda|\mathbf{x})$ as a $\text{Gamma}(\alpha + \sum x_i, \beta + n)$ distribution!

Example 10.2.4. Let $\mathbf{x} = (x_1, \dots, x_n)$ be a random sample from a $N(\mu, \sigma^2)$ distribution where σ^2 is known. Suppose we use a $N(\alpha, \beta)$ prior for μ .

Derive the posterior:

$$\pi(\mu|\mathbf{x}) \propto \pi(\mu) \times L(\mu; \mathbf{x})$$

Note that our posterior mean for μ is a weighted average of the prior mean α and the observed sample mean \bar{x} . We can see how the choice of prior affects the posterior mean:

- Increasing n shifts the weight on to the observed sample mean \bar{x} .
- Increasing β (so that our prior is more *diffuse or vague*) shifts the weight on to \bar{x} .
- Decreasing β , so that we are more certain about μ in our prior belief, shifts the weight on to α .

Conjugacy: Examples 10.2.1-4 are all examples where the prior distribution and posterior distributions come from the same family. Identifying the posterior distribution was particularly easy in general, as the sample size and data enter into the expressions for the parameters of the posterior distribution via a straight forward expression! In these circumstances we say that the prior in question is the *conjugate* prior for the parameter/distribution.

- Example 10.2.1. The Beta distribution is the conjugate prior for p in Binomial(n, p).
- Example 10.2.2. The Gamma distribution is the conjugate prior for λ in Exp(λ).
- Example 10.2.3. The Gamma distribution is the conjugate prior for λ in Poi(λ).
- Example 10.2.4. The Normal distribution is the conjugate prior for μ in $N(\mu, \sigma^2)$ where σ^2 is known.

10.2 Reporting conclusions from a Bayesian analysis

Any Bayesian analysis produces a posterior density, $\pi(\theta|y)$, for a parameter of interest. As well as graphical representations of the $\pi(\theta|y)$ (e.g. a plot of the posterior density function) we can summarise it by reporting:

- the posterior mean;
- the posterior variance or standard deviation.

In examples 10.2.1-10.2.4 the posterior comes from a family whose mean and variance can be written down easily from standard formulae.

In the case where you have a lot of data (e.g. a large value of the sample size n) theoretical results (not described) show that the posterior density will typically be approximately Normal, so that the posterior mean and variance (more or less) completely characterise the distribution, since these quantities specify precisely the Normal distribution.

10.2.1 Deriving Bayesian estimators

Throughout this section we consider the case where our parameter θ is continuous and our belief about θ is expressed as a posterior density $\pi(\theta|\mathbf{x})$. **What value $\hat{\theta}$ should you report as estimate or summary of $\theta|\mathbf{x}$?**

The *Bayes Principle* says we should choose $\hat{\theta}$ so as to minimise the expected estimation error:

$$E_{\theta}\{\ell(\hat{\theta}, \theta)\} = \int \ell(\hat{\theta}, \theta) \pi(\theta | \mathbf{x}) d\theta,$$

where ℓ defines how we measure the estimation error.

The most widely used choices of ℓ , and the associated estimators, are:

1. Bayes estimator under the quadratic loss:

$$\ell(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2,$$

$$E_{\theta}\{\ell(\hat{\theta}, \theta)\} = \int (\hat{\theta} - \theta)^2 \pi(\theta | \mathbf{x}) d\theta,$$

and

$$\frac{d}{d\hat{\theta}} E_{\theta}\{\ell(\hat{\theta}, \theta)\} = 2 \int (\hat{\theta} - \theta) \pi(\theta | \mathbf{x}) d\theta.$$

Equating this to zero for minimum leads to

$$\hat{\theta} \int \pi(\theta | \mathbf{x}) d\theta = \hat{\theta} = \int \theta \pi(\theta | \mathbf{x}) d\theta.$$

So the Bayes' estimator $\hat{\theta}$ under this loss is the *mean* of the posterior distribution.

2. Bayes estimator under the absolute error loss:

$$\ell(\hat{\theta}, \theta) = |\hat{\theta} - \theta|,$$

$$\begin{aligned} E_{\theta}\{\ell(\hat{\theta}, \theta)\} &= \int |\hat{\theta} - \theta| \pi(\theta | \mathbf{x}) d\theta \\ &= \int_{-\infty}^{\hat{\theta}} (\hat{\theta} - \theta) \pi(\theta | \mathbf{x}) d\theta - \int_{\hat{\theta}}^{\infty} (\hat{\theta} - \theta) \pi(\theta | \mathbf{x}) d\theta \end{aligned}$$

Then

$$\frac{d}{d\hat{\theta}} E_{\theta}\{\ell(\hat{\theta}, \theta)\} = \int_{-\infty}^{\hat{\theta}} \pi(\theta | \mathbf{x}) d\theta - \int_{\hat{\theta}}^{\infty} \pi(\theta | \mathbf{x}) d\theta.$$

Equate this to zero for minimum:

$$\int_{-\infty}^{\hat{\theta}} \pi(\theta | \mathbf{x}) d\theta = \int_{\hat{\theta}}^{\infty} \pi(\theta | \mathbf{x}) d\theta.$$

So $\hat{\theta}$ is the *median* of the posterior distribution.

3. Bayes estimator under the 0 – 1 loss.

$$\ell(\hat{\theta}, \theta) = 0 \quad \text{iff} \quad \hat{\theta} = \theta, \quad \text{otherwise} \quad \ell(\hat{\theta}, \theta) = 1.$$

$$E_{\theta}\{\ell(\hat{\theta}, \theta)\} = \int \ell(\hat{\theta}, \theta) \pi(\theta | \mathbf{x}) d\theta.$$

Consider instead a small interval of length 2ε and a function $\ell_{\varepsilon}(\hat{\theta}, \theta) = 0$ if $\hat{\theta} - \varepsilon < \theta < \hat{\theta} + \varepsilon$, and $\ell_{\varepsilon}(\hat{\theta}, \theta) = 1$, otherwise. [Note that this becomes the required $\ell(\hat{\theta}, \theta)$ as $\varepsilon \rightarrow 0$]. Then

$$E_{\theta}\{\ell(\hat{\theta}, \theta)\} = 1 - \int_{\hat{\theta}-\varepsilon}^{\hat{\theta}+\varepsilon} \pi(\theta | \mathbf{x}) d\theta \approx 1 - 2\varepsilon \pi(\hat{\theta} | \mathbf{x}), \quad \text{for small } \varepsilon.$$

Clearly this is minimised by taking $\hat{\theta}$ to maximise $\pi(\hat{\theta} | \mathbf{x})$. So, $\hat{\theta}$ is the *mode* of the posterior distribution.

10.2.2 Some examples

Bernoulli/Binomial

Let $X = x$ be a single observation from $B(n, p)$. Let the prior distribution for an unknown parameter p be $Beta(\alpha_1, \alpha_2)$. Investigate the Bayesian estimation of p .

- Likelihood $\propto p^x(1-p)^{n-x}$.
- Prior $\propto p^{\alpha_1-1}(1-p)^{\alpha_2-1}$.
- Posterior $\propto p^{x+\alpha_1-1}(1-p)^{n-x+\alpha_2-1}$.

So, we may conclude that the posterior distribution is $Beta(x + \alpha_1, n - x + \alpha_2)$.

(a) Under quadratic loss, the Bayesian estimator is the mean, i.e.

$$\hat{p} = \frac{X + \alpha_1}{n + \alpha_1 + \alpha_2}.$$

(b) Under absolute error loss, the Bayesian estimator is the median. Unfortunately, it may be calculated numerically only.

(c) Under zero-one loss, the Bayesian estimator is the mode. It can be easily shown that it is

$$\hat{p} = \frac{X + \alpha_1 - 1}{n + \alpha_1 + \alpha_2 - 2}.$$

Proceed further with the quadratic loss estimator \hat{p} . It can be re-expressed as

$$\frac{n}{n + \alpha_1 + \alpha_2} \cdot \frac{X}{n} + \frac{\alpha_1 + \alpha_2}{n + \alpha_1 + \alpha_2} \cdot \frac{\alpha_1}{\alpha_1 + \alpha_2}$$

which is a *weighted average* of the *classical MLE* $\frac{X}{n}$ and the prior mean $\frac{\alpha_1}{\alpha_1 + \alpha_2}$.

Comments:

(1) Here X/n depends on the data only and not on the prior; and $\alpha_1/(\alpha_1 + \alpha_2)$ depends on the prior only, and not on the data.

(2) The weights are in the ratio $n : (\alpha_1 + \alpha_2)$.

This gives us an indication of the ‘value’ of the prior information in terms of a pseudo sample size, *i.e.*, if we felt our prior information suggested that p would be about $1/4$ and that this information was worth as much as having a sample of size 20, then we would take

$$\alpha_1 + \alpha_2 = 20, \quad \frac{\alpha_1}{\alpha_1 + \alpha_2} = \frac{1}{4}$$

which leads to $\alpha_1 = 5, \alpha_2 = 15$, *i.e.* to the prior $Beta(5, 15)$.

Normal distribution

Let \mathbf{x} be a random sample from $N(\mu, \sigma^2)$ distribution, with known σ^2 . Let the prior distribution for μ be $N(\mu_0, \sigma_0^2)$, with known μ_0 and σ_0^2 . Investigate the Bayesian estimation of μ .

Likelihood $\propto \exp\left(-\frac{n}{2\sigma^2}(\mu - \bar{x})^2\right)$.

Prior $\propto \exp\left(-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right)$.

We know that the posterior density is proportional to the product of the prior density and the likelihood.

After some simple manipulations, one can show that the posterior is

$$\propto \exp\left(-\frac{1}{2\sigma_1^2}(\mu - \mu_1)^2\right)$$

where

$$\mu_1 = \frac{\sigma_0^2 \bar{X} + \frac{\sigma^2}{n} \mu_0}{\sigma_0^2 + \frac{\sigma^2}{n}} \quad \text{and} \quad \sigma_1^2 = \frac{\frac{\sigma^2}{n} \cdot \sigma_0^2}{\frac{\sigma^2}{n} + \sigma_0^2}.$$

So, the posterior distribution for the parameter μ is $N(\mu_1, \sigma_1^2)$. We write this as

$$\mu | \mathbf{x} \sim N(\mu_1, \sigma_1^2).$$

Then, due to the symmetry of the normal distribution around its mean, all three loss functions give the same estimator:

$$\hat{\mu} = \frac{\sigma_0^2 \bar{X} + \frac{\sigma^2}{n} \mu_0}{\sigma_0^2 + \frac{\sigma^2}{n}}.$$

Again we can notice that this estimator is a *weighted average* of \bar{X} (this is the classical MLE which depends on data only) and of μ_0 (this comes from the prior), with weights in the ratio $\sigma_0^2 : \frac{\sigma^2}{n}$.

One can rewrite this in as:

$$\hat{\mu} = \frac{\frac{1}{\sigma^2/n} \cdot \bar{X} + \frac{1}{\sigma_0^2} \cdot \mu_0}{\frac{1}{\sigma^2/n} + \frac{1}{\sigma_0^2}}.$$

This form is better (why?).

10.2.3 Reporting an interval

Recall that confidence intervals could be used to report an interval for a given quantity, but that their interpretation was slightly problematic in the classical setting. Remember that in the classical setting, the parameter θ is fixed and it is the *interval* that is the random outcome. Therefore, the level of confidence relates to the frequency with which the random interval contains the true parameter over repeated sampling, rather than a probability that conditions on the particular data observed.

By contrast the Bayesian approach lends itself very nicely to reporting and interpreting intervals since we are treating θ as a random variable from the outset. We can therefore use $\pi(\theta|y)$ to make probability statements about θ that are expressed in terms of *conditional probabilities* given y .

Since $\pi(\theta|y)$ represents my posterior belief about θ having observed y , then I can express my posterior belief that θ lies in a particular interval (θ_L, θ_U) as

$$\mathbb{P}(\theta \in (\theta_L, \theta_U) | \mathbf{x}) = \int_{\theta_L}^{\theta_U} \pi(\theta | \mathbf{x}) d\theta.$$

Now if we select (θ_L, θ_U) so that this probability is $(1 - \alpha)$, then (θ_L, θ_U) is what is known as a $100(1-\alpha)\%$ *credible interval* for θ .

There are two natural and fairly obvious ways to do this (both best illustrated graphically).

- (A) *Equal-tailed interval*. To construct this interval we simply exclude an area of $\frac{\alpha}{2}$ from each tail of the distribution. The quantities θ_L and θ_U are respectively the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ quantiles of the posterior distribution.

- (B) *Highest posterior density (HPD) interval*. This approach forms an interval from those parameter values for which the posterior density is highest. Having selected α we seek (θ_L, θ_U) such that

$$\int_{\theta_L}^{\theta_U} \pi(\theta|\mathbf{x})d\theta = 1 - \alpha,$$

with the added constraint that

$$\pi(\theta_1|y) \geq \pi(\theta_2|y), \theta_1 \in (\theta_L, \theta_U), \theta_2 \notin (\theta_L, \theta_U).$$

We can envisage the construction of the HPD interval via the following diagram.

It is usually easier to calculate the intervals using method (A) above and the approach can be easily generalised to allow unequal tail exclusion probabilities. For example, in 10.2.1, our posterior density for p was Beta(29, 81). We can get an equal tailed 95%-credible interval for p using R as follows:

```
> pL = qbeta(0.025, 29, 81)
> pU = qbeta(0.975, 29, 81)
> pL
[1] 0.1859729
> pU
[1] 0.3494381
```

giving (0.186, 0.349) as the 95% interval. Note that since $\pi(p|y)$ is reasonably symmetric (see graph), then an HPD interval would not look that much different.

Moreover, the posterior mean of p is $29/110 = 0.263$ and the posterior standard deviation is 0.042. Appealing to a normal approximation we could get an approximate 95% credible interval by computing the posterior mean ± 1.96 standard deviations. This results in the interval (0.182, 0.346), which corresponds quite closely with the interval above.

Further comments on HPD intervals:

- They are generally the shortest possible interval for a given level of credibility.
- They are not so straightforward to calculate!
- They are not preserved by 1-1 transformations of the parameters in general. If φ is a 1-1 mapping of the parameter space, and (θ_L, θ_U) is a $100(1-\alpha)$ HPD interval, then $(\varphi(\theta_L), \varphi(\theta_U))$ is not necessarily a $100(1-\alpha)$ HPD interval for the parametrisation φ .
- If the posterior density is multi-modal than you might end up with an HPD interval which is formed from several distinct sub-intervals.

10.2.4 More on selecting priors

Informative priors

When we have a lot of prior information on the parameter of interest θ (or strong beliefs about it) then we can represent this in our prior distribution.

Example 1: For any given coin we should expect $P(H)$ to be pretty close to 0.5 (but perhaps not exactly equal to 0.5). Before carrying out a coin-tossing experiment we might assign a prior to p , $\pi(p) \sim \text{Beta}(100, 100)$ to reflect this strong belief. In this case, the prior mean is 0.5 and the prior standard deviation is 0.035. This would be an example where we use an *informative prior*.

After an experiment in which we observe r heads in n tosses, say,

$$\pi(p|\mathbf{x}) \sim \text{Beta}(100 + r, 100 + n - r).$$

Clearly r and $n - r$ must be sufficiently large if the mean and variance of the posterior are to look substantially different from those of the prior. (i.e. we need to collect a lot of data to change our strong prior belief). In general we can assign an informative $\text{Beta}(\alpha, \beta)$ prior for a probability p by selecting α and β to be large.

Example 2: When using a $\text{Gamma}(\alpha, \beta)$ prior for λ in $\text{Poi}(\lambda)$ or $\text{Exp}(\lambda)$ we obtain an informative prior if the value of α is large, since the ratio of the prior mean to prior standard deviation is $\alpha^{\frac{1}{2}}$.

Vague or non-informative priors

What do we do if we don't have a lot of prior information? Then we use a *vague prior*. Essentially this is one which supports a wide range of values of the parameter and doesn't rule out *a priori* all but a strong region of the parameter space. Some examples:

- For p in $\text{Bin}(n, p)$ we obtain a vague prior by setting $\pi(p) \sim \text{Beta}(\alpha, \beta)$ where α and β are small.
- For λ in $\text{Poi}(\lambda)$ or $\text{Exp}(\lambda)$ a vague prior for λ is given by $\pi(\lambda) \sim \text{Gamma}(\alpha, \beta)$ where α and β are small.
- For μ in $N(\mu, \sigma^2)$ where σ^2 is known, a vague prior for μ could be e.g. $\pi(\mu) \sim N(0, \gamma^2)$ where γ^2 is large.

Ignorance priors

Many statisticians have tried to identify prior distributions that can be used to represent total ignorance about a parameter.

Example. For p in $\text{Bin}(n, p)$ it is tempting to think that a uniform prior represents ignorance, but it is not preserved under 1-1 parameter transformation in the following sense.

If we were ‘totally ignorant’ about p then we are ‘totally ignorant’ about p^2 . (*Do you agree with this suggestion?*) However, if $p \sim U(0, 1)$ then

$$p^2 \sim \text{Beta}(\frac{1}{2}, 1).$$

This shows that proposing a uniform distribution for a parameter is not equivalent to total ignorance.

Ignorance priors for location parameters. For a location parameter whose range is $(-\infty, \infty)$, e.g. μ in $N(\mu, \sigma^2)$ where σ^2 is known, we use a flat prior $\pi(\mu) = c$.

Remark. Since the parameter space for $N(\mu, \sigma^2)$ is $(-\infty, \infty)$ then $\pi(\mu) = c$ isn’t a proper distribution (why?). However, recall our general rule:

$$\pi(\theta|\mathbf{x}) = \frac{\pi(\theta)L(\theta; \mathbf{x})}{\int \pi(\theta')L(\theta'; \mathbf{x})d\theta'}, \quad (**)$$

We will obtain a proper posterior distribution so long as the integral on the denominator is a finite quantity.

If we set $\pi(\mu) = c$ and observe even just a single observation y then the integral is finite so we nevertheless get a proper posterior distribution even though the prior $\pi(\mu)$ is improper.

Notation: We use the language of proportionality and write $\pi(\mu) \propto 1$ to indicate a constant flat prior on the parameter space.

Ignorance priors for scale parameters. For a scale parameter whose range is $(0, \infty)$ e.g. λ in $\text{Exp}(\lambda)$, or σ in $N(0, \sigma^2)$ we obtain appropriate non-informative priors as

$$\pi(\lambda) \propto \frac{1}{\lambda}, \quad \pi(\sigma) \propto 1/\sigma.$$

Note that if $\varphi(\sigma) = \sigma^2$ then the above prior on σ leads to the prior $\pi(\varphi) \propto \frac{1}{\varphi}$.

A general approach to derive an ignorance prior for θ is the so-called *Jeffreys’* prior given by

$$\pi(\theta) \propto I(\theta)^{\frac{1}{2}}.$$

where we recall that $I(\theta)$ is the Fisher information.

10.3 Predictive distributions

In statistical inference, once we have fitted a model to a set of data we usually want to use the model to make predictions about future outcomes. Bayesian inference allows you to translate the uncertainty in parameters into the predictions you make. We illustrate this with a simple example.

Example: You are in charge of recruitment in a company. At the start you have no knowledge of the quality of job applicants from university X. In the first year you receive 10 applications of which 3 turn out to be suitable. During the second year you hear that you will shortly receive 3 further applications from university X. Let Z denote the (as yet unknown) number of these applicants who will turn out to be suitable. How can you identify a probability distribution to describe your beliefs regarding Z ?

Solution: In Bayesian statistics the solution to this problem is one of determining a predictive distribution for Z given the observation. We can proceed as follows:

Let p denote the unknown probability that an applicant from university X is suitable. At the beginning you were ignorant about p so let's suppose you represented that using a uniform prior for p , that is

$$p \sim U(0, 1) \sim \text{Beta}(1, 1)$$

During the first year you obtain data \mathbf{x} (3 suitable out of 10). Assuming a binomial model and using Bayesian reasoning you update your prior to obtain a posterior distribution for p :

$$\pi(p|\mathbf{y}) \sim \text{Beta}(4, 8).$$

Consider now the distribution of Z . Conditional on p you know $Z \sim \text{Bin}(3, p)$, but you don't know p . However you have a distribution for p , provided by the posterior above. Therefore it is possible to identify a probability function for p and z jointly as follows:

$$\pi(p, z|\mathbf{x}) = \pi(p|\mathbf{x})\mathbb{P}(Z = z|p)$$

where $\pi(p|\mathbf{y})$ is the Beta(4, 8) density and $\mathbb{P}(Z = z|p)$ is the binomial probability mass function. We now obtain the marginal distribution for Z by integrating with respect to p . This gives what is known as the predictive probability mass function for Z given observation \mathbf{x} . This gives the equation

$$f_Z(z|\mathbf{x}) = \pi(z|\mathbf{x}) = \int \pi(p, z|\mathbf{x})dp = \int \pi(p|\mathbf{x})\mathbb{P}(Z = z|p)dp.$$

Notice that $f_Z(z|\mathbf{y})$ is simply the posterior expectation of the function $g(p) = \mathbb{P}(Z = z|p)$. Therefore we are simply averaging the value of $\mathbb{P}(Z = z|p)$ over the posterior. Now work this out for the above situation. You will need to know about some properties of the Beta function!

This is an example of a beta-binomial distribution. You can think of it as a particular kind of mixture distribution where a random draw is obtained by first drawing a probability p from a beta

distribution and then drawing a sample from a binomial distribution with that particular value of p . To generate a random sample of size n from it, first draw a random sample p_1, p_2, \dots, p_n from $\text{Beta}(4, 8)$ and then draw $z_i \sim \text{Bin}(3, p_i), i = 1, 2, 3, n$. Using R you could easily investigate its properties using simulation. However we have an analytic form for the probability mass function which we can compute using R.

Now consider the values of $f_Z(z|\mathbf{x})$ for $z = 0, 1, 2, 3$.

```
> z = c(0:3)
> fz = 6*beta(4+z, 11-z)/(beta(4, 8)*gamma(z+1)*gamma(3-z+1))
> fz
[1] 0.32967033 0.39560440 0.21978022 0.05494505
```

The predictive probability that we get no suitable candidates is 0.33.

Now compare what would we would have predicted had we assumed that p was equal to its posterior mean $1/3$. (This is not fully Bayesian!)

```
> fz2 = dbinom(z, 3, 1/3)
> fz2
[1] 0.29629630 0.44444444 0.22222222 0.03703704
```

We find that conditioning on the posterior mean leads to smaller estimates for the probability of extreme values ($z = 0$ and $z = 3$). This illustrates the fact that by taking account of the posterior uncertainty in p , the fully Bayesian approach predicts that extreme events are more likely in comparison to approaches that use a point estimate for p . This tends to hold in general. That is, if you take account of parameter uncertainty when making predictions, you will recognise a greater possibility that extreme events may occur.

F20SA / F21SA Statistical Modelling and Analysis

Chapter 11: Non Parametric Methods

Contents

11.1 Introduction	11-1
11.1.1 Example data set	11-1
11.2 Permutation and Randomization Test	11-2
11.2.1 Setup	11-2
11.2.2 Permutation Method	11-2
11.2.3 Randomization Method	11-4
11.3 Bootstrap Methods	11-8
11.3.1 Bootstrap Confidence Interval	11-8
11.3.2 Smoothed Bootstrap	11-10

11.1 Introduction

Through out the statistical analysis we have carried out in this course we have made assumptions about the distribution of our data. For example:

- The likelihood function when finding MLE,
- The normally distributed data when carry out t-tests,
- The errors for linear regression being normally distributed,
- The density function of the data when carry out Bayesian analysis.

Unfortunately in many circumstances these assumptions are no well founded and the reality may lie far away from the truth. In these circumstances applying the standard tests can provide misleading results, for example not capturing the true variability of the data.

This problem has led to the development of a large range of statistical methodologies which do away with the need for an underlying model for the distribution of the data. These are called non-parametric methods.

In this chapter we will give a brief introduction to this field by introducing two methodologies for exploring the mean of a data set.

11.1.1 Example data set

A study of 12 students has been carried out to explore the relative time spent on studying using a range of online resources. The records of students behaviour were used to produce a resource utilization index for different resource types. The index was constructed in such a way that a value of zero would be expected if each resource was used equally with no preference. A positive result indicates a preference for online resources and visa versa.

The following data was collected:

0.13, -0.01, -0.01, 0.42, -0.02, 0.01, 0.09, 0.03, 0.04, 0.06, 0.12, 0.03

These values suggest that most students do not have a strong preference for online resources but that a small number may have a stronger preference for example student 4.

We are now interested in testing the null hypothesis that the mean value for this index score is zero and the precision of this estimate.

Remarks:

- The classical approach would be to conduct a one-sided t-test, make use of a normal assumption for the data.
- There are a range of non-parametric tests for this problem including the Wilcoxon's signed rank test. These often discard the actual values.
- We are going to consider two approaches:
 - Permutation and randomization test
 - Bootstrap method

11.2 Permutation and Randomization Test

11.2.1 Setup

- We have observation x_1, \dots, x_n for which we are interested in understand the population mean.
- We are going to do away with the assumption that the observations are normally distributed.
- To make things easier we are going to make the assumption that the values are symmetric about their mean, μ . (Not necessarily normal).
- Under this assumption the sign of $x_i - \mu$ is random.
- We are interested exploring the null hypothesis $H_0 : \mu = 0$.

11.2.2 Permutation Method

Permutation Test

Idea: We consider all permutations of the signs of the data and determine which of these have a more extreme result.

Under $H_0 : \mu = 0$ all possible permutations of $\{\pm x_i\}$ are equally likely. We can enumerate them and get:

0.13, 0.01, 0.01, 0.42, 0.02, 0.01, 0.09, 0.03, 0.04, 0.06, 0.12, 0.03
 -0.13, 0.01, 0.01, 0.42, 0.02, 0.01, 0.09, 0.03, 0.04, 0.06, 0.12, 0.03
 0.13, -0.01, 0.01, 0.42, 0.02, 0.01, 0.09, 0.03, 0.04, 0.06, 0.12, 0.03
 ⋮
 ⋮
 -0.13, -0.01, -0.01, -0.42, -0.02, -0.01, -0.09, -0.03, -0.04, -0.06, -0.12, -0.03

- There are 2^n different permutations. Here $n = 12$ so there are $2^{12} = 4096$ permutations.
- We need to decide how many of these are at least as extreme as the observed permutation: 0.13, -0.01, -0.01, 0.42, -0.02, 0.01, 0.09, 0.03, 0.04, 0.06, 0.12, 0.03
- To do this we need to select a suitable test statistic to use to make this comparison. Given we are interested in the population mean, μ , a natural choice is the sample mean \bar{x} .
- Under the null hypothesis, the farther this is from zero the more extreme the permutation.

We can now use this to find a p-value for our observations as the p-value is defined to be the probability that a value is at least as extreme as the observed statistics, given the null hypothesis is true.

- Under H_0 all the permutations are equally likely.
- Simply need to count the proportion of permutations that yield a sample mean greater than \bar{x} .
- One-tailed ($H_1 : \mu > 0$): we only count $\geq \bar{x}$
- Two-tailed ($H_1 : \mu \neq 0$): we also count $\leq -\bar{x}$.
- In practice normally double the one sided for the two sided case.
- In this case it does not matter.

Here the sample mean is 0.074, so we are counting the number of permutations with a sample mean larger than this. For the above data we find that 24 permutations out of the 4096 yield a mean ≥ 0.074 . So if the alternative hypothesis is two tailed the p-values is thus $2 \times 24/4096 = 0.012$.

Permutation Interval

In addition to the p-value we can use a similar methodology to find a confidence interval for the mean.

Consider now $H_0 : \mu = \mu_0$:

- We form differences $x_i - \mu_0$
- Enumerate all permutations $\{\pm(x_i - \mu_0)\}$.
- We can use these to calculate a one-sided p-value against the alternative $H_1 : \mu > \mu_0$.

Remember we reject the null hypothesis if the p-value is less than 0.025 (for a 2.5% level test). So to find the lower limit of 95% we look for lowest μ_0 such that the p-value is just over 0.025, i.e. the smallest value such that we would not reject the null hypothesis.

For the upper limit we can use the similar idea but instead use the alternative hypothesis $H_1 : \mu < \mu_0$. For this alternative hypothesis we are now interested in permutations where the sample mean is less than the original mean.

Due to the symmetric nature of the permutation distribution in this example, the 95% permutation interval we have found is made of all values of μ_0 such that we would not reject the null hypothesis $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ at the 95% level. To find these upper and lower limits we can use standard search methods, eg method of bisection.

For this data we have the permutation interval is (0.013, 0.150).

General comments

- We have not need to assume normality to do calculate these p-values and interval.
- We did assume symmetry and means it is not much more robustness than the t-test.
- We used the sample mean but we could have used a large range of test statistics instead.
- Possible examples are sample median, trimmed means, the number of positive observations.
- The optimal choice of statistics depends on the null and alternative hypothesis of interest.

Advantages:

1. The method is exact.
2. No specific distribution is assumed for the data.
3. The distribution of the test statistic is not needed.

Disadvantages:

1. Each permutation of the data must be equally likely under the null hypothesis. If this is not true we need to know the probability of each permutation and sample according to it.
2. In general these ideas are not useful for complex example (see Bootstrap)

11.2.3 Randomization Method

In general generating all possible permutations is a computationally difficult task and it is often impractical to generate every possible permutation when carrying out the permutation test. For example if we had 20 data points we would need to consider $2^{20} = 1073741824$ permutations. This is difficult for the permutation test but not impossible for the more computationally intensive permutation interval.

Instead we can generate our permutations at random and use this to calculate our p-value and later for interval integrate this with a stochastic search algorithm.

Randomization Test

For the one sample example we are considering we can generate a random permutation by listing the values of $|x_i - \mu|$ (or $|x_i - M|$ if we are interested in the median) and then randomly assigning a sign to each value ($Pr(+) = Pr(-) = 0.5$) independently. Then we calculate the test statistic for each of our randomizations and count the proportion of these which are at least as extreme as the test statistic for the real data to give us our p-value.

We can use the following code to carry out the randomization test with 1000 samples:

```
perm <- function (sdata , signs , nrand)
{ rmean <- NULL
  n <- length(data)
  for (i in 1:nrand) {
    rsigns <- sample(signs , n , replace=TRUE)
    rdata <- data * rsigns
    rmean[i] <- mean(rdata) }
  rmean
}
```

```

sdata <- c(0.13, 0.01, 0.01, 0.42, 0.02, 0.01, 0.09, 0.03, 0.04, 0.06,
0.12, 0.03)
meandata <- mean()
signs <- c(1, 1)
nrand <- 999
rmean <- perm(sdata, signs, nrand)
rmean[nrand+1] <- meandata
rabsmean <- abs(rmean)
extreme <- (rabsmean >= meandata)
p <- sum(extreme)/length(rmean)
p
hist(rmean, br=20, xlim=c(0.1, 0.1))

```

For this data with a 1000 randomizations for this data I obtained a p-value of 0.015 which compares well to the value we obtain previously with all permutations of 0.012.

Randomization Interval

Obtaining the exact permutation interval is very computational intensive and substantially more than a single permutation test. So we need to use an efficient randomization method for approximating the permutation interval. Such a scheme is based on the Robbins-Monro stochastic approximation method.

Assume we are seeking a 95% confidence interval for the mean index value. We focus on the upper limit but analogous method works for the lower limit.

- We can initialize our upper and lower bounds by values found by making the assumption the data is normally distributed and using the t-statistic. In this case the interval would be $(-0.002, 0.150)$
- Since we are interested in the upper limit we care about the hypothesis $H_0 : \mu = \mu_U$ vs $H_1 : \mu < \mu_U$ at the 2.5% level.
- Let U_j be the current estimate of μ_U at the j^{th} step.
- Subtract U_j from each observation, randomly generate a permutation using these values and calculate \bar{y}_j the sample mean for this permutation.
- We update U_j using the following:

$$U_{j+1} = \begin{cases} U_j - c\alpha/j, & \text{if } \bar{y}_j + U_j > \bar{x} \\ U_j - c(1-\alpha)/j, & \text{if } \bar{y}_j + U_j \leq \bar{x} \end{cases}$$

where \bar{x} is the sample mean of the original data, $\alpha = 0.025$, and c is a step length constant.

- To prevent wild oscillations for j small we arbitrarily start at $j = 40$.
- The optimal choice of the step length constant depends on the true distribution, here we will take it as $c = 16(U_j - \bar{x})$. This is based on the optimal value if the data followed a normal distribution.
- Under very general conditions, we have U_j converges to μ_U as $j \rightarrow \infty$.

Example code to this:

```

# Data are assumed to be in the vector data
rmean < NULL
lows < NULL
highs < NULL
stepno < NULL
# Calculate sample size and test statistic = sample mean
n < length(data)
meandata < mean(data)

#values from classical t test
loest < 0.002
hiest < 0.15

lows[1] < loest
highs[1] < hiest
stepno[1] < 1
signs < c(1,1)
# now start Robbins Monro search at step 40; 1000 steps in all
for (i in 40:1000) {
  dlo < data loest
  dhi < data hiest
  # lower limit first.
  rsigns < sample(signs,n,replace=TRUE)
  rmean < mean(dlo*rsigns)
  steplength < 16*(meandata loest)
  if(rmean+loest < meandata){
    loest < loest+steplength*0.025/(i 1)
  }else{
    loest < loest steplength*0.975/(i 1))
  }
  lows[i 38] < loest
  # now upper limit
  rsigns < sample(signs,n,replace=TRUE)
  rmean < mean(dhi*rsigns)
  steplength < 16*(hiest meandata)
  if(rmean+hiest > meandata){
    hiest < hiest steplength*0.025/(i 1)
  }else{
    hiest < hiest+steplength*0.975/(i 1))
  }
  highs[i 38] < hiest
  stepno[i 38] < i 38 }

loest
hiest
par(mfrow = c(1,2))
plot(lows,stepno)
plot(highs,stepno)

```

We have considered the permutation and randomization method for just a single sample of data but these ideas can easily be extended to more general settings. For example:

- Comparing the difference in means between two samples. Here we can permute the data in each sample.
- For linear regression testing the null hypothesis that the slope is 0. Here we would permute the y values, i.e. randomly pairing up the y and x values.

11.3 Bootstrap Methods

The second class of methods of interest are that of bootstrap methods, a form of re-sampling data. First, we can resample our observations with replacement. Of course if we resampled n observations from n without replacement, we would simply get the same sample every time. By sampling with replacement, we seek to mimic the variation we would get if we could draw new independent samples. We then use that variation to quantify the precision of quantities estimated from our original sample. This is the nonparametric bootstrap. This method is primarily used produce confidence for test statistics of interest.

Idea: Assuming the data is independently identically distributed we can use the data to produce an empirical distribution which is ‘close’ to the original distribution. So sampling from it will produce samples with similar properties to a sample from the original distribution.

11.3.1 Bootstrap Confidence Interval

The bootstrap methodology is particularly useful in producing confidence intervals for a wide range of parameters. This is especially useful in the case where we want to produce confidence for multiple parameters together. The permutations and randomization confidence we calculated in the previous section are exact when there is only a single parameter to estimate and the data is independent identically distributed. Unfortunately, as we move to more parameters such confidence intervals become approximate and are difficult to setup for more complicated parameter estimates.

In comparison the bootstrap confidence is easy to carry out and gives an approximate confidence interval on all the parameters of interest from a single set of resamplings of the data.

We again consider the example from the previous section:

0.13, -0.01, -0.01, 0.42, -0.02, 0.01, 0.09, 0.03, 0.04, 0.06, 0.12, 0.03

with sample mean $\bar{x} = 0.074$.

Now we want to estimate the variance of the sample mean and to find a 95% confidence interval for the mean value of the population mean.

- We assume the observations are i.i.d.
- We approximate the true distribution by the empirical distribution and generate a new sample.
- Note that sampling from the empirical distribution is the same as sampling with replacement from the original data set.
- Generally we create new samples with the same number of data points as the original data set.
- We then repeat this b times.

For example:

Sample	Observations												Mean
1	-0.01	0.42	0.12	0.09	0.13	0.01	-0.02	0.12	-0.01	-0.01	0.12	-0.02	0.078
2	0.06	0.01	0.13	0.04	0.06	-0.01	0.03	-0.01	-0.02	0.03	0.01	-0.01	0.027
⋮													
⋮													
b	-0.01	0.04	-0.02	0.42	0.03	0.12	0.13	-0.02	0.42	0.09	0.01	0.06	0.106

We assume that the means of the bootstrap resamples represent the variability we would observe in the sample mean if we were to obtain many samples of size 12 from the population. The sample variance of the bootstrap means is an estimate of the variance of \bar{x} , and approximate confidence intervals may be found using the percentile method.

Percentile method for $100(1 - 2\alpha)\%$ confidence interval:

1. Order the bootstrap means from smallest to largest.
2. Find the r^{th} value for lower limit, where $r = \alpha(b + 1)$.
3. Find the s^{th} value for lower limit, where $s = (1 - \alpha)(b + 1)$.

For example if we generate 999 samples and want the 95% confidence interval:

- Lower limit is the 25th smallest value.
- Upper limit is the 25th largest value.

Here is example code to generate the bootstrap confidence interval:

```
# Data are assumed to be in the vector data.
# Calculate sample size and sample mean.
meandata <- mean(data)
n <- length(data)
nboot <- 999
alpha <- 0.025 # i.e. 95% confidence interval
bootmean <- NULL
for (i in 1:nboot) {
  rdata <- sample(data, n, replace=TRUE)
  bootmean[i] <- mean(rdata) }
bootse <- sd(bootmean)
nlo <- round((nboot+1)*alpha)
nhi <- round((nboot+1)*(1-alpha))
bootmean <- sort(bootmean)
low <- bootmean[nlo]
high <- bootmean[nhi]
# print bootstrap s.e. and 95% percentile confidence limits
results <- c(bootse, low, high)
results
```

We can carry use the same idea for other quantities of interest. For example:

- Median, evaluate the sample median for each resample.
- Variance, evaluate the sample variance for each resample.
- Skewness etc.

Advantages:

1. It is a simple technique to apply.
2. It is a general and robust method for setting confidence limits.

Disadvantages:

1. It assumes the samples are independent identically distributed.
2. In some cases the units used for resampling is not clear. For example, multiple regression, should we resample by observation or the errors.
3. The method is generally only asymptotically exact as both b and n tend to infinity.

11.3.2 Smoothed Bootstrap

In the original bootstrap we approximate the true distribution by a the empirical distribution function which places a point mass at each data point. Especially when sample size n is small, this bootstrap methodology is limited because the same n observations are reused repeatedly. The smoothed bootstrap is sometimes used to overcome this problem. The main idea is to smooth the observed sample and then resample from the smoothed density instead of the empirical density.

This smoothing is generally carried out by replacing each observation by a density of a given form. Two standard cases are a triangular density or a normal density with standard deviation with h . We will focus on the later here. If the data is y_1, \dots, y_n , then the density we sample from is given by:

$$\hat{f}(t; h) = \frac{1}{nh} \sum_{i=1}^n \psi\left(\frac{t - y_i}{h}\right),$$

where $\psi(t)$ is the standard normal density and h is called the window size; the larger the value of h , the greater the degree of smoothing.

Sampling from this density is very easy, we sample as before for the standard bootstrap but then to each sample we independently add a normally distributed noise with standard deviation h to each sample.

Example R code to do this:

```
# Data are assumed to be in the vector data.
# Calculate sample size and sample mean.
meandata <- mean(data)
n <- length(data)
h <- 0.0001
nboot <- 999
alpha <- 0.025 # i.e. 95% confidence interval
bootmean <- NULL
for (i in 1:nboot) {
  rdata <- sample(data, n, replace=TRUE)+rnorm(n, sd=h)
  bootmean[i] <- mean(rdata) }
bootse <- sd(bootmean)
nlo <- round((nboot+1)*alpha)
nhi <- round((nboot+1)*(1-alpha))
bootmean <- sort(bootmean)
low <- bootmean[nlo]
high <- bootmean[nhi]
# print bootstrap s.e. and 95% percentile confidence limits
results <- c(bootse, low, high)
results
```


F20SA / F21SA Statistical Modelling and Analysis

Chapter 12: Principal component analysis and factor analysis

Contents

12.1 Principal component analysis	12–1
12.2 Factor analysis	12–3

12.1 Principal component analysis

Principal component analysis (PCA) is a multivariate data analysis technique that is often used to perform dimensionality reduction for exploratory data analysis, data restoration, and prediction problems.

Suppose that we have a sample of size n , and that for each element of the sample we record p features (e.g., a population of n individuals for which we have recorded p relevant genetic markers). We organise this information in the form of a data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, where each row $\mathbf{x}_i = (x_{i,1}, \dots, x_{i,p})^\top$, for $i = 1, \dots, n$, has the information associated with one sample element.

PCA is a strategy for identifying patterns and structure in our sample \mathbf{X} . Knowledge of this patterns and structure then allows representing \mathbf{X} more compactly, using a number of features $p' \ll p$, with little loss of information. These new features provide an efficient and meaningful representation, and are called the “principal components” of \mathbf{X} . In order to identify these principal components we proceed as follows:

1. Normalise (centre-and-scale) \mathbf{X} such that each one of its columns has sample mean zero and unit sample variance; we denote this normalised dataset by $\tilde{\mathbf{X}}$.
2. Identify the vector or direction in \mathbb{R}^p along which the row vectors of $\tilde{\mathbf{X}}$ have highest variance. This produces the 1st and most important principal component $\mathbf{v}_1 \in \mathbb{R}^p$.
3. From all the vectors or direction in \mathbb{R}^p that are orthogonal (perpendicular) to \mathbf{v}_1 , we identify that along which the row vectors of $\tilde{\mathbf{X}}$ have highest variance. This produces the 2nd principal component $\mathbf{v}_2 \in \mathbb{R}^p$, orthogonal to \mathbf{v}_1 by construction.

4. From all the vectors or direction in \mathbb{R}^p that are orthogonal to both \mathbf{v}_1 and \mathbf{v}_2 , we identify that in which the rows of $\tilde{\mathbf{X}}$ have highest variance. This produces the 3rd principal component $\mathbf{v}_3 \in \mathbb{R}^p$.
5. We repeat this analysis to identify the remaining $p - 3$ principal components, producing a total of p principal component vectors that are orthogonal to each other and that hence perfectly represent $\tilde{\mathbf{X}}$ (and \mathbf{X} after carefully reversing the scale and centre transformation).

Once the p principal components are available we calculate the relative contribution of each component (as measured by the variance along the directions specified by the components). This allows identifying dominant components that carry most of the information of $\tilde{\mathbf{X}}$ and that are useful for visualisation, exploratory analysis, restoration, and prediction. The weak components carry little information and can be omitted to reduce the dimension of the data and obtain a more compact and meaningful representation. Usually, the first $p/5$ components suffice to provide a reliable representation.

For example, by conducting a PCA of genetic data from European populations, and then displaying the data points in the coordinate system of $(\mathbf{v}_1, \mathbf{v}_2)$, the scatter plot of Fig 12.1 is obtained. Notice that $\mathbf{v}_1, \mathbf{v}_2$ nicely capture the genetic similarities and differences of the different European populations, which are closely related to their spatial proximity.

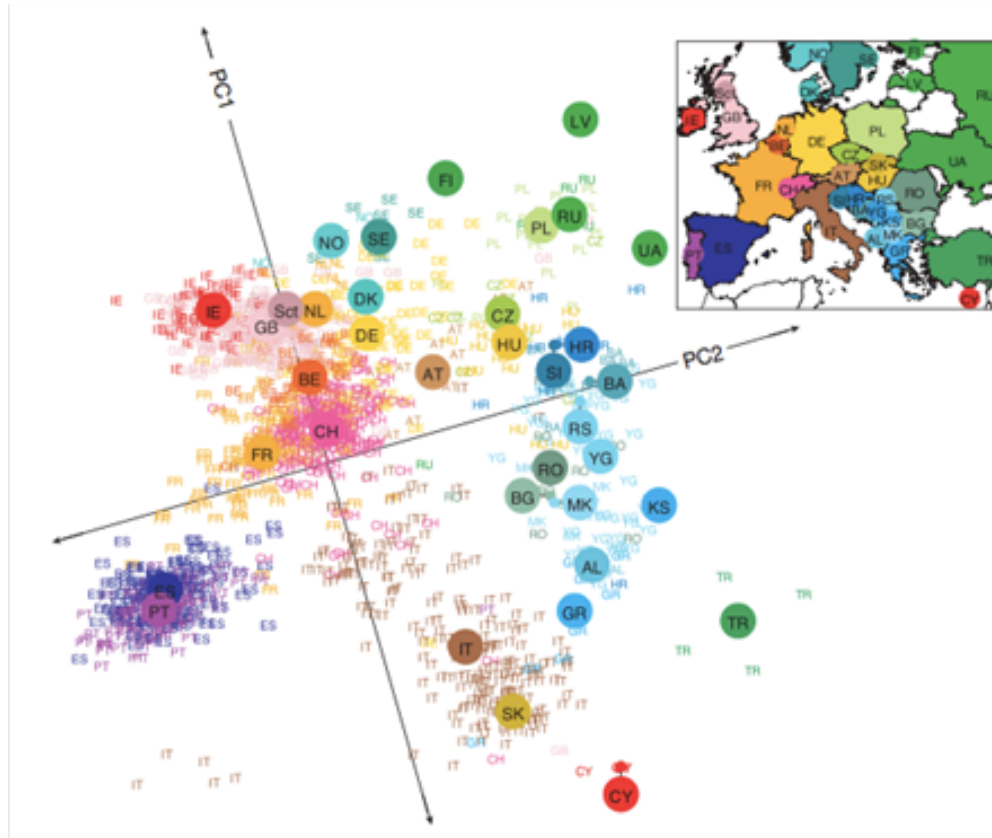


Figure 12.1: A statistical summary of genetic data from $n=1,387$ European individual ($p=197,146$ features per individual) based on PCA (axis PC1 corresponds to \mathbf{v}_1 , and axis PC2 to \mathbf{v}_2). Small coloured labels represent individuals and large coloured points represent median PC1 and PC2 values for each country. The PC axes are rotated to emphasise the similarity to the geographic map of Europe. Source: Novembre et al., Genes mirror geography within Europe, *Nature* 456, pp. 98–101 (06 November 2008)

From an computational viewpoint, PCA is efficiently performed by calculating a so-called singular value decomposition of $\tilde{\mathbf{X}}$, i.e., calculating the decomposition

$$\tilde{\mathbf{X}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$$

where $\mathbf{U} \in \mathbb{R}^{n \times n}$ and $\mathbf{V} \in \mathbb{R}^{p \times p}$ are matrices with orthogonal column vectors of unit length, and $\mathbf{\Sigma} \in \mathbb{R}^{n \times p}$ is a rectangular diagonal matrix with non-negative diagonal elements satisfying the order $\Sigma_{1,1} \geq \Sigma_{2,2} \geq \dots \Sigma_{p,p} \geq 0$. The resulting matrix $\mathbf{V} \in \mathbb{R}^{p \times p}$ has the principal components as column vectors, and the diagonal elements of $\mathbf{\Sigma}$ represent the importance of each of these components in $\tilde{\mathbf{X}}$.

To perform PCA in R we use the function `prcomp`. For example, a PCA of the `mtcars` dataset, which contains data on 32 car models, produces the following result (we conduct the PCA on the columns 1 – 7, 10 and 11 of the dataset, so \mathbf{X} has dimension 32×9 in this example):

```
mtcars.pca <- prcomp(mtcars[,c(1:7,10,11)], center = TRUE, scale. = TRUE)
```

```
summary(mtcars.pca)
```

```
## Importance of components:
##               PC1      PC2      PC3      PC4      PC5      PC6
## Standard deviation  2.3782  1.4429  0.71008  0.51481  0.42797  0.35184
## Proportion of Variance 0.6284  0.2313  0.05602  0.02945  0.02035  0.01375
## Cumulative Proportion 0.6284  0.8598  0.91581  0.94525  0.96560  0.97936
##               PC7      PC8      PC9
## Standard deviation  0.32413  0.2419  0.14896
## Proportion of Variance 0.01167  0.0065  0.00247
## Cumulative Proportion 0.99103  0.9975  1.00000
```

Notice that PC1 (that is v_1) explains 63% of the total variance, which means that nearly two-thirds of the information in the dataset (9 variables) can be encapsulated by just that one principal component. PC2 explains 23% of the variance. So, by knowing the position of a sample in relation to just PC1 and PC2, you can get a very accurate view on where it stands in relation to other samples, as just PC1 and PC2 can explain 86% of the variance. To visualise the data on the coordinate system of PC1 and PC2 we can use the `ggbiplot` function provided by the `ggbiplot` library.

12.2 Factor analysis

Factor analysis (FA) is an alternative approach to represent correlated variables (e.g., the rows of \mathbf{X} corresponding to the different elements in our sample) in terms of a lower number of unobserved variables called factors. Similarly to PCA, the observed variables are modelled as linear combinations of the potential factors, plus "error" terms. A main difference with PCA is that FA will produce an error or residual with diagonal covariance matrix, hence efficiently capturing in the factors all the correlations in the data. However, the residual values can be relatively large. Conversely, low-dimensional representations obtained via PCA will produce an error or residual covariance matrix that is optimal in a mean squared sense, but with a active non-diagonal elements. Use the `factanal` function to perform FA in R. The `nFactors` package provides tools to choose the number of factors used in the analysis.