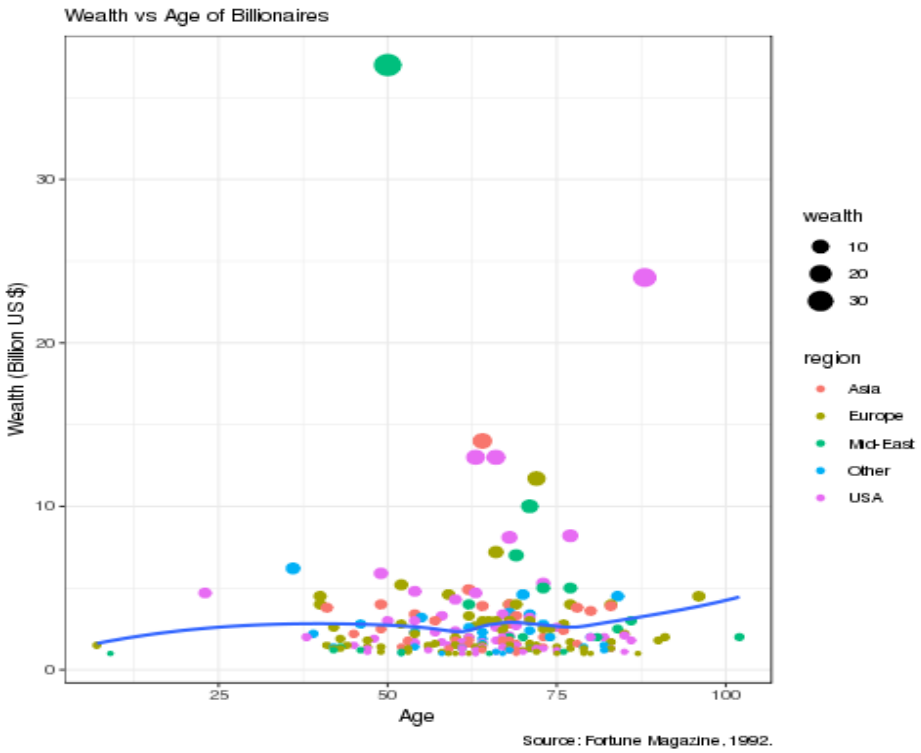
Mathematical Sciences

UNIVERSITY OF Southampton

MATH1024: Introduction to Probability and Statistics

Prof Sujit Sahu

Autumn 2019



Contents

1 Introduction to Statistics **9**

1.1 Lecture 1: What is statistics? . . . 9
1.1.1 Early and modern definitions . . . 9
1.1.2 Uncertainty: the main obstacle to decision making . . . 10
1.1.3 Statistics tames uncertainty . . . 10
1.1.4 Why should I study statistics as part of my degree? . . . 10
1.1.5 What's in this module? . . . 11
1.1.6 Take home points: . . . 11
1.2 Lecture 2: Guessing ages . . . 12
1.2.1 Lecture mission . . . 12
1.2.2 Experiment details . . . 12
1.2.3 Revealing the truth and error calculation . . . 12
1.2.4 Protecting your privacy . . . 13
1.2.5 An example data set . . . 13
1.2.6 Take home points . . . 13
1.3 Lecture 3: Basic statistics . . . 14
1.3.1 Lecture mission . . . 14
1.3.2 How do I obtain data? . . . 14
1.3.3 Summarising data . . . 14
1.3.4 Take home points . . . 17
1.4 Lecture 4: Data visualisation with `R` . . . 17
1.4.1 Lecture mission . . . 17
1.4.2 Get into R . . . 18
1.4.3 Keeping and saving commands in a script file . . . 19
1.4.4 How do I get my data into `R`? . . . 19
1.4.5 Working with data in `R` . . . 20
1.4.6 Summary statistics from `R` . . . 21
1.4.7 Graphical exploration using `R` . . . 21
1.4.8 Take home points . . . 22
1.5 Lecture 5: Help with report writing . . . 22
1.5.1 Lecture mission . . . 22
1.5.2 Guidance on data inspection, analysis and conclusions . . . 22
1.5.3 Some advice on writing your report . . . 23
1.5.4 Take home points . . . 24

2 Introduction to Probability **25**
2.1 Lecture 6: Definitions of probability . . . 25
2.1.1 Why should we study probability? . . . 25
2.1.2 Two types of probabilities: subjective and objective . . . 25
2.1.3 Union, intersection, mutually exclusive and complementary events . . . 26
2.1.4 Axioms of probability . . . 28
2.1.5 Application to an experiment with equally likely outcomes . . . 29
2.1.6 Take home points . . . 29
2.2 Lecture 7: Using combinatorics to find probability . . . 29
2.2.1 Lecture mission . . . 29
2.2.2 Multiplication rule of counting . . . 30
2.2.3 Calculation of probabilities of events under sampling 'at random' . . . 31
2.2.4 A general 'urn problem' . . . 31
2.2.5 Take home points . . . 33
2.3 Lecture 8: Conditional probability and the Bayes Theorem . . . 33
2.3.1 Lecture mission . . . 33
2.3.2 Definition of conditional probability . . . 34
2.3.3 Multiplication rule of conditional probability . . . 34
2.3.4 Total probability formula . . . 35
2.3.5 The Bayes theorem . . . 37
2.3.6 Take home points . . . 37
2.4 Lecture 9: Independent events . . . 38
2.4.1 Lecture mission . . . 38
2.4.2 Definition . . . 38
2.4.3 Take home points . . . 40
2.5 Lecture 10: Fun probability calculation for independent events . . . 41
2.5.1 Lecture mission . . . 41
2.5.2 System reliability . . . 41
2.5.3 The randomised response technique . . . 42
2.5.4 Take home points . . . 43

3 Random Variables and Their Probability Distributions **45**
3.1 Lecture 11: Definition of a random variable . . . 45
3.1.1 Lecture mission . . . 45
3.1.2 Introduction . . . 45
3.1.3 Discrete or continuous random variable . . . 46
3.1.4 Probability distribution of a random variable . . . 46
3.1.5 Cumulative distribution function (cdf) . . . 48
3.1.6 Take home points . . . 49
3.2 Lecture 12: Expectation and variance of a random variable . . . 50
3.2.1 Lecture mission . . . 50
3.2.2 Mean or expectation . . . 50
3.2.3 Take home points . . . 52
3.3 Lecture 13: Standard discrete distributions . . . 52
3.3.1 Lecture mission . . . 52

3.3.2 Bernoulli distribution . . . 52
3.3.3 Binomial distribution . . . 53
3.3.4 Geometric distribution . . . 55
3.3.5 Hypergeometric distribution . . . 56
3.3.6 Take home points . . . 57
3.4 Lecture 14: Further standard discrete distributions . . . 57
3.4.1 Lecture mission . . . 57
3.4.2 Negative binomial distribution . . . 57
3.4.3 Poisson distribution . . . 58
3.4.4 Take home points . . . 59
3.5 Lecture 15: Standard continuous distributions . . . 60
3.5.1 Lecture mission . . . 60
3.5.2 Exponential distribution . . . 60
3.5.3 Take home points . . . 63
3.6 Lecture 16: The normal distribution . . . 64
3.6.1 Lecture mission . . . 64
3.6.2 The pdf, mean and variance of the normal distribution . . . 64
3.6.3 Take home points . . . 66
3.7 Lecture 17: The standard normal distribution . . . 66
3.7.1 Lecture mission . . . 66
3.7.2 Standard normal distribution . . . 66
3.7.3 Take home points . . . 69
3.8 Lecture 18: Joint distributions . . . 69
3.8.1 Lecture mission . . . 69
3.8.2 Joint distribution of discrete random variables . . . 70
3.8.3 Covariance and correlation . . . 72
3.8.4 Independence . . . 73
3.8.5 Take home points . . . 74
3.9 Lecture 19: Properties of the sample sum and mean . . . 74
3.9.1 Lecture mission . . . 74
3.9.2 Introduction . . . 75
3.9.3 Take home points . . . 77
3.10 Lecture 20: The Central Limit Theorem . . . 78
3.10.1 Lecture mission . . . 78
3.10.2 Statement of the Central Limit Theorem (CLT) . . . 78
3.10.3 Application of CLT to binomial distribution . . . 79
3.10.4 Take home points . . . 81

4 Statistical Inference **83**
4.1 Lecture 21: Foundations of statistical inference . . . 83
4.1.1 Statistical models . . . 84
4.1.2 A fully specified model . . . 84
4.1.3 A parametric statistical model . . . 85
4.1.4 A nonparametric statistical model . . . 85
4.1.5 Should we prefer parametric or nonparametric and why? . . . 86

4.1.6 Take home points . . . 86
4.2 Lecture 22: Estimation . . . 86
4.2.1 Lecture mission . . . 86
4.2.2 Population and sample . . . 87
4.2.3 Statistic and estimator . . . 87
4.2.4 Bias and mean square error . . . 88
4.2.5 Take home points . . . 90
4.3 Lecture 23: Estimation of mean and variance and standard error . . . 90
4.3.1 Lecture mission . . . 90
4.3.2 Estimation of a population mean . . . 90
4.3.3 Standard deviation and standard error . . . 92
4.3.4 Take home points . . . 93
4.4 Lecture 24: Interval estimation . . . 93
4.4.1 Lecture mission . . . 93
4.4.2 Basics . . . 93
4.4.3 Confidence interval for a normal mean . . . 94
4.4.4 Take home points . . . 96
4.5 Lecture 25: Confidence intervals using the CLT . . . 97
4.5.1 Lecture mission . . . 97
4.5.2 Confidence intervals for μ using the CLT . . . 97
4.5.3 Confidence interval for a Bernoulli p by quadratic inversion . . . 98
4.5.4 Confidence interval for a Poisson λ by quadratic inversion . . . 100
4.5.5 Take home points . . . 100
4.6 Lecture 26: Exact confidence interval for the normal mean . . . 101
4.6.1 Lecture mission . . . 101
4.6.2 Obtaining an exact confidence interval for the normal mean . . . 101
4.6.3 Take home points . . . 103
4.7 Lecture 27: Hypothesis testing I . . . 103
4.7.1 Lecture mission . . . 103
4.7.2 Introduction . . . 104
4.7.3 Hypothesis testing procedure . . . 104
4.7.4 The test statistic . . . 107
4.7.5 Testing a normal mean μ . . . 107
4.7.6 Take home points . . . 108
4.8 Lecture 28: Hypothesis testing II . . . 109
4.8.1 Lecture mission . . . 109
4.8.2 The significance level . . . 109
4.8.3 Rejection region for the t-test . . . 109
4.8.4 t-test summary . . . 110
4.8.5 p-values . . . 111
4.8.6 p-value examples . . . 111
4.8.7 Equivalence of testing and interval estimation . . . 112
4.8.8 Take home points . . . 112
4.9 Lecture 29: Two sample t-tests . . . 112

4.9.1 Lecture mission . . . 112
4.9.2 Two sample t-test statistic . . . 113
4.9.3 Paired t-test . . . 114
4.9.4 Take home points . . . 115
4.10 Lecture 30: Data collection and design of experiments . . . 115
4.10.1 Lecture mission . . . 115
4.10.2 Simple random sampling . . . 115
4.10.3 Design of experiments . . . 116
4.10.4 Take home points . . . 120

A Mathematical Concepts Needed in MATH1024 **121**
A.1 Discrete sums . . . 121
A.2 Counting and combinatorics . . . 122
A.3 Binomial theorem . . . 122
A.4 Negative binomial series . . . 124
A.5 Logarithm and the exponential function . . . 124
A.5.1 Logarithm . . . 124
A.5.2 The exponential function . . . 125
A.6 Integration . . . 125
A.6.1 Fundamental theorem of calculus . . . 125
A.6.2 Even and odd functions . . . 125
A.6.3 Improper integral and the gamma function . . . 126

B Worked Examples **129**
B.1 Probability examples . . . 129
B.2 Solutions: Probability examples . . . 131
B.3 Statistics examples . . . 138
B.4 Solutions: Statistics examples . . . 140

C Homework Sheets **145**

Chapter 1

Introduction to Statistics

1.1 Lecture 1: What is statistics?

1.1.1 Early and modern definitions

- The word *statistics* has its roots in the Latin word *status* which means the state, and in the middle of the 18th century was intended to mean:

  collection, processing and use of data by the state.

- With the rapid industrialization of Europe in the first half of the 19th century, statistics became established as a discipline. This led to the formation of the Royal Statistical Society, the premier professional association of statisticians in the UK and also world-wide, in 1834.

- During this 19th century growth period, statistics acquired a new meaning as the *interpretation of data or methods of extracting information from data for decision making.* Thus statistics has its modern meaning as the methods for:

  collection, analysis and interpretation of data.

- Indeed, the Oxford English Dictionary defines *statistics* as: "*The practice or science of collecting and analysing numerical data in large quantities, especially for the purpose of inferring proportions in a whole from those in a representative sample.*"

- Note that the word 'state' has gone from its definition. Instead, statistical methods are now essential for **everyone** wanting to answer questions using data.

For example, will it rain tomorrow? Does eating red meat make us live longer? Is smoking harmful during pregnancy? Is the new shampoo better than the old? Will the UK economy get better after Brexit? At a more personal level: What degree classification will I get at graduation? How long will I live for? What prospects do I have in the career I have chosen? How do I invest my money to maximise the return? Will the stock market crash tomorrow?

1.1.2 Uncertainty: the main obstacle to decision making

The main obstacle to answering the types of questions above is *uncertainty*, which means **lack of one-to-one correspondence between cause and effect**. For example, having a diet of (well-cooked) red meat for a period of time is not going to kill me immediately. The effect of smoking during pregnancy is difficult to judge because of the presence of other factors, e.g. diet and lifestyle; such effects will not be known for a long time, e.g. at least until the birth. Thus it seems:

Uncertainty is the only certainty!

1.1.3 Statistics tames uncertainty

- It is clear that we may never be able to get to the bottom of every case to learn the full truth and so will have to make a decision under uncertainty; thus mistakes cannot be avoided!
- If mistakes cannot be avoided, it is better to know how often we make mistakes (which provides knowledge of the amount of uncertainty) by following a particular rule of decision making.
- Such knowledge could be put to use in finding a rule of decision making which does not betray us too often, or which minimises the frequency of wrong decisions, or which minimises the loss due to wrong decisions.

Thus we have the equation:

Uncertain knowledge	+	Knowledge of the extent of uncertainty in it	=	Usable knowledge

Researchers often make guesses about scientific quantities. For example, try to guess my age: 65 or 45? These predictions are meaningless without the associated uncertainties. Instead, appropriate data collection and correct application of statistical methods may enable us to make statements like: I am 97% certain that the correct age is between 47 and 54 years. Remember, "to guess is cheap, to guess incorrectly is expensive" – old Chinese proverb.

1.1.4 Why should I study statistics as part of my degree?

- Studying statistics will equip you with the basic skills in data analysis and doing science with data.
- A decent level of statistical knowledge is required no matter what branch of mathematics, engineering, science and social science you will be studying.
- Learning statistical theories gives you the opportunity to practice your deductive mathematical skills on real life problems. In this way, you will improve at mathematical methods while studying statistical methods.

"All knowledge is, in final analysis, history.
All sciences are, in the abstract, mathematics.
All judgements are, in their rationale, statistics."

Prof. C. R. Rao

1.1.5 What's in this module?

- **Chapter 1:** We will start with the basic statistics used in everyday life, e.g. mean, median, mode, standard deviation, etc. Statistical analysis and report writing will be discussed. We will also learn how to explore data using graphical methods.

  - For this we will use the `R` statistical package. `R` is freely available to download. Search `download R` or go to: `https://cran.r-project.org/`. We will use it as a calculator and also as a graphics package to explore data, perform statistical analysis, illustrate theorems and calculate probabilities. **You do not need to learn any programming language.** You will be instructed to learn basic commands like: `2+2; mean(x); plot(x,y)`.
  - In this module we will demonstrate using the `R` package. A nicer experience is provided by the commercial, but still freely available, `R Studio` software. Please feel free to use that.

- **Chapter 2: Introduction to Probability.** We will define and interpret probability as a measure of uncertainty. We will learn the rules of probability and then explore **fun examples of probability**, e.g. the probability of winning the National Lottery.

- **Chapter 3: Random variables.** We will learn that the results of different random experiments lead to different random variables following distributions such as the binomial, normal, etc. We will learn their basic properties, e.g. mean and variance.

- **Chapter 4: Statistical Inference.** We will discuss basic ideas of statistical inference, including techniques of point and interval estimation and hypothesis testing.

1.1.6 Take home points:

- We apply statistical methods whenever there is uncertainty and complete enumeration is not possible.

- This module will provide a very gentle introduction to statistics and probability together with the software package `R` for data analysis.

- Statistical knowledge is essential for any scientific career in academia, industry and government.

- Read the New York Times article **For Today's Graduate, Just One Word: Statistics** (search on Google).

- Watch the YouTube video **Joy of Statistics** before attending the next lecture.

1.2 Lecture 2: Guessing ages

1.2.1 Lecture mission

In this lecture, we will illustrate the concept of data collection and variation by guessing the ages of ten Southampton mathematicians from their photos. Every student needs to get involved and will have the opportunity to try their skill and luck in guessing the ages of people. This is a competition to find out which group or groups can best estimate the ages. We will then use the data for the statistical report writing part of the assessment, and also to illustrate various statistical concepts with easy-to-understand numbers and diagrams.

1.2.2 Experiment details

Here are the sequential steps that will be followed in the class:

1. All students are asked to be seated promptly and to form a group of four students each starting from the left most person in a row and then continuing.
2. Students left over at the end of a row (which could be at most three) are asked to move elsewhere so that a group of four can be formed. It does not matter if you are seated next to your friend as it is only a scientific data collection exercise.
3. In the extreme cases groups of size three will be allowed to ease the problem of accommodating everyone. It is compulsory that every student must be a member of a group - no students should be left out.
4. Each group should nominate an anchor person who will fill in the form and communicate the guesses.
5. At this point classroom assistants will distribute identical photo-cards and a form to fill in.
6. Each group will have up to 15 minutes to guess the ages of the people appearing in the photographs. The anchor person must then note down the guesses and fill in the rest of the form.

1.2.3 Revealing the truth and error calculation

1. Once every group has filled in their form, the true ages will be revealed on the data projector.
2. The anchor person is then asked to note down the errors and calculate average **absolute** errors.
3. Once this has been done the anchor person should hand over the form to one of the classroom assistants. *You are allowed to take a picture of the filled-in form before handing it in.*
4. If time permits we will enter the absolute errors only in a live displayed spreadsheet so that the groups can see the variation of the errors and see for themselves how they fared in the guessing game. The individual group errors will not be revealed in the class to avoid possible embarrassment! Below is how we will protect your privacy.

1.2.4 Protecting your privacy

1. Do not write your names anywhere on the form.

2. The form will not collect any personal information.

3. For the interest of science and curiosity, the form will collect aggregate information on:

   (a) The number of students in the group.

   (b) Sex. This should be either male, female or mixed. If there is at least one male and one female student in each group, please choose the mixed option for your group. This option can also be chosen if you would rather not reveal the sex of your group.

1.2.5 An example data set

♡ **Example 1 Errors in age guessing** Table 1.1 provides an example data set taken from the book *Teaching Statistics: A bag of tricks* by Andrew Gelman and Deborah Nolan, publisher Oxford University Press.

Table 1.1: Errors (estimated age minus actual age) for 10 photographed individuals (numbered 1 to 10) by 10 groups of students, A-J. The group sizes and sexes of the groups are given in the last two columns of the table.

	Photograph number										Average	# in	
Group	1	2	3	4	5	6	7	8	9	10	abs error	group	Sex
A	+14	–6	+5	+19	0	–5	–9	–1	–7	–7	7.3	3	F
B	+8	0	+5	0	+5	–1	–8	–1	–8	0	3.7	4	F
C	+6	–5	+6	+3	+1	–8	–18	+1	–9	–6	6.1	4	M
D	+10	–7	+3	+3	+2	–2	–13	+6	–7	–7	6.0	2	M/F
E	+11	–3	+4	+2	–1	0	–17	0	–14	+3	5.5	2	F
F	+13	–3	+3	+5	–2	–8	–9	–1	–7	0	5.1	3	F
G	+9	–4	+3	0	+4	–13	–15	+6	–7	+5	6.6	4	M
H	+11	0	+2	+8	+3	+3	–15	+1	–7	0	5.0	4	M
I	+6	–2	+2	+8	+3	–8	–7	–1	+1	–2	4.0	4	F
J	+11	+2	+3	+11	+1	–8	–14	–2	–1	0	5.3	4	F
Truth	29	30	14	42	37	57	73	24	82	45			

1.2.6 Take home points

We have generated an interesting data set on age-guessing. In future lectures we will learn about methods of data exploration to reveal the student's skills in age-guessing. The generated data will be made available to you.

1.3 Lecture 3: Basic statistics

1.3.1 Lecture mission

In Lecture 1 we got a glimpse of the nature of uncertainty and statistics. In this lecture we get our hands dirty with data and learn a bit more about it.

How do I obtain data? How do I summarise it? Which is the best measure among mean, median and mode? What do I make of the spread of the data?

1.3.2 How do I obtain data?

How should we collect data in the first place? Unless we can ask everyone in the population we should select individuals randomly (haphazardly) in order to get a representative sample. Otherwise we may introduce bias. For example, in order to gauge student opinion in this class I should not only survey the international students. But there are cases when systematic sampling may be preferable. For example, selecting every third caller in a radio phone-in show for a prize, or sampling air pollution hourly or daily. There is a whole branch of statistics called *survey methods* or *sample surveys*, where these issues are studied.

As well as randomness, we need to pay attention to the **design** of the study. In a *designed experiment* the investigator controls the values of certain experimental variables and then measures a corresponding output or response variable. In *designed surveys* an investigator collects data on a randomly selected sample of a well-defined population. Designed studies can often be more effective at providing reliable conclusions, but are frequently not possible because of difficulties in the study.

We will return to the topics of survey methods and designed surveys later in Lecture 30. Until then we assume that we have data from n randomly selected sampling units, which we will conveniently denote by $x_1, x_2, \ldots, x_n$. We will assume that these values are numeric, either discrete like counts, e.g. number of road accidents, or continuous, e.g. heights of 4-year-olds, marks obtained in an examination. We will consider the following example:

♡ **Example 2 Fast food service time** The service times (in seconds) of customers at a fast-food restaurant. The first row is for customers who were served from 9–10AM and the second row is for customers who who were served from 2–3PM on the same day.

AM	38	100	64	43	63	59	107	52	86	77
PM	45	62	52	72	81	88	64	75	59	70

How can we explore the data?

1.3.3 Summarising data

- We summarise categorical (not numeric) data by tables. For example: 5 reds, 6 blacks etc.
- For numeric data $x_1, x_2, \ldots, x_n$, we would like to know the centre (measures of location or central tendency) and the spread or variability.

Measures of location

- We are seeking a representative value for the data $x_1, x_2, \ldots, x_n$ which should be a function of the data. If a is that representative value then how much error is associated with it?
- The total error could be the sum of squares of the deviations from a, SSE $= \sum_{i=1}^{n}(x_i - a)^2$ or the sum of the absolute deviations from a, SSA $= \sum_{i=1}^{n}|x_i - a|$.
- What value of a will minimise the SSE or the SSA? For SSE the answer is the sample mean and for SSA the answer is the sample median.

The sample mean minimises the SSE

- Let us define the sample mean by:

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- How can we prove the above assertion? Use the derivative method. Set $\frac{\partial}{\partial a}$ SSE $= 0$ and solve for a. Check the second derivative condition that it is positive at the solution for a. Try this at home.
- Following is an alternative proof that establishes a very important result in statistics.

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - a)^2 &= \sum_{i=1}^{n}(x_i - \bar{x} + \bar{x} - a)^2 \quad \{\text{Add and subtract } \bar{x}\} \\
&= \sum_{i=1}^{n}\left\{(x_i - \bar{x})^2 + 2(x_i - \bar{x})(\bar{x} - a) + (\bar{x} - a)^2\right\} \\
&= \sum_{i=1}^{n}(x_i - \bar{x})^2 + 2(\bar{x} - a)\sum_{i=1}^{n}(x_i - \bar{x}) + \sum_{i=1}^{n}(\bar{x} - a)^2 \\
&= \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - a)^2,
\end{aligned}$$

since $\sum_{i=1}^{n}(x_i - \bar{x}) = n\bar{x} - n\bar{x} = 0$.

  - Now note that: the first term is free of a; the second term is non-negative for any value of a. Hence the minimum occurs when the second term is zero, i.e. when $a = \bar{x}$.
  - This establishes the fact that

    the sum of (or mean) squares of the deviations from any number a is minimised when a is the mean.

  - In the proof we also noted that $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$. This is stated as:

    the sum of the deviations of a set of numbers from their mean is zero.

  - In statistics and in this module, you will come across these two facts again and again!
- The above justifies why we often use the mean as a representative value. For the service time data, the mean time in AM is 68.9 seconds and for PM the mean is 66.8 seconds.

The sample median minimises the SSA

- Here the derivative approach does not work since the derivative does not exist for the absolute function.
- Instead we use the following argument. First, order the observations:

  $$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}.$$

  For the AM service time data: $38 < 43 < 52 < 59 < 63 < 64 < 77 < 86 < 100 < 107$.
- Now note that:

  $$\text{SSA} = \sum_{i=1}^{n} |x_i - a| = \sum_{i=1}^{n} |x_{(i)} - a| = |x_{(1)} - a| + |x_{(n)} - a| + |x_{(2)} - a| + |x_{(n-1)} - a| + \cdots$$

- Easy to argue that $|x_{(1)} - a| + |x_{(n)} - a|$ is minimised when a is such that $x_{(1)} \leq a \leq x_{(n)}$.
- Easy to argue that $|x_{(2)} - a| + |x_{(n-1)} - a|$ is minimised when a is such that $x_{(2)} \leq a \leq x_{(n-1)}$.
- Finally, when n is odd, the last term $|x_{(\frac{n+1}{2})} - a|$ is minimised when $a = x_{(\frac{n+1}{2})}$ or the middle value in the ordered list.
- If however, n is even, the last pair of terms will be $|x_{(\frac{n}{2})} - a| + |x_{(\frac{n}{2}+1)} - a|$. This will be minimised when a is any value between $x_{(\frac{n}{2})}$ and $x_{(\frac{n}{2}+1)}$. For convenience, we often take the mean of these as the middle value.
- Hence the middle value, popularly known as the median, minimises the SSA. Hence the median is also often used as a representative value or a measure of central tendency. This establishes the fact that:

the sum of (or mean) of the absolute deviations from any number a is minimised when a is the median.

- To recap: the median is defined as the observation ranked $\frac{1}{2}(n+1)$ in the ordered list if n is odd. If n is even, the median is any value between $\frac{n}{2}$th and $(\frac{n}{2}+1)$th in the ordered list. For example, for the AM service times, $n = 10$ and $38 < 43 < 52 < 59 < 63 < 64 < 77 < 86 < 100 < 107$. So the median is any value between 63 and 64. For convenience, we often take the mean of these. So the median is 63.5 seconds. Note that we use the unit of the observations when reporting any measure of location.

The sample mode minimises the average of a 0-1 error function.
The mode or the most frequent (or the most likely) value in the data is taken as the most representative value if we consider a 0-1 error function instead of the SSA or SSE above. Here, one assumes that the error is 0 if our guess a is the correct answer and 1 if it is not. It can then be proved that (proof not required) the best guess a will be the mode of the data.

Which of the three (mean, median and mode) would you prefer?
The mean gets more affected by extreme observations while the median does not. For example for the AM service times, suppose the next observation is 190. The median will be 64 instead of 63.5 but the mean will shoot up to 79.9.

Measures of spread

- A quick measure of the spread is the *range*, which is defined as the difference between the maximum and minimum observations. For the AM service times the range is 69 $(107 - 38)$ seconds.
- Standard deviation: square root of variance $= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right) = \sum_{i=1}^{n} x_i^2 - 2\bar{x}(n\bar{x}) + n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2.$$

Hence we calculate variance by the formula:

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

- Sometimes the variance is defined with the divisor n instead of $n-1$. We have chosen $n-1$ since this is the default in `R`. We will return to this in Chapter 4.
- The standard deviation (sd)for the AM service times is 23.2 seconds. Note that it has the same unit as the observations.
- The interquartile range (IQR) is the difference between the third, Q_3 and first, Q_1 quartiles, which are respectively the observations ranked $\frac{1}{4}(3n+1)$ and $\frac{1}{4}(n+3)$ in the ordered list. Note that the median is the second quartile, Q_2. When n is even, definitions of Q_3 and Q_1 are similar to that of the median, Q_2. The IQR for the AM service times is $83.75 - 53.75 = 30$ seconds.

1.3.4 Take home points

- As a measure of location we have three choices: mean, median and mode, each of which is optimal under a different consideration.
- We have discussed three measures of spread: range, sd and the IQR.
- Additional examples and mathematical details are provided in Section A.1.

1.4 Lecture 4: Data visualisation with `R`

1.4.1 Lecture mission

How do I graphically explore the data? (See two data sets below). Which of the data points are outliers? Hence, we have two urgent needs:

1. Need a computer-based calculator to calculate the summary statistics!
2. Need to use a computer software package to visualise our data!

Our mission in this lecture is to get started with `R`. We will learn the basic `R` commands (`mean, var, summary, table, hist` and `boxplot`) to explore data sets.

♡ **Example 3 Computer failures**
Weekly failures of a university computer system over a period of two years: $4, 0, 0, 0, \ldots, 4, 2, 13$.

4	0	0	0	3	2	0	0	6	7
6	2	1	11	6	1	2	1	1	2
0	2	2	1	0	12	8	4	5	0
5	4	1	0	8	2	5	2	1	12
8	9	10	17	2	3	4	8	1	2
5	1	2	2	3	1	2	0	2	1
6	3	3	6	11	10	4	3	0	2
4	2	1	5	3	3	2	5	3	4
1	3	6	4	4	5	2	10	4	1
5	6	9	7	3	1	3	0	2	2
1	4	2	13						

♡ **Example 4 Weight gain of students**
Is it true that students tend to gain weight during their first year in college? Cornell Professor of Nutrition, David Levitsky, recruited students from two large sections of an introductory health course. Although they were volunteers, they appeared to match the rest of the freshman class in terms of demographic variables such as sex and ethnicity. 68 students were weighed during the first week of the semester, then again 12 weeks later.

student number	initial weight (kg)	final weight (kg)
1	77.56423	76.20346
2	49.89512	50.34871
⋮	⋮	⋮
67	75.74986	77.11064
68	59.42055	59.42055

1.4.2 Get into R

- It is very strongly recommended that you download and install `R` (and `Rstudio` optionally) on your own computer.
- On a university workstation go to: **All Programs** → **Statistics** → `R`
- A small window (*the R console*) will appear, and you can type commands at the prompt >
- `R` functions are typically of the form `function(arguments)`
- Example: `mean(c(38, 100, 64, 43, 63, 59, 107, 52, 86, 77))` and hitting the Enter button computes the mean of the numbers entered.
- The letter `c()` is also a command that concatenates (collects) the input numbers into a vector

- Use `help(mean)` or simply `?mean` for more information.
- Even when an `R` function has no arguments we still need to use the brackets, such as in `ls()` which gives a list of objects in the current workspace.
- The assignment operator is `<-` or `=`.
- Example: `x <- 2+2` means that `x` is assigned the value of the expression 2+2.
- Comments are inserted after a `#` sign. For example, `# I love R`.

1.4.3 Keeping and saving commands in a script file

- To easily modify (edit) previous commands we can use the up (↑) and down (↓) arrow keys.
- However, we almost never type the long `R` commands at the `R` prompt > as we are prone to making mistakes and we may need to modify the commands for improved functionality.
- That is why we prefer to simply write down the commands one after another in a script file and save those future use.
  - **File** → **New File** → **R Script** (for a new script).
  - **File** → **Open File** (for an existing script).
- You can either execute the entire script or only parts by highlighting the respective commands and then clicking the Run button or Ctrl + R to execute.

1.4.4 How do I get my data into `R`?

`R` allows many different ways to read data.

- To read just a vector of numbers separated by tab or space use `scan("filename.txt")`.
- To read a tab-delimited text file of data with the first row giving the column headers, the command is: `read.table("filename.txt", head=TRUE)`.
- For comma-separated files (such as the ones exported by EXCEL), the command is

  `read.table("filename.csv", head=TRUE, sep=",")` or simply

  `read.csv("filename.csv", head=TRUE)`.
- The option `head=TRUE` tells that the first row of the data file contains the column headers.
- For these commands to work we need to tell `R` where in the computer these data files are stored.
- In the university workstations ther is a drive called `H:` which is permanent (will be there for you to use throughout your degree programme.
- From Windows File Explorer navigate to `H:` and create a sub-folder `math1024` and then a sub-folder `data` within `math1024`.

- Throughout this module we will assume that our data directory will be the `data` sub-folder within the `math1024` folder. Hence the directory path is: `H:/math1024/data`
- We use the set working directory command by: `setwd("H:/math1024/data")`
- To get (e.g. to print) the current working directory, simply type `getwd()` and press Enter.
- *You are reminded that the following data reading commands will fail if you have not set the working directory correctly.*
- Assuming that you have set the working directory to where your data files are saved, simply type and enter
  - `cfail <- scan("compfail.txt")`
  - `ffood <- read.csv("servicetime.csv", head=T)`
  - `wgain <- read.table("wtgain.txt", head=T)`
  - `head=T` tells that the first line in the file contains the header (names) of the columns.
- That's all the hard work needed to get the data into `R`!

1.4.5 Working with data in `R`

- The data read by the `read.table` and `read.csv` commands are saved as `data frames` in `R`. These are versatile matrices which can hold numeric as well as character data. You will see this in the billionaires data set later.
- Just type `ffood` and hit the `Enter` button or the `Run` icon. See what happens.
- A convenient way to see the data is to see either the head or the tail of the data. For example, type `head(ffood)` and Enter or `tail(ffood)` and Enter.
- To know the dimension (how many rows and columns) issue `dim(ffood)`.
- To access elements of a data frame we can use square brackets, e.g. `ffood[1, 2]` gives the first row second column element, `ffood[1, ]` gives everything in the first row and `ffood[, 1]` gives everything in the first column.
- The named columns in a data frame are often accessed by using the $ operator. For example, `ffood$AM` prints the column whose header is `AM`.
- So, what do you think `mean(ffood$AM)` will give?
- There are many `R` functions with intuitive names, e.g. `mean, median, var, min, max, sum, prod, summary, seq, rep` etc. We will explain them as we need them.

1.4.6 Summary statistics from R

- Use `summary(ffood)`; `summary(cfail)`; `summary(wgain)`; to get the summaries.
- What does the command `table(cfail)` give?
- To calculate variance, try `var(ffood$AM)`. What does the command `var(c(ffood$AM, ffoood$PM)` give?

1.4.7 Graphical exploration using R

- The commands are `stem`, `hist`, `plot` and `boxplot`.
- A stem and leaf diagram is produced by the command `stem`. Issue the command `stem(ffood$AM)` and `?stem` to learn more.
- Histograms are produced by `hist(cfail)`.
- Modify the command so that it looks a bit nicer: `hist(cfail, xlab="Number of weekly computer failures")`
- To obtain a scatter plot of the before and after weights of the students, we issue the command `plot(wgain$initial, wgain$final)`
- Add a 45^o degree line by `abline(0, 1, col="red")`
- A nicer and more informative plot can be obtained by: `plot(wgain$initial, wgain$final, xlab="Wt in Week 1", ylab="Wt in Week 12", pch="*", las=1)`

  `abline(0, 1, col="red")`

  `title("A scatterplot of the weights in Week 12 against the weights in Week 1")`
- You can save the graph in any format you like using the menus.
- To draw boxplots use the boxplot command, e.g., `boxplot(cfail)`
- The default boxplot shows the median and whiskers drawn to the nearest observation from the first and third quartiles but not beyond the distance 1.5 times the inter-quartile range. Points beyond the two whiskers are suspected outliers and are plotted individually.
- `boxplot(ffood)` generates two boxplots side-by-side: one for the AM service times and the other for the PM service times.
- Various parameters of the plot are controlled by the `par` command. To learn about these type `?par`.

1.4.8 Take home points

This lecture provided an opportunity to get started with `R`. It is expected that students will install `R` (and `Rstudio` optionally) on their computer. If that is not possible then they should use a university workstation to go through the basic commands to read and graphically display data. All the commands in this lecture can be downloaded from the course content section of the blackboard site for MATH1024 in the folder `R commands`. The data files are in the `data` sub-folder. This week's exercise is to go through the commands in the file `Rfile1.R`. A challenging exercise is to draw the butterfly with different colours as instructed in the file. This may help you gain advanced understanding of the `R`-language and programming.

`R` is intuitive and easy to learn, and there are a wealth of community resources online. `R` is not a spreadsheet programme like `EXCEL` - it produces higher-quality graphics and is excellent for advanced statistical methods where `EXCEL` will fail. `R` will be used throughout this module and in all subsequent Statistics modules in years 2 and 3.

1.5 Lecture 5: Help with report writing

1.5.1 Lecture mission

In this lecture we discuss how to write a statistical report. We will discuss the main sections of a report and illustrate data analysis with the age-guessing data presented in Section 1.2.5.

1.5.2 Guidance on data inspection, analysis and conclusions

1. Check the data to look for any unusual observations. (Hopefully any major mistakes will have been identified and corrected during the experiment!)

2. Obtain simple summaries of the whole data and also by different groups as shown in the sample report.

3. Provide comments on the summaries.

4. Provide histograms of the error values and also the absolute errors. How do the shapes change and why?

5. Provide side-by-side boxplots separately for each classifier of the data, e.g, gender, true age, race.

6. Other investigations you may undertake: (i) ranking of the groups according to average absolute accuracy, (ii) identifying patterns in accuracy, for example, is it easier to guess the ages of older people? (iii) which of the variables could be the best predictor of the true ages: true age, race, sex?

7. You are also free to investigate to discover other patterns which may be present in the data. There may be other patterns lurking around

8. Based on your experience of the experiment, try to suggest an aspect of the experimental procedure which might be changed to improve the experiment in the future, or any additional

information which might usefully be included in future experiments to improve the data collection and analysis.

1.5.3 Some advice on writing your report

1. **The page limit**, 5 sides of A4 paper for the report including the `R` code written by you, is strict and is easily sufficient to receive high grades. You do not need to include the data set in the page containing code. All relevant materials, including plots and any appendices, must fall within these limits. Marks will be deducted for work which exceeds the page limits. If you have too much material then you will need to decide what is important and what can be left out. Printed output must not be smaller than 10pt font (Arial or similar).

2. **Including code and output.** You are asked to include `R` code, suitably edited, in your submission, but reports which include excessive and unnecessary `R` output will be penalised – you should extract exactly what you need and nothing else, from `R` to incorporate in your report.

3. The project involves the modelling of real data. There is not necessarily a single 'correct' form of analysis. You may explore your own ideas in analysis which you think are plausible. Submissions which demonstrate a good appreciation of statistical data analysis, together with correct application of appropriate methods will receive high marks. Careful explanation and clear presentation are also important.

4. **General advice and guidance.**

   (a) Choose an informative title.

   (b) Number the main sections so you can refer to them; for example: 'see Section 4', 'as was explained in Section 1'.

   (c) Always use a spell checker - spelling mistakes will distract the reader and undermine the impact of your report.

   (d) In scientific writing, short sentences are often clearer than longer sentences that are concerned with several points. [Try reading your report aloud to check where your sentences should end.]

   (e) Use paragraphs to group together sentences that relate to the same theme.

   (f) Include only tables and figures of particular importance in the text.

   (g) Make sure that you discuss any tables or figures that you do include in the text. What do they show?

   (h) Any table in the report should have a number and a caption which describes exactly what is given in the table. Numbers in the table should have only a modest number of significant figures (no more than 3) so that the reader can easily appreciate their size. Units of measurement should be given.

   (i) Any graph (figure) should have a number and a caption, labelled axes and units of measurement (on the axes or in the caption).

(j) You must copy-paste the `R` commands you have used to produce the summaries and graphs. You have the option of putting those sequentially as in the sample report or you can put those in an Appendix. But you cannot exceed the **page-limit of 5 sides.**

(k) Number the pages in your report.

A list of texts (with descriptions) on writing skills is available at
http://www.mantex.co.uk/reviews/writing-skills-reviews/

1.5.4 Take home points

In this lecture, we have discussed a sample report and learned various aspects of statistical report writing.

Chapter 2

Introduction to Probability

Chapter mission

Why should we study probability? What are probabilities? How do you find them? What are the main laws of probabilities? How about some fun examples where probabilities are used to solve real-life problems?

2.1 Lecture 6: Definitions of probability

2.1.1 Why should we study probability?

Probabilities are often used to express the uncertainty of events of interest happening. For example, we may say that: (i) it is highly likely that Manchester City will retain the premiership title this season or to be more specific, I think there is more than an 80% chance that Manchester City will keep the title; (ii) the probability of a tossed fair coin landing heads is 0.5. So it is clear that probabilities mean different things to different people. As we have seen in the previous chapter, there is uncertainty everywhere. Hence, probabilities are used as tools to quantify the associated uncertainty. The theory of statistics has its basis in the mathematical theory of probability. A statistician must be fully aware of what probability means to him/her and what it means to other people. In this lecture we will learn the basic definitions of probability and how to find them.

2.1.2 Two types of probabilities: subjective and objective

The two examples above, Manchester City and tossing a coin, convey two different interpretations of probability. The Manchester City probability is the commentator's own subjective belief, isn't it? The commentator certainly has not performed a large experiment involving all the 20 teams over the whole (future) season under all playing conditions, players, managers and transfers. This notion is known as subjective probability. *Subjective probability* gives a measure of the plausibility of the proposition, to the person making it, in the light of past experience (e.g. Manchester City are the current champions) and other evidence (e.g. they spent the maximum amount of money buying players). There are plenty of other examples, e.g. I think there is a 70% chance that the FTSE 100 will rise tomorrow, or according to the Met Office there is a 40% chance that we will have a white Christmas this year in Southampton. Subjective probabilities are nowadays used cleverly

in a statistical framework called Bayesian inference. Such methods allow one to combine expert opinion and evidence from data to make the best possible inferences and prediction. Unfortunately discussion of Bayesian inference methods is beyond the scope of this module, although we will talk about it when possible.

The second definition of probability comes from the long-term relative frequency of a result of a random experiment (e.g. coin tossing) which can be repeated an infinite number of times under *essentially similar conditions*. First we give some essential definitions.

Random experiments. The experiment is random because in advance we do not know exactly what *outcome* the experiment will give, even though we can write down all the possible outcomes which together are called the **sample space** (`S`). For example, in a coin tossing experiment, `S` = {head, tail}. If we toss two coins together, `S` = {HH, HT, TH, TT} where H and T denote respectively the outcome head and tail from the toss of a single coin.

Event. An event is defined as a particular result of the random experiment. For example, HH (two heads) is an event when we toss two coins together. Similarly, at least one head e.g. {HH, HT, TH} is an event as well. Events are denoted by capital letters $A, B, C, \ldots$ or A_1, B_1, A_2 etc., and a single outcome is called an *elementary event*, e.g. HH. An event which is a group of elementary events is called a composite event, e.g. at least one head. How to determine the probability of a given event A, $P\{A\}$, is the focus of probability theory.

Probability as relative frequency. Imagine we are able to repeat a random experiment under identical conditions and count how many of those repetitions result in the event A. The relative frequency of A, i.e. the ratio

$$\frac{\text{the number of repetitions resulting in } A}{\text{total number of repetitions}},$$

approaches a fixed limit value as the number of repetitions increases. This limit value is defined as $P\{A\}$.

As a simple example, in the experiment of tossing a particular coin, suppose we are interested in the event A of getting a 'head'. We can toss the coin 1000 times (i.e. do 1000 replications of the experiment) and record the number of heads out of the 1000 replications. Then the relative frequency of A out of the 1000 replications is the proportion of heads observed.

Sometimes, however, it is much easier to find $P\{A\}$ by using some 'common knowledge' about probability. For example, if the coin in the example above is fair (i.e. $P\{\text{'head'}\} = P\{\text{'tail'}\}$), then this information and the common knowledge that $P\{\text{'head'}\} + P\{\text{'tail'}\} = 1$ immediately imply that $P\{\text{'head'}\} = 0.5$ and $P\{\text{'tail'}\} = 0.5$. Next, the essential 'common knowledge' about probability will be formalized as the *axioms of probability*, which form the foundation of probability theory. But before that, we need to learn a bit more about the event space (collection of all events).

2.1.3 Union, intersection, mutually exclusive and complementary events

For us to proceed we need to establish parallels between probability theory and set theory, which is taught in calculus. The sample space S is called the whole set and it is composed of all possible

elementary events (outcomes from a single replicate).

♡ **Example 5 Die throw** Roll a six-faced die and observe the score on the uppermost face. Here $S = \{1, 2, 3, 4, 5, 6\}$, which is composed of six elementary events.

The *union of two given events A and B*, denoted as (A or B) or $A \cup B$, consists of the outcomes that are either in A or B or both. 'Event $A \cup B$ occurs' means 'either A or B occurs or both occur'.

For example, in Example 5, suppose A is the event that *an even number is observed*. This event consists of the set of outcomes 2, 4 and 6, i.e. $A = \{\text{an even number}\} = \{2, 4, 6\}$. Suppose B is the event that *a number larger than 3 is observed*. This event consists of the outcomes 4, 5 and 6, i.e. $B = \{\text{a number larger than 3}\} = \{4, 5, 6\}$. Hence the event $A \cup B = \{\text{an even number or a number larger than 3}\} = \{2, 4, 5, 6\}$. Clearly, when a 6 is observed, both A and B have occurred.

The *intersection of two given events A and B*, denoted as (A and B) or $A \cap B$, consists of the outcomes that are common to both A and B. 'Event $A \cap B$ occurs' means 'both A and B occur'. For example, in Example 5, $A \cap B = \{4, 6\}$. Additionally, if $C = \{\text{a number less than 6}\} = \{1, 2, 3, 4, 5\}$, the intersection of events A and C is the event $A \cap C = \{\text{an even number less than 6}\} = \{2, 4\}$.

The union and intersection of two events can be generalized in an obvious way to the union and intersection of more than two events.

Two events A and D are said to be *mutually exclusive* if $A \cap D = \emptyset$, where $\emptyset$ denotes the empty set, i.e. A and D have no outcomes in common. Intuitively, 'A and D are mutually exclusive' means 'A and D cannot occur simultaneously in the experiment'.

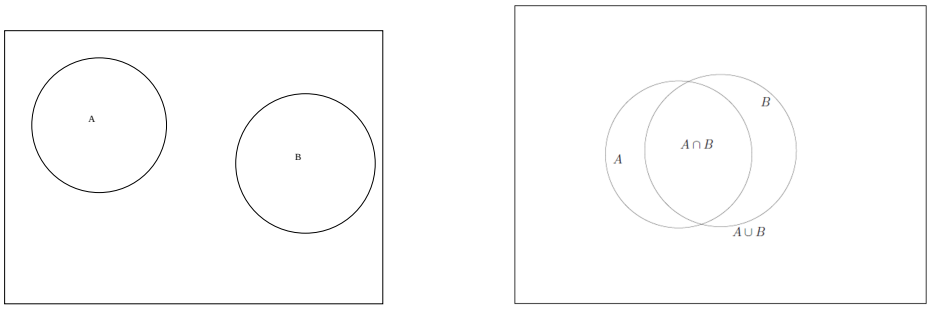


Figure 2.1: In the left plot A and B are mutually exclusive; the right plot shows $A \cup B$ and $A \cap B$.

In Example 5, if $D = \{\text{an odd number}\} = \{1, 3, 5\}$, then $A \cap D = \emptyset$ and so A and D are mutually exclusive. As expected, A and D cannot occur simultaneously in the experiment.

For a given event A, *the complement of A* is the event that consists of all the outcomes not in A and is denoted by A'. Note that $A \cup A' = S$ and $A \cap A' = \emptyset$.

Thus, we can see the parallels between Set theory and Probability theory:

	Set theory	Probability theory
(1)	Space	Sample space
(2)	Element or point	Elementary event
(3)	Set	Event

2.1.4 Axioms of probability

Here are the three axioms of probability:

A1 $P\{S\} = 1$,

A2 $0 \leq P\{A\} \leq 1$ for any event A,

A3 $P\{A \cup B\} = P\{A\} + P\{B\}$ provided that A and B are mutually exclusive events.

Here are some of the consequences of the axioms of probability:

(1) For any event A, $P\{A\} = 1 - P\{A'\}$.

(2) From (1) and Axiom **A1**, $P\{\emptyset\} = 1 - P\{S\} = 0$. Hence if A and B are mutually exclusive events, then $P\{A \cap B\} = 0$.

(3) If D is a subset of E, $D \subset E$, then $P\{E \cap D'\} = P\{E\} - P\{D\}$ which implies for arbitrary events A and B, $P\{A \cap B'\} = P\{A\} - P\{A \cap B\}$.

(4) It can be shown by mathematical induction that Axiom **A3** holds for more than two mutually exclusive events:

$$P\{A_1 \cup A_2 \cup \cdots \cup A_k\} = P\{A_1\} + P\{A_2\} + \ldots + P\{A_k\}$$

provided that $A_1, \ldots, A_k$ are mutually exclusive events.

Hence, the probability of an event A is the sum of the probabilities of the individual outcomes that make up the event.

(5) For the union of two arbitrary events, we have the *General addition rule*: For any two events A and B,

$$\boxed{P\{A \cup B\} = P\{A\} + P\{B\} - P\{A \cap B\}.}$$

Proof: We can write $A \cup B = (A \cap B') \cup (A \cap B) \cup (A' \cap B)$. All three of these are mutually exclusive events. Hence,

$$\begin{aligned} P\{A \cup B\} &= P\{A \cap B'\} + P\{A \cap B\} + P\{A' \cap B\} \\ &= P\{A\} - P\{A \cap B\} + P\{A \cap B\} + P\{B\} - P\{A \cap B\} \\ &= P\{A\} + P\{B\} - P\{A \cap B\}. \end{aligned}$$

(6) The sum of the probabilities of all the outcomes in the sample space S is 1.

2.1.5 Application to an experiment with equally likely outcomes

For an experiment with N equally likely possible outcomes, the axioms (and the consequences above) can be used to find $P\{A\}$ of any event A in the following way.

From consequence (4), we assign probability $1/N$ to each outcome.

For any event A, we find $P\{A\}$ by adding up $1/N$ for each of the outcomes in event A:

$$P\{A\} = \frac{\text{number of outcomes in } A}{\text{total number of possible outcomes of the experiment}}.$$

Return to Example 5 where a six-faced die is rolled. Suppose that one wins a bet if a 6 is rolled. Then the probability of winning the bet is 1/6 as there are six possible outcomes in the sample space and exactly one of those, 6, wins the bet. Suppose A denotes the event that an even-numbered face is rolled. Then $P\{A\} = 3/6 = 1/2$ as we can expect.

♡ **Example 6 Dice throw** Roll 2 distinguishable dice and observe the scores. Here $S = \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),\ldots,(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\}$ which consists of 36 possible outcomes or elementary events, $A_1,\ldots,A_{36}$. What is the probability of the outcome 6 in both the dice? The required probability is 1/36. What is the probability that the sum of the two dice is greater than 6? How about the probability that the sum is less than any number, e.g. 8? **Hint:** Write down the sum for each of the 36 outcomes and then find the probabilities asked just by inspection. Remember, each of the 36 outcomes has equal probability 1/36.

2.1.6 Take home points

This lecture has laid the foundation for studying probability. We discussed two types of probabilities, subjective and objective by relative frequencies. Using three axioms of probability we have derived the elementary rules for probability. We then discussed how we can use elementary laws of probability to find the probabilities of some events from the dice throw example.

The next lecture will continue to find probabilities using specialist counting techniques called permutation and combination. This will allow us to find probabilities in a number of practical situations.

2.2 Lecture 7: Using combinatorics to find probability

2.2.1 Lecture mission

We will learn common counting techniques. Suppose there are 4 boys and 6 girls available for a committee membership, but there are only 3 posts. How many possible committees can be formed? How many of those will be girls only?

The UK National Lottery selects 6 numbers at random from 1 to 49. I bought one ticket - what is the probability that I will win the jackpot?

2.2.2 Multiplication rule of counting

To complete a specific task, one has to complete $k(\geq 1)$ sub-tasks sequentially. If there are n_i different ways to complete the i-th sub-task $(i = 1, \ldots, k)$ then there are $n_1 \times n_2 \times \ldots \times n_k$ different ways to complete the task.

♡ **Example 7 Counting** Suppose there are 7 routes to London from Southampton and then there are 5 routes to Cambridge out of London. How many ways can I travel to Cambridge from Southampton via London. The answer is obviously 35.

The number of permutations of k from n: ${}^nP_k$

The task is to select $k(\geq 1)$ from the n $(n \geq k)$ available people and sit the k selected people in k (different) chairs. By considering the i-th sub-task as selecting a person to sit in the i-th chair $(i = 1, \ldots, k)$, it follows directly from the multiplication rule above that there are $n(n-1)\cdots(n-[k-1])$ ways to complete the task. The number $n(n-1)\cdots(n-[k-1])$ is called the number of permutations of k from n and denoted by

$$ {}^nP_k = n(n-1)\cdots(n-[k-1]). $$

In particular, when $k = n$ we have ${}^nP_n = n(n-1)\cdots 1$, which is called 'n factorial' and denoted as $n!$. Note that 0! is defined to be 1. It is clear that

$$ {}^nP_k = n(n-1)\cdots(n-[k-1]) = \frac{n(n-1)\cdots(n-[k-1])\times(n-k)!}{(n-k)!} = \frac{n!}{(n-k)!}. $$

♡ **Example 8 Football** How many possible rankings are there for the 20 football teams in the premier league at the end of a season? This number is given by ${}^{20}P_{20} = 20!$, which is a huge number! How many possible permutations are there for the top 4 positions who will qualify to play in Europe in the next season? This number is given by ${}^{20}P_4 = 20 \times 19 \times 18 \times 17$.

The number of combinations of k from n: ${}^nC_k$ or $\binom{n}{k}$

The task is to select $k(\geq 1)$ from the n $(n \geq k)$ available people. Note that this task does NOT involve sitting the k selected people in k (different) chairs. We want to find the number of possible ways to complete this task, which is denoted as ${}^nC_k$ or $\binom{n}{k}$.

For this, let us reconsider the task of "selecting $k(\geq 1)$ from the n $(n \geq k)$ available people and sitting the k selected people in k (different) chairs", which we already know from the discussion above has ${}^nP_k$ ways to complete.

Alternatively, to complete this task, one has to complete two sub-tasks sequentially. The first sub-task is to select $k(\geq 1)$ from the n $(n \geq k)$ available people, which has ${}^nC_k$ ways. The second sub-task is to sit the k selected people in k (different) chairs, which has $k!$ ways. It follows directly from the multiplication rule that there are ${}^nC_k \times k!$ to complete the task. Hence we have

$$ {}^nP_k = {}^nC_k \times k!, \quad i.e., \ {}^nC_k = \frac{{}^nP_k}{k!} = \frac{n!}{(n-k)!k!} $$

♡ **Example 9 Football** How many possible ways are there to choose 3 teams for the bottom positions of the premier league table at the end of a season? This number is given by ${}^{20}C_3 = 20 \times 19 \times 18/3!$, which does not take into consideration the rankings of the three bottom teams!

♡ **Example 10 Microchip** A box contains 12 microchips of which 4 are faulty. a sample of size 3 is drawn from the box *without replacement*.

- How many selections of 3 can be made? ${}^{12}C_3$.
- How many samples have all 3 chips faulty? ${}^{4}C_3$.
- How many selections have exactly 2 faulty chips? ${}^{8}C_1\,{}^{4}C_2$.
- How many samples of 3 have 2 or more faulty chips? ${}^{8}C_1\,{}^{4}C_2 + {}^{4}C_3$

More examples and details regarding the combinations are provided in Section A.2. You are strongly recommended to read that section now.

2.2.3 Calculation of probabilities of events under sampling 'at random'

For the experiment of 'selecting a sample of size n from a box of N items without replacement', a sample is said to be selected *at random* if all the possible samples of size n are equally likely to be selected. All the possible samples are then equally likely outcomes of the experiment and so assigned equal probabilities.

♡ **Example 11 Microchip continued** In Example 10 assume that 3 microchips are selected at random without replacement. Then

- each outcome (sample of size 3) has probability $1/{}^{12}C_3$.
- $P\{\text{all 3 selected microchips are faulty}\} = {}^{4}C_3/{}^{12}C_3$.
- $P\{\text{2 chips are faulty}\} = {}^{8}C_1\,{}^{4}C_2/{}^{12}C_3$.
- $P\{\text{2 or more chips are faulty}\} = ({}^{8}C_1\,{}^{4}C_2 + {}^{4}C_3)/{}^{12}C_3$.

2.2.4 A general 'urn problem'

Example 10 is one particular case of the following general urn problem which can be solved by the same technique.

A sample of size n is drawn at random without replacement from a box of N items containing a proportion p of defective items.

- How many defective items are in the box? Np. How many good items are there? $N(1-p)$. Assume these to be integers.

- The probability of exactly x number of defectives in the sample of n is
$$\frac{{}^{Np}C_x \; {}^{N(1-p)}C_{n-x}}{{}^{N}C_n}$$
- Which values of x (in terms of N, n and p) make this expression well defined?

We'll see later that these values of x and the corresponding probabilities make up what is called the *hyper-geometric distribution.*

♡ **Example 12 Selecting a committee** There are 10 students available for a committee of which 4 are boys and 6 are girls. A random sample of 3 students are chosen to form the committee - what is the probability that exactly one is a boy?

The total number of possible outcomes of the experiment is equal to the number of ways of selecting 3 students from 10 and given by ${}^{10}C_3$. The number of outcomes in the event 'exactly one is a boy' is equal to the number of ways of selecting 3 students from 10 with exactly one boy, and given by ${}^{4}C_1 \, {}^{6}C_2$.

Hence

$$\begin{aligned} P\{\text{exactly one boy}\} &= \frac{\text{number of ways of selecting one boy and two girls}}{\text{number of ways of selecting 3 students}} \\ &= \frac{{}^{4}C_1 \, {}^{6}C_2}{{}^{10}C_3} = \frac{4 \times 15}{120} = \frac{1}{2}. \end{aligned}$$

Similarly,

$$P\{\text{two boys}\} = \frac{{}^{4}C_2 \; {}^{6}C_1}{{}^{10}C_3} = \frac{6 \times 6}{120} = \frac{3}{10}.$$

♡ **Example 13 The National Lottery** In Lotto, a winning ticket has six numbers from 1 to 49 matching those on the balls drawn on a Wednesday or Saturday evening. The 'experiment' consists of drawing the balls from a box containing 49 balls. The 'randomness', the equal chance of any set of six numbers being drawn, is ensured by the spinning machine, which rotates the balls during the selection process. What is the probability of winning the jackpot?

Total number of possible selections of six balls/numbers is given by ${}^{49}C_6$.

There is only 1 selection for winning the jackpot. Hence

$$P\{\text{jackpot}\} = \frac{1}{{}^{49}C_6} = 7.15 \times 10^{-8}.$$

which is roughly 1 in 13.98 million.

Other prizes are given for fewer matches. The corresponding probabilities are:

$$P\{5 \text{ matches}\} = \frac{{}^{6}C_5 \; {}^{43}C_1}{{}^{49}C_6} = 1.84 \times 10^{-5}.$$

$$P\{4 \text{ matches}\} = \frac{{}^{6}C_4 \; {}^{43}C_2}{{}^{49}C_6} = 0.0009686197$$

$$P\{3 \text{ matches}\} = \frac{{}^{6}C_3 \; {}^{43}C_3}{{}^{49}C_6} = 0.0176504$$

There is one other way of winning by using the bonus ball – matching 5 of the selected 6 balls plus matching the bonus ball. The probability of this is given by

$$P\{5 \text{ matches } + \text{ bonus}\} = \frac{6}{{}^{49}C_6} = 4.29 \times 10^{-7}.$$

Adding all these probabilities of winning some kind of prize together gives

$$P\{\text{winning}\} \approx 0.0186 \approx 1/53.7.$$

So a player buying one ticket each week would expect to win a prize, (most likely a £10 prize for matching three numbers) about once a year

2.2.5 Take home points

We have learned the multiplication rule of counting and the number of permutations and combinations. We have applied the rules to find probabilities of interesting events, e.g. the jackpot in the UK National Lottery.

2.3 Lecture 8: Conditional probability and the Bayes Theorem

2.3.1 Lecture mission

This lecture is all about using additional information, i.e. things that have already happened, in the calculation of probabilities. For example, a person may have a certain disease, e.g. diabetes or HIV/AIDS, whether or not they show any symptoms of it. Suppose a randomly selected person is found to have the symptom. Given this additional information, what is the probability that they have the disease? Note that having the symptom does not fully guarantee that the person has the disease.

Applications of conditional probability occur naturally in actuarial science and medical studies, where conditional probabilities such as "what is the probability that a person will survive for another 20 years given that they are still alive at the age of 40?" are calculated.

In many real problems, one has to determine the probability of an event A when one already has some partial knowledge of the outcome of an experiment, i.e. another event B has already occurred. For this, one needs to find the conditional probability.

♡ **Example 14 Dice throw continued** Return to the rolling of a fair die (Example 5). Let

$$A = \{\text{a number greater than } 3\} = \{4, 5, 6\},$$
$$B = \{\text{an even number}\} = \{2, 4, 6\}.$$

It is clear that $P\{B\} = 3/6 = 1/2$. This is the unconditional probability of the event B. It is sometimes called the *prior* probability of B.

However, suppose that we are told that the event A has already occurred. What is the probability of B now given that A has already happened?

The sample space of the experiment is $S = \{1, 2, 3, 4, 5, 6\}$, which contains $n = 6$ equally likely outcomes.

Given the partial knowledge that event A has occurred, only the $n_A = 3$ outcomes in $A = \{4, 5, 6\}$ could have occurred. However, only some of the outcomes in B among these n_A outcomes in A will make event B occur; the number of such outcomes is given by the number of outcomes $n_{A\cap B}$ in both A and B, i.e., $A \cap B$, and equal to 2. Hence the probability of B, given the partial knowledge that event A has occurred, is equal to

$$\frac{2}{3} = \frac{n_{A\cap B}}{n_A} = \frac{n_{A\cap B}/n}{n_A/n} = \frac{P\{A \cap B\}}{P\{A\}}.$$

Hence we say that $P\{B|A\} = \frac{2}{3}$, which is often interpreted as the *posterior* probability of B given A. The additional knowledge that A has already occurred has helped us to revise the prior probability of 1/2 to 2/3.

This simple example leads to the following general definition of conditional probability.

2.3.2 Definition of conditional probability

For events A and B with $P\{A\} > 0$, the conditional probability of event B, given that event A has occurred, is

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

♡ **Example 15** Of all individuals buying a mobile phone, 60% include a 64GB hard disk in their purchase, 40% include a 16 MP camera and 30% include both. If a randomly selected purchase includes a 16 MP camera, what is the probability that a 64GB hard disk is also included?

The conditional probability is given by:

$$P\{\text{64GB}|\text{16 MP}\} = \frac{P\{\text{64GB} \cap \text{16 MP}\}}{P\{\text{16 MP}\}} = \frac{0.3}{0.4} = 0.75.$$

2.3.3 Multiplication rule of conditional probability

By rearranging the conditional probability definition, we obtain the multiplication rule of conditional probability as follows:

$$P\{A \cap B\} = P\{A\}P\{B|A\}.$$

Clearly the roles of A and B could be interchanged leading to:

$$P\{A \cap B\} = P\{B\}P\{A|B\}.$$

Hence the multiplication rule of conditional probability for two events is:

$$\boxed{P\{A \cap B\} = P\{B\}P\{A|B\} = P\{A\}P\{B|A\}.}$$

It is straightforward to show by mathematical induction the following multiplication rule of conditional probability for $k(\geq 2)$ events $A_1, A_2, \ldots, A_k$:

$$P\{A_1 \cap A_2 \cap \ldots \cap A_k\} = P\{A_1\}P\{A_2|A_1\}P\{A_3|A_1 \cap A_2\} \ldots P\{A_k|A_1 \cap A_2 \cap \ldots \cap A_{k-1}\}.$$

♡ **Example 16 Selecting a committee continued** Return to the committee selection example (Example 12), where there are 4 boys (B) and 6 girls (G). We want to select a 2-person committee. Find:

(i) the probability that both are boys,

(ii) the probability that one is a boy and the other is a girl.

We have already dealt with this type of urn problem by using the combinatorial method. Here, the multiplication rule is used instead.

Let B_i be the event that the i-th person is a boy, and G_i be the event that the i-th person is a girl, $i = 1, 2$. Then

$$\text{Prob in (i)} = P\{B_1 \cap B_2\} = P\{B_1\}P\{B_2|B_1\} = \frac{4}{10} \times \frac{3}{9}.$$

$$\text{Prob in (ii)} = P\{B_1 \cap G_2\} + P\{G_1 \cap B_2\} = P\{B_1\}P\{G_2|B_1\} + P\{G_1\}P\{B_2|G_1\} = \frac{4}{10} \times \frac{6}{9} + \frac{6}{10} \times \frac{4}{9}.$$

You can find the probability that 'both are girls' in a similar way.

2.3.4 Total probability formula

♡ **Example 17 Phones** Suppose that in our world there are only three phone manufacturing companies: A Pale, B Sung and C Windows, and their market shares are respectively 30, 40 and 30 percent. Suppose also that respectively 5, 8, and 10 percent of their phones become faulty within one year. If I buy a phone randomly (ignoring the manufacturer), what is the probability that my phone will develop a fault within one year? After finding the probability, suppose that my phone developed a fault in the first year - what is the probability that it was made by A Pale?

Company	Market share	Percent defective
A Pale	30%	5%
B Sung	40%	8%
C Windows	30%	10%

To answer this type of question, we derive two of the most useful results in probability theory: the total probability formula and the Bayes theorem. First, let us derive the total probability formula.

Let $B_1, B_2, \ldots, B_k$ be a set of mutually exclusive, i.e.

$$B_i \cap B_j = \emptyset \quad \text{for all } 1 \leq i \neq j \leq k,$$

and exhaustive events, i.e.:

$$B_1 \cup B_2 \cup \ldots \cup B_k = S.$$

Now any event A can be represented by

$$A = A \cap S = (A \cap B_1) \cup (A \cap B_2) \cup \ldots \cup (A \cap B_k)$$

where $(A \cap B_1), (A \cap B_2), \ldots, (A \cap B_k)$ are mutually exclusive events. Hence the Axiom **A3** of probability gives

$$\begin{aligned} P\{A\} &= P\{A \cap B_1\} + P\{A \cap B_2\} + \ldots + P\{A \cap B_k\} \\ &= P\{B_1\}P\{A|B_1\} + P\{B_2\}P\{A|B_2\} + \ldots + P\{B_k\}P\{A|B_k\}; \end{aligned}$$

this last expression is called **the total probability formula** for $P\{A\}$.

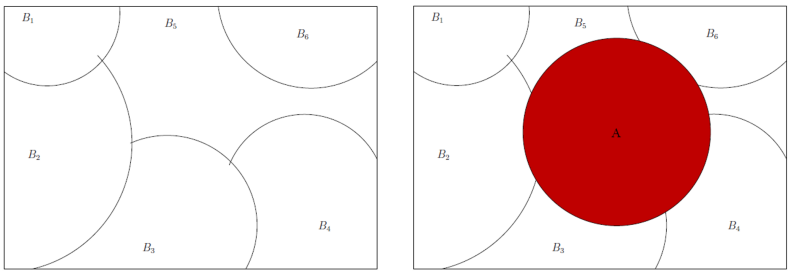


Figure 2.2: The left figure shows the mutually exclusive and exhaustive events $B_1, \ldots, B_6$ (they form a partition of the sample space); the right figure shows a possible event A.

♡ **Example 18 Phones continued** We can now find the probability of the event, say A, that a randomly selected phone develops a fault within one year. Let B_1, B_2, B_3 be the events that the phone is manufactured respectively by companies A Pale, B Sung and C Windows. Then we have:

$$\begin{aligned} P\{A\} &= P\{B_1\}P\{A|B_1\} + P\{B_2\}P\{A|B_2\} + P\{B_3\}P\{A|B_3\} \\ &= 0.30 \times 0.05 + 0.40 \times 0.08 + 0.30 \times 0.10 \\ &= = 0.077. \end{aligned}$$

Now suppose that my phone has developed a fault within one year. What is the probability that it was manufactured by A Pale? To answer this we need to introduce the Bayes Theorem.

2.3.5 The Bayes theorem

Let A be an event, and let $B_1, B_2, \ldots, B_k$ be a set of mutually exclusive and exhaustive events. Then, for $i = 1, \ldots, k$,

$$\boxed{P\{B_i|A\} = \frac{P\{B_i\}P\{A|B_i\}}{\sum_{j=1}^{k} P\{B_j\}P\{A|B_j\}}}$$

Proof: Use the multiplication rule of conditional probability twice to give

$$P\{B_i|A\} = \frac{P\{B_i \cap A\}}{P\{A\}} = \frac{P\{B_i\}P\{A|B_i\}}{P\{A\}}.$$

The Bayes theorem follows directly by substituting $P\{A\}$ by the total probability formula.
The probability, $P\{B_i|A\}$ is called the posterior probability of B_i and $P\{B_i\}$ is called the prior probability. The Bayes theorem is the rule that converts the prior probability into the posterior probability by using the additional information that some other event, A above, has already occurred.

♡ **Example 19 Phones continued** The probability that my faulty phone was manufactured by A Pale is

$$P\{B_1|A\} = \frac{P\{B_1\}P\{A|B_1\}}{P\{A\}} = \frac{0.30 \times 0.05}{0.077} = 0.1948.$$

Similarly, the probability that the faulty phone was manufactured by B Sung is 0.4156, and the probability that it was manufactured by C Windows is 1-0.1948-0.4156 = 0.3896.

The worked examples section contains further illustrations of the Bayes theorem. Note that $\sum_{i=1}^{k} P\{B_i|A\} = 1$. Why? Nowadays the Bayes theorem is used to make statistical inference as well.

2.3.6 Take home points

We have learned three important concepts: (i) conditional probability, (ii) formula for total probability and (iii) Bayes theorem. Much of statistical theory depends on these fundamental concepts. Spend extra time to learn these and try all the examples and exercises.

2.4 Lecture 9: Independent events

2.4.1 Lecture mission

The previous lecture has shown that probability of an event may change if we have additional information. However, in many situations the probabilities may not change. For example, the probability of getting an 'A' in Math1024 should not depend on the student's race and sex; the results of two coin tosses should not depend on each other; an expectant mother should not think that she must have a higher probability of having a son given that her previous three children were all girls.
In this lecture we will learn about the probabilities of independent events. Much of statistical theory relies on the concept of independence.

2.4.2 Definition

We have seen examples where prior knowledge that an event A has occurred has changed the probability that event B occurs. There are many situations where this does not happen. The events are then said to be independent.

Intuitively, events A and B are independent if the occurrence of one event does not affect the probability that the other event occurs.

This is equivalent to saying that

$$P\{B|A\} = P\{B\}, \quad \text{where } P\{A\} > 0, \quad \text{and } P\{A|B\} = P\{A\}, \quad \text{where } P\{B\} > 0.$$

These give the following formal definition.

$$\boxed{A \text{ and } B \text{ are independent events if } P\{A \cap B\} = P\{A\}P\{B\}.}$$

♡ **Example 20 Die throw**
Throw a fair die. Let A be the event that "the result is even" and B be the event that "the result is greater than 3". We want to show that A and B are not independent.

For this, we have $P\{A \cap B\} = P\{\text{either a 4 or 6 thrown}\} = 1/3$, but $P\{A\} = 1/2$ and $P\{B\} = 1/2$, so that $P\{A\}P\{B\} = 1/4 \neq 1/3 = P\{A \cap B\}$. Therefore A and B are not independent events.

Note that *independence* is not the same as the *mutually exclusive* property. When two events, A and B, are mutually exclusive, the probability of their intersection, $A \cap B$, is zero, i.e. $P\{A \cap B\} = 0$. But if the two events are independent then $P\{A \cap B\} = P\{A\} \times P\{B\}$.

Independence is often assumed on physical grounds, although sometimes incorrectly. There are serious consequences for wrongly assuming independence, e.g. the financial crisis in 2008. However, when the events are independent then the simpler product formula for joint probability is then used

instead of the formula involving more complicated conditional probabilities.

♡ **Example 21** Two fair dice when shaken together are assumed to behave independently. Hence the probability of two sixes is $1/6 \times 1/6 = 1/36$.

♡ **Example 22 Assessing risk in legal cases** In recent years there have been some disastrous miscarriages of justice as a result of incorrect assumption of independence. Please read "Incorrect use of independence – Sally Clark Case" on Blackboard.

Independence of complementary events: If A and B are independent, so are A' and B'.

Proof: Given that $P\{A \cap B\} = P\{A\}P\{B\}$, we need to show that $P\{A' \cap B'\} = P\{A'\}P\{B'\}$. This follows from

$$\begin{aligned} P\{A' \cap B'\} &= 1 - P\{A \cup B\} \\ &= 1 - [P\{A\} + P\{B\} - P\{A \cap B\}] \\ &= 1 - [P\{A\} + P\{B\} - P\{A\}P\{B\}] \\ &= [1 - P\{A\}] - P\{B\}[1 - P\{A\}] \\ &= [1 - P\{A\}][1 - P\{B\}] = P\{A'\}P\{B'\} \end{aligned}$$

The ideas of conditional probability and independence can be extended to *more than two events*.

Definition Three events A, B and C are defined to be independent if

$$P\{A \cap B\} = P\{A\}P\{B\},\ P\{A \cap C\} = P\{A\}P\{C\},\ P\{B \cap C\} = P\{B\}P\{C\}, \tag{2.1}$$

$$P\{A \cap B \cap C\} = P\{A\}P\{B\}P\{C\} \tag{2.2}$$

Note that (2.1) does NOT imply (2.2), as shown by the next example. Hence, to show the independence of A, B and C, it is necessary to show that both (2.1) and (2.2) hold.

♡ **Example 23** A box contains eight tickets, each labelled with a binary number. Two are labelled with the binary number 111, two are labelled with 100, two with 010 and two with 001. An experiment consists of drawing one ticket at random from the box.

Let A be the event "the first digit is 1", B the event "the second digit is 1" and C be the event "the third digit is 1". It is clear that $P\{A\} = P\{B\} = P\{C\} = 4/8 = 1/2$ and $P\{A \cap B\} = P\{A \cap C\} = P\{B \cap C\} = 1/4$, so the events are pairwise independent, i.e. (2.1) holds. However $P\{A \cap B \cap C\} = 2/8 \neq P\{A\}P\{B\}P\{C\} = 1/8$. So (2.2) does not hold and A, B and C are not independent.

Bernoulli trials The notion of independent events naturally leads to a set of independent trials (or random experiments, e.g. repeated coin tossing). A set of independent trials, where each trial has only two possible outcomes, conveniently called success (S) and failure (F), and the probability

of success is the same in each trial are called a set of *Bernoulli trials*. There are lots of fun examples involving Bernoulli trials.

♡ **Example 24 Feller's road crossing example** The flow of traffic at a certain street crossing is such that the probability of a car passing during any given second is p and cars arrive randomly, i.e. there is no interaction between the passing of cars at different seconds. Treating seconds as indivisible time units, and supposing that a pedestrian can cross the street only if no car is to pass during the next three seconds, find the probability that the pedestrian has to wait for exactly $k = 0, 1, 2, 3, 4$ seconds.

Let C_i denote the event that a car comes in the ith second and let N_i denote the event that no car arrives in the ith second.

1. Consider $k = 0$. The pedestrian does not have to wait if and only if there are no cars in the next three seconds, i.e. the event $N_1N_2N_3$. Now the arrival of the cars in successive seconds are independent and the probability of no car coming in any secod is $q = 1 - p$. Hence the answer is $P\{N_1N_2N_3\} = q \cdot q \cdot q = q^3$.

2. Consider $k = 1$. The person has to wait for one second if there is a car in the first second and none in the next three, i.e. the event $C_1N_2N_3N_4$. Hence the probability of that is pq^3.

3. Consider $k = 2$. The person has to wait two seconds if and only if there is a car in the 2nd second but none in the next three. It does not matter if there is a car or none in the first second. Hence:

$$P\{\text{wait 2 seconds}\} = P\{C_1C_2N_3N_4N_5\} + P\{N_1C_2N_3N_4N_5\} = p \cdot p \cdot q^3 + q \cdot p \cdot q^3 = pq^3.$$

4. Consider $k = 3$. The person has to wait for three seconds if and only if a car passes in the 3rd second but none in the next three, $C_3N_4N_5N_6$. Anything can happen in the first two seconds, i.e. C_1C_2, C_1N_2, N_1C_2, N_1N_2 − all these four cases are mutually exclusive. Hence,

$$\begin{aligned} P\{\text{wait 3 seconds}\} &= P\{C_1C_2C_3N_4N_5N_6\} + P\{N_1C_2C_3N_4N_5N_6\} \\ &\quad + P\{C_1N_2C_3N_4N_5N_6\} + P\{N_1N_2C_3N_4N_5N_6\} \\ &= p \cdot p \cdot p \cdot q^3 + p \cdot q \cdot p \cdot q^3 + q \cdot p \cdot p \cdot q^3 + q \cdot q \cdot p \cdot q^3 \\ &= pq^3. \end{aligned}$$

5. Consider $k = 4$. This is more complicated because the person has to wait exactly 4 seconds if and only if a car passes in at least one of the first 3 seconds, one passes at the 4th but none pass in the next 3 seconds. The probability that at least one passes in the first three seconds is 1 minus the probability that there is none in the first 3 seconds. This probability is $1 - q^3$. Hence the answer is $(1 - q^3)pq^3$.

2.4.3 Take home points

We have learned the concept of independent events. It is much easier to calculate probabilities when events are independent. However, there is danger in assuming events to be independent when they are not. For example, there may be serial or spatial dependence! The concept of Bernoulli trials has been introduced. More examples to follow!

2.5 Lecture 10: Fun probability calculation for independent events

2.5.1 Lecture mission

We are continuing with the notion of independent events. This lecture will discuss two substantial examples: one is called system reliability where we have to find the probability of a system, built from several independent components, functioning. For example, we want to find out the probability that a machine/phone or module-based software system will continue to function. In the second substantial example we would like to cleverly find out probabilities of sensitive events, e.g. do I have HIV/AIDS or did I take any illegal drugs during last summer's music festival?

2.5.2 System reliability

Two components in series

Suppose each component has a separate operating mechanism. This means that they operate independently.

Let A_i be the event "component i works when required" and let $P\{A_i\} = p_i$ for $i = 1, 2$. For the system of A_1 and A_2 in series, the event "the system works" is the event $\{A_1 \cap A_2\}$. Hence $P\{\text{system works}\} = P\{A_1 \cap A_2\} = P\{A_1\}P\{A_2\} = p_1 p_2$.

The reliability gets lower when components are included in series. For n components in series, $P\{\text{system works}\} = p_1 p_2 \cdots p_n$. When $p_i = p$ for all i, the reliability of a series of n components is $P\{\text{system works}\} = p^n$.

Two components in parallel

For the system of A_1 and A_2 in parallel, the event "the system works when required" is now given by the event $\{A_1 \cup A_2\}$. Hence

$$P\{\text{system works}\} = P\{A_1 \cup A_2\} = P\{A_1\} + P\{A_2\} - P\{A_1 \cap A_2\} = p_1 + p_2 - p_1 p_2.$$

This is greater than either p_1 or p_2 so that the inclusion of a (redundant) component in parallel increases the reliability of the system. Another way of arriving at this result uses complementary events:

$$\begin{aligned} P\{\text{system works}\} &= 1 - P\{\text{system fails}\} \\ &= 1 - P\{A_1' \cap A_2'\} \\ &= 1 - P\{A_1'\}P\{A_2'\} \\ &= 1 - (1 - p_1)(1 - p_2) \\ &= p_1 + p_2 - p_1 p_2. \end{aligned}$$

In general, with n components in parallel, the reliability of the system is

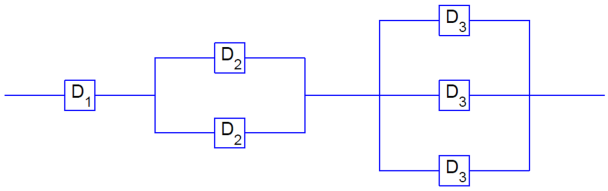
$$P\{\text{system works}\} = 1 - (1 - p_1)(1 - p_2) \cdots (1 - p_n).$$

If $p_i = p$ for all i, we have $P\{\text{system works}\} = 1 - (1 - p)^n$.

A general system

The ideas above can be combined to evaluate the reliability of more complex systems.

♡ **Example 25 Switches** Six switches make up the circuit shown in the graph.



Each has the probability $p_i = P\{D_i\}$ of closing correctly; the mechanisms are independent; all are operated by the same impulse. Then

$$P\{\text{current flows when required}\} = p_1 \times [1 - (1 - p_2)(1 - p_2)] \times [1 - (1 - p_3)(1 - p_3)(1 - p_3)].$$

There are some additional examples of reliability applications given in the "Reliability Examples" document available on Blackboard. You are advised to read through and understand these additional examples/applications.

2.5.3 The randomised response technique

This is an important application of the total probability formula - it is used to try to get honest answers to sensitive questions.

Often we wish to estimate the proportion of people in a population who would not respond 'yes' to some sensitive question such as:

- Have you taken an illegal drug during the last 12 months?
- Have you had an abortion?
- Do you have HIV/AIDs?
- Are you a racist?

It is unlikely that truthful answers will be given in an open questionnaire, even if it is stressed that the responses would be treated with anonymity. Some years ago a randomised response technique was introduced to overcome this difficulty. This is a simple application of conditional probability. It ensures that the interviewee can answer truthfully without the interviewer (or anyone else) knowing the answer to the sensitive question. How? Consider two alternative questions, for example:

Question 1: Was your mother born in January?
Question 2: Have you ever taken illegal substances in the last 12 months?

Question 1 should not be contentious and should not be such that the interviewer could find out the true answer.

The respondent answers only 1 of the two questions. Which question is answered by the respondent is determined by a randomisation device, the result of which is known only to the respondent. The interviewer records only whether the answer given was Yes or No (and he/she does not know which question has been answered). The proportion of Yes answers to the question of interest can be estimated from the total proportion of Yes answers obtained. Carry out this simple experiment:

Toss a coin - do not reveal the result of the coin toss!

If heads - answer Question 1: Was your mother born in January?

If tails - answer Question 2: Have you ever taken illegal substances in the last 12 months?

We need to record the following information for the outcome of the experiment:

Total number in sample = n;

Total answering Yes = r, so that an estimate of $P\{\text{Yes}\}$ is r/n.

This information can be used to estimate the proportion of Yes answers to the main question of interest, Question 2.

Suppose that

- Q_1 is the event that 'Q1 was answered'
- Q_2 is the event that 'Q2 was answered'

Then, assuming that the coin was unbiased, $P\{Q_1\} = 0.5$ and $P\{Q_2\} = 0.5$. Also, assuming that birthdays of mothers are evenly distributed over the months, we have that the probability that the interviewee will answer Yes to Q1 is 1/12. Let Y be the event that a 'Yes' answer is given. Then the total probability formula gives

$$P\{Y\} = P\{Q_1\}P\{Y|Q_1\} + P\{Q_2\}P\{Y|Q_2\},$$

which leads to

$$\frac{r}{n} \approx \frac{1}{2} \times \frac{1}{12} + \frac{1}{2} \times P\{Y|Q_2\}.$$

Hence

$$P\{Y|Q_2\} \approx 2 \cdot \frac{r}{n} - \frac{1}{12}.$$

2.5.4 Take home points

In this lecture we have learned a couple of further applications of elementary rules of probability. You will see many more examples in the exercise sheets. In many subsequent second and third year modules these laws of probabilities must be applied to get answers to more difficult questions.

We, however, will move on to the next chapter on random variables, which formalises the concepts of probabilities in structured practical cases. The concept of random variables allows us to calculate probabilities of random events much more easily in structured ways.

Chapter 3

Random Variables and Their Probability Distributions

Chapter mission

Last chapter's combinatorial probabilities are difficult to find and very problem-specific. Instead, in this chapter we shall find easier ways to calculate probability in structured cases. The outcomes of random experiments will be represented as values of a variable which will be random since the outcomes are random (or un-predictable with certainty). In so doing, we will make our life a lot easier in calculating probabilities in many stylised situations which represent reality. For example, we shall learn to calculate what is the probability that a computer will make fewer than 10 errors while making 10^{15} computations when it has a very tiny chance, 10^{-14}, of making an erroneous computation.

3.1 Lecture 11: Definition of a random variable

3.1.1 Lecture mission

In this lecture we will learn about the probability distribution of a random variable defined by its probability function. The probability function will be called the probability mass function for discrete random variables and the probability density function for continuous random variables.

3.1.2 Introduction

A random variable defines a one-to-one mapping of the sample space consisting of all possible outcomes of a random experiment to the set of real numbers. For example, I toss a coin. Assuming the coin is fair, there are two possible equally likely outcomes: head or tail. These two outcomes must be mapped to real numbers. For convenience, I may define the mapping which assigns the value 1 if head turns up and 0 otherwise. Hence, we have the mapping:

$$\text{Head} \to 1, \text{Tail} \to 0.$$

We can conveniently denote the random variable by X which is the number of heads obtained by tossing a single coin. Obviously, all possible values of X are 0 and 1.

You will say that this is a trivial example. Indeed it is. But it is very easy to generalise the concept of random variables. Simply define a mapping of the outcomes of a random experiment to the real number space. For example, I toss the coin n times and count the number of heads and denote that to be X. Obviously, X can take any real positive integer value between 0 and n. Among other examples, suppose I select a University of Southampton student at random and measure their height. The outcome in metres will be a number between one metre and two metres for sure. But I can't exactly tell which value it will be since I do not know which student will be selected in the first place. However, when a student has been selected I can measure their height and get a value such as 1.432 metres.

We now introduce two notations: X (or in general the capital letters Y, Z etc.) to denote the random variable, e.g. height of a randomly selected student, and the corresponding lower case letter x (y, z) to denote a particular value, e.g. 1.432 metres. We will follow this convention throughout. For a random variable, say X, we will also adopt the notation $P(X \in A)$, read probability that X belongs to A, instead of the previous $P\{A\}$ for any event A.

3.1.3 Discrete or continuous random variable

If a random variable has a *finite* or *countably infinite* set of values it is called *discrete*. For example, the number of Apple computer users among 20 randomly selected students, or the number of credit cards a randomly selected person has in their wallet.

When the random variable can take any value on the real line it is called a continuous random variable. For example, the height of a randomly selected student. A random variable can also take a mixture of discrete and continuous values, e.g. volume of precipitation collected in a day; some days it could be zero, on other days it could be a continuous measurement, e.g. 1.234 mm.

3.1.4 Probability distribution of a random variable

Recall the first axiom of probability ($P\{S\} = 1$), which means total probability equals 1. Since a random variable is merely a mapping from the outcome space to the real line, the combined probability of all possible values of the random variable must be equal to 1.

> *A probability distribution distributes the total probability 1 among the possible values of the random variable.*

For example, returning to the coin-tossing experiment, if the probability of getting a head with a coin is p (and therefore the probability of getting a tail is $1-p$), then the probability that $Y = 0$ is $1-p$ and the probability that $Y = 1$ is p. This gives us the *probability distribution* of Y, and we say that Y has the *probability function* given by:

$P(Y=0)$	$=$	$1-p$
$P(Y=1)$	$=$	p
Total	$=$	1.

This is an example of the **Bernoulli distribution** with parameter p, perhaps the simplest discrete distribution.

♡ **Example 26** Suppose we consider tossing the coin twice and again defining the random variable X to be the number of heads obtained. The values that X can take are 0, 1 and 2 with probabilities $(1-p)^2$, $2p(1-p)$ and p^2, respectively. Here the distribution is:

Value(x)	$P(X = x)$
0	$(1-p)^2$
1	$2p(1-p)$
2	p^2
Total prob	1.

This is a particular case of the Binomial distribution. We will learn about it soon.

In general, for a discrete random variable we define a function $f(x)$ to denote $P(X = x)$ (or $f(y)$ to denote $P(Y = y)$) and call the function $f(x)$ the *probability function (pf)* or *probability mass function (pmf)* of the random variable X. Arbitrary functions cannot be a pmf since the total probability must be 1 and all probabilities are non-negative. Hence, for $f(x)$ to be the pmf of a random variable X, we require:

1. $f(x) \geq 0$ for all possible values of x.
2. $\sum_{\text{all } x} f(x) = 1$.

So for the binomial example, we have the following probability distribution.

x	$f(x)$	general form
0	$(1-p)^2$	${}^2C_x\, p^x(1-p)^{2-x}$
1	$2p(1-p)$	${}^2C_x\, p^x(1-p)^{2-x}$
2	p^2	${}^2C_x\, p^x(1-p)^{2-x}$
Total	1	1

Note that $f(x) = 0$ for any other value of x and thus $f(x)$ is a discrete function of x.

Continuous random variable

In many situations (both theoretical and practical) we often encounter random variables that are inherently continuous because they are measured on a continuum (such as time, length, weight) or can be conveniently well-approximated by considering them as continuous (such as the annual income of adults in a population, closing share prices).

For a continuous random variable, $P(X = x)$ is defined to be zero since we assume that the measurements are continuous and there is zero probability of observing a particular value, e.g. 1.2. The argument goes that a finer measuring instrument will give us an even more precise measurement than 1.2 and so on. Thus for a continuous random variable we adopt the convention that

$P(X = x) = 0$ for any particular value x on the real line. But we define probabilities for positive length intervals, e.g. $P(1.2 < X < 1.9)$.

For a continuous random variable X we define its probability by using a continuous function $f(x)$ which we call its *probability density function*, abbreviated as its pdf. With the pdf we define probabilities as integrals, e.g.

$$P(a < X < b) = \int_a^b f(u)\, du,$$

which is naturally interpreted as the area under the curve $f(x)$ inside the interval (a, b). Recall that we do not use $f(x) = P(X = x)$ for any x as by convention we set $P(X = x) = 0$.

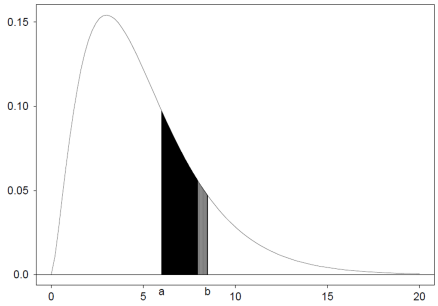


Figure 3.1: The shaded area is $P(a < X < b)$ if the pdf of X is the drawn curve.

Since we are dealing with probabilities which are always between 0 and 1, just any arbitrary function $f(x)$ cannot be a pdf of some random variable. For $f(x)$ to be a pdf, as in the discrete case, we must have:

1. $f(x) \geq 0$ for all possible values of x, i.e. $-\infty < x < \infty$.
2. $\int_{-\infty}^{\infty} f(u)du = 1$.

It is very simple to describe the above two requirements:

(i) the probabilities are non-negative, and (ii) the total probability must be 1,

(recall $P\{S\} = 1$), where S is the sample space.

3.1.5 Cumulative distribution function (cdf)

Along with the pdf we also frequently make use of another function which is called the cumulative distribution function, abbreviated as the cdf. The cdf simply calculates the probability of the random variable up to its argument. For a discrete random variable, the cdf is the cumulative sum of the pmf $f(u)$ up to (and including) $u = x$. That is,

$$P(X \leq x) \equiv F(x) = \sum_{u \leq x} f(u)$$

when X is a discrete random variable with pmf $f(x)$.

♡ **Example 27** Let X be the number of heads in the experiment of tossing two fair coins. Then the probability function is

$$P(X = 0) = 1/4, \quad P(X = 1) = 1/2, \quad P(X = 2) = 1/4.$$

From the definition, the CDF is given by

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1/4 & \text{if } 0 \leq x < 1 \\ 3/4 & \text{if } 1 \leq x < 2 \\ 1 & \text{if } 2 \leq x \end{cases}.$$

Note that the cdf for a discrete random variable is a step function. The jump-points are the possible values of the random variable (r.v.), and the height of a jump gives the probability of the random variable taking that value. It is clear that the probability mass function is uniquely determined by the cdf.

For a continuous random variable X, the cdf if defined as:

$$P(X \leq x) \equiv F(x) = \int_{-\infty}^{x} f(u)du.$$

The fundamental theorem of calculus then tells us:

$$\boxed{f(x) = \frac{dF(x)}{dx}}$$

that is, for a continuous random variable the pdf is the derivative of the cdf. Also for any random variable X, $P(c < X < d) = F(d) - F(c)$. Let us consider an example.

♡ **Example 28 Uniform distribution** Suppose,

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a < x < b \\ 0 & \text{otherwise} \end{cases}.$$

We now have the cdf $F(x) = \int_a^x \frac{du}{b-a} = \frac{x-a}{b-a}, a < x < b$. A quick check confirms that $F'(x) = f(x)$. If $a = 0, b = 1$ and then $P(0.5 < X < 0.75) = F(0.75) - F(0.5) = 0.25$. We shall see many more examples later.

3.1.6 Take home points

In this lecture, we have learnt what the pmf, pdf and cdf of a random variable are. We know the interrelationships and how to use them to calculate probabilities of interest.

3.2 Lecture 12: Expectation and variance of a random variable

3.2.1 Lecture mission

We will discover many more different variables which simply have different probability functions. In this lecture we will learn about the two most important properties of random variables, i.e. the mean and the variance.

3.2.2 Mean or expectation

In Chapter 1, we defined the mean and variance of sample data $x_1, \ldots, x_n$. The random variables with either a pmf $f(x)$ or a pdf $f(x)$ also have their own mean, which can be called expectation (central tendency), and variance as well. The mean is called an expectation since it is a value we can 'expect'! The expectation is defined as:

$$E(X) = \begin{cases} \sum_{\text{all } x} x f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x f(x) dx & \text{if } X \text{ is continuous} \end{cases}.$$

Thus, roughly speaking:

the expected value is either sum or integral of value times probability.

We use the $E(\cdot)$ notation to denote expectation. The argument is in upper case since it is the expected value of the random variable which is denoted by an upper case letter. We often use the Greek letter μ to denote $E(X)$.

♡ **Example 29 Discrete** Consider the fair-die tossing experiment, with each of the six sides having a probability of 1/6 of landing face up. Let X be the number on the up-face of the die. Then

$$E(X) = \sum_{x=1}^{6} x P(X = x) = \sum_{x=1}^{6} x/6 = 3.5.$$

♡ **Example 30 Continuous** Consider the uniform distribution which has the pdf $f(x) = \frac{1}{b-a}, a < x < b$.

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x\, f(x) dx \\ &= \int_a^b \frac{x}{b-a} dx \\ &= \frac{b^2-a^2}{2(b-a)} = \frac{b+a}{2}, \end{aligned}$$

the mid-point of the interval (a, b).

If $Y = g(X)$ for any function $g(\cdot)$, then Y is a random variable as well. To find $E(Y)$ we simply use the value times probability rule, i.e. the expected value of Y is either sum or integral of its value, $g(x)$ times probability $f(x)$.

$$E(Y) = E(g(X)) = \begin{cases} \sum_{\text{all } x} g(x) f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} g(x) f(x) dx & \text{if } X \text{ is continuous} \end{cases}.$$

For example, if X is continuous, then $E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)dx$. We prove one important property of expectation, namely expectation is a linear operator.

$$\boxed{\text{Suppose } Y = g(X) = aX+b; \text{ then } E(Y) = aE(X)+b.}$$

The proof of this is simple and given below for the continuous case. In the discrete case replace integral ($\int$) by summation ($\sum$).

$$\begin{aligned} E(Y) &= \int_{-\infty}^{\infty}(ax+b)f(x)dx \\ &= a\int_{-\infty}^{\infty} xf(x)dx + b\int_{-\infty}^{\infty} f(x)dx \\ &= aE(X)+b, \end{aligned}$$

using the value times probability definition of the expectation and the total probability is 1 property ($\int_{-\infty}^{\infty} f(x)dx = 1$) in the last integral. This is very convenient, e.g. suppose $E(X) = 5$ and $Y = -2X + 549$; then $E(Y) = 539$.

Variance of a random variable

The variance measures the variability of a random variable and is defined by:

$$\text{Var}(X) = E(X-\mu)^2 = \begin{cases} \sum_{\text{all } x}(x-\mu)^2 f(x) & \text{if X is discrete} \\ \int_{-\infty}^{\infty}(x-\mu)^2 f(x)dx & \text{if X is continuous} \end{cases},$$

where $\mu = E(X)$, and when the sum or integral exists. They can't always be assumed to exist! When the variance exists, it is the expectation of $(X-\mu)^2$ where μ is the mean of X. We now derive an easy formula to calculate the variance:

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

The proof is given below:

$$\begin{aligned} \text{Var}(X) &= E(X-\mu)^2 \\ &= E(X^2 - 2X\mu + \mu^2) \\ &= E(X^2) - 2\mu E(X) + \mu^2 \\ &= E(X^2) - 2\mu\mu + \mu^2 \\ &= E(X^2) - \mu^2. \end{aligned}$$

Thus:

the variance of a random variable is the expected value of its square minus the square of its expected value.

We usually denote the variance by σ^2. The square is there to emphasise that the variance of any random variable is always non-negative. When can the variance be zero? When there is no variation at all in the random variable, i.e. it takes only a single value μ with probability 1. Hence, there is nothing random about the random variable – we can predict its outcome with certainty.

The square root of the variance is called the standard deviation of the random variable.

♡ **Example 31 Uniform** Consider the uniform distribution which has the pdf $f(x) = \frac{1}{b-a}, a < x < b$.

$$\begin{aligned} E(X^2) &= \int_a^b \frac{x^2}{b-a} dx \\ &= \frac{b^3-a^3}{3(b-a)} = \frac{b^2+ab+a^2}{3}, \end{aligned}$$

Hence

$$\text{Var}(X) = \frac{b^2+ab+a^2}{3} - \left(\frac{b+a}{2}\right)^2 = \frac{(b-a)^2}{12},$$

after simplification.

We prove one important property of the variance.

Suppose $Y = aX + b$; then $\text{Var}(Y) = a^2\text{Var}(X)$.

The proof of this is simple and is given below for the continuous case. In the discrete case replace integral ($\int$) by summation ($\sum$).

$$\begin{aligned} \text{Var}(Y) &= E(Y - E(Y))^2 \\ &= \int_{-\infty}^{\infty} (ax + b - a\mu - b)^2 f(x) dx \\ &= a^2 \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx \\ &= a^2 \text{Var}(X). \end{aligned}$$

This is a very useful result, e.g. suppose $\text{Var}(X) = 25$ and $Y = -X + 5,000,000$; then $\text{Var}(Y) = \text{Var}(X) = 25$ and the standard deviation, $\sigma = 5$. In words a location shift, b, does not change variance but a multiplicative constant, a say, gets squared in variance, a^2.

3.2.3 Take home points

In this lecture we have learned what are called the expectation and variance of a random variable. We have also learned that the expectation operator distributes linearly and the formula for the variance of a linear function of a random variable.

3.3 Lecture 13: Standard discrete distributions

3.3.1 Lecture mission

In this lecture we will learn about the Bernoulli, Binomial, Hypergeometric and Geometric distributions and their properties.

3.3.2 Bernoulli distribution

The Bernoulli distribution has pmf $f(x) = p^x(1-p)^{1-x}, x = 0, 1$. Hence $E(X) = 0 \cdot (1-p) + 1 \cdot p = p$, $E(X^2) = 0^2 \cdot (1-p) + 1^2 \cdot p = p$ and $\text{Var}(X) = E(X^2) - (E(X))^2 = p - p^2 = p(1-p)$. Hence $\text{Var}(X) < E(X)$.

3.3.3 Binomial distribution

Suppose that we have a sequence of n Bernoulli trials (defined in Lecture 9, e.g. coin tosses) such that we get a success (S) or failure (F) with probabilities $P\{S\} = p$ and $P\{F\} = 1-p$ respectively. Let X be the number of successs in the n trials. Then X is called a binomial random variable with parameters n and p.

An outcome of the experiment (of carrying out n such independent trials) is represented by a sequence of S's and F's (such as $SS...FS...SF$) that comprises x S's, and $(n-x)$ F's.

The probability associated with this outcome is

$$P\{SS...FS...SF\} = pp\cdots(1-p)p\cdots p(1-p) = p^x(1-p)^{n-x}.$$

For this sequence, $X = x$, but there are many other sequences which will also give $X = x$. In fact there are $\binom{n}{x}$ such sequences. Hence

$$P(X = x) = \binom{n}{x}p^x(1-p)^{n-x}, x = 0, 1, \ldots, n.$$

This is the pmf of the Binomial Distribution with parameters n and p, often written as $\text{Bin}(n, p)$.

How can we guarantee that $\sum_{x=0}^{n} P(X = x) = 1$? This guarantee is provided by the binomial theorem:

$$(a+b)^n = b^n + \binom{n}{1}ab^{n-1} + \cdots + \binom{n}{x}a^xb^{n-x} + \cdots + a^n.$$

To prove, $\sum_{x=0}^{n} P(X = x) = 1$, i.e. to prove, $\sum_{x=0}^{n}\binom{n}{x}p^x(1-p)^{n-x} = 1$, choose $a = p$ and $b = 1-p$ in the binomial theorem.

♡ **Example 32** Suppose that widgets are manufactured in a mass production process with 1% defective. The widgets are packaged in bags of 10 with a money-back guarantee if more than 1 widget per bag is defective. For what proportion of bags would the company have to provide a refund?

Firstly, we want to find the probability that a randomly selected bag has at most 1 defective widget. Note that the number of defective widgets in a bag X, $X \sim \text{Bin}(n = 10, p = 0.01)$. So, this probability is equal to

$$P(X = 0) + P(X = 1) = (0.99)^{10} + 10(0.01)^1(0.99)^9 = 0.9957.$$

Hence the probability that a refund is required is $1 - 0.9957 = 0.0043$, i.e. only just over 4 in 1000 bags will incur the refund on average.

Using `R` to calculate probabilities

Probabilities under all the standard distributions have been calculated in `R` and will be used throughout **MATH1024**. You will not be required to use any tables. For the binomial distribution the command `dbinom(x=3, size=5, prob=0.34)` calculates the pmf of $\text{Bin}(n = 5, p = 0.34)$ at

$x = 3$. That is, the command `dbinom(x=3, size=5, prob=0.34)` will return the value $P(X = 3) = \binom{5}{3}(0.34)^3(1-0.34)^{5-3}$. The command `pbinom` returns the cdf or the probability up to and including the argument. Thus `pbinom(q=3, size=5, prob=0.34)` will return the value of $P(X \leq 3)$ when $X \sim \text{Bin}(n = 5, p = 0.34)$. As a check, in the above example the command is `pbinom(q=1, size=10, prob=0.01)`, which returns 0.9957338.

♡ **Example 33** A binomial random variable can also be described using the urn model. Suppose we have an urn (population) containing N individuals, a proportion p of which are of type S and a proportion $1-p$ of type F. If we select a sample of n individuals at random **with replacement**, then the number, X, of type S individuals in the sample follows the binomial distribution with parameters n and p.

Mean of the Binomial distribution
Let $X \sim \text{Bin}(n, p)$. We have

$$E(X) = \sum_{x=0}^{n} xP(X = x) = \sum_{x=0}^{n} x\binom{n}{x}p^x(1-p)^{n-x}.$$

Below we prove that $E(X) = np$. Recall that $k! = k(k-1)!$ for any $k > 0$.

$$\begin{aligned}
E(X) &= \textstyle\sum_{x=0}^{n} x\binom{n}{x}p^x(1-p)^{n-x} \\
&= \textstyle\sum_{x=1}^{n} x\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x} \\
&= \textstyle\sum_{x=1}^{n} \frac{n!}{(x-1)!(n-x)!}p^x(1-p)^{n-x} \\
&= np\textstyle\sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-1-x+1)!}p^{x-1}(1-p)^{n-1-x+1} \\
&= np\textstyle\sum_{y=0}^{n-1} \frac{(n-1)!}{(y)!(n-1-y)!}p^{y}(1-p)^{n-1-y} \\
&= np(p+1-p)^{n-1} = np,
\end{aligned}$$

where we used the substitution $y = x - 1$ and then the binomial theorem to conclude that the last sum is equal to 1.

Variance of the Binomial distribution
Let $X \sim \text{Bin}(n, p)$. Then $\text{Var}(X) = np(1-p)$. It is difficult to find $E(X^2)$ directly, but the factorial structure allows us to find $E[X(X-1)]$. Recall that $k! = k(k-1)(k-2)!$ for any $k > 1$.

$$\begin{aligned}
E[(X(X-1)] &= \textstyle\sum_{x=0}^{n} x(x-1)\binom{n}{x}p^x(1-p)^{n-x} \\
&= \textstyle\sum_{x=2}^{n} x(x-1)\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x} \\
&= \textstyle\sum_{x=2}^{n} \frac{n!}{(x-2)!(n-x)!}p^x(1-p)^{n-x} \\
&= n(n-1)p^2\textstyle\sum_{x=2}^{n} \frac{(n-2)!}{(x-2)!(n-2-x+2)!}p^{x-2}(1-p)^{n-2-x+2} \\
&= n(n-1)p^2\textstyle\sum_{y=0}^{n-2} \frac{(n-2)!}{(y)!(n-2-y)!}p^{y}(1-p)^{n-2-y} \\
&= n(n-1)p^2(p+1-p)^{n-2}.
\end{aligned}$$

Now, $E(X^2) = E[X(X-1)] + E(X) = n(n-1)p^2 + np$. Hence,

$$\text{Var}(X) = E(X^2) - (E(X))^2 = n(n-1)p^2 + np - (np)^2 = np(1-p).$$

It is illuminating to see these direct proofs. Later on we shall apply statistical theory to directly prove these! Notice that the binomial theorem is used repeatedly to prove the results.

3.3.4 Geometric distribution

Suppose that we have the same situation as for the binomial distribution but we consider a different r.v. X, which is defined as the number of trials that lead to the first success. The outcomes for this experiment are:

$$\begin{array}{rll} S & X=1, & P(X=1)=p \\ FS & X=2, & P(X=2)=(1-p)p \\ FFS & X=3, & P(X=3)=(1-p)^2p \\ FFFS & X=4, & P(X=4)=(1-p)^3p \\ \vdots & \vdots & \end{array}$$

In general we have

$$P(X=x)=(1-p)^{x-1}p, x=1,2,\ldots$$

This is called the *geometric distribution*, and it has a (countably) infinite domain starting at 1 not 0. We write $X \sim \text{Geo}(p)$.

Let us check that the probability function has the required property:

$$\begin{aligned} \sum_{x=1}^{\infty} P(X=x) &= \sum_{x=1}^{\infty}(1-p)^{x-1}p \\ &= p\sum_{y=0}^{\infty}(1-p)^{y} \quad [\text{substitute } y=x-1] \\ &= p\,\frac{1}{1-(1-p)} \quad [\text{see Section A.4}] \\ &= 1. \end{aligned}$$

We can also find the probability that $X > k$ for some given natural number k:

$$\begin{aligned} \sum_{x=k+1}^{\infty} P(X=x) &= \sum_{x=k+1}^{\infty}(1-p)^{x-1}p \\ &= p[(1-p)^{k+1-1}+(1-p)^{k+2-1}+(1-p)^{k+3-1}+\ldots \\ &= p(1-p)^k\sum_{y=0}^{\infty}(1-p)^{y} \\ &= (1-p)^k. \end{aligned}$$

Memoryless property of the geometric distribution.
Let X follow the geometric distribution and suppose that s and k are positive integers. We then have

$$P(X > s+k|X>k) = P(X>s).$$

The proof is given below. In practice this means that the random variable does not remember its age (denoted by k) to determine how long more (denoted by s) it will survive! The proof below uses the definition of conditional probability

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

Now the proof,

$$\begin{aligned} P(X > s+k|X > k) &= \tfrac{P(X>s+k, X>k)}{P(X>k)} \\ &= \tfrac{P(X>s+k)}{P(X>k)} \\ &= \tfrac{(1-p)^{s+k}}{(1-p)^k} \\ &= (1-p)^s, \end{aligned}$$

which does not depend on k. Note that the event $X > s+k$ and $X > k$ implies and is implied by $X > s+k$ since $s > 0$.

Mean and variance of the Geometric distribution
Let $X \sim$Geo(p). We can show that $E(X) = p$ using the negative binomial series, see Section A.4, as follows:

$$\begin{aligned} E(X) &= \textstyle\sum_{x=1}^{\infty} xP(X=x) \\ &= \textstyle\sum_{x=1}^{\infty} xp(1-p)^{x-1} \\ &= p\left[1 + 2(1-p) + 3(1-p)^2 + 4(1-p)^3 + \ldots\right] \end{aligned}$$

For $n > 0$ and $|x| < 1$, the negative binomial series is given by:

$$(1-x)^{-n} = 1 + nx + \frac{1}{2}n(n+1)x^2 + \frac{1}{6}n(n+1)(n+2)x^3 + \cdots + \frac{n(n+1)(n+2)\cdots(n+k-1)}{k!}x^k + \cdots$$

With $n = 2$ and $x = 1-p$ the general term is given by:

$$\frac{n(n+1)(n+2)(n+k-1)}{k!} = \frac{2 \times 3 \times 4 \times \cdots \times (2+k-1)}{k!} = k+1.$$

Thus $E(X) = p(1-1+p)^{-2} = 1/p$. It can be shown that $\text{Var}(X) = (1-p)/p^2$ using negative binomial series. But this is more complicated and is not required. The second-year module MATH2011 will provide an alternative proof.

3.3.5 Hypergeometric distribution

Suppose we have an urn (population) containing N individuals, a proportion p of which are of type S and a proportion $1-p$ of type F. If we select a sample of n individuals at random **without replacement**, then the number, X, of type S individuals in the sample has the hypergeometric distribution:

$$P(X = x) = \frac{\binom{Np}{x}\binom{N(1-p)}{n-x}}{\binom{N}{n}}, x = 0, 1, \ldots, n,$$

assuming that $x \leq Np$ and $n - x \leq N(1-p)$ so that the above combinations are well defined. The mean and variance of the hypergeometric distribution are given by

$$E(X) = np, \text{Var}(X) = npq\frac{N-n}{N-1}.$$

The proofs of the above results use complicated finite summation and so are omitted. But note that when $N \to \infty$ the variance converges to the variance of the binomial distribution. Indeed, the hypergeometric distribution is a finite population analogue of the binomial distribution.

3.3.6 Take home points

We have learned properties of the Bernoulli, Binomial, Geometric and Hypergeometric distributions.

3.4 Lecture 14: Further standard discrete distributions

3.4.1 Lecture mission

In this lecture we introduce the negative binomial distribution as a generalisation of the geometric distribution, and the Poisson distribution.

3.4.2 Negative binomial distribution

Still in the Bernoulli trials set-up, we define the random variable X to be the total number of trials until the r-th success occurs, where r is a given natural number. This is known as the negative binomial distribution with parameters p and r.

[Note: if $r = 1$, the negative binomial distribution is just the geometric distribution.]

Firstly we need to identify the possible values of X. Possible values for X are $x = r, r+1, r+2, \ldots$. Secondly, the probability mass function is given by

$$\begin{aligned} P(X = x) &= \binom{x-1}{r-1} p^{r-1}(1-p)^{(x-1)-(r-1)} \times p \\ &= \binom{x-1}{r-1} p^r (1-p)^{x-r}, x = r, r+1, \ldots \end{aligned}$$

♡ **Example 34** In a board game that uses a single fair die, a player cannot start until they have rolled a six. Let X be the number of rolls needed until they get a six. Then X is a Geometric random variable with success probability $p = 1/6$.

♡ **Example 35** A man plays roulette, betting on red each time. He decides to keep playing until he achieves his second win. The success probability for each game is 18/37 and the results of games are independent. Let X be the number of games played until he gets his second win. Then X is a Negative Binomial random variable with $r = 2$ and $p = 18/37$. What is the probability he plays more than 3 games? i.e. find $P(X > 3)$.

Derivation of the mean and variance of the negative binomial distribution involves complicated negative binomial series and will be skipped for now, but will be proved in Lecture 19. For completeness we note down the mean and variance:

$$E(X) = \frac{r}{p}, \quad \text{Var}(X) = r\frac{1-p}{p^2}.$$

Thus when $r = 1$, the mean and variance of the negative binomial distribution are equal to those of the geometric distribution.

3.4.3 Poisson distribution

The Poisson distribution can be obtained as the limit of the binomial distribution with parameters n and p when $n \to \infty$ and $p \to 0$ simultaneously, but the product $\lambda = np$ remains finite. In practice this means that the Poisson distribution counts rare events (since $p \to 0$) in an infinite population (since $n \to \infty$). Theoretically, a random variable following the Poisson distribution can take any integer value from 0 to ∞. Examples of the Poisson distribution include: the number of breast cancer patients in Southampton; the number of text messages sent (or received) per day by a randomly selected first-year student; the number of credit cards a randomly selected person has in their wallet.

Let us find the pmf of the Poisson distribution as the limit of the pmf of the binomial distribution. Recall that if $X \sim \text{Bin}(n, p)$ then $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$. Now:

$$\begin{aligned} P(X = x) &= \binom{n}{x} p^x (1-p)^{n-x} \\ &= \binom{n}{x} \frac{n^n}{n^n} p^x (1-p)^{n-x} \\ &= \frac{n(n-1)\cdots(n-x+1)}{n^x x!} (np)^x \left(n(1-p)\right)^{n-x} \frac{1}{n^{n-x}} \\ &= \frac{n}{n} \frac{(n-1)}{n} \cdots \frac{(n-x+1)}{n} \frac{\lambda^x}{x!} \left(1 - \tfrac{\lambda}{n}\right)^{n-x}. \\ &= \frac{n}{n} \frac{(n-1)}{n} \cdots \frac{(n-x+1)}{n} \frac{\lambda^x}{x!} \left(1 - \tfrac{\lambda}{n}\right)^{n} \left(1 - \tfrac{\lambda}{n}\right)^{-x}. \end{aligned}$$

Now it is easy to see that the above tends to

$$e^{-\lambda} \frac{\lambda^x}{x!}$$

as $n \to \infty$ for any fixed value of x in the range 0, 1, 2, ... Note that we have used the exponential limit:

$$e^{-\lambda} = \lim_{n \to \infty} \left(1 - \frac{\lambda}{n}\right)^n,$$

and

$$\lim_{n \to \infty} \left(1 - \frac{\lambda}{n}\right)^{-x} = 1$$

and

$$\lim_{n \to \infty} \frac{n}{n} \frac{(n-1)}{n} \cdots \frac{(n-x+1)}{n} = 1.$$

A random variable X has the Poisson distribution with parameter λ if it has the pmf:

$$P(X = x) = e^{-\lambda} \frac{\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots$$

We write $X \sim \text{Poisson}(\lambda)$. It is trivial to show $\sum_{x=0}^{\infty} P(X = x) = 1$, i.e. $\sum_{x=0}^{\infty} e^{-\lambda} \frac{\lambda^x}{x!} = 1$. The identity you need is simply the expansion of e^{λ}.

Mean of the Poisson distribution
Let $X \sim \text{Poisson}(\lambda)$. Then

$$\begin{aligned} E(X) &= \sum_{x=0}^{\infty} xP(X=x) \\ &= \sum_{x=0}^{\infty} xe^{-\lambda}\frac{\lambda^x}{x!} \\ &= e^{-\lambda}\sum_{x=1}^{\infty} x\frac{\lambda^x}{x!} \\ &= e^{-\lambda}\sum_{x=1}^{\infty} \frac{\lambda\cdot\lambda^{(x-1)}}{(x-1)!} \\ &= \lambda e^{-\lambda}\sum_{x=1}^{\infty} \frac{\lambda^{(x-1)}}{(x-1)!} \\ &= \lambda e^{-\lambda}\sum_{y=0}^{\infty} \frac{\lambda^y}{y!} \quad [y=x-1] \\ &= \lambda e^{-\lambda}e^{\lambda} \quad [\text{using the expansion of } e^{\lambda}] \\ &= \lambda. \end{aligned}$$

Variance of the Poisson distribution
Let $X \sim \text{Poisson}(\lambda)$. Then

$$\begin{aligned} E[X(X-1)] &= \sum_{x=0}^{\infty} x(x-1)P(X=x) \\ &= \sum_{x=0}^{\infty} x(x-1)e^{-\lambda}\frac{\lambda^x}{x!} \\ &= e^{-\lambda}\sum_{x=2}^{\infty} x(x-1)\frac{\lambda^x}{x!} \\ &= e^{-\lambda}\sum_{x=2}^{\infty} \lambda^2\frac{\lambda^{x-2}}{(x-2)!} \\ &= \lambda^2 e^{-\lambda}\sum_{y=0}^{\infty} \frac{\lambda^y}{y!} \quad [y=x-2] \\ &= \lambda^2 e^{-\lambda}e^{\lambda} = \lambda^2 \quad [\text{using the expansion of } e^{\lambda}] \end{aligned}$$

Now, $E(X^2) = E[X(X-1)] + E(X) = \lambda^2 + \lambda$. Hence,

$$\text{Var}(X) = E(X^2) - (E(X))^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

Hence, the mean and variance are the same for the Poisson distribution.

The Poisson distribution can be derived from another consideration when we are waiting for events to occur, e.g. waiting for a bus to arrive or to be served at a supermarket till. The number of occurrences in a given time interval can sometimes be modelled by the Poisson distribution. Here the assumption is that the probability of an event (arrival) is proportional to the length of the waiting time for small time intervals. Such a process is called a Poisson process, and it can be shown that the waiting time between successive events can be modelled by the exponential distribution which is discussed in the next lecture.

Using `R` to calculate probabilities

For the Poisson distribution the command `dpois(x=3, lambda=5)` calculates the pmf of Poisson($\lambda = 5$) at $x = 3$. That is, the command will return the value $P(X = 3) = e^{-5}\frac{5^3}{3!}$. The command `ppois` returns the cdf or the probability up to and including the argument. Thus `ppois(q=3, lambda=5)` will return the value of $P(X \leq 3)$ when $X \sim \text{Poisson}(\lambda = 5)$.

3.4.4 Take home points

In this lecture we have learned about the Negative Binomial and Poisson distributions.

3.5 Lecture 15: Standard continuous distributions

3.5.1 Lecture mission

In this lecture we will learn about the Exponential distribution and its properties.

3.5.2 Exponential distribution

A continuous random variable X is said to follow the exponential distribution if its pdf is of the form:

$$f(x) = \begin{cases} \theta e^{-\theta x} & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases}$$

where $\theta > 0$ is a parameter. We write $X \sim Exponential(\theta)$. The distribution only resides in the positive half of the real line, and the tail goes down to zero exponentially as $x \to \infty$. The rate at which that happens is the parameter θ. Hence θ is known as the *rate parameter*.

It is easy to prove that $\int_0^\infty f(x)dx = 1$. This is left as an exercise. To find the mean and variance of the distribution we need the gamma function as discussed in Section A.6.3.

Definition: Gamma function:
For any positive number a,

$$\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}dx$$

is defined to be the gamma function and it has a finite real value. Moreover, we have the following facts:

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}; \quad \Gamma(1) = 1; \quad \Gamma(a) = (a-1)\Gamma(a-1) \quad \text{if } a > 1.$$

These last two facts imply that $\Gamma(k) = (k-1)!$ when k is a positive integer. Find $\Gamma\left(\frac{3}{2}\right)$.

Mean and variance of the exponential distribution

By definition,

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} xf(x)dx \\ &= \int_0^\infty x\theta e^{-\theta x}dx \\ &= \int_0^\infty ye^{-y}\frac{dy}{\theta} \quad \{\text{substitute } y = \theta x\} \\ &= \frac{1}{\theta}\int_0^\infty y^{2-1}e^{-y}dy \\ &= \frac{1}{\theta}\Gamma(2) \\ &= \frac{1}{\theta} \quad \{\text{since } \Gamma(2) = 1! = 1\}. \end{aligned}$$

Now,

$$\begin{aligned} E(X^2) &= \int_{-\infty}^{\infty} x^2f(x)dx \\ &= \int_0^\infty x^2\theta e^{-\theta x}dx \\ &= \theta\int_0^\infty \left(\frac{y}{\theta}\right)^2 e^{-y}\frac{dy}{\theta} \quad \{\text{substitute } y = \theta x\} \\ &= \frac{1}{\theta^2}\int_0^\infty y^{3-1}e^{-y}dy \\ &= \frac{1}{\theta^2}\Gamma(3) \\ &= \frac{2}{\theta^2} \quad \{\text{since } \Gamma(3) = 2! = 2\}, \end{aligned}$$

and so $\text{Var}(X) = E(X^2) - [E(X)]^2 = 2/\theta^2 - 1/\theta^2 = 1/\theta^2$. Note that for this random variable the mean is equal to the standard deviation.

It is easy to find the cdf of the exponential distribution. For $x > 0$,

$$F(x) = P(X \leq x) = \int_0^x \theta e^{-\theta u} du = 1 - e^{-\theta x}.$$

We have $F(0) = 0$ and $F(x) \to 1$ when $x \to \infty$ and $F(x)$ is non-decreasing in x. The cdf can be used to solve many problems. A few examples follow.

Using `R` to calculate probabilities

For the exponential distribution the command `dexp(x=3, rate=1/2)` calculates the pdf at $x = 3$. The rate parameter to be supplied is the θ parameter here. The command `pexp` returns the cdf or the probability up to and including the argument. Thus `pexp(q=3, rate=1/2)` will return the value of $P(X \leq 3)$ when $X \sim Exponential(\theta = 0.5)$.

♡ **Example 36 Mobile phone** Suppose that the lifetime of a phone (e.g. the time until the phone does not function even after repairs), denoted by X, manufactured by the company A Pale, is exponentially distributed with mean 550 days.

1. Find the probability that a randomly selected phone will still function after two years, i.e. $X > 730$? [Assume there is no leap year in the two years].
2. What are the times by which 25%, 50%, 75% and 90% of the manufactured phones will have failed?

Here the mean $1/\theta = 550$. Hence $\theta = 1/550$ is the rate parameter. The solution to the first problem is

$$P(X > 730) = 1 - P(X \leq 730) = 1 - (1 - e^{-730/550}) = e^{-730/550} = 0.2652.$$

The `R` command to find this is `1-pexp(q=730, rate=1/550)`.

For the second problem we are given the probabilities of failure (0.25, 0.50 etc.). We will have to invert the probabilities to find the value of the random variable. In other words, we will have to find a q such that $F(q) = p$, where p is the given probability. For example, what value of q will give us $F(q) = 0.25$, so that 25% of the phones will have failed by time q?

For a given $0 < p < 1$, the pth quantile (or $100p$ percentile) of the random variable X with cdf $F(x)$ is defined to be the value q for which $F(q) = p$.

The 50th percentile is called the median. The 25th and 75th percentiles are called the quartiles.

♡ **Example 37 Uniform distribution** Consider the uniform distribution $U(a, b)$ in the interval (a, b). Here $F(x) = \frac{x-a}{b-a}$. So for a given p, $F(q) = p$ implies $q = a + p(b - a)$.

For the uniform $U(a,b)$ distribution the median is $\frac{b+a}{2}$, and the quartiles are: $\frac{b+3a}{4}$ and $\frac{3b+a}{4}$.

Returning to the exponential distribution example, we have $p = F(q) = 1 - e^{-\theta q}$. Find q when p is given.

$$\begin{array}{rrcl} & p & = & 1 - e^{-\theta q} \\ \Rightarrow & e^{-\theta q} & = & 1 - p \\ \Rightarrow & -\theta q & = & \log(1-p) \\ \Rightarrow & q & = & \frac{-\log(1-p)}{\theta} \\ \Rightarrow & q & = & -550 \times \log(1-p). \end{array}$$

Review the rules of log in Section A.5. Now we have the following table:

p	$q = -550 \times \log(1-p)$
0.25	158.22
0.50	381.23
0.75	762.46
0.90	1266.422

In `R` you can find these values by `qexp(p=0.25, rate=1/550)`, `qexp(p=0.50, rate=1/550)`, etc. For fun, you can find `qexp(p=0.99, rate=1/550)` = 6 years and 343 days! The function `qexp(p, rate)` calculates the $100p$ percentile of the exponential distribution with parameter `rate`.

♡ **Example 38 Survival function** The exponential distribution is sometimes used to model the survival times in different experiments. For example, an exponential random variable T may be assumed to model the number of days a cancer patient survives after chemotherapy. In such a situation, the function $S(t) = 1 - F(t) = e^{-\theta t}$ is called the survival function. See Figure 3.2 for an example plot.

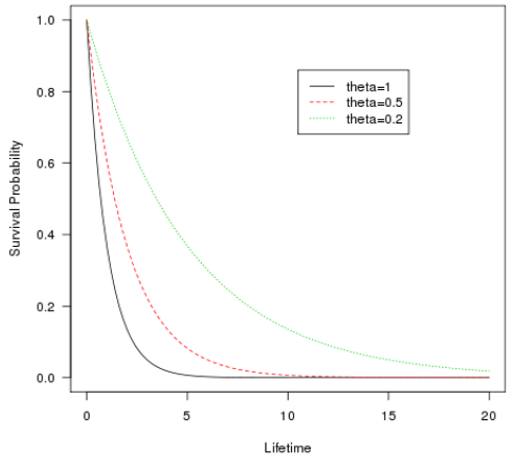


Figure 3.2: $S(t)$ for $\theta = 1, 0.5, 0.2$.

Assuming the mean survival time to be 100 days for a fatal late detected cancer, we can expect that half of the patients survive 69.3 days after chemo since `qexp(0.50, rate=1/100)` = 69.3. You will learn more about this in a third-year module, Math3085: Survival models, important in actuary.

♡ **Example 39 Memoryless property** Like the geometric distribution, the exponential distribution also has the memoryless property. In simple terms, it means that the probability that the system will survive an additional period $s > 0$ given that it has survived up to time t is the same as the probability that the system survives the period s to begin with. That is, it forgets that it has survived up to a particular time when it is thinking of its future remaining life time.

The proof is exactly as in the case of the geometric distribution, reproduced below. Recall the definition of conditional probability:

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

Now the proof,

$$\begin{aligned} P(X > s+t|X > t) &= \frac{P(X>s+t, X>t)}{P(X>t)} \\ &= \frac{P(X>s+t)}{P(X>t)} \\ &= \frac{e^{-\theta(s+t)}}{e^{-\theta t}} \\ &= e^{-\theta s} \\ &= P(X > s). \end{aligned}$$

Note that the event $X > s+t$ and $X > t$ implies and is implied by $X > s+t$ since $s > 0$.

♡ **Example 40** The time T between any two successive arrivals in a hospital emergency department has probability density function:

$$f(t) = \begin{cases} \lambda e^{-\lambda t} & \text{if } t \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Historically, on average the mean of these inter-arrival times is 5 minutes. Calculate (i) $P(0 < T < 5)$, (ii) $P(T < 10|T > 5)$.

An estimate of $E(T)$ is 5. As $E(T) = \frac{1}{\lambda}$ we take $\frac{1}{5}$ as the estimate of λ.

(i) $P(0 < T < 5) = \int_0^5 \frac{1}{5} e^{-t/5}\, dt = [-e^{-t/5}]_0^5 = 1 - e^{-1} = 0.63212.$

(ii)

$$\begin{aligned} P(T < 10|T > 5) &= \frac{P(5 < T < 10)}{P(T > 5)} \\ &= \frac{\int_5^{10} \frac{1}{5} e^{-t/5}\, dt}{\int_5^{\infty} \frac{1}{5} e^{-t/5}\, dt} = \frac{[-e^{-t/5}]_5^{10}}{[-e^{-t/5}]_5^{\infty}} \\ &= 1 - e^{-1} = 0.63212. \end{aligned}$$

3.5.3 Take home points

In this lecture we have learned many properties of the Exponential distribution.

3.6 Lecture 16: The normal distribution

3.6.1 Lecture mission

The normal distribution is the most commonly encountered continuous distribution in statistics and in science in general. This lecture will be entirely devoted to learning many properties of this distribution.

3.6.2 The pdf, mean and variance of the normal distribution

A random variable X is said to have the normal distribution with parameters μ and σ^2 if it has the following pdf:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, -\infty < x < \infty \tag{3.1}$$

where $-\infty < \mu < \infty$ and $\sigma > 0$ are two given constants. It is easy to see that $f(x) > 0$ for all x. We will now prove that

R1 $\int_{-\infty}^{\infty} f(x)dx = 1$ or total probability equals 1, so that $f(x)$ defines a valid pdf.

R2 $E(X) = \mu$, i.e. the mean is μ.

R3 $\mathrm{Var}(X) = \sigma^2$, i.e. the variance is σ^2.

We denote the normal distribution by the notation $N(\mu, \sigma^2)$.

Suppose all of these hold (since they are proved below). Then it is easy to remember the pdf of the normal distribution:

$$f(\text{variable}) = \frac{1}{\sqrt{2\pi\ \text{variance}}} \exp\left\{-\frac{(\text{variable}-\text{mean})^2}{2\ \text{variance}}\right\}$$

where variable denotes the random variable. The density (pdf) is much easier to remember and work with when the mean $\mu = 0$ and variance $\sigma^2 = 1$. In this case, we simply write:

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} \quad \text{or} \quad f(\text{variable}) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{\text{variable}^2}{2}\right\}.$$

Now let us prove the 3 assertions, **R1**, **R2** and **R3**. **R1** is proved as follows:

$$\begin{aligned}
\int_{-\infty}^{\infty} f(x)dx &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{-\frac{z^2}{2}\right\} dz \quad [\text{substitute } z = \tfrac{x-\mu}{\sigma} \text{ so that } dx = \sigma dz] \\
&= \frac{1}{\sqrt{2\pi}} 2 \int_{0}^{\infty} \exp\left\{-\frac{z^2}{2}\right\} dz \quad [\text{since the integrand is an even function}] \\
&= \frac{1}{\sqrt{2\pi}} 2 \int_{0}^{\infty} \exp\{-u\} \frac{du}{\sqrt{2u}} \quad [\text{substitute } u = \tfrac{z^2}{2} \text{ so that } z = \sqrt{2u} \text{ and } dz = \tfrac{du}{\sqrt{2u}}] \\
&= \frac{1}{2\sqrt{\pi}} 2 \int_{0}^{\infty} u^{\frac{1}{2}-1} \exp\{-u\} du \quad [\text{rearrange the terms}] \\
&= \frac{1}{\sqrt{\pi}} \Gamma\left(\tfrac{1}{2}\right) \quad [\text{recall the definition of the Gamma function}] \\
&= \frac{1}{\sqrt{\pi}} \sqrt{\pi} = 1 \quad [\text{as } \Gamma\left(\tfrac{1}{2}\right) = \sqrt{\pi}].
\end{aligned}$$

To prove **R2**, i.e. $E(X) = \mu$, we prove the following two results:

$$(i) X \sim N(\mu, \sigma^2) \longleftrightarrow Z \equiv \frac{X-\mu}{\sigma} \sim N(0,1) \tag{3.2}$$

$$(ii) \quad E(Z) = 0. \tag{3.3}$$

Then by the linearity of expectations, i.e. if $X = \mu + \sigma Z$ for constants μ and σ then $E(X) = \mu + \sigma E(Z) = \mu$, the result follows. To prove (3.2), we first calculate the cdf, given by:

$$\begin{aligned}
\Phi(z) &= P(Z \le z) \\
&= P\left(\frac{X-\mu}{\sigma} \le z\right) \\
&= P(X \le \mu + z\sigma) \\
&= \int_{-\infty}^{\mu+z\sigma} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx \\
&= \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{u^2}{2}\right\} du, \quad [u = (x-\mu)/\sigma]
\end{aligned}$$

and so the pdf of Z is

$$\frac{d\Phi(z)}{dz} = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\} \quad \text{for } -\infty < z < \infty,$$

by the fundamental theorem of calculus. This proves that $Z \sim N(0,1)$. The converse is proved just by reversing the steps. Thus we have proved (i) above. We use the $\Phi(\cdot)$ notation to denote the cdf of the standard normal distribution. Now:

$$\begin{aligned}
E(Z) &= \int_{-\infty}^{\infty} z f(z) dz \\
&= \int_{-\infty}^{\infty} z \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\} dz \\
&= \frac{1}{\sqrt{2\pi}} \times 0 = 0,
\end{aligned}$$

since the integrand $g(z) = z \exp\left\{-\frac{z^2}{2}\right\}$ is an odd function, i.e. $g(z) = -g(-z)$; for an odd function $g(z)$, $\int_{-a}^{a} g(z)dz = 0$ for any a. Therefore we have also proved (3.3) and hence **R2**.

To prove **R3**, i.e. $\mathrm{Var}(X) = \sigma^2$, we show that $\mathrm{Var}(Z) = 1$ where $Z = \frac{X-\mu}{\sigma}$ and then claim that $\mathrm{Var}(X) = \sigma^2 \mathrm{Var}(Z) = \sigma^2$ from our earlier result. Since $E(Z) = 0$, $\mathrm{Var}(Z) = E(Z^2)$, which is

calculated below:

$$\begin{aligned}
E(Z^2) &= \int_{-\infty}^{\infty} z^2 f(z) dz \\
&= \int_{-\infty}^{\infty} z^2 \tfrac{1}{\sqrt{2\pi}} \exp\left\{-\tfrac{z^2}{2}\right\} dz \\
&= \tfrac{2}{\sqrt{2\pi}} \int_0^{\infty} z^2 \exp\left\{-\tfrac{z^2}{2}\right\} dz \quad \text{[since the integrand is an even function]} \\
&= \tfrac{2}{\sqrt{2\pi}} \int_0^{\infty} 2u \exp\{-u\} \tfrac{du}{\sqrt{2u}} \quad \text{[substituted } u = \tfrac{z^2}{2} \text{ so that } z = \sqrt{2u} \text{ and } dz = \tfrac{du}{\sqrt{2u}}\text{]} \\
&= \tfrac{4}{2\sqrt{\pi}} \int_0^{\infty} u^{\frac{1}{2}} \exp\{-u\} du \\
&= \tfrac{2}{\sqrt{\pi}} \int_0^{\infty} u^{\frac{3}{2}-1} \exp\{-u\} du \\
&= \tfrac{2}{\sqrt{\pi}} \Gamma\left(\tfrac{3}{2}\right) \quad \text{[definition of the gamma function]} \\
&= \tfrac{2}{\sqrt{\pi}} \left(\tfrac{3}{2}-1\right) \Gamma\left(\tfrac{3}{2}-1\right) \quad \text{[reduction property of the gamma function]} \\
&= \tfrac{2}{\sqrt{\pi}} \tfrac{1}{2} \sqrt{\pi} \quad \text{[since } \Gamma\left(\tfrac{1}{2}\right) = \sqrt{\pi}\text{]} \\
&= 1,
\end{aligned}$$

as we hoped for! This proves **R3**.

Linear transformation of a Normal random variable
Suppose $X \sim N(\mu, \sigma^2)$ and a and b are constants. Then the distribution of $Y = aX + b$ is $N(a\mu + b, a^2\sigma^2)$.

Proof: The result that Y has mean $a\mu + b$ and variance $a^2\sigma^2$ can already be claimed from the linearity of the expectations and the variance result for linear functions. What remains to be proved is that the normality of Y, i.e. how can we claim that Y will follow the normal distribution too? For this, we note that

$$Y = aX + b = a(\mu + \sigma Z) + b = (a\mu + b) + a\sigma Z$$

since $X = \mu + \sigma Z$. Now we use (3.2) to claim the normality of Y.

3.6.3 Take home points

In this lecture we have learned the most important properties of the normal distribution.

3.7 Lecture 17: The standard normal distribution

3.7.1 Lecture mission

In this lecture we learn how to calculate probabilities under the normal distribution. We also learn how to calculate these probabilities using `R`.

3.7.2 Standard normal distribution

Now we can claim that the normal pdf (3.1) is symmetric about the mean μ. The spread of the pdf is determined by σ, the standard deviation of the distribution. When $\mu = 0$ and $\sigma = 1$, the normal distribution $N(0, 1)$ is called is called the standard normal distribution. The standard normal distribution, often denoted by Z, is used to calculate probabilities of interest for any normal

distribution because of the following reasons. Suppose $X \sim N(\mu, \sigma^2)$ and we are interested in finding $P(a \leq X \leq b)$ for two constants a and b.

$$\begin{aligned}
P(a \leq X \leq b) &= \int_a^b f(x)dx \\
&= \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} \exp\left\{-\frac{z^2}{2}\right\} dz \quad [\text{substituted } z = \tfrac{x-\mu}{\sigma} \text{ so that } dx = \sigma dz] \\
&= \int_{-\infty}^{\frac{b-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\} dz - \int_{-\infty}^{\frac{a-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\} dz \\
&= P\left(Z \leq \tfrac{b-\mu}{\sigma}\right) - P\left(Z \leq \tfrac{a-\mu}{\sigma}\right) \\
&= \text{cdf of } Z \text{ at } \tfrac{b-\mu}{\sigma} - \text{ cdf of } Z \text{ at } \tfrac{a-\mu}{\sigma} \\
&= \Phi\left(\tfrac{b-\mu}{\sigma}\right) - \Phi\left(\tfrac{a-\mu}{\sigma}\right)
\end{aligned}$$

where we use the notation $\Phi(\cdot)$ to denote the cdf of Z, i.e.

$$P(Z \leq z) = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{u^2}{2}\right\} du.$$

This result allows us to find the probabilities about a normal random variable X of any mean μ and variance σ^2 through the probabilities of the standard normal random variable Z. For this reason, only $\Phi(z)$ is tabulated. Further more, due to the symmetry of the pdf of Z, $\Phi(z)$ is tabulated only for positive z values. Suppose $a > 0$, then

$$\begin{aligned}
\Phi(-a) = P(Z \leq -a) &= P(Z > a) \\
&= 1 - P(Z \leq a) \\
&= 1 - \Phi(a).
\end{aligned}$$

In `R`, we use the function `pnorm` to calculate the probabilities. The general function is: `pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)`. So, we use the command `pnorm(1)` to calculate $\Phi(1) = P(Z \leq 1)$. We can also use the command `pnorm(15, mean=10, sd=2)` to calculate $P(X \leq 15)$ when $X \sim N(\mu = 10, \sigma^2 = 4)$ directly.

1. $P(-1 < Z < 1) = \Phi(1) - \Phi(-1) = 0.6827$. This means that 68.27% of the probability lies within 1 standard deviation of the mean.

2. $P(-2 < Z < 2) = \Phi(2) - \Phi(-2) = 0.9545$. This means that 95.45% of the probability lies within 2 standard deviations of the mean.

3. $P(-3 < Z < 3) = \Phi(3) - \Phi(-3) = 0.9973$. This means that 99.73% of the probability lies within 3 standard deviations of the mean.

We are often interested in the quantiles (inverse-cdf of probability, $\Phi^{-1}(\cdot)$ of the normal distribution for various reasons. We find the pth quantile by issuing the `R` command `qnorm(p)`.

1. `qnorm(0.95)` $= \Phi^{-1}(0.95) = 1.645$. This means that the 95th percentile of the standard normal distribution is 1.645. This also means that $P(-1.645 < Z < 1.645) = \Phi(1.645) - \Phi(-1.645) = 0.90$.

2. `qnorm(0.975)` $= \Phi^{-1}(0.975) = 1.96$. This means that the 97.5th percentile of the standard normal distribution is 1.96. This also means that $P(-1.96 < Z < 1.96) = \Phi(1.96) - \Phi(-1.96) = 0.95$.

♡ **Example 41** Historically, the marks in MATH1024 follow the normal distribution with mean 58 and standard deviation 32.25.

1. What percentage of students will fail (i.e. score less than 40) in MATH1024? Answer: `pnorm(40, mean=58, sd=32.25)` $= 28.84\%$.

2. What percentage of students will get an A result (score greater than 70)? Answer: `1- pnorm(70, mean=58, sd=32.25)` $= 35.49\%$.

3. What is the probability that a randomly selected student will score more than 90? Answer: `1- pnorm(90, mean=58, sd=32.25)` $= 0.1605$.

4. What is the probability that a randomly selected student will score less than 25? Answer: `pnorm(25, mean=58, sd=32.25)` $= 0.1531$. Ouch!

5. What is the probability that a randomly selected student scores a 2:1, (i.e. a mark between 60 and 70)? Left as an exercise.

♡ **Example 42** A lecturer set and marked an examination and found that the distribution of marks was $N(42, 14^2)$. The school's policy is to present scaled marks whose distribution is $N(50, 15^2)$. What linear transformation should the lecturer apply to the raw marks to accomplish this and what would the raw mark of 40 be transformed to?

Suppose $X \sim N(\mu_x = 42, \sigma_x^2 = 14^2)$ and $Y \sim N(\mu_y = 50, \sigma_y^2 = 15^2)$. Hence, we should have

$$Z = \frac{X - \mu_x}{\sigma_x} = \frac{Y - \mu_y}{\sigma_y},$$

giving us:

$$Y = \mu_y + \frac{\sigma_y}{\sigma_x}(X - \mu_x) = 50 + \frac{15}{14}(X - 42).$$

Now at raw mark $X = 40$, the transformed mark would be:

$$Y = 50 + \frac{15}{14}(40 - 42) = 47.86.$$

♡ **Example 43 Log-normal distribution**

If $X \sim N(\mu, \sigma^2)$ then the random variable $Y = \exp(X)$ is called a log-normal random variable and its distribution is called a log-normal distribution with parameters μ and σ^2.

The mean of the random variable Y is given by

$$\begin{aligned}
E(Y) &= E[\exp(X)] \\
&= \int_{-\infty}^{\infty} \exp(x) \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx \\
&= \exp\left\{-\frac{\mu^2-(\mu+\sigma^2)^2}{2\sigma^2}\right\} \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{x^2-2(\mu+\sigma^2)x+(\mu+\sigma^2)^2}{2\sigma^2}\right\} dx \\
&= \exp\left\{-\frac{\mu^2-(\mu+\sigma^2)^2}{2\sigma^2}\right\} \text{[integrating a } N(\mu+\sigma^2, \sigma^2) \text{ r.v. over its domain]} \\
&= \exp\left\{\mu + \sigma^2/2\right\}
\end{aligned}$$

Similarly, one can show that

$$\begin{aligned}
E(Y^2) &= E[\exp(2X)] \\
&= \int_{-\infty}^{\infty} \exp(2x) \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx \\
&= \cdots \\
&= \exp\left\{2\mu + 2\sigma^2\right\}.
\end{aligned}$$

Hence, the variance is given by

$$\text{Var}(Y) = E(Y^2) - (E(Y))^2 = \exp\left\{2\mu + 2\sigma^2\right\} - \exp\left\{2\mu + \sigma^2\right\}.$$

3.7.3 Take home points

In this lecture we have learned more properties of the normal distribution. These properties will be required for calculating probabilities and making inference. We have also introduced the log-normal distribution which is often used in practice for modelling economic variables of interest in business and finance, e.g. volume of sales, income of individuals. You do not need to remember the mean and variance of the log-normal distribution.

3.8 Lecture 18: Joint distributions

3.8.1 Lecture mission

Often we need to study more than one random variable, e.g. height and weight, simultaneously, so that we can exploit the relationship between them to make inferences about their properties. Multiple random variables are studied through their joint probability distribution. In this lecture we will study covariance and correlation and then discuss when random variables are independent.

3.8.2 Joint distribution of discrete random variables

If X and Y are discrete, the quantity $f(x,y) = P(X = x \cap Y = y)$ is called the *joint probability mass function* (joint pmf) of X and Y. To be a joint pmf, $f(x,y)$ needs to satisfy two conditions:

$$(i) \quad f(x,y) \geq 0$$

for all x and y and

$$(ii) \quad \sum_{\text{All } x} \sum_{\text{All } y} f(x,y) = 1.$$

The marginal probability mass functions (marginal pmf's) of X and Y are respectively

$$f_X(x) = \sum_y f(x,y), \quad f_Y(y) = \sum_x f(x,y).$$

Use the identity $\sum_x \sum_y f(x,y) = 1$ to prove that $f_X(x)$ and $f_Y(y)$ are really pmf's.

♡ **Example 44** Suppose that two fair dice are tossed independently one after the other. Let

$$X = \begin{cases} -1 & \text{if the result from die 1 is larger} \\ 0 & \text{if the results are equal} \\ 1 & \text{if the result from die 1 is smaller.} \end{cases}$$

Let Y = |difference between the two dice|. There are 36 possible outcomes. Each of them gives a pair of values of X and Y. Y can take any of the values 0, 1, 2, 3, 4, 5. Construct the joint probability table for X and Y.

Results		x	y	Results		x	y	Results		x	y
1	1	0	0	3	1	–1	2	5	1	–1	4
1	2	1	1	3	2	–1	1	5	2	–1	3
1	3	1	2	3	3	0	0	5	3	–1	2
1	4	1	3	3	4	1	1	5	4	–1	1
1	5	1	4	3	5	1	2	5	5	0	0
1	6	1	5	3	6	1	3	5	6	1	1
2	1	–1	1	4	1	–1	3	6	1	–1	5
2	2	0	0	4	2	–1	2	6	2	–1	4
2	3	1	1	4	3	–1	1	6	3	–1	3
2	4	1	2	4	4	0	0	6	4	–1	2
2	5	1	3	4	5	1	1	6	5	–1	1
2	6	1	4	4	6	1	2	6	6	0	0

Each pair of results above (and hence pair of values of X and Y) has the same probability 1/36. Hence the joint probability table is given in Table 3.1

The marginal probability distributions are just the row totals or column totals depending on whether you want the marginal distribution of X or Y. For example, the marginal distribution of X is given in Table 3.2.

Table 3.1: Joint probability distribution of X and Y

		y						
		0	1	2	3	4	5	Total
x	-1	0	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	$\frac{15}{36}$
	0	$\frac{6}{36}$	0	0	0	0	0	$\frac{6}{36}$
	1	0	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	$\frac{15}{36}$
	Total	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{8}{36}$	$\frac{6}{36}$	$\frac{4}{36}$	$\frac{2}{36}$	1

Table 3.2: Marginal probability distribution of X.

x	$P(X = x)$
-1	$\frac{15}{36}$
0	$\frac{6}{36}$
1	$\frac{15}{36}$
Total	1

Exercises: Write down the marginal distribution of Y and hence find the mean and variance of Y.

Bivariate continuous distributions

If X and Y are continuous, a non-negative real-valued function $f(x, y)$ is called the *joint probability density function* (joint pdf) of X and Y if

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)dxdy = 1.$$

The marginal pdf's of X and Y are respectively

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)dy, \quad f_Y(y) = \int_{-\infty}^{\infty} f(x, y)dx.$$

♡ **Example 45** Define a joint pdf by

$$f(x, y) = \begin{cases} 6xy^2 & \text{if } 0 < x < 1 \text{ and } 0 < y < 1 \\ 0 & \text{otherwise.} \end{cases}$$

How can we show that the above is a pdf? It is non-negative for all x and y values. But does it integrate to 1? We are going to use the following rule.

Result Suppose that a real-valued function $f(x, y)$ is continuous in a region D where $a < x < b$ and $c < y < d$, then

$$\int\int_D f(x, y)dxdy = \int_c^d dy \int_a^b f(x, y)dx.$$

Here a and b may depend upon y but c and d should be free of x and y. When we evaluate the inner integral $\int_a^b f(x,y)dx$, we treat y as constant.

Notes: To evaluate a bivariate integral over a region A we:

- Draw a picture of A whenever possible.
- Rewrite the region A as an intersection of two one-dimensional intervals. The first interval is obtained by treating one variable as constant.
- Perform two one-dimensional integrals.

♡ **Example 46 Continued**

$$\begin{aligned}\int_0^1\int_0^1 f(x,y)dxdy &= \int_0^1\int_0^1 6xy^2\,dxdy \\ &= 6\int_0^1 y^2dy\int_0^1 x\,dx \\ &= 3\int_0^1 y^2dy \quad [\text{as } \int_0^1 x\,dx=\tfrac{1}{2}] \\ &= 1. \quad [\text{as } \int_0^1 y^2\,dy=\tfrac{1}{3}]\end{aligned}$$

Now we can find the marginal pdf's as well.

$$f_X(x)=2x, 0<x<1 \text{ and } f_Y(y)=3y^2, 0<y<1.$$

The probability of any event in the two-dimensional space can be found by integration and again more details will be provided in a second-year module. You will come across multivariate integrals in a second semester module. **You will not be asked to do bivariate integration in this module.**

3.8.3 Covariance and correlation

We first define the expectation of a real-valued scalar function $g(X,Y)$ of X and Y:

$$E[g(X,Y)]=\begin{cases}\sum_x\sum_y g(x,y)f(x,y) & \text{if } X \text{ and } Y \text{ are discrete}\\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)dxdy & \text{if } X \text{ and } Y \text{ are continuous.}\end{cases}$$

♡ **Example 47 Example 44 continued** Let $g(x,y)=xy$.

$$E(XY)=(-1)(0)0+(-1)(1)\frac{5}{36}+\cdots+(1)(5)\frac{1}{36}=0.$$

Exercises: Try $g(x,y)=x$. It will be the same thing as $E(X)=\sum_x xf_X(x)$.

We will not consider any continuous examples as the second-year module MATH2011 will study them in detail.

Suppose that two random variables X and Y have joint pmf or pdf $f(x,y)$ and let $E(X)=\mu_x$ and $E(Y)=\mu_y$. The covariance between X and Y is defined by

$$\boxed{\text{Cov}(X,Y) = E\left[(X-\mu_x)(Y-\mu_y)\right] = E(XY) - \mu_x\mu_y.}$$

Let $\sigma_x^2 = \text{Var}(X) = E(X^2) - \mu_x^2$ and $\sigma_y^2 = \text{Var}(Y) = E(Y^2) - \mu_y^2$. The correlation coefficient between X and Y is defined by:

$$\boxed{\text{Corr}(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}} = \frac{E(XY) - \mu_x\mu_y}{\sigma_x\,\sigma_y}.}$$

It can be proved that for any two random variables, $-1 \leq \text{Corr}(X,Y) \leq 1$. The correlation $\text{Corr}(X,Y)$ is a measure of linear dependency between two random variables X and Y, and it is free of the measuring units of X and Y as the units cancel in the ratio.

3.8.4 Independence

Independence is an important concept. Recall that we say two events A and B are independent if $P(A \cap B) = P(A) \times P(B)$. We use the same idea here. Two random variables X and Y having the joint pdf or pmf $f(x,y)$ are said to be independent if and only if

$$\boxed{f(x,y) = f_X(x) \times f_Y(y) \text{ for } \textbf{ALL } x \text{ and } y.}$$

♡ **Example 48 Discrete Case** X and Y are independent if *each* cell probability, $f(x,y)$, is the product of the corresponding row and column totals. In our very first dice example (Example 44) X and Y are not independent. Verify that in the following example X and Y are independent. We need to check all 9 cells.

		y			
		1	2	3	Total
x	0	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{3}$
	1	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{2}$
	2	$\frac{1}{12}$	$\frac{1}{24}$	$\frac{1}{24}$	$\frac{1}{6}$
	Total	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	1

♡ **Example 49** Let $f(x,y) = 6xy^2,\ 0 < x < 1,\ 0 < y < 1$. Check that X and Y are independent.

♡ **Example 50** Let $f(x,y) = 2x,\ 0 \leq x \leq 1,\ 0 \leq y \leq 1$. Check that X and Y are independent.

♡ **Example 51 Deceptive**

The joint pdf may look like something you can factorise. But X and Y may not be independent because they may be related in the domain.

1. $f(x,y) = \frac{21}{4}x^2 y,\ x^2 \leq y \leq 1$. Not independent!
2. $f(x,y) = e^{-y}, 0 < x < y < \infty$. Not independent!

Consequences of Independence

- Suppose that X and Y are independent random variables. Then

$$P(X \in A, Y \in B) = P(X \in A) \times P(Y \in B)$$

for any events A and B. That is, the joint probability can be obtained as the product of the marginal probabilities. We will use this result in the next lecture. For example, suppose Jack and Jess are two randomly selected students. Let X denote the height of Jack and Y denote the height of Jess. Then we have,

$$P(X < 182 \text{ and } Y > 165) = P(X < 182) \times P(Y > 165).$$

Obviously this has to be true for any numbers other than the example numbers 182 and 165, and for any inequalities.

- Further, let $g(x)$ be a function of x only and $h(y)$ be a function of y only. Then, if X and Y are independent, it is easy to prove that

$$E[g(X)h(Y)] = E[g(X)] \times E[h(Y)].$$

As a special case, let $g(x) = x$ and $h(y) = y$. Then we have

$$E(XY) = E(X) \times E(Y).$$

Consequently, for independent random variables X and Y, $\text{Cov}(X,Y) = 0$ and $\text{Corr}(X,Y) = 0$. But the converse is not true in general. That is, merely having $\text{Corr}(X,Y) = 0$ does not imply that X and Y are independent random variables.

3.8.5 Take home points

We have discussed the joint distribution of two random variables. The discrete case is easy to conceptualise and analyse. The discussion of the continuous case requires bivariate integration and hence is postponed to the second year. We have also introduced covariance and correlation. We have the very important result that if two random variables are independent, their joint probability distribution factorises and their correlation is 0.

3.9 Lecture 19: Properties of the sample sum and mean

3.9.1 Lecture mission

In this lecture we consider sums of random variables, which arise frequently in both practice and theoretical results. For example, the mark achieved in an exam is the sum of the marks for each

question, and the sample mean is proportional to the sum of the sample values. By doing this, in the next lecture we will introduce the widely-used central limit theorem, the normal approximation to the binomial distribution and so on. In this lecture we will also use this theory to reproduce some of the results we obtained before, e.g. finding the mean and variance of the binomial and negative binomial distributions.

3.9.2 Introduction

Suppose we have obtained a random sample from a distribution with pmf or pdf $f(x)$, so that X can either be a discrete or a continuous random variable. We will learn more about random sampling in the next chapter. Let $X_1, \ldots, X_n$ denote the random sample of size n where n is a positive integer. We use upper case letters since each member of the random sample is a random variable. For example, I toss a fair coin n times and let X_i take the value 1 if a head appears in the ith trial and 0 otherwise. Now I have a random sample $X_1, \ldots, X_n$ from the Bernoulli distribution with probability of success equal to 0.5 since the coin is assumed to be fair.

We can get a random sample from a continuous random variable as well. Suppose it is known that the distribution of the heights of first-year students is normal with mean 175 centimetres and standard deviation 8 centimetres. I can randomly select a number of first-year students and record each student's height.

Suppose $X_1, \ldots, X_n$ is a random sample from a population with distribution $f(x)$. Then it can be shown that the random variables $X_1, \ldots, X_n$ are mutually independent, i.e.

$$P(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n) = P(X_1 \in A_1) \times P(X_2 \in A_2) \times \cdots P(X_n \in A_n)$$

for any set of events, $A_1, A_2, \ldots A_n$. That is, the joint probability can be obtained as the product of individual probabilities. An example of this for $n = 2$ was given in the previous lecture; see the discussion just below the paragraph **Consequences of independence**.

♡ **Example 52 Distribution of the sum of independent binomial random variables** Suppose $X \sim \text{Bin}(m, p)$ and $Y \sim \text{Bin}(n, p)$ independently. Note that p is the same in both distributions. Using the above fact that joint probability is the multiplication of individual probabilities, we can conclude that $Z = X + Y$ has the binomial distribution. It is intuitively clear that this should happen since X comes from m Bernoulli trials and Y comes from n Bernoulli trials independently, so Z comes from $m + n$ Bernoulli trials with common success probability p. We can prove the result mathematically as well, by finding the probability mass function of $Z = X + Y$ directly and observing that it is of the appropriate form. First, note that

$$P(Z = z) = P(X = x, Y = y)$$

subject to the constraint that $x + y = z, 0 \leq x \leq m$, $0 \leq y \leq n$. Thus,

$$\begin{aligned} P(Z = z) &= \textstyle\sum_{x+y=z} P(X = x, Y = y) \\ &= \textstyle\sum_{x+y=z} \binom{m}{x} p^x (1-p)^{m-x} \binom{n}{y} p^y (1-p)^{n-y} \\ &= \textstyle\sum_{x+y=z} \binom{m}{x}\binom{n}{y} p^z (1-p)^{m+n-z} \\ &= p^z (1-p)^{m+n-z} \textstyle\sum_{x+y=z} \binom{m}{x}\binom{n}{y} \\ &= \binom{m+n}{z} p^z (1-p)^{m+n-z}, \end{aligned}$$

using a result stated in Section A.3. Thus, we have proved that the sum of independent binomial random variables with common probability is binomial as well. This is called the reproductive property of random variables. You are asked to prove this for the Poisson distribution in an exercise sheet.

Now we will state two main results without proof. The proofs will presented in the second-year distribution theory module MATH2011. Suppose that $X_1, \ldots, X_n$ is a random sample from a population distribution with finite variance, and suppose that $E(X_i) = \mu_i$ and $\text{Var}(X_i) = \sigma_i^2$. Define a new random variable

$$Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n$$

where $a_1, a_2, \ldots, a_n$ are constants. Then:

1. $E(Y) = a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n$.
2. $\text{Var}(Y) = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_n^2\sigma_n^2$.

For example, if $a_i = 1$ for all $i = 1, \ldots, n$, the two results above imply that:

The expectation of the sum of independent random variables is the sum of the expectations of the individual random variables

and

the variance of the sum of independent random variables is the sum of the variances of the individual random variables.

The second result is only true for independent random variables, e.g. random samples. Now we will consider many examples.

♡ **Example 53 Mean and variance of binomial distribution**
Suppose $Y \sim \text{Bin}(n, p)$. Then we can write:

$$Y = X_1 + X_2 + \ldots + X_n$$

where each X_i is an independent Bernoulli trial with success probability p. We have shown before that, $E(X_i) = p$ and $\text{Var}(X_i) = p(1-p)$ by direct calculation. Now the above two results imply that:

$$E(Y) = E\left(\sum_{i=1}^{n} X_i\right) = p + p + \ldots + p = np.$$

$$\text{Var}(Y) = \text{Var}(X_1) + \cdots + \text{Var}(X_n) = p(1-p) + \ldots + p(1-p) = np(1-p).$$

Thus we avoided the complicated sums used to derive $E(X)$ and $\text{Var}(X)$ in Section 3.3.3.

♡ **Example 54 Mean and variance of negative binomial distribution**
Recall that the negative binomial random variable Y is the number of trials needed to obtain the

r-th success in a sequence of independent Bernoulli trials, each with success probability p. Let X_i be the number of trials needed after the $(i-1)$-th success to obtain the i-th success. It is easy to see that each X_i is a geometric random variable and $Y = X_1 + \cdots + X_r$. Hence,

$$E(Y) = E(X_1) + \cdots + E(X_r) = 1/p + \cdots + 1/p = r/p$$

and

$$\text{Var}(Y) = \text{Var}(X_1) + \cdots + \text{Var}(X_r) = (1-p)/p^2 + \cdots + (1-p)/p^2 = r(1-p)/p^2.$$

♡ **Example 55 Sum of independent Normal random variables**
Suppose that $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, 2, \ldots, k$ are independent random variables. Let $a_1, a_2, \ldots, a_k$ be constants and suppose that

$$Y = a_1 X_1 + \cdots + a_k X_k.$$

Then we can prove that:

$$Y \sim N\left(\sum_{i=1}^{k} a_i \mu_i, \sum_{i=1}^{k} a_i^2 \sigma_i^2\right).$$

It is clear that $E(Y) = \sum_{i=1}^{k} a_i \mu_i$ and $\text{Var}(Y) = \sum_{i=1}^{k} a_i^2 \sigma_i^2$. But that Y has the normal distribution cannot yet be proved with the theory we know. This proof will be provided in the second-year distribution theory module MATH2011.

As a consequence of the stated result we can easily see the following. Suppose X_1 and X_2 are independent $N(\mu, \sigma^2)$ random variables. Then $2X_1 \sim N(2\mu, 4\sigma^2)$, $X_1 + X_2 \sim N(2\mu, 2\sigma^2)$, and $X_1 - X_2 \sim N(0, 2\sigma^2)$. Note that $2X_1$ and $X_1 + X_2$ have different distributions.

Suppose that $X_i \sim N(\mu, \sigma^2)$, $i = 1, \ldots, n$ are independent. Then

$$X_1 + \cdots + X_n \sim N(n\mu, n\sigma^2),$$

and consequently,

$$\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n) \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

3.9.3 Take home points

In this lecture we learned that the sample sum or the mean are random variables in their own right. Also, we have obtained the distribution of the sample sum for the binomial random variable. We have stated two important results regarding the mean and variance of the sample sum. Moreover, we have stated without proof that the distribution of the sample mean is normal if the samples are from the normal distribution itself. This is also an example of the reproductive property of the distributions. You will learn more about this in the second-year module MATH2011. In this module, we will use these facts to introduce the central limit theorem – perhaps the most widely-used result in statistics.

3.10 Lecture 20: The Central Limit Theorem

3.10.1 Lecture mission

The sum (and average) of independent random variables show a remarkable behaviour in practice which is captured by the Central Limit Theorem (CLT). These random variables do not even have to be continuous, all we require is that they are independent and each of them has a finite mean and a finite variance. A version of the CLT follows.

3.10.2 Statement of the Central Limit Theorem (CLT)

Let $X_1, \ldots, X_n$ be independent random variables with finite $E(X_i) = \mu_i$ and finite $\text{Var}(X_i) = \sigma_i^2$. Define $Y = \sum_{i=1}^n X_i$. Then, for a sufficiently large n, *the central limit theorem states that* Y is approximately normally distributed with

$$E(Y) = \sum_{i=1}^n \mu_i, \quad \text{Var}(Y) = \sum_{i=1}^n \sigma_i^2.$$

This also implies that $\bar{X} = \frac{1}{n}Y$ also follows the normal distribution approximately, as the sample size $n \to \infty$. In particular, if $\mu_i = \mu$ and $\sigma_i^2 = \sigma^2$, i.e. all means are equal and all variances are equal, then the CLT states that, as $n \to \infty$,

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Equivalently,

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

as $n \to \infty$. The notion of convergence is explained by the convergence of distribution of $\bar{X}$ to that of the normal distribution with the appropriate mean and variance. It means that the cdf of the left hand side, $\sqrt{n}\frac{(\bar{X}-\mu)}{\sigma}$, converges to the cdf of the standard normal random variable, $\Phi(\cdot)$. In other words,

$$\lim_{n\to\infty} P\left(\sqrt{n}\frac{(\bar{X} - \mu)}{\sigma} \leq z\right) = \Phi(z), \quad -\infty < z < \infty.$$

So for "large samples", we can use $N(0, 1)$ as an approximation to the sampling distribution of $\sqrt{n}(\bar{X} - \mu)/\sigma$. This result is 'exact', i.e. no approximation is required, if the distribution of the X_i's are normal in the first place – this was discussed in the previous lecture.

How large does n have to be before this approximation becomes usable? There is no definitive answer to this, as it depends on how "close to normal" the distribution of X is. However, it is often a pretty good approximation for sample sizes as small as 20, or even smaller. It also depends on the skewness of the distribution of X; if the X-variables are highly skewed, then n will usually need to be larger than for corresponding symmetric X-variables for the approximation to be good. We will investigate this numerically using `R`.

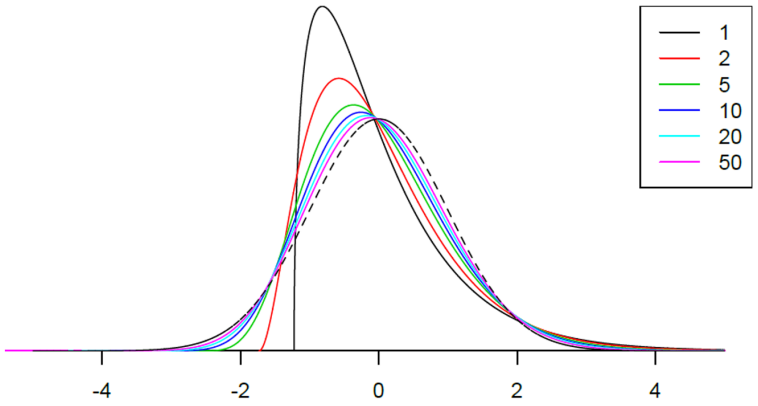


Figure 3.3: Distribution of normalised sample means for samples of different sizes. Initially very skew (original distribution, $n = 1$) becoming rapidly closer to standard normal (dashed line) with increasing n.

3.10.3 Application of CLT to binomial distribution

We know that a binomial random variable Y with parameters n and p is the number of successes in a set of n independent Bernoulli trials, each with success probability p. We have also learnt that

$$Y = X_1 + X_2 + \cdots + X_n,$$

where $X_1, \ldots, X_n$ are independent Bernoulli random variables with success probability p. It follows from the CLT that, for a sufficiently large n, Y is approximately normally distributed with expectation $E(Y) = np$ and variance $\mathrm{Var}(Y) = np(1-p)$.

Hence, for given integers y_1 and y_2 between 0 and n and a suitably large n, we have

$$\begin{aligned} P(y_1 \le Y \le y_2) &= P\left\{ \frac{y_1 - np}{\sqrt{np(1-p)}} \le \frac{Y - np}{\sqrt{np(1-p)}} \le \frac{y_2 - np}{\sqrt{np(1-p)}} \right\} \\ &\approx P\left\{ \frac{y_1 - np}{\sqrt{np(1-p)}} \le Z \le \frac{y_2 - np}{\sqrt{np(1-p)}} \right\}, \end{aligned}$$

where $Z \sim N(0,1)$.

We should take account of the fact that the binomial random variable Y is integer-valued, and so $P(y_1 \le Y \le y_2) = P(y_1 - f_1 \le Y \le y_2 + f_2)$ for any two fractions $0 < f_1, f_2 < 1$. This is called

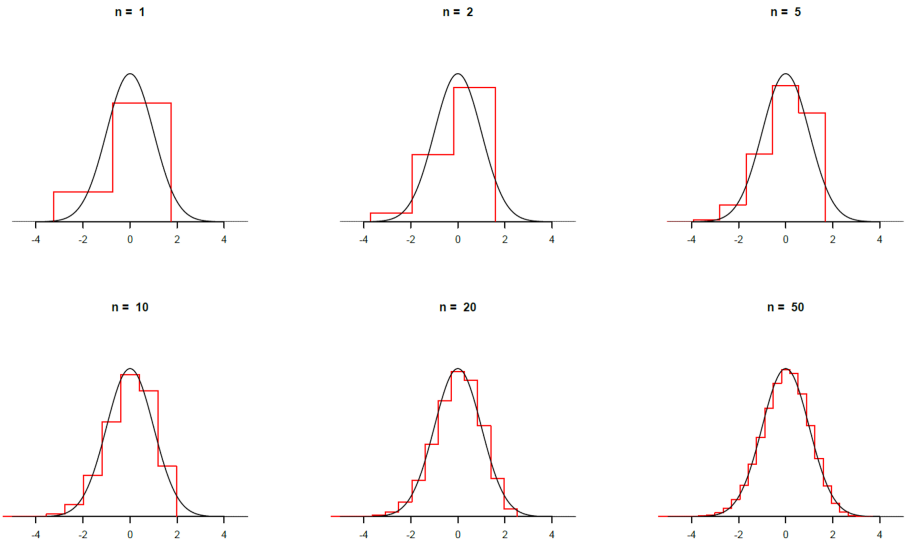


Figure 3.4: Histograms of normalised sample means for Bernoulli ($p = 0.8$) samples of different sizes. – converging to standard normal.

continuity correction and we take $f_1 = f_2 = 0.5$ in practice.

$$\begin{aligned} P(y_1 \leq Y \leq y_2) &= P(y_1 - 0.5 \leq Y \leq y_2 + 0.5) \\ &= P\left\{ \frac{y_1 - 0.5 - np}{\sqrt{np(1-p)}} \leq \frac{Y - np}{\sqrt{np(1-p)}} \leq \frac{y_2 + 0.5 - np}{\sqrt{np(1-p)}} \right\} \\ &\approx P\left\{ \frac{y_1 - 0.5 - np}{\sqrt{np(1-p)}} \leq Z \leq \frac{y_2 + 0.5 - np}{\sqrt{np(1-p)}} \right\}. \end{aligned}$$

What do we mean by a suitably large n? A commonly-used guideline is that the approximation is adequate if $np \geq 5$ and $n(1-p) \geq 5$.

♡ **Example 56** A producer of natural yoghurt believed that the market share of their brand was 10%. To investigate this, a survey of 2500 yoghurt consumers was carried out. It was observed that only 205 of the people surveyed expressed a preference for their brand. Should the producer be concerned that they might be losing market share?

Assume that the conjecture about market share is true. Then the number of people Y who prefer this product follows a binomial distribution with $p = 0.1$ and $n = 2500$. So the mean is $np = 250$, the variance is $np(1-p) = 225$, and the standard deviation is 15. The exact probability of observing $(Y \leq 205)$ is given by the sum of the binomial probabilities up to and including 205,

which is difficult to compute. However, this can be approximated by using the CLT:

$$\begin{aligned} P(Y \leq 205) &= P(Y \leq 205.5) \\ &= P\left\{ \frac{Y - np}{\sqrt{np(1-p)}} \leq \frac{205.5 - np}{\sqrt{np(1-p)}} \right\} \\ &\approx P\left\{ Z \leq \frac{205.5 - np}{\sqrt{np(1-p)}} \right\} \\ &= P\left\{ Z \leq \frac{205.5 - 250}{15} \right\} \\ &= \Phi(-2.967) = 0.0015. \end{aligned}$$

This probability is so small that it casts doubt on the validity of the assumption that the market share is 10%.

3.10.4 Take home points

In this lecture we have learned about the central limit theorem. This basically states that the sampling distribution of the sample sum (and also the mean) is an approximate normal distribution regardless of the probability distribution of the original random variables, provided that those random variables have finite means and variances.

Chapter 4

Statistical Inference

Chapter mission

In the last chapter we learned the probability distributions of common random variables that we use in practice. We learned how to calculate the probabilities based on our assumption of a probability distribution with known parameter values. Statistical inference is the process by which we try to learn about those probability distributions using only random observations. Hence, if our aim is to learn about some typical characteristics of the population of Southampton students, we simply randomly select few students, observe their characteristics and then try to generalise, as discussed in Lecture 1. For example, suppose we are interested in learning what proportion of Southampton students are of Indian origin. We may then select a number of students at random and observe the sample proportion of Indian origin students. We will then claim that the sample proportion is really our guess for the population proportion. But obviously we may be making grave errors since we are inferring about some unknown based on only a tiny fraction of total information. Statistical inference methods formalise these aspects. We will learn some of these methods here.

4.1 Lecture 21: Foundations of statistical inference

Statistical analysis (or inference) involves drawing conclusions, and making predictions and decisions, using the evidence provided to us by observed data. To do this we use probability distributions, often called *statistical models*, to describe the process by which the observed data were generated. For example, we may suppose that the true proportion of Indian origin students is p, $0 < p < 1$, and if we have selected n students at random, that each of those students gives rise to a Bernoulli distribution which takes the value 1 if the student is of Indian origin and 0 otherwise. The success probability of the Bernoulli distribution will be the unknown p. The underlying statistical model is then the Bernoulli distribution.

To illustrate with another example, suppose we have observed fast food waiting times in the morning and afternoon. If we assume time (number of whole seconds) to be discrete, then a suitable model for the random variable X = "the number of seconds waited" would be the Poisson distribution. However, if we treat time as continuous then the random variable X = "the waiting time" could be modelled as a normal random variable. Now, in general, it is clear that:

- The form of the assumed model helps us to understand the real-world process by which the data were generated.
- If the model explains the observed data well, then it should also inform us about future (or unobserved) data, and hence help us to make predictions (and decisions contingent on unobserved data).
- The use of statistical models, together with a carefully constructed methodology for their analysis, also allows us to quantify the uncertainty associated with any conclusions, predictions or decisions we make.

As we have noted in Lecture 3, we will use the notation $x_1, x_2, \ldots, x_n$ to denote n observations of the random variables $X_1, X_2, \ldots, X_n$ (corresponding capital letters). For the fast food waiting time example, we have $n = 20$, $x_1 = 38, x_2 = 100, \ldots, x_{20} = 70$, and X_i is the waiting time for the ith person in the sample.

4.1.1 Statistical models

Suppose we denote the complete data by the vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ and use $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ for the corresponding random variables. A statistical model specifies a probability distribution for the random variables $\boldsymbol{X}$ corresponding to the data observations $\boldsymbol{x}$. Providing a specification for the distribution of n jointly varying random variables can be a daunting task, particularly if n is large. However, this task is made much easier if we can make some *simplifying assumptions*, such as

1. $X_1, X_2, \ldots, X_n$ are *independent* random variables,
2. $X_1, X_2, \ldots, X_n$ have the same probability distribution (so $x_1, x_2, \ldots, x_n$ are observations of a single random variable X).

Assumption 1 depends on the sampling mechanism and is very common in practice. If we are to make this assumption for the Southampton student sampling experiment, we need to select randomly among all possible students. We should not get the sample from an event in the Indian or Chinese Student Association as that will give us a biased result. The assumption will be violated when samples are correlated either in time or in space, e.g. the daily air pollution level in Southampton for the last year or the air pollution levels in two nearby locations in Southampton. In this module we will only consider data sets where Assumption 1 is valid. Assumption 2 is not always appropriate, but is often reasonable when we are modelling a single variable. In the fast food waiting time example, if we assume that there are no differences between the AM and PM waiting times, then we can say that $X_1, \ldots, X_{20}$ are *independent and identically distributed* (or i.i.d. for short).

4.1.2 A fully specified model

Sometimes a model completely specifies the probability distribution of $X_1, X_2, \ldots, X_n$. For example, if we assume that the waiting time $X \sim N(\mu, \sigma^2)$ where $\mu = 100$, and $\sigma^2 = 100$, then this is a fully specified model. In this case, there is no need to collect any data as there is no need

to make any inference about any unknown quantities, although we may use the data to judge the plausibility of the model.

However, a fully specified model would be appropriate when for example, there is some external (to the data) theory as to why the model (in particular the values of μ and σ^2) was appropriate. Fully specified models such as this are uncommon as we rarely have external theory which allows us to specify a model so precisely.

4.1.3 A parametric statistical model

A parametric statistical model specifies a probability distribution for a random sample apart from the value of a number of parameters in that distribution. This could be confusing in the first instance - a parametric model does not specify parameters! Here the word parametric signifies the fact that the probability distribution is completely specified by a few parameters in the first place. For example, the Poisson distribution is parameterised by the parameter λ which happens to be the mean of the distribution; the normal distribution is parameterised by two parameters, the mean μ and the variance σ^2.

When a parametric statistical model is assumed with some unknown parameters, statistical inference methods use data to *estimate* the unknown parameters, e.g. λ, μ, σ^2. Estimation will be discussed in more detail in the following lectures.

4.1.4 A nonparametric statistical model

Sometimes it is not appropriate, or we want to avoid, making a precise specification for the distribution which generated $X_1, X_2, \ldots, X_n$. For example, when the data histogram does not show a bell-shaped distribution, it would be wrong to assume a normal distribution for the data. In such a case, although we can attempt to use some other non-bell-shaped parametric model, we can decide altogether to abandon parametric models. We may then still assume that $X_1, X_2, \ldots, X_n$ are i.i.d. random variables, but from a nonparametric statistical model which cannot be written down, having a probability function which only depends on a *finite* number of parameters. Such analysis approaches are also called distribution-free methods.

♡ **Example 57 Return to the computer failure example**

Let X denote the count of computer failures per week. We want to estimate how often will the computer system fail at least once per week in the next year? The answer is $52 \times (1 - P(X = 0))$. But how would you estimate $P(X = 0)$? Consider two approaches.

1. **Nonparametric**. Estimate $P(X = 0)$ by the relative frequency of number of zeros in the above sample, which is 12 out of 104. Thus our estimate of $P(X = 0)$ is $12/104$. Hence, our estimate of the number of weeks when there will be at least one computer failure is $52 \times (1 - 12/104) = 46$.

2. **Parametric**. Suppose we assume that X follows the Poisson distribution with parameter λ.

Then the answer to the above question is

$$\begin{aligned} 52 \times (1 - P(X = 0)) &= 52 \times \left(1 - e^{-\lambda} \frac{\lambda^0}{0!}\right) \\ &= 52 \times \left(1 - e^{-\lambda}\right) \end{aligned}$$

which involves the unknown parameter λ. For the Poisson distribution we know that $E(X) = \lambda$. Hence we could use the *sample mean* $\bar{X}$ to estimate $E(X) = \lambda$. Thus our estimate $\hat{\lambda} = \bar{x} = 3.75$. This type of estimator is called a *moment estimator*. Now our answer is $52 \times \left(1 - e^{-3.75}\right) =$`52 * (1- exp(-3.75))` $= 50.78 \approx 51$, which is very different compared to our answer of 46 from the nonparametric approach.

4.1.5 Should we prefer parametric or nonparametric and why?

The parametric approach should be preferred if the assumption of the Poisson distribution can be justified for the data. For example, we can look at the data histogram or compare the fitted probabilities of different values of X, i.e. $\hat{P}(X = x) = e^{-\hat{\lambda}} \frac{\hat{\lambda}^x}{x!}$, with the relative frequencies from the sample. In general, often model-based analysis is preferred because it is more precise and accurate, and we can find estimates of uncertainty in such analysis based on the structure of the model. We shall see this later.

The nonparametric approach should be preferred if the model cannot be justified for the data, as in this case the parametric approach will provide incorrect answers.

4.1.6 Take home points

We have discussed the foundations of statistical inference through many examples. We have also discussed two broad approaches for making statistical inference: parametric and nonparametric.

4.2 Lecture 22: Estimation

4.2.1 Lecture mission

Once we have collected data and proposed a statistical model for our data, the initial statistical analysis usually involves *estimation*.

- For a parametric model, we need to estimate the unknown (unspecified) parameter λ. For example, if our model for the computer failure data is that they are i.i.d. Poisson, we need to estimate the mean (λ) of the Poisson distribution.
- For a nonparametric model, we may want to estimate the properties of the data-generating distribution. For example, if our model for the computer failure data is that they are i.i.d., following the distribution of an unspecified common random variable X, then we may want to estimate $\mu = E(X)$ or $\sigma^2 = \text{Var}(X)$.

In the following, we use the generic notation θ to denote the *estimand* (what we want to estimate or the parameter). For example, θ is the parameter λ in the first example, and θ may be either μ or σ^2 or both in the second example.

4.2.2 Population and sample

Recall that a statistical model specifies a probability distribution for the random variables $\boldsymbol{X}$ corresponding to the data observations $\boldsymbol{x}$.

- The observations $\boldsymbol{x} = (x_1, \ldots, x_n)$ are called the *sample*, and quantities derived from the sample are sample quantities. For example, as in Chapter 1, we call

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

  the *sample mean*.

- The probability distribution for $\boldsymbol{X}$ specified in our model represents all possible observations which might have been observed in our sample, and is therefore sometimes referred to as the *population*. Quantities derived from this distribution are population quantities.

  For example, if our model is that $X_1, \ldots, X_n$ are i.i.d., following the common distribution of a random variable X, then we call $E(X)$ the *population mean*.

4.2.3 Statistic and estimator

A *statistic* $T(\boldsymbol{x})$ is any function of the observed data $x_1, \ldots, x_n$ alone (and therefore does not depend on any parameters or other unknowns).

An *estimate* of θ is any statistic which is used to estimate θ under a particular statistical model. We will use $\tilde{\theta}(\boldsymbol{x})$ (sometimes shortened to $\tilde{\theta}$) to denote an estimate of θ.

An estimate $\tilde{\theta}(\boldsymbol{x})$ is an observation of a corresponding random variable $\tilde{\theta}(\boldsymbol{X})$ which is called an *estimator*. Thus an estimate is a particular observed value, e.g. 1.2, but an estimator is a random variable which can take values which are called estimates.

An estimate is a particular numerical value, e.g. $\bar{x}$*; an estimator is a random variable, e.g.* $\bar{X}$*.*

The probability distribution of any estimator $\tilde{\theta}(\boldsymbol{X})$ is called its *sampling distribution*. The estimate $\tilde{\theta}(\boldsymbol{x})$ is an observed value (a number), and is a single observation from the sampling distribution of $\tilde{\theta}(\boldsymbol{X})$.

♡ **Example 58** Suppose that we have a random sample $X_1, \ldots, X_n$ from the uniform distribution on the interval $[0, \theta]$ where $\theta > 0$ is unknown. Suppose that $n = 5$ and we have the sample observations $x_1 = 2.3, x_2 = 3.6, x_3 = 20.2$, $x_4 = 0.9, x_5 = 17.2$. Our objective is to estimate θ. How can we proceed?

Here the pdf $f(x) = \frac{1}{\theta}$ for $0 \leq x \leq \theta$ and 0 otherwise. Hence $E(X) = \int_0^\theta \frac{1}{\theta} x dx = \frac{\theta}{2}$. There are many possible estimators for θ, e.g. $\hat{\theta}_1(\boldsymbol{X}) = 2\,\bar{X}$, which is motivated by the method of moments because $\theta = 2E(X)$. A second estimator is $\hat{\theta}_2(\boldsymbol{X}) = \max\{X_1, X_2, \ldots, X_n\}$, which is intuitive since

θ must be greater than or equal to all observed values and thus the maximum of the sample value will be closest to θ. This is also the maximum likelihood estimate of θ, which you will learn in MATH3044.

How could we choose between the two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$? This is where we need to learn the sampling distribution of an estimator to determine which estimator will be unbiased, i.e. correct on average, and which will have minimum variability. We will formally define these in a minute, but first let us derive the sampling distribution, i.e. the pdf, of $\hat{\theta}_2$. Note that $\hat{\theta}_2$ is a random variable since the sample $X_1, \ldots, X_n$ is random. We will first find its cdf and then differentiate the cdf to get the pdf. For ease of notation, suppose $Y = \hat{\theta}_2(\boldsymbol{X}) = \max\{X_1, X_2, \ldots, X_n\}$. For any $0 < y < \theta$, the cdf of Y, $F(y)$ is given by:

$$\begin{aligned} P(Y \leq y) &= P\left(\max\{X_1, X_2, \ldots, X_n\} \leq y\right) \\ &= P\left(X_1 \leq y, X_2 \leq y, \ldots, X_n \leq y\right)) \quad [\max \leq y \text{ if and only if each } \leq y] \\ &= P\left(X_1 \leq y\right) P(X_2 \leq y) \cdots P(X_n \leq y) \quad [\text{since the } X\text{'s are independent}] \\ &= \tfrac{y}{\theta}\,\tfrac{y}{\theta} \cdots \tfrac{y}{\theta} \\ &= \left(\tfrac{y}{\theta}\right)^n. \end{aligned}$$

Now the pdf of Y is $f(y) = \frac{dF(y)}{dy} = n\frac{y^{n-1}}{\theta^n}$ for $0 \leq y \leq \theta$. We can plot this as a function of y to see the pdf. Now $E(\hat{\theta}_2) = E(Y) = \frac{n}{n+1}\theta$ and $\text{Var}(\hat{\theta}_2) = \frac{n\theta^2}{(n+2)(n+1)^2}$. You can prove this by easy integration.

4.2.4 Bias and mean square error

In the uniform distribution example we saw that the estimator $\hat{\theta}_2 = Y = \max\{X_1, X_2, \ldots, X_n\}$ is a random variable and its pdf is given by $f(y) = n\frac{y^{n-1}}{\theta^n}$ for $0 \leq y \leq \theta$. This probability distribution is called the sampling distribution of $\hat{\theta}_2$. For this we have seen that $E(\hat{\theta}_2) = \frac{n}{n+1}\theta$.

In general, we define the *bias* of an estimator $\tilde{\theta}(\boldsymbol{X})$ of θ to be

$$\text{bias}(\tilde{\theta}) = E(\tilde{\theta}) - \theta.$$

An estimator $\tilde{\theta}(\boldsymbol{X})$ is said to be *unbiased* if

$$\text{bias}(\tilde{\theta}) = 0, \quad \text{i.e. if} \quad E(\tilde{\theta}) = \theta.$$

So an estimator is unbiased if the expectation of its sampling distribution is equal to the quantity we are trying to estimate. Unbiased means "getting it right on average", i.e. under repeated sampling (relative frequency interpretation of probability).

Thus for the uniform distribution example, $\hat{\theta}_2$ is a biased estimator of θ and

$$\text{bias}(\hat{\theta}_2) = E(\hat{\theta}_2) - \theta = \frac{n}{n+1}\theta - \theta = -\frac{1}{n+1}\theta,$$

which goes to zero as $n \to \infty$. However, $\hat{\theta}_1 = 2\bar{X}$ is unbiased since $E(\hat{\theta}_1) = 2E(\bar{X}) = 2\frac{\theta}{2} = \theta$.

Unbiased estimators are "correct on average", but that does not mean that they are guaranteed to provide estimates which are close to the estimand θ. A better measure of the quality of an estimator than bias is the *mean squared error* (or m.s.e.), defined as

$$\text{m.s.e.}(\tilde{\theta}) = E\left[(\tilde{\theta} - \theta)^2\right].$$

Therefore, if $\tilde{\theta}$ is unbiased for θ, i.e. if $E(\tilde{\theta}) = \theta$, then $\text{m.s.e.}(\tilde{\theta}) = \text{Var}(\tilde{\theta})$. In general, we have the following result:

$$\text{m.s.e.}(\tilde{\theta}) = \text{Var}(\tilde{\theta}) + \text{bias}(\tilde{\theta})^2.$$

The proof is similar to the one we did in Lecture 3.

$$\begin{aligned}
\text{m.s.e.}(\tilde{\theta}) &= E\left[(\tilde{\theta} - \theta)^2\right] \\
&= E\left[\left(\tilde{\theta} - E\left(\tilde{\theta}\right) + E\left(\tilde{\theta}\right) - \theta\right)^2\right] \\
&= E\left[\left(\tilde{\theta} - E\left(\tilde{\theta}\right)\right)^2 + \left(E\left(\tilde{\theta}\right) - \theta\right)^2 + 2\left(\tilde{\theta} - E\left(\tilde{\theta}\right)\right)\left(E\left(\tilde{\theta}\right) - \theta\right)\right] \\
&= E\left[\tilde{\theta} - E\left(\tilde{\theta}\right)\right]^2 + E\left[E\left(\tilde{\theta}\right) - \theta\right]^2 + 2E\left[\left(\tilde{\theta} - E\left(\tilde{\theta}\right)\right)\left(E\left(\tilde{\theta}\right) - \theta\right)\right] \\
&= \text{Var}\left(\tilde{\theta}\right) + \left[E\left(\tilde{\theta}\right) - \theta\right]^2 + 2\left(E\left(\tilde{\theta}\right) - \theta\right)E\left[\left(\tilde{\theta} - E\left(\tilde{\theta}\right)\right)\right] \\
&= \text{Var}\left(\tilde{\theta}\right) + \text{bias}(\tilde{\theta})^2 + 2\left(E\left(\tilde{\theta}\right) - \theta\right)\left[E\left(\tilde{\theta}\right) - E\left(\tilde{\theta}\right)\right] \\
&= \text{Var}\left(\tilde{\theta}\right) + \text{bias}(\tilde{\theta})^2.
\end{aligned}$$

Hence, the mean squared error incorporates both the bias and the variability (sampling variance) of $\tilde{\theta}$. We are then faced with the bias-variance trade-off when selecting an optimal estimator. We may allow the estimator to have a little bit of bias if we can ensure that the variance of the biased estimator will be much smaller than that of any unbiased estimator.

♡ **Example 59 Uniform distribution** Continuing with the uniform distribution $U[0, \theta]$ example, we have seen that $\hat{\theta}_1 = 2\bar{X}$ is unbiased for θ but $\text{bias}(\hat{\theta}_2) = -\frac{1}{n+1}\theta$. How do these estimators compare with respect to the m.s.e? Since $\hat{\theta}_1$ is unbiased, its m.s.e is its variance. In the next lecture, we will prove that for random sampling from any population

$$\text{Var}(\bar{X}) = \frac{\text{Var}(X)}{n},$$

where $\text{Var}(X)$ is the variance of the population sampled from. Returning to our example, we know that if $X \sim U[0, \theta]$ then $\text{Var}(X) = \frac{\theta^2}{12}$. Therefore we have:

$$\text{m.s.e.}(\hat{\theta}_1) = \text{Var}\left(\hat{\theta}_1\right) = \text{Var}\left(2\bar{X}\right) = 4\text{Var}\left(\bar{X}\right) = 4\frac{\theta^2}{12n} = \frac{\theta^2}{3n}.$$

Now, for $\hat{\theta}_2$ we know that:

1. $\text{Var}(\hat{\theta}_2) = \frac{n\theta^2}{(n+2)(n+1)^2}$;

2. $\text{bias}(\hat{\theta}_2) = -\frac{1}{n+1}\theta$.

Now

$$\begin{aligned} \text{m.s.e.}(\hat{\theta}_2) &= \text{Var}\left(\hat{\theta}_2\right) + \text{bias}(\hat{\theta}_2)^2 \\ &= \frac{n\theta^2}{(n+2)(n+1)^2} + \frac{\theta^2}{(n+1)^2} \\ &= \frac{\theta^2}{(n+1)^2}\left(\frac{n}{n+2} + 1\right) \\ &= \frac{\theta^2}{(n+1)^2}\frac{2n+2}{n+2}. \end{aligned}$$

Clearly, the m.s.e of $\hat{\theta}_2$ is an order of magnitude (of order n^2 rather than n) smaller than the m.s.e of $\hat{\theta}_1$, providing justification for the preference of $\hat{\theta}_2 = \max\{X_1, X_2, \ldots, X_n\}$ as an estimator of θ.

4.2.5 Take home points

In this lecture we have learned the basics of estimation. We have learned that estimates are particular values and estimators have probability distributions. We have also learned the concepts of the bias and variance of an estimator. We have proved a key fact that the mean squared error of an estimator is composed of two pieces, namely bias and variance. Sometimes there may be bias-variance trade-off where a little bias can lead to much lower variance. We have illustrated this with an example.

4.3 Lecture 23: Estimation of mean and variance and standard error

4.3.1 Lecture mission

Often, one of the main tasks of a statistician is to estimate a population average or mean. However the estimates, using whatever procedure, will not be usable or scientifically meaningful if we do not know their associated uncertainties. For example, a statement such as: "the Arctic ocean will be completely ice-free in the summer in the next few decades" provides little information as it does not communicate the extent or the nature of the uncertainty in it. Perhaps a more precise statement could be: "the Arctic ocean will be completely ice-free in the summer some time in the next 20-30 years". This last statement not only gives a numerical value for the number of years for complete ice-melt in the summer, but also acknowledges the uncertainty of ± 5 years in the estimate. A statistician's main job is to estimate such uncertainties. In this lecture, we will get started with estimating uncertainties when we estimate a population mean. We will introduce the standard error of an estimator.

4.3.2 Estimation of a population mean

Suppose that $x_1, \ldots, x_n$ is a random sample from any probability distribution $f(x)$, which may be discrete or continuous. Suppose that we want to estimate the unknown population mean $E(X) = \mu$ and variance, $\text{Var}(X) = \sigma^2$. In order to do this, it is not necessary to make any assumptions about $f(x)$, so this may be thought of as *nonparametric* inference.

We have the following results:

R1 the sample mean

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is an unbiased estimator of $\mu = E(X)$, i.e. $E(\bar{X}) = \mu$,

R2 $\mathrm{Var}(\bar{X}) = \sigma^2/n$,

R3 the sample variance with divisor $n-1$

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

is an unbiased estimator of σ^2, i.e. $E(S^2) = \sigma^2$.

We prove **R1** as follows.

$$E[\bar{X}] = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} E(X) = E(X),$$

so $\bar{X}$ is an unbiased estimator of $E(X)$.

We prove **R2** using the result that for independent random variables the variance of the sum is the sum of the variances from Lecture 19. Thus,

$$\mathrm{Var}[\bar{X}] = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X) = \frac{n}{n^2}\mathrm{Var}(X) = \frac{\sigma^2}{n},$$

so the m.s.e. of $\bar{X}$ is $\mathrm{Var}(X)/n$. This proves the following assertion we made earlier:

Variance of the sample mean = Population Variance divided by the sample size.

We now want to prove **R3**, i.e. show that the sample variance with divisor $n-1$ is an unbiased estimator of the population variance σ^2, i.e. $E(S^2) = \sigma^2$. We have

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

To evaluate the expectation of the above, we need $E(X_i^2)$ and $E(\bar{X}^2)$. In general, we know for any random variable,

$$\mathrm{Var}(Y) = E(Y^2) - (E(Y))^2 \quad \Rightarrow E(Y^2) = \mathrm{Var}(Y) + (E(Y))^2.$$

Thus, we have

$$E(X_i^2) = \mathrm{Var}(X_i) + (E(X_i))^2 = \sigma^2 + \mu^2,$$

and

$$E(\bar{X}^2) = \mathrm{Var}(\bar{X}) + (E(\bar{X}))^2 = \sigma^2/n + \mu^2,$$

from R1 and R2. Now

$$\begin{aligned} E(S^2) &= E\left\{\tfrac{1}{n-1}\left[\textstyle\sum_{i=1}^n X_i^2 - n\bar{X}^2\right]\right\} \\ &= \tfrac{1}{n-1}\left[\textstyle\sum_{i=1}^n E(X_i^2) - nE(\bar{X}^2)\right] \\ &= \tfrac{1}{n-1}\left[\textstyle\sum_{i=1}^n (\sigma^2+\mu^2) - n(\sigma^2/n+\mu^2)\right] \\ &= \tfrac{1}{n-1}\left[n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2)\right] \\ &= \sigma^2 \equiv \mathrm{Var}(X). \end{aligned}$$

In words, this proves that

The sample variance is an unbiased estimator of the population variance.

4.3.3 Standard deviation and standard error

It follows that, for an unbiased (or close to unbiased) estimator $\tilde{\theta}$,

$$\text{m.s.e.}(\tilde{\theta}) = \mathrm{Var}(\tilde{\theta})$$

and therefore the sampling variance of the estimator is an important summary of its quality.

We usually prefer to focus on the standard deviation of the sampling distribution of $\tilde{\theta}$,

$$\text{s.d.}(\tilde{\theta}) = \sqrt{\mathrm{Var}(\tilde{\theta})}.$$

In practice we will not know s.d.$(\tilde{\theta})$, as it will typically depend on unknown features of the distribution of $X_1, \ldots, X_n$. However, we may be able to estimate *s.d.*$(\tilde{\theta})$ using the observed sample $x_1, \ldots, x_n$. We define the *standard error*, s.e.$(\tilde{\theta})$, of an estimator $\tilde{\theta}$ to be *an estimate of the standard deviation of its sampling distribution, s.d.*$(\tilde{\theta})$.

Standard error of an estimator is an **estimate** *of the standard deviation of its sampling distribution*

We proved that

$$\mathrm{Var}[\bar{X}] = \frac{\sigma^2}{n} \quad \Rightarrow \quad \text{s.d.}(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

As σ is unknown, we cannot calculate this standard deviation. However, we know that $E(S^2) = \sigma^2$, i.e. that the sample variance is an unbiased estimator of the population variance. Hence S^2/n is an unbiased estimator for $\mathrm{Var}(\bar{X})$. Therefore we obtain the *standard error of the mean*, s.e.$(\bar{X})$, by plugging in the estimate

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

of σ into s.d.$(\bar{X})$ to obtain

$$\text{s.e.}(\bar{X}) = \frac{s}{\sqrt{n}}.$$

Therefore, for the computer failure data, our estimate, $\bar{x} = 3.75$, for the population mean is associated with a standard error

$$\text{s.e.}(\bar{X}) = \frac{3.381}{\sqrt{104}} = 0.332.$$

Note that this is 'a' standard error, so other standard errors may be available. Indeed, for parametric inference, where we make assumptions about $f(x)$, alternative standard errors are available. For example, $X_1, \ldots, X_n$ are i.i.d. Poisson(λ) random variables. $E(X) = \lambda$, so $\bar{X}$ is an unbiased estimator of λ. $\text{Var}(X) = \lambda$, so another $\text{s.e.}(\bar{X}) = \sqrt{\hat{\lambda}/n} = \sqrt{\bar{x}/n}$. In the computer failure data example, this is $\sqrt{\frac{3.75}{104}} = 0.19$.

4.3.4 Take home points

In this lecture we have defined the standard error of an estimator. This is very important in practice, as the standard error tells us how precise our estimate is through how concentrated the sampling distribution of the estimator is. For example, in the age guessing example a standard error of 15 years indicates hugely inaccurate guesses. We have learned three key results: the sample mean is an unbiased estimate of the population mean; the variance of the sample mean is the population variance divided by the sample size; and the sample variance with divisor $n - 1$ is an unbiased estimator of the population variance.

4.4 Lecture 24: Interval estimation

4.4.1 Lecture mission

In any estimation problem it is very hard to guess the exact true value, but it is often much better (and easier?) to provide an interval where the true value is very likely to fall. For example, think of guessing the age of a stranger. In this lecture we will learn to use the results of the previous lecture to obtain confidence intervals for a mean parameter of interest. The methods are important to learn so that we can make probability statements about the random intervals as opposed to just pure guesses, e.g. estimating my age to be somewhere between 30 and 60. Statistical methods allow us to be much more precise by harnessing the power of the data.

4.4.2 Basics

An estimate $\tilde{\theta}$ of a parameter θ is sometimes referred to as a *point estimate*. The usefulness of a point estimate is enhanced if some kind of measure of its precision can also be provided. Usually, for an unbiased estimator, this will be a standard error, an estimate of the standard deviation of the associated estimator, as we have discussed previously. An alternative summary of the information provided by the observed data about the location of a parameter θ and the associated precision is an *interval estimate* or *confidence interval*.

Suppose that $x_1, \ldots, x_n$ are observations of random variables $X_1, \ldots, X_n$ whose joint pdf is specified apart from a single parameter θ. To construct a confidence interval for θ, we need to find

a random variable $T(\mathbf{X}, \theta)$ whose distribution does not depend on θ and is therefore known. *This random variable $T(\mathbf{X}, \theta)$ is called a pivot for θ.* Hence we can find numbers h_1 and h_2 such that

$$P(h_1 \leq T(\mathbf{X}, \theta) \leq h_2) = 1 - \alpha \quad (1),$$

where $1 - \alpha$ is any specified probability. If (1) can be 'inverted' (or manipulated), we can write it as

$$P[g_1(\mathbf{X}) \leq \theta \leq g_2(\mathbf{X})] = 1 - \alpha. \quad (2)$$

Hence with probability $1 - \alpha$, the parameter θ will lie between the random variables $g_1(\mathbf{X})$ and $g_2(\mathbf{X})$. Alternatively, the random interval $[g_1(\mathbf{X}), g_2(\mathbf{X})]$ includes θ with probability $1 - \alpha$. Now, when we observe $x_1, \ldots, x_n$, we observe a single observation of the random interval $[g_1(\mathbf{X}), g_2(\mathbf{X})]$, which can be evaluated as $[g_1(\mathbf{x}), g_2(\mathbf{x})]$. We do not know if θ lies inside or outside this interval, but we do know that if we observed repeated samples, then $100(1 - \alpha)\%$ of the resulting intervals would contain θ. Hence, if $1 - \alpha$ is high, we can be reasonably confident that our observed interval contains θ. We call the observed interval $[g_1(\mathbf{x}), g_2(\mathbf{x})]$ a $100(1 - \alpha)\%$ confidence interval for θ. It is common to present intervals with high confidence levels, usually 90%, 95% or 99%, so that $\alpha = 0.1, 0.05$ or 0.01 respectively.

4.4.3 Confidence interval for a normal mean

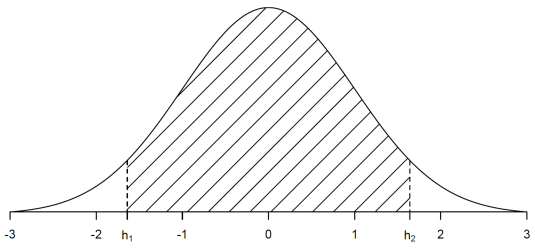
Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$ random variables. We know that from CLT Lecture 16

$$\bar{X} \sim N(\mu, \sigma^2/n) \quad \Rightarrow \quad \sqrt{n}\frac{(\bar{X} - \mu)}{\sigma} \sim N(0, 1).$$

Suppose we know that $\sigma = 10$, so $\sqrt{n}(\bar{X} - \mu)/\sigma$ is a pivot for μ. Then we can use the distribution function of the standard normal distribution to find values h_1 and h_2 such that

$$P\left(h_1 \leq \sqrt{n}\frac{(\bar{X} - \mu)}{\sigma} \leq h_2\right) = 1 - \alpha$$

for a chosen value of $1 - \alpha$ which is called the *confidence level*. So h_1 and h_2 are chosen so that the shaded area in the figure is equal to the confidence level $1 - \alpha$.



It is common practice to make the interval symmetric, so that the two unshaded areas are equal (to $\alpha/2$), in which case

$$-h_1 = h_2 \equiv h \qquad \text{and} \qquad \Phi(h) = 1 - \frac{\alpha}{2}.$$

The most common choice of confidence level is $1 - \alpha = 0.95$, in which case $h = 1.96 =$ `qnorm(0.975)`. You may also occasionally see 90% ($h = 1.645 =$ `qnorm(0.95)`) or 99% ($h =$

2.58=`qnorm(0.995)`) intervals. We discussed these values in Lecture 17. We generally use the 95% intervals for a reasonably high level of confidence without making the interval unnecessarily wide.

Therefore we have

$$P\left(-1.96 \leq \sqrt{n}\frac{(\bar{X}-\mu)}{\sigma} \leq 1.96\right) = 0.95$$
$$\Rightarrow \quad P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95.$$

Hence, $\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}$ and $\bar{X} + 1.96\frac{\sigma}{\sqrt{n}}$ are the endpoints of a random interval which includes μ with probability 0.95. The observed value of this interval, $\left(\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}\right)$, is called a *95% confidence interval* for μ.

♡ **Example 60** For the fast food waiting time data, we have $n = 20$ data points combined from the morning and afternoon data sets. We have $\bar{x} = 67.85$ and $n = 20$. Hence, under the normal model assuming (just for the sake of illustration) $\sigma = 18$, a 95% confidence interval for μ is

$$\begin{aligned} 67.85 - 1.96(18/\sqrt{20}) \leq \quad & \mu \quad \leq 67.85 + 1.96(18/\sqrt{20}) \\ \Rightarrow \quad 59.96 \leq \quad & \mu \quad \leq 75.74 \end{aligned}$$

The `R` command is `mean(a) + c(-1, 1) * qnorm(0.975) * 18/sqrt(20)`, assuming `a` is the vector containing 20 waiting times. If σ is unknown, we need to seek alternative methods for finding the confidence intervals.

Some important remarks about confidence intervals.

1. Notice that $\bar{x}$ is an unbiased estimate of μ, $\sigma/\sqrt{n}$ is the standard error of the estimate and 1.96 (in general h in the above discussion) is a critical value from the associated known sampling distribution. The formula $(\bar{x} \pm 1.96\,\sigma/\sqrt{n})$ for the confidence interval is then generalised as:

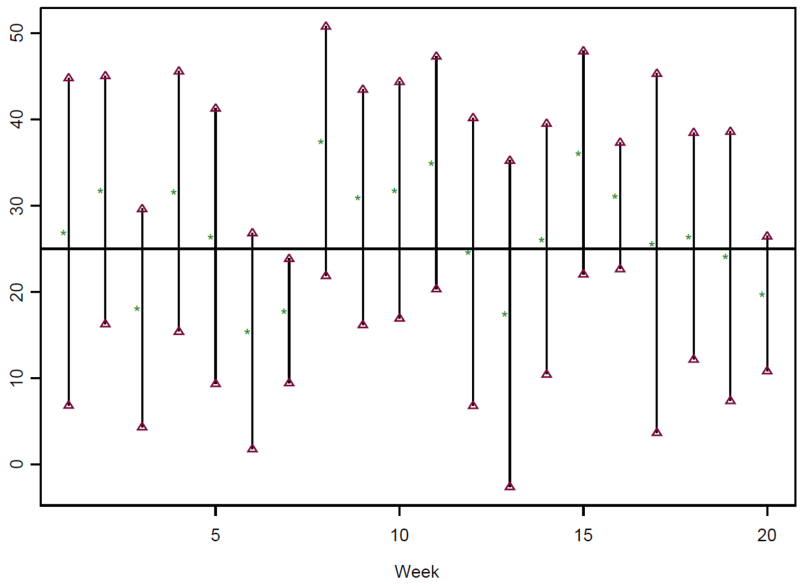
Estimate ± Critical value × Standard error

   where the estimate is $\bar{x}$, the critical value is 1.96 and the standard error is $\sigma/\sqrt{n}$. This is so much easier to remember. We will see that this formula holds in many of the following examples, but not all.

2. Confidence intervals are frequently used, but also frequently misinterpreted. A $100(1-\alpha)\%$ confidence interval for θ is a single observation of a random interval which, under repeated sampling, would include θ $100(1-\alpha)\%$ of the time.

   The following example from the National Lottery in the UK clarifies the interpretation. We collected 6 chosen lottery numbers (sampled at random from 1 to 49) for 20 weeks and then constructed 95% confidence intervals for the population mean $\mu = 25$ and plotted the intervals along with the observed sample means in the following figure. It can be seen that exactly

one out of 20 (5%) of the intervals do not contain the true population mean 25. Although this is a coincidence, it explains the main point that if we construct the random intervals with $100(1-\alpha)\%$ confidence levels again and again for hypothetical repetition of the data, on average $100(1-\alpha)\%$ of them will contain the true parameter.



3. A confidence interval is not a probability interval. You should avoid making statements like $P(1.3 < \theta < 2.2) = 0.95$. In the classical approach to statistics you can only make probability statements about random variables, and θ is assumed to be a constant.

4. If a confidence interval is interpreted as a probability interval, this may lead to problems. For example, suppose that X_1 and X_2 are i.i.d. $U[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$ random variables. Then $P[\min(X_1, X_2) < \theta < \max(X_1, X_2)] = \frac{1}{2}$ so $[\min(x_1, x_2), \max(x_1, x_2)]$ is a 50% confidence interval for θ, where x_1 and x_2 are the observed values of X_1 and X_2. Now suppose that $x_1 = 0.3$ and $x_2 = 0.9$. What is $P(0.3 < \theta < 0.9)$?

4.4.4 Take home points

In this lecture we have learned to obtain confidence intervals by using an appropriate statistic in the pivoting technique. The main task is then to invert the inequality so that the unknown parameter is in the middle by itself and the two end points are functions of the sample observations. The most difficult task is to correctly interpret confidence intervals, which are not probability intervals but have long-run properties. That is, the interval will contain the true parameter with the stipulated confidence level only under infinitely repeated sampling.

4.5 Lecture 25: Confidence intervals using the CLT

4.5.1 Lecture mission

Confidence intervals are generally difficult to find. The difficulty lies in finding a pivot, i.e. a statistic $T(\mathbf{X}, \theta)$ such that

$$P(h_1 \leq T(\mathbf{X}, \theta) \leq h_2) = 1 - \alpha$$

for two suitable numbers h_1 and h_2, and also that the above can be inverted to put the unknown θ in the middle of the inequality inside the probability statement. One solution to this problem is to use the powerful Central Limit Theorem (CLT) to claim normality, and then basically follow the above normal example for known variance.

4.5.2 Confidence intervals for μ using the CLT

The CLT allows us to assume the large sample approximation

$$\sqrt{n}\frac{(\bar{X} - \mu)}{\sigma} \overset{approx}{\sim} N(0,1) \text{ as } n \to \infty.$$

So a general confidence interval for μ can be constructed, just as before in Section 4.4.3. Thus a 95% confidence interval (CI) for μ is given by $\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$. But note that σ is unknown so this CI cannot be used unless we can estimate σ, i.e. replace the unknown s.d. of $\bar{X}$ by its estimated standard error. In this case, we get the CI in the familiar form:

Estimate ± Critical value × Standard error

Suppose that we do not assume any distribution for the sampled random variable X but assume only that $X_1, \ldots, X_n$ are i.i.d, following the distribution of X where $E(X) = \mu$ and $\text{Var}(X) = \sigma^2$. We know that the standard error of $\bar{X}$ is $s/\sqrt{n}$ where s is the sample standard deviation with divisor $n - 1$. Then the following provides a 95% CI for μ:

$$\bar{x} \pm 1.96\frac{s}{\sqrt{n}}.$$

♡ **Example 61** For the computer failure data, $\bar{x} = 3.75$, $s = 3.381$ and $n = 104$. Under the model that the data are observations of i.i.d. random variables with population mean μ (but no other assumptions about the underlying distribution), we compute a 95% confidence interval for μ to be

$$\left(3.75 - 1.96\frac{3.381}{\sqrt{104}}, 3.75 + 1.96\frac{3.381}{\sqrt{104}}\right) = (3.10, 4.40).$$

If we can assume a distribution for X, i.e. a parametric model for X, then we can do slightly better in estimating the standard error of $\bar{X}$ and as a result we can improve upon the previously obtained 95% CI. Two examples follow.

♡ **Example 62 Poisson** If $X_1, \ldots, X_n$ are modelled as i.i.d. Poisson(λ) random variables, then $\mu = \lambda$ and $\sigma^2 = \lambda$. We know $\text{Var}(\bar{X}) = \sigma^2/n = \lambda/n$. Hence a standard error is $\sqrt{\hat{\lambda}/n} = \sqrt{\bar{x}/n}$

since $\hat{\lambda} = \bar{X}$ is an unbiased estimator of λ. Thus a 95% CI for $\mu = \lambda$ is given by

$$\bar{x} \pm 1.96\sqrt{\frac{\bar{x}}{n}}.$$

For the computer failure data, $\bar{x} = 3.75$, $s = 3.381$ and $n = 104$. Under the model that the data are observations of i.i.d. random variables following a Poisson distribution with population mean λ, we compute a 95% confidence interval for λ as

$$\bar{x} \pm 1.96\sqrt{\frac{\bar{x}}{n}} = 3.75 \pm 1.96\sqrt{3.75/104} = (3.38, 4.12).$$

We see that this interval is narrower ($0.74 = 4.12 - 3.38$) than the earlier interval (3.10,4.40), which has a length of 1.3. We prefer narrower confidence intervals as they facilitate more accurate inference regarding the unknown parameter.

♡ **Example 63 Bernoulli** If $X_1, \ldots, X_n$ are modelled as i.i.d. Bernoulli(p) random variables, then $\mu = p$ and $\sigma^2 = p(1-p)$. We know $\text{Var}(\bar{X}) = \sigma^2/n = p(1-p)/n$. Hence a standard error is $\sqrt{\hat{p}(1-\hat{p})/n} = \sqrt{\bar{x}(1-\bar{x})/n}$, since $\hat{p} = \bar{X}$ is an unbiased estimator of p. Thus a 95% CI for $\mu = p$ is given by

$$\bar{x} \pm 1.96\sqrt{\frac{\bar{x}(1-\bar{x})}{n}}.$$

For the example, suppose $\bar{x} = 0.2$ and $n = 10$. Then we obtain the 95% CI as

$$0.2 \pm 1.96\sqrt{(0.2 \times 0.8)/10} = (-0.048, 0.448).$$

This is **wrong** as n is too small for the large sample approximation to be accurate. Hence we need to look for other alternatives which may work better.

4.5.3 Confidence interval for a Bernoulli p by quadratic inversion

It turns out that for the Bernoulli and Poisson distributions we can find alternative confidence intervals *without using the approximation for standard error* but still using the CLT. This is more complicated and requires us to solve a quadratic equation. We consider the two distributions separately.

We start with the CLT and obtain the following statement:

$$\begin{aligned}
& P\left(-1.96 \le \sqrt{n}\tfrac{(\bar{X}-p)}{\sqrt{p(1-p)}} \le 1.96\right) = 0.95 \\
\Leftrightarrow \quad & P\left(-1.96\sqrt{p(1-p)} \le \sqrt{n}(\bar{X}-p) \le 1.96\sqrt{p(1-p)}\right) = 0.95 \\
\Leftrightarrow \quad & P\left(-1.96\sqrt{p(1-p)/n} \le (\bar{X}-p) \le 1.96\sqrt{p(1-p)/n}\right) = 0.95 \\
\Leftrightarrow \quad & P\left(p - 1.96\sqrt{p(1-p)/n} \le \bar{X} \le p + 1.96\sqrt{p(1-p)/n}\right) = 0.95 \\
\Leftrightarrow \quad & P\left(L(p) \le \bar{X} \le R(p)\right) = 0.95,
\end{aligned}$$

where $L(p) = p - h\sqrt{p(1-p)/n}, R(p) = p + h\sqrt{p(1-p)/n}, h = 1.96$. Now, consider the inverse mappings $L^{-1}(x)$ and $R^{-1}(x)$ so that:

$$\begin{aligned} & P\left[L(p) \leq \bar{X} \leq R(p)\right] = 0.95 \\ \Leftrightarrow\ & P\left[R^{-1}(\bar{X}) \leq p \leq L^{-1}(\bar{X})\right] = 0.95 \end{aligned}$$

which now defines our confidence interval $(R^{-1}(\bar{X}), L^{-1}(\bar{X}))$ for p. We can obtain $R^{-1}(\bar{x})$ and $L^{-1}(\bar{x})$ by solving the equations $R(p) = \bar{x}$ and $L(p) = \bar{x}$ for p, treating n and $\bar{x}$ as known quantities. Thus we have,

$$\begin{aligned} R(p) = \bar{x}, \quad & L(p) = \bar{x} \\ \Leftrightarrow \quad & (\bar{x} - p)^2 = h^2 p(1-p)/n, \qquad \text{where } h = 1.96 \\ \Leftrightarrow \quad & p^2(1 + h^2/n) - p(2\bar{x} + h^2/n) + \bar{x}^2 = 0 \end{aligned}$$

The endpoints of the confidence interval are the roots of the quadratic. Hence, the endpoints of the 95% confidence interval for p are:

$$\begin{aligned} & \frac{\left(2\bar{x} + \frac{h^2}{n}\right) \pm \left[\left(2\bar{x} + \frac{h^2}{n}\right)^2 - 4\bar{x}^2\left(1 + \frac{h^2}{n}\right)\right]^{1/2}}{2\left(1 + \frac{h^2}{n}\right)} \\ = & \frac{\left(\bar{x} + \frac{h^2}{2n}\right) \pm \left[\left(\bar{x} + \frac{h^2}{2n}\right)^2 - \bar{x}^2\left(1 + \frac{h^2}{n}\right)\right]^{1/2}}{\left(1 + \frac{h^2}{n}\right)} \\ = & \frac{\bar{x} + \frac{h^2}{2n} \pm \frac{h}{\sqrt{n}}\left[\frac{h^2}{4n} + \bar{x}(1-\bar{x})\right]^{1/2}}{\left(1 + \frac{h^2}{n}\right)}. \end{aligned}$$

This is sometimes called the *Wilson Score Interval*. The following `R` code calculates this for given n, $\bar{x}$ and confidence level α which determines the value of h. Returning to the previous example, $n = 10$ and $\bar{x} = 0.2$, the 95% CI obtained from this method is (0.057, 0.510) compared to the previous illegitimate one $(-0.048, 0.448)$. In fact you can see that the intervals obtained by quadratic inversion are more symmetric and narrower as n increases, and are also more symmetric for $\bar{x}$ closer to 0.5. See the table below:

n	$\bar{x}$	Quadratic inversion		Plug-in s.e. estimation	
		Lower end	Upper end	Lower end	Upper end
10	0.2	0.057	0.510	–0.048	0.448
10	0.5	0.237	0.763	0.190	0.810
20	0.1	0.028	0.301	–0.031	0.231
20	0.2	0.081	0.416	0.025	0.375
20	0.5	0.299	0.701	0.281	0.719
50	0.1	0.043	0.214	0.017	0.183
50	0.2	0.112	0.330	0.089	0.311
50	0.5	0.366	0.634	0.361	0.639

For smaller n and $\bar{x}$ closer to 0 (or 1), the approximation required for the plug-in estimate of the standard error is insufficiently reliable. However, for larger n it is adequate.

4.5.4 Confidence interval for a Poisson λ by quadratic inversion

Here we proceed as in the Bernoulli case and using the CLT claim that a 95% CI for λ is given by:

$$P\left(-1.96 \leq \sqrt{n}\frac{(\bar{X}-\lambda)}{\sqrt{\lambda}} \leq 1.96\right) = 0.95 \quad \Rightarrow P\left(n\frac{(\bar{X}-\lambda)^2}{\lambda} \leq 1.96^2\right) = 0.95.$$

Now the confidence interval for λ is found by solving the (quadratic) equality for λ by treating $n, \bar{x}$ and h to be known:

$$\begin{aligned} & n\frac{(\bar{x}-\lambda)^2}{\lambda} = h^2, \qquad \text{where } h = 1.96 \\ \Rightarrow \quad & \bar{x}^2 - 2\lambda\bar{x} + \lambda^2 = h^2\lambda/n \\ \Rightarrow \quad & \lambda^2 - \lambda(2\bar{x} + h^2/n) + \bar{x}^2 \ = 0. \end{aligned}$$

Hence, the endpoints of the 95% confidence interval for λ are:

$$\frac{\left(2\bar{x} + \frac{h^2}{n}\right) \pm \left[\left(2\bar{x} + \frac{h^2}{n}\right)^2 - 4\bar{x}^2\right]^{1/2}}{2} = \bar{x} + \frac{h^2}{2n} \pm \frac{h}{n^{1/2}}\left[\frac{h^2}{4n} + \bar{x}\right]^{1/2}.$$

♡ **Example 64** For the computer failure data, $\bar{x} = 3.75$ and $n = 104$. For a 95% confidence interval (CI), $h = 1.96$. Hence, we calculate the above CI using the `R` commands:

```
x <- scan("compfail.txt")
n <- length(x)
h <- qnorm(0.975)
mean(x) + (h*h)/(2*n) + c(-1, 1) * h/sqrt(n) * sqrt(h*h/(4*n) + mean(x))
```

The result is (3.40, 4.14), which compares well with the earlier interval (3.38, 4.12).

4.5.5 Take home points

In this lecture we learned how to find confidence intervals using the CLT, when the sample size is large. We have seen that we can make more accurate inferences if we can assume a model, e.g. the Poisson model for the computer failure data. However, we have also encountered problems when applying the method for a small sample size. In such cases we should use alternative methods for calculating confidence intervals. For example, we learned a technique of finding confidence intervals which does not require us to approximately estimate the standard errors for Bernoulli and Poisson distributions. In the next lecture, we will learn how to find an exact confidence interval for the normal mean μ using the t-distribution.

4.6 Lecture 26: Exact confidence interval for the normal mean

4.6.1 Lecture mission

Recall that we can obtain better quality inferences if we can justify a precise model for the data. This saying is analogous to the claim that a person can better predict and infer in a situation when there are established rules and regulations, i.e. the analogue of a statistical model. In this lecture, we will discuss a procedure for finding confidence intervals based on the statistical modelling assumption that the data are from a normal distribution. This assumption will enable us to find an exact confidence interval for the mean rather than an approximate one using the central limit theorem.

4.6.2 Obtaining an exact confidence interval for the normal mean

For normal models we do not have to rely on large sample approximations, because it turns out that the distribution of

$$T = \frac{\sqrt{n}(\bar{X} - \mu)}{S},$$

where S^2 is the sample variance with divisor $n-1$, is standard (easily calculated) and thus the statistic $T = T(\mathbf{X}, \mu)$ can be an exact pivot for any sample size $n > 1$. The point about easy calculation is that for any given $1-\alpha$, e.g. $1-\alpha = 0.95$, we can calculate the critical value h such that $P(-h < T < h) = 1-\alpha$. Note also that the pivot T does not involve the other unknown parameter of the normal model, namely the variance σ^2. If indeed, we can find h for any given $1-\alpha$, we can proceed as follows to find the exact CI for μ:

$$\begin{aligned} & P(-h \leq T \leq h) = 1-\alpha \\ \text{i.e.} \quad & P\left(-h \leq \sqrt{n}\frac{(\bar{X}-\mu)}{S} \leq h\right) = 0.95 \\ \Rightarrow \quad & P\left(\bar{X} - h\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + h\frac{S}{\sqrt{n}}\right) = 0.95 \end{aligned}$$

The observed value of this interval, $(\bar{x} \pm h\frac{s}{\sqrt{n}})$, is the 95% confidence interval for μ. Remarkably, this also of the general form, Estimate $\pm$ Critical value $\times$ Standard error, where the Critical value is h and the standard error of the sample mean is $\frac{s}{\sqrt{n}}$. Now, how do we find the critical value h for a given $1-\alpha$? We need to introduce the t-distribution.

Let $X_1, \ldots, X_n$ be i.i.d $N(\mu, \sigma^2)$ random variables. Define $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and

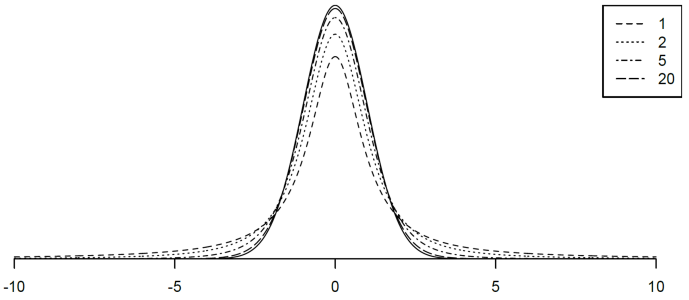
$$S^2 = \frac{1}{n-1}\left(\sum_{i=1}^n X_i^2 - n\bar{X}^2\right).$$

Then, it can be shown (and will be in MATH2011) that

$$\sqrt{n}\frac{(\bar{X}-\mu)}{S} \sim t_{n-1},$$

where t_{n-1} denotes the standard t *distribution* with $n-1$ *degrees of freedom*. The standard t distribution is a family of distributions which depend on one parameter called the degrees-of-freedom (df) which is $n-1$ here. The concept of degrees of freedom is that it is usually the number of independent random samples, n here, minus the number of linear parameters estimated, 1 here for μ. Hence the df is $n-1$.

The probability density function of the t_k distribution is similar to a standard normal, in that it is symmetric around zero and 'bell-shaped', but the t-distribution is more *heavy-tailed*, giving greater probability to observations further away from zero. The figure below illustrates the t_k density function for $k = 1, 2, 5, 20$ together with the standard normal pdf (solid line).



The values of h for a given $1-\alpha$ have been tabulated using the standard t-distribution and can be obtained using the `R` command `qt` (abbreviation for quantile of t). For example, if we want to find h for $1-\alpha = 0.95$ and $n = 20$ then we issue the command: `qt(0.975, df=19)` $= 2.093$. Note that it should be 0.975 so that we are splitting 0.05 probability between the two tails equally and the df should be $n-1 = 19$. Indeed, using the above command repeatedly, we obtain the following critical values for the 95% interval for different values of the sample size n.

n	2	5	10	15	20	30	50	100	∞
h	12.71	2.78	2.26	2.14	2.09	2.05	2.01	1.98	1.96

Note that the critical value approaches 1.96 (which is the critical value for the normal distribution) as $n \to \infty$, since the t-distribution itself approaches the normal distribution for large values of its df parameter.

> If you can justify that the underlying distribution is normal then you can use the t-distribution-based confidence interval.

♡ **Example 65 Fast food waiting time revisited** We would like to find a confidence interval for the true mean waiting time. If X denotes the waiting time in seconds, we have $n = 20$, $\bar{x} = 67.85$, $s = 18.36$. Hence, recalling that the critical value $h = 2.093$, from the command `qt(0.975, df=19)`, a 95% confidence interval for μ is

$$\begin{aligned} 67.85 - 2.093 \times 18.36/\sqrt{20} \leq \quad \mu \quad &\leq 67.85 + 2.093 \times 18.36/\sqrt{20} \\ \Rightarrow \quad 59.26 \leq \quad \mu \quad &\leq 76.44. \end{aligned}$$

In `R` we issue the commands:

```
ffood <- read.csv("servicetime.csv", head=T)
a <- c(ffood$AM, ffood$PM)
mean(a) + c(-1, 1) * qt(0.975, df=19) * sqrt(var(a))/sqrt(20)
```

does the job and it gives the result (59.25, 76.45).

If we want a 90% confidence interval then we issue the command:

```
mean(a) + c(-1, 1) * qt(0.95, df=19) * sqrt(var(a))/sqrt(20),
```

which give (60.75, 74.95).

If we want a 99% confidence interval then we issue the command:

```
mean(a) + c(-1, 1) * qt(0.995, df=19) * sqrt(var(a))/sqrt(20),
```

which gives (56.10, 79.60). We can see clearly that the interval is getting wider as the level of confidence is getting higher.

♡ **Example 66 Weight gain revisited** We would like to find a confidence interval for the true average weight gain (final weight – initial weight). Here $n = 68, \bar{x} = 0.8672$ and $s = 0.9653$. Hence, a 95% confidence interval for μ is

$$\begin{aligned} 0.8672 - 1.996 \times 0.9653/\sqrt{68} \leq \quad \mu \quad &\leq 0.8672 + 1.996 \times 0.9653/\sqrt{68} \\ \Rightarrow \quad 0.6335 \leq \quad \mu \quad &\leq 1.1008 \end{aligned}$$

[In R, we obtain the critical value 1.996 by `qt(0.975, df=67)` or `-qt(0.025, df=67)`]

In R the command is: `mean(x) + c(-1, 1) * qt(0.975, df=67) * sqrt(var(x)/68)` if the vector `x` contains the 68 weight gain differences. You may obtain this by issuing the commands:

```
wgain <- read.table("wtgain.txt", head=T)
x <- wgain$final -wgain$initial
```

Note that the interval here does not include the value 0, so it is very likely that the weight gain is significantly positive, which we will justfy using what is called testing of hypothesis.

4.6.3 Take home points

In this lecture we have learned how to find an exact confidence interval for a population mean based on the assumption that the population is normal. The confidence interval is based on the t-distribution which is a very important distribution in statistics. The t-distribution converges to the normal distribution when its only parameter, called the degrees of freedom, becomes very large. If the assumption of the normal distribution for the data can be justified, then the method of inference based on the t-distribution is best when the variance parameter, sometimes called the nuisance parameter, is unknown.

4.7 Lecture 27: Hypothesis testing I

4.7.1 Lecture mission

The manager of a new fast food chain claims that the average waiting time to be served in their restaurant is less than a minute. The marketing department of a mobile phone company claims that their phones never break down in the first three years of their lifetime. A professor of nutrition

claims that students gain significant weight in the first year of their life in college away fom home. How can we verify these claims? We will learn the procedures of hypothesis testing for such problems.

4.7.2 Introduction

In statistical inference, we use observations $x_1, \ldots, x_n$ of univariate random variables $X_1, \ldots, X_n$ in order to draw inferences about the probability distribution $f(x)$ of the underlying random variable X. So far, we have mainly been concerned with estimating features (usually unknown parameters) of $f(x)$. It is often of interest to compare alternative specifications for $f(x)$. If we have a set of competing probability models which might have generated the observed data, we may want to determine which of the models is most appropriate. A proposed (hypothesised) model for $X_1, \ldots, X_n$ is then referred to as a *hypothesis*, and pairs of models are compared using hypothesis tests.

For example, we may have two competing alternatives, $f^{(0)}(x)$ (model H_0) and $f^{(1)}(x)$ (model H_1) for $f(x)$, both of which completely specify the joint distribution of the sample $X_1, \ldots, X_n$. Completely specified statistical models are called *simple* hypotheses. Usually, H_0 and H_1 both take the same parametric form $f(x, \theta)$, but with different values $\theta^{(0)}$ and $\theta^{(1)}$ of θ. Thus the joint distribution of the sample given by $f(\mathbf{X})$ is completely specified apart from the values of the unknown parameter θ and $\theta^{(0)} \neq \theta^{(1)}$ are specified alternative values.

More generally, competing hypotheses often do not completely specify the joint distribution of $X_1, \ldots, X_n$. For example, a hypothesis may state that $X_1, \ldots, X_n$ is a random sample from the probability distribution $f(x; \theta)$ where $\theta < 0$. This is not a completely specified hypothesis, since it is not possible to calculate probabilities such as $P(X_1 < 2)$ when the hypothesis is true, as we do not know the exact value of θ. Such an hypothesis is called a *composite* hypothesis.

Examples of hypotheses:
$X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ with $\mu = 0$, $\sigma^2 = 2$.
$X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ with $\mu = 0$, $\sigma^2 \in \mathcal{R}_+$.
$X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ with $\mu \neq 0$, $\sigma^2 \in \mathcal{R}_+$.
$X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ with $p = \frac{1}{2}$.
$X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ with $p \neq \frac{1}{2}$.
$X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ with $p > \frac{1}{2}$.
$X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$ with $\lambda = 1$.
$X_1, \ldots, X_n \sim \text{Poisson}(\theta)$ with $\theta > 1$.

4.7.3 Hypothesis testing procedure

A hypothesis test provides a mechanism for comparing two competing statistical models, H_0 and H_1. A hypothesis test does not treat the two hypotheses (models) symmetrically. One hypothesis, H_0, is given special status, and referred to as the *null hypothesis*. The null hypothesis is the reference model, and is assumed to be appropriate unless the observed data strongly indicate that H_0 is inappropriate, and that H_1 (the *alternative* hypothesis) should be preferred.

Hence, the fact that a hypothesis test does not reject H_0 should not be taken as evidence that H_0 is true and H_1 is not, or that H_0 is better-supported by the data than H_1, merely that the data does not provide significant evidence to reject H_0 in favour of H_1.

A hypothesis test is defined by its *critical region* or *rejection region*, which we shall denote by C. C is a subset of $\mathcal{R}^n$ and is the set of possible observed values of $\mathbf{X}$ which, if observed, would lead to rejection of H_0 in favour of H_1, *i.e.*

$$\begin{array}{ll} \text{If } \mathbf{x} \in C & H_0 \text{ is rejected in favour of } H_1 \\ \text{If } \mathbf{x} \notin C & H_0 \text{ is not rejected} \end{array}$$

As $\mathbf{X}$ is a random variable, there remains the possibility that a hypothesis test will give an erroneous result. We define two types of error:

Type I error: H_0 is rejected when it is true
Type II error: H_0 is not rejected when it is false

The following table helps to understand further:

	H_0 true	H_0 false
Reject H_0	Type I error	Correct decision
Do not reject H_0	Correct decision	Type II error

When H_0 and H_1 are simple hypotheses, we can define

$$\begin{array}{rcll} \alpha = P(\text{Type I error}) & = & P(\mathbf{X} \in C) & \text{if } H_0 \text{ is true} \\ \beta = P(\text{Type II error}) & = & P(\mathbf{X} \notin C) & \text{if } H_1 \text{ is true} \end{array}$$

♡ **Example 67 Uniform** Suppose that we have **one** observation from the uniform distribution on the range $(0, \theta)$. In this case, $f(x) = 1/\theta$ if $0 < x < \theta$ and $P(X \leq x) = \frac{x}{\theta}$ for $0 < x < \theta$. We want to test $H_0 : \theta = 1$ against the alternative $H_1 : \theta = 2$. Suppose we decide arbitrarily that we will reject H_0 if $X > 0.75$. Then

$$\begin{array}{rcll} \alpha = P(\text{Type I error}) & = & P(X > 0.75) & \text{if } H_0 \text{ is true} \\ \beta = P(\text{Type II error}) & = & P(X < 0.75) & \text{if } H_1 \text{ is true} \end{array}$$

which will imply:

$$\alpha = P(X > 0.75 | \theta = 1) = 1 - 0.75 = \frac{1}{4},$$
$$\beta = P(X < 0.75 | \theta = 2) = 0.75/2 = \frac{3}{8}.$$

Here the notation | means given that.

♡ **Example 68 Poisson** The daily demand for a product has a Poisson distribution with mean λ, the demands on different days being statistically independent. It is desired to test the hypotheses

$H_0 : \lambda = 0.7, H_1 : \lambda = 0.3$. The null hypothesis is to be accepted if in 20 days the number of days with no demand is less than 15. Calculate the Type I and Type II error probabilities.

Let p denote the probability that the demand on a given day is zero.
Then

$$p = e^{-\lambda} = \begin{cases} e^{-0.7} & \text{under } H_0 \\ e^{-0.3} & \text{under } H_1. \end{cases}$$

If X denotes the number of days out of 20 with zero demand, it follows that

$$X \sim B(20, e^{-0.7}) \text{ under } H_0,$$
$$X \sim B(20, e^{-0.3}) \text{ under } H_1.$$

Thus

$$\begin{aligned} \alpha &= P(\text{Reject } H_0 | H_0 \text{ true}) \\ &= P(X \geq 15 | X \sim B(20, e^{-0.7})) \\ &= 1 - P(X \leq 14 | X \sim B(20, 0.4966)) \\ &= 1 - 0.98028 \\ &= 0.01923 \text{ (`1-pbinom(14,size=20,prob=0.4966)` in `R`).} \end{aligned}$$

Furthermore

$$\begin{aligned} \beta &= P(\text{Accept } H_0 | H_1 \text{ true}) \\ &= P(X \leq 14 | X \sim B(20, e^{-0.3})) \\ &= P(X \leq 14 | X \sim B(20, 0.7408)) \\ &= P(Y \geq 6 | Y \sim B(20, 0.2592)) \\ &= 1 - P(Y \leq 5 | Y \sim B(20, 0.2592)) \\ &= 1 - 0.58022 \\ &= 0.42023 \text{ (`1-pbinom(5,size=20,prob=0.2592)` in `R`).} \end{aligned}$$

Sometimes α is called the *size* (or *significance level*) of the test and $\omega \equiv 1 - \beta$ is called the *power* of the test. Ideally, we would like to avoid error so we would like to make both α and β as small as possible. In other words, a good test will have small size, but large power. However, it is not possible to make α and β both arbitrarily small. For example if $C = \emptyset$ then $\alpha = 0$, but $\beta = 1$. On the other hand if $C = \mathbf{S} = \mathcal{R}^n$ then $\beta = 0$, but $\alpha = 1$.

The general hypothesis testing procedure is to fix α to be some small value (often 0.05), so that the probability of a Type I error is limited. In doing this, we are giving H_0 precedence over H_1, and acknowledging that Type I error is potentially more serious than Type II error. (Note that for discrete random variables, it may be difficult to find C so that the test has exactly the required size). Given our specified α, we try to choose a test, defined by its rejection region C, to make β as small as possible, *i.e.* we try to find the most powerful test of a specified size. Where H_0 and

H_1 are *simple* hypotheses this can be achieved easily.

Note that tests are usually based on a one-dimensional *test statistic* $T(\mathbf{X})$ whose sample space is some subset of $\mathcal{R}$. The rejection region is then a set of possible values for $T(\mathbf{X})$, so we also think of C as a subset of $\mathcal{R}$. In order to be able to ensure the test has size α, the distribution of the test statistic under H_0 should be known.

4.7.4 The test statistic

We perform a hypothesis test by computing a *test statistic*, $T(\boldsymbol{X})$. A test statistic must (obviously) be a statistic (i.e. a function of $\boldsymbol{X}$ and other known quantities only). Furthermore, the random variable $T(\boldsymbol{X})$ must have a distribution which is known under the null hypothesis. The easiest way to construct a test statistic is to obtain a pivot for θ. If $T(\boldsymbol{X}, \theta)$ is a pivot for θ then its sampling distribution is known and, therefore, under the null hypothesis ($\theta = \theta_0$) the sampling distribution of $T(\boldsymbol{X}, \theta_0)$ is known. Hence $T(\boldsymbol{x}, \theta_0)$ is a test statistic, as it depends on observed data $\boldsymbol{x}$ and the hypothesised value θ_0 only. We then assess the plausibility of H_0 by evaluating whether $T(\boldsymbol{x}, \theta_0)$ seems like a *reasonable observation from its (known) distribution*. This is all rather abstract. How does it work in a concrete example?

4.7.5 Testing a normal mean μ

Suppose that we observe data $x_1, \ldots, x_n$ which are modelled as observations of i.i.d. $N(\mu, \sigma^2)$ random variables $X_1, \ldots, X_n$, and we want to test the null hypothesis

$$H_0 : \mu = \mu_0$$

against the alternative hypothesis

$$H_1 : \mu \neq \mu_0.$$

We recall that

$$\sqrt{n}\frac{(\bar{X} - \mu)}{S} \sim t_{n-1}$$

and therefore, when H_0 is true, often written as under H_0,

$$\sqrt{n}\frac{(\bar{X} - \mu_0)}{S} \sim t_{n-1}$$

so $\sqrt{n}(\bar{X} - \mu_0)/s$ is a test statistic for this test. The sampling distribution of the test statistic when the null hypothesis is true is called the null distribution of the test statistic. In this example, the null distribution is the t-distribution with $n - 1$ degrees of freedom.

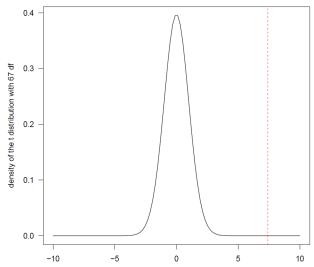
This test is called a *t-test*. We reject the null hypothesis H_0 in favour of the alternative H_1 if the observed test statistic *seems unlikely to have been generated by the null distribution*.

♡ **Example 69 Weight gain data**

For the weight gain data, if x denotes the differences in weight gain, we have $\bar{x} = 0.8672$, $s = 0.9653$ and $n = 68$. Hence our test statistic for the null hypothesis $H_0 : \mu = \mu_0 = 0$ is

$$\sqrt{n}\frac{(\bar{x} - \mu_0)}{s} = 7.41.$$

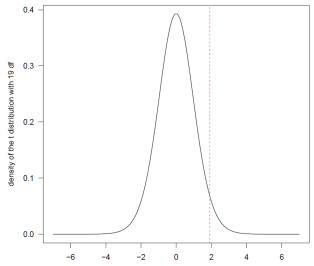
The observed value of 7.41 does not seem reasonable from the graph below. The graph has plotted the density of the t-distribution with 67 degrees of freedom, and a vertical line is drawn at the observed value of 7.41. So there may be evidence here to reject $\mathrm{H}_0 : \mu = 0$.



♡ **Example 70 Fast food waiting time revisited** Suppose the manager of the fast food outlet claims that the average waiting time is only 60 seconds. So, we want to test $H_0 : \mu = 60$. We have $n = 20, \bar{x} = 67.85, s = 18.36$. Hence our test statistic for the null hypothesis $\mathrm{H}_0 : \mu = \mu_0 = 60$ is

$$\sqrt{n}\frac{(\bar{x} - \mu_0)}{s} = \sqrt{20}\frac{(67.85 - 60)}{18.36} = 1.91.$$

The observed value of 1.91 may or may not be reasonable from the graph below. The graph has plotted the density of the t-distribution with 19 degrees of freedom and a vertical line is drawn at the observed value of 1.91. This value is a bit out in the tail but we are not sure, unlike in the previous weight gain example. So how can we decide whether to reject the null hypothesis?



4.7.6 Take home points

In this lecture we have learned the concepts of hypothesis testing such as: simple and composite hypotheses, null and alternative hypotheses, type I error and type II error and their probabilities, test statistic and critical region. We have also introduced the t-test statistic for testing hypotheses regarding a normal mean. We have not yet learned when to reject a null hypothesis, which will be discussed in the next lecture.

4.8 Lecture 28: Hypothesis testing II

4.8.1 Lecture mission

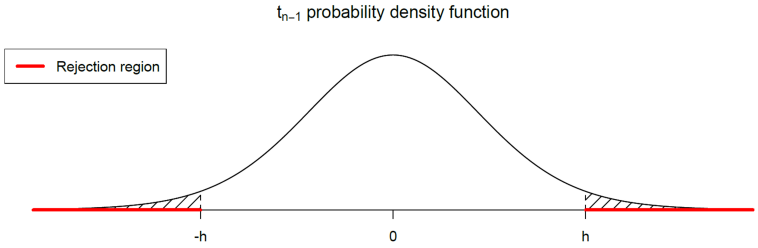
This lecture will discuss the rejection region of a hypothesis test with an example. We will learn the key concepts of the level of significance, rejection region and the p-value associated with a hypothesis test. We will also learn the equivalence of testing and interval estimation.

4.8.2 The significance level

In the weight gain example, it seems clear that there is no evidence to reject H_0, but how extreme (far from the mean of the null distribution) should the test statistic be in order for H_0 to be rejected? The *significance level* of the test, α, is the probability that we will erroneously reject H_0 (called *Type I error* as discussed before). Clearly we would like α to be small, but making it too small risks failing to reject H_0 even when it provides a poor model for the observed data (*Type II error*). Conventionally, α is usually set to a value of 0.05, or 5%. Therefore we reject H_0 when the test statistic lies in a *rejection region* which has probability $\alpha = 0.05$ under the null distribution.

4.8.3 Rejection region for the t-test

For the t-test, the null distribution is t_{n-1} where n is the sample size, so the rejection region for the test corresponds to a region of total probability $\alpha = 0.05$ comprising the 'most extreme' values in the direction of the alternative hypothesis. If the alternative hypothesis is two-sided, e.g. $H_1 : \mu \neq \mu_0$, then this is obtained as below, where the two shaded regions both have area (probability) $\alpha/2 = 0.025$.



The value of h depends on the sample size n and can be found by issuing the `qt` command. Here are few examples obtained from `qt(0.975, df=c(1, 4, 9, 14, 19, 29, 49, 99))`:

n	2	5	10	15	20	30	50	100	∞
h	12.71	2.78	2.26	2.14	2.09	2.05	2.01	1.98	1.96

Note that we need to put $n - 1$ in the df argument of `qt` and the last value for $n = \infty$ is obtained from the normal distribution.

However, if the alternative hypothesis is one-sided, e.g. $H_1 : \mu > \mu_0$, then the critical region will only be in the right tail. Consequently, we need to leave an area α on the right and as a result the critical values will be from a command such as:

```
qt(0.95, df=c(1, 4, 9, 14, 19, 29, 49, 99))
```

n	2	5	10	15	20	30	50	100	∞
h	6.31	2.13	1.83	1.76	1.73	1.70	1.68	1.66	1.64

4.8.4 t-test summary

Suppose that we observe data $x_1, \ldots, x_n$ which are modelled as observations of i.i.d. $N(\mu, \sigma^2)$ random variables $X_1, \ldots, X_n$ and we want to test the null hypothesis $\mathrm{H}_0 : \mu = \mu_0$ against the alternative hypothesis $\mathrm{H}_1 : \mu \neq \mu_0$:

1. Compute the test statistic
$$t = \sqrt{n}\frac{(\bar{x} - \mu_0)}{s}.$$
2. For chosen significance level α (usually 0.05) calculate the rejection region for t, which is of the form $|t| > h$ where $-h$ is the $\alpha/2$ percentile of the null distribution, t_{n-1}.
3. If your computed t lies in the rejection region, i.e. $|t| > h$, you report that H_0 is rejected in favour of H_1 at the chosen level of significance. If t does not lie in the rejection region, you report that H_0 is not rejected. [Never refer to 'accepting' a hypothesis.]

♡ **Example 71 Fast food waiting time** We would like to test $H_0 : \mu = 60$ against the alternative $H_1 : \mu > 60$, as this alternative will refute the claim of the store manager that customers only wait for a maximum of one minute. We calculated the observed value to be 1.91. This is a one-sided test and for a 5% level of significance, the critical value h will come from `qt(0.95, df=19)`=1.73. Thus the observed value is higher than the critical value so we will reject the null hypothesis, disputing the manager's claim regarding a minute wait.

♡ **Example 72 Weight gain**

For the weight gain example $\bar{x} = 0.8671$, $s = 0.9653$, $n = 68$. Then, we would be interested in testing $\mathrm{H}_0 : \mu = 0$ against the alternative hypothesis $\mathrm{H}_1 : \mu \neq 0$ in the model that the data are observations of i.i.d. $N(\mu, \sigma^2)$ random variables.

- We obtain the test statistic
$$t = \sqrt{n}\frac{(\bar{x} - \mu_0)}{s} = \sqrt{68}\frac{(0.8671 - 0)}{0.9653} = 7.41.$$
- Under H_0 this is an observation from a t_{67} distribution. For significance level $\alpha = 0.05$ the rejection region is $|t| > 1.996$.
- Our computed test statistic lies in the rejection region, i.e. $|t| > 1.996$, so H_0 is rejected in favour of H_1 at the 5% level of significance.

In `R` we can perform the test as follows:

```
wgain <- read.table("wtgain.txt", head=T)
x <- wgain$final - wgain$initial
t.test(x)
```

This gives the results: t = 7.4074, and df = 67.

4.8.5 p-values

The result of a test is most commonly summarised by rejection or non-rejection of H_0 at the stated level of significance. An alternative, which you may see in practice, is the computation of a *p-value*. This is the probability that the reference distribution would have generated the actual observed value of the statistic *or something more extreme*. A small p-value is evidence against the null hypothesis, as it indicates that the observed data were unlikely to have been generated by the reference distribution. In many examples a threshold of 0.05 is used, below which the null hypothesis is rejected as being insufficiently well-supported by the observed data. Hence for the t-test with a two-sided alternative, the p-value is given by:

$$p = P(|T| > |t_{\text{obs}}|) = 2P(T > |t_{\text{obs}}|),$$

where T has a t_{n-1} distribution and t_{obs} is the observed sample value.

However, if the alternative is one-sided and to the right then the p-value is given by:

$$p = P(T > t_{\text{obs}}),$$

where T has a t_{n-1} distribution and t_{obs} is the observed sample value.

A small p-value corresponds to an observation of T that is improbable (since it is far out in the low probability tail area) under H_0 and hence provides evidence against H_0. The p-value should *not* be misinterpreted as the probability that H_0 is true. H_0 is not a random event (under our models) and so cannot be assigned a probability. The null hypothesis is rejected at significance level α if the p-value for the test is less than α.

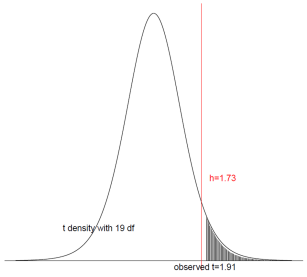
Reject H_0 if p-value $< \alpha$.

4.8.6 p-value examples

In the fast food example, a test of $H_0 : \mu = 60$ resulted in a test statistic $t = 1.91$. Then the p-value is given by:

$$p = P(T > 1.91) = 0.036, \quad \text{when } T \sim t_{19}.$$

This is the area of the shaded region in the figure overleaf. In `R` it is: `1 - pt(1.91, df=19)`. The p-value 0.036 indicates some evidence against the manager's claim at the 5% level of significance but not the 1% level of significance. In the graph, what would be the area under the curve to the right of of the red line?



When the alternative hypothesis is two-sided the p-value has to be calculated from $P(|T| > t_{\text{obs}})$, where t_{obs} is the observed value and T follows the t-distribution with $n-1$ df. For the weight gain example, because the alternative is two-sided, the p-value is given by:

$$p = P(|T| > 7.41) = 2.78 \times 10^{-10} \approx 0.0, \quad \text{when } T \sim t_{67}.$$

This very small p-value for the second example indicates very strong evidence against the null hypothesis of no weight gain in the first year of university.

4.8.7 Equivalence of testing and interval estimation

Note that the 95% confidence interval for μ in the weight gain example has previously been calculated to be (0.6335, 1.1008) in Section 4.6.2. This interval does not include the hypothesised value 0 of μ. Hence we can conclude that the hypothesis test at the 5% level of significance will reject the null hypothesis $H_0 : \mu = 0$. This is because $|T_{obs} = \frac{\sqrt{n}(\bar{x}-\mu_0)}{s}| > h$ implies and is implied by μ_0 being outside the interval $(\bar{x} - hs/\sqrt{n}, \bar{x} + hs/\sqrt{n})$. Notice that h is the same in both. For this reason we often just calculate the confidence interval and take the reject/do not reject decision merely by inspection.

4.8.8 Take home points

This lecture has introduced the key concepts for hypothesis testing. We have defined the p-value of a test and learned that we reject the null hypothesis if the p-value of the test is less than a given level of significance. We have also learned that hypothesis testing and interval estimation are equivalent concepts.

4.9 Lecture 29: Two sample t-tests

4.9.1 Lecture mission

Suppose that we observe two samples of data, $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$, and that we propose to model them as observations of

$$X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu_X, \sigma_X^2)$$

and

$$Y_1, \ldots, Y_m \overset{i.i.d.}{\sim} N(\mu_Y, \sigma_Y^2)$$

respectively, where it is also assumed that the X and Y variables are independent of each other. Suppose that we want to test the hypothesis that the distributions of X and Y are identical, that is

$$\mathrm{H}_0 : \mu_X = \mu_Y, \quad \sigma_X = \sigma_Y = \sigma$$

against the alternative hypothesis

$$\mathrm{H}_1 : \mu_X \neq \mu_Y.$$

4.9.2 Two sample t-test statistic

In the probability lectures we proved that

$$\bar{X} \sim N(\mu_X, \sigma_X^2/n) \quad \text{and} \quad \bar{Y} \sim N(\mu_Y, \sigma_Y^2/m)$$

and therefore

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

Hence, under H_0,

$$\bar{X} - \bar{Y} \sim N\left(0, \sigma^2\left[\frac{1}{n} + \frac{1}{m}\right]\right) \quad \Rightarrow \quad \sqrt{\frac{nm}{n+m}}\frac{(\bar{X} - \bar{Y})}{\sigma} \sim N(0,1).$$

The involvement of the (unknown) σ above means that this is not a pivotal test statistic. It will be proved in MATH2011 that if σ is replaced by its unbiased estimator S, which here is the *two-sample estimator of the common standard deviation*, given by

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{i=1}^m (Y_i - \bar{Y})^2}{n+m-2},$$

then

$$\sqrt{\frac{nm}{n+m}}\frac{(\bar{X} - \bar{Y})}{S} \sim t_{n+m-2}.$$

Hence

$$t = \sqrt{\frac{nm}{n+m}}\frac{(\bar{x} - \bar{y})}{s}$$

is a test statistic for this test. The rejection region is $|t| > h$ where $-h$ is the $\alpha/2$ (usually 0.025) percentile of t_{n+m-2}.

Confidence interval for $\mu_X - \mu_Y$.
From the hypothesis testing, a $100(1-\alpha)\%$ confidence interval is given by

$$\bar{x} - \bar{y} \pm h\sqrt{\frac{n+m}{nm}}\, s,$$

where $-h$ is the $\alpha/2$ (usually 0.025) percentile of t_{n+m-2}.

♡ **Example 73 Fast food waiting time as a two sample t-test**
In this example, we would like to know if there are significant differences between the AM and PM waiting times. Here the 10 morning waiting times (x) are: 38, 100, 64, 43, 63, 59, 107, 52, 86, 77 and the 10 afternoon waiting times (y) are: 45, 62, 52, 72, 81, 88, 64, 75, 59, 70. Here $n = m = 10$, $\bar{x} = 68.9, \bar{y} = 66.8$, $s_x^2 = 538.22$ and $s_y^2 = 171.29$. From this we calculate,

$$s^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2} = 354.8,$$

$$t_{\text{obs}} = \sqrt{\frac{nm}{n+m}}\frac{(\bar{x}-\bar{y})}{s} = 0.25.$$

This is not significant as the critical value $h =$ `qt(0.975,18)`$= 2.10$ is larger in absolute value than 0.25. This can be achieved by calling the `R` function `t.test` as follows:

```
y <- read.csv("servicetime.csv", head=T)
t.test(y$AM, y$PM)
```

It automatically calculates the test statistic as 0.249 and a p-value of 0.8067. It also obtains the 95% CI given by (–15.94, 20.14).

4.9.3 Paired t-test

Sometimes the assumption that the X and Y variables are independent of each other is unlikely to be valid, due to the design of the study. The most common example of this is where $n = m$ and data are *paired*. For example, a measurement has been made on patients before treatment (X) and then again on the same set of patients after treatment (Y). Recall the weight gain example is exactly of this type. In such examples, we proceed by computing data on the differences

$$z_i = x_i - y_i, \quad i = 1, \ldots, n$$

and modelling these differences as observations of i.i.d. $N(\mu_z, \sigma_Z^2)$ variables $Z_1, \ldots, Z_n$. Then, a test of the hypothesis $\mu_X = \mu_Y$ is achieved by testing $\mu_Z = 0$, which is just a standard (one sample) t-test, as described previously.

♡ **Example 74 Paired t-test** Water-quality researchers wish to measure the biomass to chlorophyll ratio for phytoplankton (in milligrams per litre of water). There are two possible tests, one less expensive than the other. To see whether the two tests give the same results, ten water samples were taken and each was measured both ways. The results are as follows:

Test 1 (x)	45.9	57.6	54.9	38.7	35.7	39.2	45.9	43.2	45.4	54.8
Test 2 (y)	48.2	64.2	56.8	47.2	43.7	45.7	53.0	52.0	45.1	57.5

To test the null-hypothesis

$$\text{H}_0 : \mu_Z = 0 \quad \text{against} \quad \text{H}_1 : \mu_Z \neq 0$$

we use the test statistic $t = \sqrt{n}\frac{\bar{z}}{s_z}$, where $s_z^2 = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})^2$.

Confidence interval for μ_Z.

From the hypothesis testing, a $100(1-\alpha)\%$ confidence interval is given by $\bar{z} \pm h\frac{s_z}{\sqrt{n}}$, where h is the critical value of the t distribution with $n-1$ degrees of freedom. In `R` we perform the test as follows:

```
x <- c(45.9, 57.6, 54.9, 38.7, 35.7, 39.2, 45.9, 43.2, 45.4, 54.8)
y <- c(48.2, 64.2, 56.8, 47.2, 43.7, 45.7, 53.0, 52.0, 45.1, 57.5)
t.test(x, y, paired=T)
```

This gives the test statistic $t_{\text{obs}} = -5.0778$ with a df of 9 and a p-value $= 0.0006649$. Thus we reject the null hypothesis. The associated 95% CI is (–7.53, –2.89), printed by `R`.

Interpretation: The values of the second test are significantly higher than the ones of the first test, and so the second test cannot be considered as a replacement for the first.

4.9.4 Take home points

We have introduced the two-sample t-test for testing whether the distributions of two independent samples of data are identical. We have also learned about the paired t-test, which we use in circumstances where the assumption of independence is not valid.

4.10 Lecture 30: Data collection and design of experiments

4.10.1 Lecture mission

The primary purpose of this lecture is to consider issues around the sampling of data in scientific experiments. We return to the question of how should we collect our data? For example, suppose I am interested in estimating the percentage of grammar school students in Southampton and I do not have access to student's personal data (due to data protection), but assume that I do have a list of all students and their university contact information. How can I sample effectively to estimate the proportion, which is a population characteristic? In a medical experiment, suppose the aim is to prove that the new drug is better than the existing drug for treating a mental health disorder. How should we select patients to allocate the drugs, often called treatments in statistical jargon? Obviously it would be wrong to administer the new drug to the male individuals and the old to the females as any difference between the effects of the new and existing drugs will be completely mixed-up with the difference between the effect of the sexes. Hence, effective sampling techniques and optimal design of the data collection experiments are necessary to make valid statistical inferences. This lecture will discuss several methods for statistical data collection.

4.10.2 Simple random sampling

Returning to the problem of estimating the proportion of grammar school students, we must aim to give some positive probability to selection of the population of Southampton students so that my sample is representative of the population. This is called *probability sampling*. A sampling scheme is called *simple random sampling*, SRS, if it gives the same probability of selection for each member of the population. There are two kinds of SRS depending on whether we select the individuals one by one with or without returning the sampled individuals to the population. When we return the selected individuals immediately back to the population, we perform *simple random sampling with replacement* or SRSWR, and if we do not return the sampled individuals back, we perform *simple random sampling without replacement* or SRSWOR. The UK National Lottery draw of six numbers each week is an example of SRSWOR.

Suppose there are N individuals in the population and we are drawing a sample of n individuals. In SRSWR, the same unit of the population may occur more than once in the sample; there are N^n possible samples (using multiplication rules of counting), and each of these samples has equal chance of $1/N^n$ to materialise. In the case of SRSWOR, at the rth drawing ($r = 1, \ldots, n$) there are $N - r + 1$ individuals in the population to sample from. All of these individuals are given equal probability of inclusion in the sample. Here no member of the population can occur more than once in the sample. There are ${}^{N}C_{n}$ possible samples and each has equal probability of inclusion $1/{}^{N}C_{n}$. This is also justified as at the rth stage one is to choose from $N - r + 1$ individuals one

of the $n - r + 1$ individuals to be included in the sample which have not yet been chosen in earlier drawings. In this case too, the probability that any specified individual, say the ith, is selected at any drawing, say the kth drawing, is:

$$\frac{N-1}{N} \times \frac{N-2}{N-1} \times \cdots \times \frac{N-k+1}{N-k+2} \frac{1}{N-k+1} = \frac{1}{N}$$

as in the case of the SRSWR. It is obvious that if one takes n individuals all at a time from the population, giving equal probability to each of the ${}^{N}C_{n}$ combinations of n members out of the N members in the population, one will still have SRSWOR.

How can we draw random samples?
A quick method is drawing numbers out of hat. But this is cumbersome and manual. In practice, we use random number series to draw samples at random. A random sampling number series is an arrangement, which may be looked upon either as linear or rectangular, in which each place has been filled in with one of the digits $0, 1, \ldots, 9$. The digit occupying any place is selected at random from these 10 digits and independently of the digits occurring in other positions. Different random number series are available in books and computers. In `R` we can easily use the `sample` command to draw random samples either using SRSWR or SRSWOR. For example, suppose the problem is to select 50 students out of the 200 in this class. In this experiment, I shall number the students 1 to 200 in any way possible, e.g. alphabetically by surname. I will then issue the command `sample(200, size=50)` for SRSWOR and `sample(200, size=50, replace=T)` for SRSWR.

There are a huge number of considerations and concepts to design good surveys avoiding bias. There may be response bias, observational bias, biases from non-response, interviewer bias, bias due to defective sampling technique, bias due to substitution, bias due to faulty differentiation of sampling units and so on. However, discussion of such topics is beyond the scope and syllabus of this module.

4.10.3 Design of experiments

An *experiment* is a means of getting an answer to the question that the experimenter has in mind. This may be to decide which of the several pain-relieving tablets that are available over the counter is the most effective or equally effective. An experiment may be conducted to study and compare the British and Chinese methods of teaching mathematics in schools. In planning an experiment, we clearly state our objectives and formulate the hypotheses we want to test. We now give some key definitions.

Treatment. The different procedures under comparison in an experiment are the different treatments. For example, in a chemical engineering experiment different factors such as Temperature (T), Concentration (C) and Catalyst (K) may affect the yield value from the experiment.

Experimental unit. An experimental unit is the material to which the treatment is applied and on which the variable under study is measured. In a human experiment in which the treatment affects the individual, the individual will be the experimental unit.

Design of experiments is a systematic and rigorous approach to problem-solving in many disciplines such as engineering, medicine etc. that applies principles and techniques at the data collection stage, so as to ensure valid inferences for the hypotheses of interest.

Three principles of experimental design

1. *Randomisation.* This is necessary to draw valid conclusions and minimise bias. In an experiment to compare two pain-relief tablets we should allocate the tablets randomly among participants – not one tablet to the boys and the other to the girls.

2. *Replication.* A treatment is repeated a number of times in order to obtain a more reliable estimate than a single observation. In an experiment to compare two diets for children, we can plan the experiment so that no particular diet is favoured in the experiment, i.e. each diet is applied approximately equally among all types of children (boys, girls, their ethnicity etc.).

   The most effective way to increase the precision of an experiment is to increase the number of replications. Remember, $\text{Var}(\bar{X}) = \sigma^2/n$, which says that the standard deviation decreases proportional to the square root of the number of replications. However, replication beyond a limit may be impractical due to cost and other considerations.

3. *Local control.* In the simplest case of local control, the experimental units are divided into homogeneous groups or blocks. The variation among these blocks is eliminated from the error and thereby efficiency is increased. These considerations lead to the topic of construction of block designs, where random allocation of treatments to the experimental units may be restricted in different ways in order to control experimental error. Another means of controlling error is through the use of confounded designs where the number of treatment combinations is very large, e.g. in factorial experiments.

Factorial experiment A thorough discussion of construction of block designs and factorial experiments is beyond the scope of this module. However, these topics are studied in the third-year module *MATH3014: Design of Experiments.* In the remainder of this lecture, we simply discuss an example of a factorial experiment and how to estimate different effects.

♡ **Example 75 A three factor experiment** Chemical engineers wanted to investigate the yield, the value of the outcome of the experiment which is often called the *response*, from a chemical process. They identified three factors that might affect the yield: Temperature (T), Concentration (C) and catalyst (K), with levels as follows:

Temperature (T)	160C	180C
Concentration (C)	20%	40%
Catalyst (K)	A	B

To investigate how factors jointly influence the response, they should be investigated in an experiment in which they are all varied. Even when there are no factors that interact, a factorial experiment gives greater accuracy. Hence they are widely used in science, agriculture and industry.

Here we will consider factorial experiments in which each factor is used at only two levels. This is a very common form of experiment, especially when many factors are to be investigated. We will *code* the levels of each factor as 0 (low) and 1 (high). Each of the 8 combinations of the factor levels were used in the experiment. Thus the treatments in standard order were:

000, 001, 010, 011, 100, 101, 110, 111.

Each treatment was used in the manufacture of one batch of the chemical and the yield (amount in grams of chemical produced) was recorded. Before the experiment was run, a decision had to be made on the order in which the treatments would be run. To avoid any unknown feature that changes with time being confounded with the effects of interest, a random ordering was used; see below. The response data are also shown in the table.

Standard order	Randomised order	T	C	K	Yield	Code
1	6	0	0	0	60	000
2	2	0	0	1	52	001
3	5	0	1	0	54	010
4	8	0	1	1	45	011
5	7	1	0	0	72	100
6	4	1	0	1	83	101
7	3	1	1	0	68	110
8	1	1	1	1	80	111

Questions of interest

1. How much is the response changed when the level of one factor is changed from high to low?
2. Does this change in response depend on the level of another factor?

For simplicity, we first consider the case of two factors only and call them factors A and B, each having two levels, 'low' (0) and 'high' (1). The four treatments in the experiment are then 00, 01, 10, 11, and suppose that we have just one response measured for each treatment combination. We denote the four response values by $yield_{00}$, $yield_{01}$, $yield_{10}$ and $yield_{11}$.

Main effects

For this particular experiment, we can answer the first question by measuring the difference between the average yields at the two levels of A:

The average yield at the high level of A is $\frac{1}{2}(yield_{11} + yield_{10})$.
The average yield at the low level of A is $\frac{1}{2}(yield_{01} + yield_{00})$.

These are represented by the open stars in the Figure 4.1. The main effect of A is defined as the difference between these two averages, that is

$$\begin{aligned} A &= \frac{1}{2}(yield_{11} + yield_{10}) - \frac{1}{2}(yield_{01} + yield_{00}) \\ &= \frac{1}{2}(yield_{11} + yield_{10} - yield_{01} - yield_{00}), \end{aligned}$$

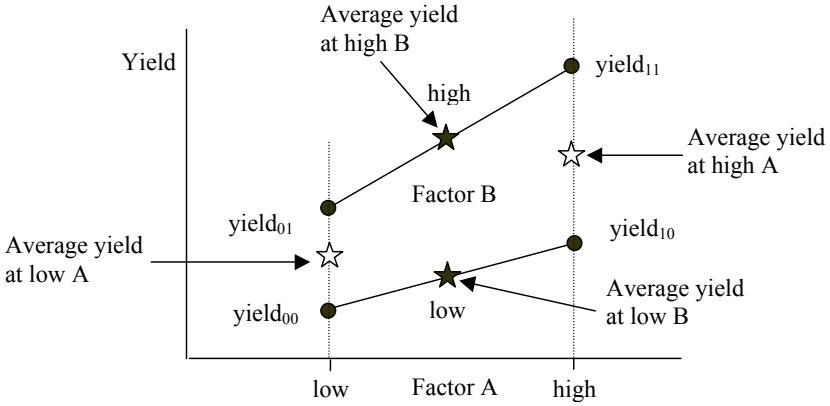


Figure 4.1: Figure showing factorial effects.

which is represented by the difference between the two open stars in Figure 4.1. Notice that A is used to denote the main effect of a factor as well as its name. This is a common practice. This quantity measures how much the response changes when factor A is changed from its low to its high level, averaged over the levels of factor B.

Similarly, the main effect of B is given by

$$\begin{aligned} B &= \frac{1}{2}(yield_{11} + yield_{01}) - \frac{1}{2}(yield_{10} + yield_{00}) \\ &= \frac{1}{2}(yield_{11} - yield_{10} + yield_{01} - yield_{00}), \end{aligned}$$

which is represented by the difference between the two black stars in Figure 4.1.

We now consider question 2, that is, whether this change in response is consistent across the two levels of factor A.

Interaction between factors A and B

Case 1: No Interaction

When the effect of factor B at a given level of A (difference between the two black stars) is the same, regardless of the level of A, the two factors A and B are said not to interact with each other. The response lines are parallel in Figure 4.2.

When the effect of factor B (difference between the two black stars) is different from the corresponding differences for different levels of A, the two factors A and B are said to interact with each other. The response lines are not parallel in Figure 4.3

Computation of Interaction Effect

We define the interaction between factors A and B as one half of the differences between

- the effect of changing B at the high level of A, $(yield_{11} - yield_{10})$, and

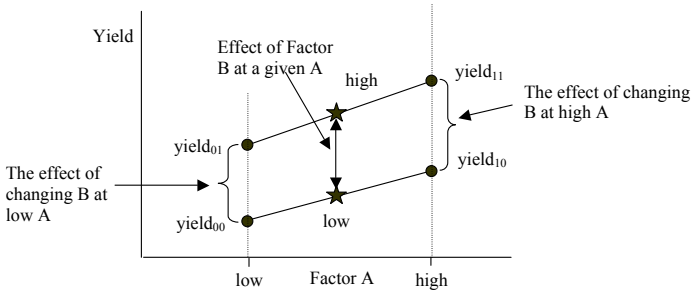


Figure 4.2: Plot showing no interaction effects.

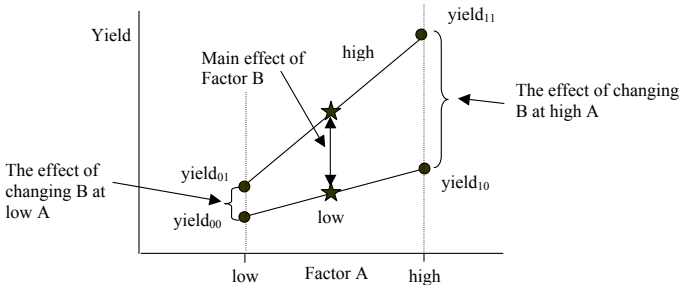


Figure 4.3: Plot showing interaction effects.

- the effect of changing B at the low level of A, $(yield_{01} - yield_{00})$, that is

$$AB = \frac{1}{2}(yield_{11} - yield_{10} - yield_{01} + yield_{00}).$$

- If the lines are parallel then this interaction, AB, will be small.

- If we interchange the roles of A and B in this expression we obtain the same formula.
- Definition: The main effects and interactions are known collectively as the *factorial effects*.
- Important note: When there is a large interaction between two factors, the two main effects cannot be interpreted separately.

4.10.4 Take home points

In this lecture we have discussed techniques of random sampling and the main ideas of design of experiments. A three factor experiment example has demonstrated estimation of the main effects of the factors and their interactions. Further analysis using `R` can be performed by following the `yield` example online. These ideas will be followed up in the third year module Math3014: Design of Experiments.

Appendix A

Mathematical Concepts Needed in MATH1024

During the first week's workshop and problem class you are asked to go through this. Please try the proofs/exercises as well and verify that your solutions are correct by talking to a workshop assistant. Solutions to some of the exercises are given at the end of this chapter and some others are discussed in lectures.

A.1 Discrete sums

1. We can prove by induction that:

$$1 + 2 + \cdots + n = \frac{n(n+1)}{2}$$

and

$$1^2 + 2^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6}$$

for a positive natural number n. You do not need to prove these, but you should try to remember them.

2. Consider a set of numbers $x_1, \ldots, x_n$ and

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

(a) Prove that $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$.

(b) Prove that $\sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2$.

(c) Prove that for any number a,

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

Hence argue that $\sum_{i=1}^{n}(x_i - a)^2$ is minimised at $a = \overline{x}$.

A.2 Counting and combinatorics

For non-negative integers n and k such that $k \leq n$, the number of combinations

$$^nC_k \equiv \binom{n}{k} = \frac{n!}{k!\,(n-k)!}.$$

Prove that

$$\binom{n}{k} = \frac{n \times (n-1) \times \cdots \times (n-k+1)}{1 \times 2 \times 3 \times \cdots \times k.}$$

Proof We have:

$$\begin{aligned} \tbinom{n}{k} &= \tfrac{n!}{k!\,(n-k)!} \\ &= \tfrac{1}{k!}\tfrac{1\times2\times3\times\cdots\times(n-k)\times(n-k+1)\times\cdots\times(n-1)\times n}{1\times2\times3\times\cdots\times(n-k)} \\ &= \tfrac{1}{k!}\left[(n-k+1)\times\cdots\times(n-1)\times n\right]. \end{aligned}$$

Hence the proof is complete. This enables us to calculate $\binom{6}{2} = \frac{6\times5}{1\times2} = 15$. In general for calculating $\binom{n}{k}$:

the numerator is the multiplication of k terms starting with n and counting down, and the denominator is the multiplication of the first k positive integers.

With $0! = 1$ and $k! = 1 \times 2 \times 3 \times \cdots \times k$, and the above formula, prove the following:

1. $^nC_k = {}^nC_{n-k}$. [This means number of ways of choosing k items out of n items is same as the number of ways of choosing $n-k$ items out of n items. Why is this meaningful?]
2. $\binom{n+1}{k} = \binom{n}{k} + \binom{n}{k-1}$.
3. For each of (1) and (2), state the meaning of these equalities in terms of the numbers of selections of k items without replacement.

A.3 Binomial theorem

By mathematical induction it can be proved that:

$$(a+b)^n = b^n + \binom{n}{1}ab^{n-1} + \cdots + \binom{n}{x}a^xb^{n-x} + \cdots + a^n$$

for any numbers a and b and a positive integer n. This is called the binomial theorem. This can be used to prove the following:

$$(1-p)^n + \binom{n}{1}p(1-p)^{n-1} + \cdots + \binom{n}{x}p^x(1-p)^{n-x} + \cdots + p^n = (p+1-p)^n = 1.$$

Thus

$$\sum_{x=0}^{n}\binom{n}{x}p^x(1-p)^{n-x} = 1,$$

which also implies:

$$\sum_{x=0}^{n-1} \binom{n-1}{x} p^x (1-p)^{n-1-x} = 1,$$

for $n > 1$.

Exercise 1. Hard Show that

$$\sum_{x+y=z} \binom{m}{x}\binom{n}{y} = \binom{m+n}{z},$$

where the above sum is also over all possible integer values of x and y such that $0 \le x \le m$ and $0 \le y \le n$.

Hint Consider the identity

$$(1+t)^m (1+t)^n = (1+t)^{m+n}$$

and compare the coefficients of t^{m+n} on both sides. If this is hard, please try small values of m and n, e.g. 2, 3 and see what happens.

Exercise 2. Hard Show that if $X \sim \text{Poisson}(\lambda)$, $Y \sim \text{Poisson}(\mu)$ and X and Y are independent random variables then

$$X + Y \sim \text{Poisson}(\lambda + \mu).$$

Suppose $X \sim \text{Poisson}(\lambda)$, $Y \sim \text{Poisson}(\mu)$, and let $Z = X + Y$. To determine the distribution of Z, we need to find $P(Z = z)$ for all $z \ge 0$.

$$\begin{aligned}
P\{Z = z\} &= P\{X + Y = z\} \\
&= \sum_{x=0}^{\infty} P\{X + Y = z \,|\, X = x\} P\{X = x\} \\
&= \sum_{x=0}^{z} P\{Y = z - x\} P\{X = x\} \\
&= \sum_{x=0}^{z} \left(e^{-\mu}\mu^{z-x}/(z-x)!\right)\left(e^{-\lambda}\lambda^x/x!\right) \\
&= e^{-\mu-\lambda} \sum_{x=0}^{z} \frac{1}{x!\,(z-x)!} \lambda^x \mu^{z-x} \\
&= \frac{e^{-(\lambda+\mu)}}{z!} \sum_{x=0}^{z} \frac{z!}{x!\,(z-x)!} \lambda^x \mu^{z-x} \\
&= \frac{e^{-(\lambda+\mu)}}{z!} \sum_{x=0}^{z} \binom{z}{x} \lambda^x \mu^{z-x} \\
&= \frac{e^{-(\lambda+\mu)}}{z!} (\mu + \lambda)^z \text{ (binomial sum).}
\end{aligned}$$

Thus $Z \sim \text{Poisson}(\lambda+\mu)$ since the above is the probability mass function of the Poisson distribution with parameter $\lambda + \mu$.

A.4 Negative binomial series

1. Sum of a finite geometric series with common ratio r:

$$\sum_{x=0}^{k} r^x = 1 + r + r^2 + \cdots + r^k = \frac{1 - r^{k+1}}{1 - r}.$$

The power of r, $k+1$, in the formula for the sum is the number of terms.

2. When $k \to \infty$ we can evaluate the sum only when $|r| < 1$. In that case

$$\sum_{x=0}^{\infty} r^x = \frac{1}{1 - r} \ [r^{k+1} \to 0 \text{ as } k \to \infty \text{ for } |r| < 1].$$

3. For a positive n and $|x| < 1$, the negative binomial series is given by:

$$(1-x)^{-n} = 1 + nx + \frac{1}{2}n(n+1)x^2 + \frac{1}{6}n(n+1)(n+2)x^3 + \cdots + \frac{n(n+1)(n+2)\cdots(n+k-1)}{k!}x^k + \cdots$$

With $n = 2$ the general term is given by:

$$\frac{n(n+1)(n+2)(n+k-1)}{k!} = \frac{2 \times 3 \times 4 \times \cdots \times (2+k-1)}{k!} = k + 1.$$

Using this and by taking $x = q$ we prove that:

$$1 + 2q + 3q^2 + 4q^3 + \cdots = (1 - q)^{-2}.$$

A.5 Logarithm and the exponential function

A.5.1 Logarithm

From this point onwards, in this module the symbol log will always mean the natural logarithm or the log to the base e. We just need to remember the following basic formulae for two suitable numbers a and b.

1. $\log(ab) = \log(a) + \log(b)$ [Log of the product is the sum of the logs]
2. $\log\left(\frac{a}{b}\right) = \log(a) - \log(b)$ [Log of the ratio is the difference of the logs]
3. $\log\left(a^b\right) = b\log(a)$ or equivalently $a^b = e^{b\log(a)}$.

There is no simple formula for $\log(a + b)$ or $\log(a - b)$. Now try the following exercises:

1. Show that $\log\left(xe^{ax^3+3x+b}\right) = \log(x) + ax^3 + 3x + b$.
2. Show that $e^{2\log(x)} = x^2$.
3. Satisfy yourself that $\log(x_1 x_2 \cdots x_n) = \sum_{i=1}^{n} \log(x_i)$.

A.5.2 The exponential function

The exponential function is defined for any finite number a:

$$\exp(a) \equiv e^a = \lim_{n\to\infty}\left(1+\frac{a}{n}\right)^n = 1 + a + \frac{a^2}{2!} + \frac{a^3}{3!} + \cdots$$

where for a positive integer k, $k! = 1 \times 2 \times 3 \times \cdots \times k$. Hence, $2! = 2$, $3! = 6$ and so on. By convention we use $0! = 1$. Now try the following exercises:

1. Satisfy yourself that $\lim_{n\to\infty}\left(1-\frac{x^2}{n}\right)^n = e^{-x^2}$.
2. Satisfy yourself that $\sum_{x=0}^{\infty} e^{-\lambda}\frac{\lambda^x}{x!} = 1$.

A.6 Integration

A.6.1 Fundamental theorem of calculus

We need to remember (but not prove) the **fundamental theorem of calculus**:

$$F(x) = \int_{-\infty}^{x} f(u)du \quad \text{implies} \quad f(x) = \frac{dF(x)}{dx}$$

For example, consider the pair

$$f(x) = \lambda e^{-\lambda x}, \quad F(x) = 1 - e^{-\lambda x}.$$

Satisfy yourself that the fundamental theorem of calculus holds.

A.6.2 Even and odd functions

A function $f(x)$ is said to be an *even function* if

$$f(x) = f(-x)$$

for all possible values of x. For example, $f(x) = e^{-\frac{x^2}{2}}$ is an even function for real x.

A function $f(x)$ is said to be an *odd function* if

$$f(x) = -f(-x)$$

for all possible values of x. For example, $f(x) = xe^{-\frac{x^2}{2}}$ is an odd function for real x. It can be proved that:

for any real positive value of a,

$$\int_{-a}^{a} f(x)dx = \begin{cases} 2\int_0^a f(x)dx & \text{if } f(x) \text{ is an even function of } x \\ 0 & \text{if } f(x) \text{ is an odd function of } x. \end{cases}$$

The proof of this is not required. Here is an example of each.

$$\begin{aligned}\int_{-a}^{a} f(x)dx &= \int_{-a}^{a} x^2 dx \\ &= \left.\frac{x^3}{3}\right|_{-a}^{a} \\ &= \frac{1}{3}(a^3 + a^3) \\ &= \frac{2}{3}a^3 \\ &= 2\int_0^a x^2 dx.\end{aligned}$$

Similarly, $f(x) = x$ is an odd function of x and consequently, $\int_{-a}^{a} f(x)dx = 0$ for any a.

A.6.3 Improper integral and the gamma function

An integral such as $\int_0^\infty f(x)dx$ is called an improper integral and it is defined as

$$\int_0^\infty f(x)dx = \lim_{a\to\infty} \int_0^a f(x)dx,$$

if the limit exists. For example, it is easy to see that

$$\int_0^\infty e^{-x}dx = 1.$$

The *gamma function* is an improper integral defined by

$$\boxed{\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx,\ \alpha > 0.}$$

It can be shown that $\Gamma(\alpha)$ exists and is finite for all real values of $\alpha > 0$. Obviously it is non-negative since it is the integral of a non-negative function. It is easy to see that

$$\Gamma(1) = 1$$

since $\int_0^\infty e^{-x}dx = 1$. The argument α enters only through the power of the dummy (x). Remember this, as we will have to recognise many gamma integrals. Important points are:

1. It is an integral from 0 to ∞.
2. The integrand must be of the form dummy (x) to the power of the parameter (α) minus one ($x^{\alpha-1}$) multiplied by e to the power of the negative dummy (e^{-x}).
3. The parameter (α) must be greater than zero.

Using integration by parts, we can prove the **reduction formula**:

$$\boxed{\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1), \quad \text{for } \alpha > 1.}$$

This formula reads:

Gamma(parameter) = (parameter − 1) Gamma(parameter − 1)

provided **parameter** > 1. The condition $\alpha > 1$ is required to ensure that $\Gamma(\alpha - 1)$ exists. The proof of this is not required, but can be proved easily by integration by parts by integrating the function e^{-x} and differentiating the function $x^{\alpha-1}$. For an integer n, by repeatedly applying the reduction formula and $\Gamma(1) = 1$, show that

$$\Gamma(n) = (n-1)!.$$

Thus $\Gamma(5) = 4! = 24$. You can guess how rapidly the gamma function increases! The last formula we need to remember for our purposes is:

$$\Gamma\left(\tfrac{1}{2}\right) = \sqrt{\pi}.$$

Proof of this is complicated and not required for this module. Using this we can calculate

$$\Gamma\left(\frac{3}{2}\right) = \left(\frac{3}{2} - 1\right)\Gamma\left(\tfrac{1}{2}\right) = \frac{\sqrt{\pi}}{2}.$$

Now we can easily tackle integrals such as the following:

1. For fun try to evaluate $\int_0^\infty x^{\alpha-1}e^{-\beta x}\,dx$ for $\alpha > 0$ and $\beta > 0$.
2. Prove that $\int_0^\infty xe^{-\lambda x}\,dx = \frac{1}{\lambda^2}$.
3. Prove that $\int_0^\infty x^2e^{-\lambda x}\,dx = \frac{2}{\lambda^3}$.

Solution to the combinatorics problems.

1. $\binom{n}{n-k} = \frac{n!}{(n-k)!(n-[n-k])!} = \frac{n!}{(n-k)!k!} = \binom{n}{k}$
2. $$\begin{aligned} RHS &= \frac{n!}{k!(n-k)!} + \frac{n!}{(k-1)!(n-[k-1])!} \\ &= \frac{n!}{k!(n-[k-1])!}[n-(k-1)+k] \\ &= \frac{n![n+1]}{k!(n-[k-1])!} \\ &= \frac{(n+1)!}{k!(n+1-k)!} \\ &= LHS \end{aligned}$$
3. (a) Number of selections (without replacement) of k objects from n is exactly the same as the number of selections of $(n-k)$ objects from n.

   (b) The number of selections of k items from $(n+1)$ consists of:

   - The number of selections that include the $(n+1)^{th}$ item. There are $\binom{n}{k-1}$ of these.
   - The number of selections that exclude the $(n+1)^{th}$ item. There are $\binom{n}{k}$ of these.

Appendix B

Worked Examples

B.1 Probability examples

76. [Conditional probability] A chest has three drawers. The first contains two gold coins, the second contains a gold and silver coin and the third has two silver coins. A drawer is chosen at random and from it a coin is chosen at random. What is the probability that the coin still remaining in the chosen drawer is gold given that the coin chosen is silver?

77. [Conditional probability] Consider a family with two children. If each child is as likely to be a boy as a girl, what is the probability that both children are boys

    (a) given that the older child is a boy,

    (b) given that at least one of the children is a boy?

78. [Conditional probability] 10% of the boys in a school are left-handed. Of those who are left-handed, 80% are left-footed; of those who are right-handed, 15% are left-footed. If a boy, selected at random, is left-footed, use Bayes Theorem to calculate the probability that he is left-handed.

79. [Bayes Theorem] A car insurance company classifies each driver as a low risk, a medium risk or a high risk. Of those currently insured, 30% are low risks, 50% are medium risks and 20% are high risks. In any given year, the probability that a driver will have at least one accident is 0.1 for a low risk, 0.3 for a medium risk, and 0.5 for a high risk. What is the probability that a randomly selected driver insured by this company has at least one accident during the next year? What is the probability that a driver who had an accident (already occurred) was previously classified as a low risk?

    Let

    $$B_1 = \{\text{a randomly selected driver is a low risk}\},$$
    $$B_2 = \{\text{a randomly selected driver is a medium risk}\},$$
    $$B_3 = \{\text{a randomly selected driver is a high risk}\},$$
    $$A = \{\text{a randomly selected driver has at least one accident during the year}\}.$$

Note that B_1, B_2, B_3 are mutually exclusive and exhaustive. Find $P\{A\}$ and $P\{B_1|A\}$.

80. [Independent events] The probability that Jane can solve a certain problem is 0.4 and that Alice can solve it is 0.3. If they both try independently, what is the probability that it is solved?

81. [Random variable] A fair die is tossed twice. Let X equal the first score plus the second score. Determine

    (a) the probability function of X,

    (b) the cumulative distribution function of X and draw its graph.

82. [Random variable] A coin is tossed three times. If X denotes the number of heads minus the number of tails, find the probability function of X and draw a graph of its cumulative distribution function when

    (a) the coin is fair,

    (b) the coin is biased so that $P\{H\} = \frac{3}{5}$ and $P\{T\} = \frac{2}{5}$.

83. [Expectation and variance] The random variable X has probability function

$$p_x = \begin{cases} \frac{1}{14}(1+x) & \text{if } x = 1, 2, 3, 4 \\ 0 & \text{otherwise.} \end{cases}$$

    Find the mean and variance of X.

84. [Expectation and variance] Let X denote the score when a fair die is thrown. Determine the probability function of X and find its mean and variance.

85. [Expectation and variance] Two fair dice are tossed and X equals the larger of the two scores obtained. Find the probability function of X and determine $E(X)$.

86. [Expectation and variance] The random variable X is uniformly distributed on the integers $0, \pm 1, \pm 2, \ldots, \pm n$, i.e.

$$p_x = \begin{cases} \frac{1}{2n+1} & \text{if } x = 0, \pm 1, \pm 2, \ldots, \pm n \\ 0 & \text{otherwise.} \end{cases}$$

    Obtain expressions for the mean and variance in terms of n. Given that the variance is 10, find n.

87. [Poisson distribution] The number of incoming calls at a switchboard in one hour is Poisson distributed with mean $\lambda = 8$. The numbers arriving in non-overlapping time intervals are statistically independent. Find the probability that in 10 non-overlapping one hour periods at least two of the periods have at least 15 calls.

88. [Continuous distribution] The random variable X has probability density function

$$f(x) = \begin{cases} kx^2(1-x) & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

(a) Find the value of k.
(b) Find the probability that X lies in the range $(0, \frac{1}{2})$.
(c) In 100 independent observations of X, how many on average will fall in the range $(0, \frac{1}{2})$?

89. [Exponential distribution] The random variable X has probability density function

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Find expressions for

(a) the mean,
(b) the standard deviation σ,
(c) the mode,
(d) the median.

Show that the interquartile range equals $\sigma \log(3)$.

90. [Cauchy distribution] A random variable X is said to have a *Cauchy* distribution if its probability density function is given by

$$f(x) = \frac{1}{\pi(1+x^2)}, -\infty < x < +\infty.$$

(a) Verify that it is a valid probability density function and sketch its graph.
(b) Find the cumulative distribution function $F(x)$.
(c) Find $P(-1 \leq X \leq 1)$.

91. [Continuous distribution] The probability density function of X has the form

$$f(x) = \begin{cases} a + bx + cx^2 & \text{if } 0 \leq x \leq 4 \\ 0 & \text{otherwise.} \end{cases}$$

If $E(X) = 2$ and $\text{Var}(X) = \frac{12}{5}$, determine the values of a, b and c.

B.2 Solutions: Probability examples

76. Define events

GG : the chosen drawer has two gold coins,
GS : the chosen drawer has one gold and one silver coin,
SS : the chosen drawer has two silver coins,
S : the coin chosen is silver.

We require $P\{GS|S\}$. By the Bayes Theorem

$$P\{GS|S\} = \frac{P\{S|GS\}P\{GS\}}{P\{S|GG\}P\{GG\} + P\{S|GS\}P\{GS\} + P\{S|SS\}P\{SS\}}$$
$$= \frac{\frac{1}{2} \times \frac{1}{3}}{0 \times \frac{1}{3} + \frac{1}{2} \times \frac{1}{3} + 1 \times \frac{1}{3}} = \frac{1}{3}.$$

77. We assume that the sexes of the children are independent.

(a)

$$\begin{aligned} P\{\text{both boys}|\text{older is a boy}\} &= \frac{P\{\text{both boys and older is a boy}\}}{P\{\text{older is a boy}\}} \\ &= \frac{P\{\text{both boys}\}}{P\{\text{older is a boy}\}} = \frac{1/4}{1/2} = \frac{1}{2}. \end{aligned}$$

(b)

$$\begin{aligned} P\{\text{both boys}|\text{at least one boy}\} &= \frac{P\{\text{both boys and at least one boy}\}}{P\{\text{at least one boy}\}} \\ &= \frac{P\{\text{both boys}\}}{1 - P\{\text{both girls}\}} = \frac{1/4}{1 - 1/4} = \frac{1}{3}. \end{aligned}$$

78. In the obvious notation

$$\begin{aligned} P\{LH|LF\} &= \frac{P\{LF|LH\}P(LH)}{P\{LF|LH\}P(LH) + P\{LF|RH\}P\{RH\}} \\ &= \frac{0.8 \times 0.1}{0.8 \times 0.1 + 0.15 \times 0.9} = \frac{16}{43}. \end{aligned}$$

79. Here we have

$P\{B_1\} = 0.3$	$P\{A\|B_1\} = 0.1$
$P\{B_2\} = 0.5$	$P\{A\|B_2\} = 0.3$
$P\{B_3\} = 0.2$	$P\{A\|B_3\} = 0.5$

Hence

$$\begin{aligned} P\{A\} &= P\{B_1\}P\{A|B_1\} + P\{B_2\}P\{A|B_2\} + P\{B_3\}P\{A|B_3\} \\ &= 0.3 \times 0.1 + 0.5 \times 0.3 + 0.2 \times 0.5 \\ &= 0.31 \end{aligned}$$

Now by the Bayes theorem,

$$\begin{aligned} P\{B_1|A\} &= \frac{P\{B_1\}P\{A|B_1\}}{P\{A\}} \\ &= \frac{0.3 \times 0.1}{0.31} \\ &= \frac{3}{31}. \end{aligned}$$

80.

$$\begin{aligned} P\{\text{problem not solved}\} &= P\{\text{Jane fails and Alice fails}\} \\ &= P\{\text{Jane fails}\}P\{\text{Alice fails}\} \text{ (by independence)} \\ &= (1 - 0.4)(1 - 0.3) \\ &= 0.42. \end{aligned}$$

Hence $P\{\text{problem solved}\} = 0.58$.

81. The sample space is

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

(a) Working along the cross-diagonals we find by enumeration that X has the following probability function

x	2	3	4	5	6	7	8	9	10	11	12
p_x	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

More concisely,

$$p_x = \begin{cases} \frac{6-|x-7|}{36} & \text{if } x = 2, \ldots, 12 \\ 0 & \text{otherwise.} \end{cases}$$

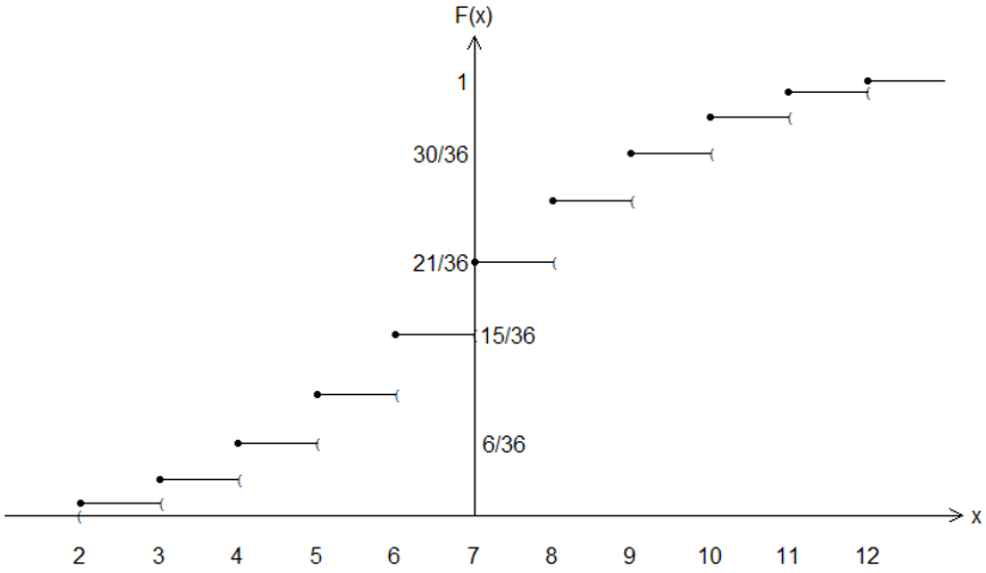
(b)

$$F(x) = \begin{cases} 0 & \text{if } x < 2 \\ \frac{1}{36} & \text{if } 2 \le x < 3 \\ \frac{3}{36} & \text{if } 3 \le x < 4, \text{ etc.} \end{cases}$$

Using the formula $\sum_{i=1}^{n} i = \frac{1}{2}n(n+1)$ we find that the cumulative distribution function $F(x)$ can be written concisely in the form

$$F(x) = \begin{cases} 0 & \text{if } x < 2 \\ \frac{(6+[x-7])(7+[x-7])}{72} & \text{if } 2 \le x < 7 \\ \frac{21}{36} + \frac{[x-7](11-[x-7])}{72} & \text{if } 7 \le x < 12 \\ 1 & \text{if } x \ge 12, \end{cases}$$

where $[x]$ denotes the integral part of x.
For example, $F(3.5) = \frac{(6-4)(7-4)}{72} = \frac{3}{36}$, $F(10.5) = \frac{21}{36} + \frac{3(11-3)}{72} = \frac{33}{36}$.



82. The possibilities are

Sequence	X	Prob. in (a)	Prob. in (b)
HHH	3	1/8	$(3/5)^3$
HHT	1	1/8	$(3/5)^2(2/5)$
HTH	1	1/8	$(3/5)^2(2/5)$
THH	1	1/8	$(3/5)^2(2/5)$
HTT	-1	1/8	$(3/5)(2/5)^2$
THT	-1	1/8	$(3/5)(2/5)^2$
TTH	-1	1/8	$(3/5)(2/5)^2$
TTT	-3	1/8	$(2/5)^3$

The probability functions of the two cases are therefore

(a)

x	-3	-1	1	3
p_x	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

(b)

x	-3	-1	1	3
p_x	$\frac{8}{125}$	$\frac{36}{125}$	$\frac{54}{125}$	$\frac{27}{125}$

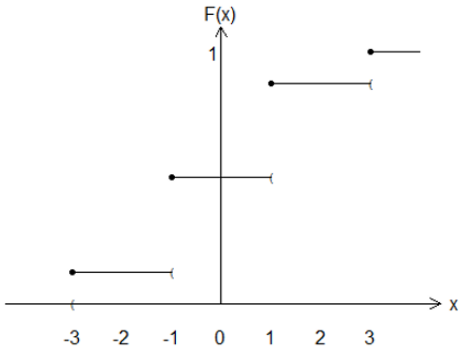
The cumulative distribution functions are

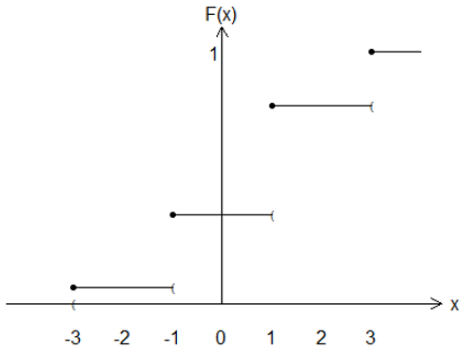
(a)

$$F(x) = \begin{cases} 0 & \text{if } x < -3 \\ \frac{1}{8} & \text{if } -3 \le x < -1 \\ \frac{1}{2} & \text{if } -1 \le x < 1 \\ \frac{7}{8} & \text{if } 1 \le x < 3 \\ 1 & \text{if } x \ge 3. \end{cases}$$

(b)

$$F(x) = \begin{cases} 0 & \text{if } x < -3 \\ \frac{8}{125} & \text{if } -3 \le x < -1 \\ \frac{44}{125} & \text{if } -1 \le x < 1 \\ \frac{98}{125} & \text{if } 1 \le x < 3 \\ 1 & \text{if } x \ge 3. \end{cases}$$





83. $p_1 = \frac{2}{14}$, $p_2 = \frac{3}{14}$, $p_3 = \frac{4}{14}$, $p_4 = \frac{5}{14}$.

$$\begin{aligned} E(X) &= \sum_x xp_x \\ &= 1 \times \frac{2}{14} + 2 \times \frac{3}{14} + 3 \times \frac{4}{14} + 4 \times \frac{5}{14} = \frac{20}{7}. \\ E(X^2) &= 1 \times \frac{2}{14} + 4 \times \frac{3}{14} + 9 \times \frac{4}{14} + 16 \times \frac{5}{14} = \frac{65}{7}. \\ \text{Therefore } \mathrm{Var}(X) &= E(X^2) - [E(X)]^2 \\ &= \frac{65}{7} - \frac{400}{49} = \frac{55}{49}. \end{aligned}$$

84. The probability function of X is

x	1	2	3	4	5	6
p_x	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

$$\begin{aligned} E(X) &= \sum_x xp_x \\ &= \frac{1}{6} \times 1 + \frac{1}{6} \times 2 + \frac{1}{6} \times 3 + \frac{1}{6} \times 4 + \frac{1}{6} \times 5 + \frac{1}{6} \times 6 \\ &= \frac{7}{2}. \\ \mathrm{Var}(X) &= E(X^2) - [E(X)]^2 \\ &= \frac{1}{6} \times 1 + \frac{1}{6} \times 4 + \frac{1}{6} \times 9 + \frac{1}{6} \times 16 + \frac{1}{6} \times 25 + \frac{1}{6} \times 36 - \left(\frac{7}{2}\right)^2 \\ &= \frac{35}{12}. \end{aligned}$$

85. Using the sample space for Question 5, we find that X has probability function

x	1	2	3	4	5	6
p_x	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{5}{36}$	$\frac{7}{36}$	$\frac{9}{36}$	$\frac{11}{36}$

$$E(X) = \frac{1}{36} \times 1 + \frac{3}{36} \times 2 + \frac{5}{36} \times 3 + \frac{7}{36} \times 4 + \frac{9}{36} \times 5 + \frac{11}{36} \times 6$$
$$= \frac{161}{36}.$$

86.

$$E(X) = \sum_x x p_x = \frac{1}{2n+1} \sum_{x=-n}^{n} x = 0.$$
$$E(X^2) = \sum_x x^2 p_x = \frac{1}{2n+1} \sum_{x=-n}^{n} x^2$$
$$= \frac{2}{(2n+1)} \frac{n(n+1)(2n+1)}{6} = \frac{n(n+1)}{3}.$$
$$\text{Therefore } \mathrm{Var}(X) = E(X^2) - [E(X)]^2 = \frac{n(n+1)}{3}.$$
$$\text{If } \mathrm{Var}(X) = 10,$$
$$\text{then } \frac{n(n+1)}{3} = 10$$
$$n^2 + n - 30 = 0.$$
$$\text{Therefore } n = 5 \text{ (rejecting } -6).$$

87. Let X be the number of calls arriving in an hour and let $P(X \geq 15) = p$.
Then Y, the number of times out of 10 that $X \geq 15$, is $B(n, p)$ with $n = 10$ and $p = 1 - 0.98274 = 0.01726$.

$$\text{Therefore } P(Y \geq 2) = 1 - P(Y \leq 1)$$
$$= 1 - ((0.98274)^{10} + 10(0.01726)(0.98274)^9)$$
$$= 0.01223.$$

88. (a) $k \int_0^1 x^2(1-x)\,dx = 1$, which implies that $k = 12$.
(b) $P(0 < X < \frac{1}{2}) = 12 \int_0^{1/2} x^2(1-x)\,dx = \frac{5}{16}$.
(c) The number of observations lying in the interval $(0, \frac{1}{2})$ is binomially distributed with parameters $n = 100$ and $p = \frac{5}{16}$, so that

$$\text{Mean number of observations in } \left(0, \frac{1}{2}\right) = np = 31.25.$$

89. (a)

$$E(X) = \lambda \int_0^\infty x e^{-\lambda x}\,dx$$
$$= [-x e^{-\lambda x}]_0^\infty + \int_0^\infty e^{-\lambda x}\,dx \text{ (integrating by parts)}$$
$$= 0 - \frac{1}{\lambda}[e^{-\lambda x}]_0^\infty = \frac{1}{\lambda}.$$

(b)

$$\begin{aligned} E(X^2) &= \lambda \int_0^\infty x^2 e^{-\lambda x}\, dx \\ &= [-x^2 e^{-\lambda x}]_0^\infty + 2\int_0^\infty x e^{-\lambda x}\, dx \\ &= 0 + \frac{2}{\lambda^2} \text{ (using the first result)} \\ &= \frac{2}{\lambda^2} \end{aligned}$$

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

(c) The mode is $x = 0$ since this value maximises $f(x)$.

(d) The median m is given by

$$\int_0^m \lambda e^{-\lambda x}\, dx = \frac{1}{2}, \text{ which implies that } m = \frac{1}{\lambda}\log(2).$$

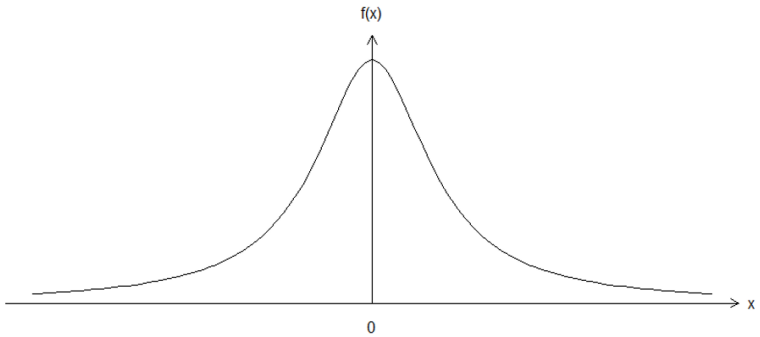
If u and l denote the upper and lower quartiles,

$$\int_0^u \lambda e^{-\lambda x}\, dx = \frac{3}{4} \text{ and } \int_0^l \lambda e^{-\lambda x}\, dx = \frac{1}{4},$$

which implies that $u = \frac{1}{\lambda}\log(4)$ and $l = \frac{1}{\lambda}\log\left(\frac{4}{3}\right)$.
The interquartile range equals

$$\frac{1}{\lambda}\log(4) - \frac{1}{\lambda}\log\left(\frac{4}{3}\right) = \frac{1}{\lambda}\log(3) = \sigma\log(3).$$

90. (a) We must show that $\int_{-\infty}^\infty f(x)\, dx = 1$ and $f(x) \geq 0$ for all x.
Now $\int_{-\infty}^\infty \frac{dx}{\pi(1+x^2)} = \frac{1}{\pi}[\tan^{-1}(x)]_{-\infty}^\infty = \frac{1}{\pi}(\frac{\pi}{2} - (\frac{-\pi}{2})) = 1$.
Also $\frac{1}{\pi(1+x^2)} \geq 0$ for all x.



(b) $F(x) = \frac{1}{\pi}\int_{-\infty}^x \frac{dy}{1+y^2} = \frac{1}{\pi}[\tan^{-1}(y)]_{-\infty}^x = \frac{1}{\pi}(\tan^{-1}(x) + \frac{\pi}{2})$.

(c) $P(-1 \leq X \leq 1) = F(1) - F(-1) = \frac{1}{\pi}(\tan^{-1}(1) + \frac{\pi}{2}) - \frac{1}{\pi}(\tan^{-1}(-1) + \frac{\pi}{2}) = \frac{1}{2}$.

91. We are given that

$$\int_{-\infty}^{\infty} xf(x)\,dx = \int_0^4 (ax + bx^2 + cx^3)\,dx = 8a + \frac{64b}{3} + 64c = E(X) = 2$$

$$\text{and } \int_{-\infty}^{\infty} x^2 f(x)\,dx = \int_0^4 (ax^2 + bx^3 + cx^4)\,dx = \frac{64a}{3} + 64b + \frac{1024c}{5}$$

$$= \text{Var}(X) + [E(X)]^2 = \frac{32}{5}.$$

$$\text{Also } \int_{-\infty}^{\infty} f(x)\,dx = \int_0^4 (a + bx + cx^2)\,dx = 4a + 8b + \frac{64c}{3} = 1.$$

Solving these equations gives $a = \frac{3}{4}, b = -\frac{3}{4}$ and $c = \frac{3}{16}$.
Therefore $f(x) = \frac{3}{16}(x-2)^2, 0 \le x \le 4$.

B.3 Statistics examples

92. [Estimation] A random sample of 10 boys and 10 girls from a large sixth form college were weighed with the following results.

Boy's weight (kg)	77	67	65	60	71	62	67	58	65	81
Girl's weight (kg)	42	57	46	49	64	61	52	50	44	59

Find

(a) unbiased estimates of μ_b and σ_b^2, the mean and variance of the weights of the boys;
(b) unbiased estimates of μ_g and σ_g^2, the mean and variance of the weights of the girls;
(c) an unbiased estimate of $\mu_b - \mu_g$.

Assuming that $\sigma_b^2 = \sigma_g^2 = \sigma^2$, calculate an unbiased estimate of σ^2 using both sets of weights.

93. [Estimation] The time that a customer has to wait for service in a restaurant has the probability density function

$$f(x) = \begin{cases} \frac{3\theta^3}{(x+\theta)^4} & \text{if } x \ge 0 \\ 0 & \text{otherwise,} \end{cases}$$

where θ is an unknown positive constant. Let $X_1, X_2, \ldots, X_n$ denote a random sample from this distribution. Show that

$$\hat{\theta} = \frac{2}{n}\sum_{i=1}^{n} X_i$$

is an unbiased estimator for θ. Find the standard error of $\hat{\theta}$.

94. [Confidence interval] In an experiment, 100 observations were taken from a normal distribution with variance 16. The experimenter quoted $[1.545, 2.861]$ as the confidence interval for μ. What level of confidence was used?

95. [Confidence interval] At the end of a severe winter a certain insurance company found that of 972 policy holders living in a large city who had insured their homes with the company, 357 had suffered more than £500-worth of snow and frost damage. Calculate an approximate 95% confidence interval for the proportion of all homeowners in the city who suffered more than £500-worth of damage. State any assumptions that you make.

96. [Confidence interval] The heights of n randomly selected seven-year-old children were measured. The sample mean and standard deviation were found to be 121 cm and 5 cm respectively. Assuming that height is normally distributed, calculate the following confidence intervals for the mean height of seven-year-old children:

    (a) 90% with $n = 16$,

    (b) 99% with $n = 16$,

    (c) 95% with $n = 16, 25, 100, 225, 400$.

97. [Confidence interval] A random variable is known to be normally distributed, but its mean μ and variance σ^2 are unknown. A 95% confidence interval for μ based on 9 observations was found to be $[22.4, 25.6]$. Calculate unbiased estimates of μ and σ^2.

98. [Confidence interval] The wavelength of radiation from a certain source is 1.372 microns. The following 10 independent measurements of the wavelength were obtained using a measuring device:

$$1.359, 1.368, 1.360, 1.374, 1.375, 1.372, 1.362, 1.372, 1.363, 1.371.$$

    Assuming that the measurements are normally distributed, calculate 95% confidence limits for the mean error in measurements obtained with this device and comment on your result.

99. [Confidence interval] In five independent attempts, a girl completed a Rubik's cube in 135.4, 152.1, 146.7, 143.5 and 146.0 seconds. In five further attempts, made two weeks later, she completed the cube in 133.1, 126.9, 129.0, 139.6 and 144.0 seconds. Find a 90% confidence interval for the change in the mean time taken to complete the cube. State your assumptions.

100. [Confidence interval] In an experiment to study the effect of a certain concentration of insulin on blood glucose levels in rats, each member of a random sample of 10 rats was treated with insulin. The blood glucose level of each rat was measured both before and after treatment. The results, in suitable units, were as follows.

Rat	1	2	3	4	5	6	7	8	9	10
Level before	2.30	2.01	1.92	1.89	2.15	1.93	2.32	1.98	2.21	1.78
Level after	1.98	1.85	2.10	1.78	1.93	1.93	1.85	1.67	1.72	1.90

    Let μ_1 and μ_2 denote respectively the mean blood glucose levels of a randomly selected rat before and after treatment with insulin. By considering the differences of the measurements on each rat and assuming that they are normally distributed, find a 95% confidence interval for $\mu_1 - \mu_2$.

101. [Confidence interval] The heights (in metres) of 10 fifteen-year-old boys were as follows:

$$1.59, 1.67, 1.55, 1.63, 1.69, 1.58, 1.66, 1.62, 1.64, 1.61.$$

Assuming that heights are normally distributed, find a 99% confidence interval for the mean height of fifteen-year-old boys.

If you were told that the true mean height of boys of this age was 1.67 m, what would you conclude?

B.4 Solutions: Statistics examples

92. (a) An unbiased estimate of μ_b is given by the mean weight of the boys,

$$\hat{\mu}_b = \frac{1}{10}(77 + 67 + \ldots + 81) = 67.3.$$

An unbiased estimate of σ_b^2 is the sample variance of the weights of the boys,

$$\hat{\sigma}_b^2 = ((77^2 + 67^2 + \ldots + 81^2) - 10\hat{\mu}_b^2)/9 = 52.6\dot{7}.$$

(b) Similarly, unbiased estimates of μ_g and σ_g^2 are

$$\hat{\mu}_g = \frac{1}{10}(42 + 57 + \ldots + 59) = 52.4,$$
$$\hat{\sigma}_g^2 = ((42^2 + 57^2 + \ldots + 59^2) - 10\hat{\mu}_g^2)/9 = 56.7\dot{1}.$$

(c) An unbiased estimate of $\mu_b - \mu_g$ is

$$\hat{\mu}_b - \hat{\mu}_g = 67.3 - 52.4 = 14.9.$$

$E(\hat{\sigma}_b^2) = E(\hat{\sigma}_g^2) = \sigma^2$ and so

$$E\left(\frac{\hat{\sigma}_b^2 + \hat{\sigma}_g^2}{2}\right) = \sigma^2.$$

Therefore an unbiased estimate of σ^2 which uses both sets of weights is

$$\begin{aligned}\frac{1}{2}(\hat{\sigma}_b^2 + \hat{\sigma}_g^2) &= \frac{1}{2}(52.6\dot{7} + 56.7\dot{1}) \\ &= 54.69\dot{4}.\end{aligned}$$

93. The mean of the distribution is

$$\begin{aligned} E(X) &= \int_0^\infty \frac{3\theta^3 x}{(x+\theta)^4}\,dx \\ &= 3\theta^3 \int_\theta^\infty \frac{y-\theta}{y^4}\,dy \text{ (where } y = x+\theta) \\ &= 3\theta^3 \left[-\frac{1}{2y^2} + \frac{\theta}{3y^3} \right]_\theta^\infty \\ &= 3\theta^3 \left(\frac{1}{2\theta^2} - \frac{1}{3\theta^2} \right) = \theta/2 \end{aligned}$$

Hence $E(X_i) = \theta/2, i = 1, 2, \ldots, n$, and $E(\hat{\theta}) = \theta$.
Thus $\hat{\theta}$ is an unbiased estimator for θ.

$$\begin{aligned} E(X^2) = \int_0^\infty \frac{3\theta^3 x^2}{(x+\theta)^4}\,dx &= 3\theta^3 \int_\theta^\infty \frac{(y-\theta)^2}{y^4}\,dy \\ &= 3\theta^3 \left[-\frac{1}{y} + \frac{\theta}{y^2} - \frac{\theta^2}{3y^3} \right]_\theta^\infty \\ &= \theta^2. \end{aligned}$$

So $\text{Var}(X) = \theta^2 - (\theta/2)^2 = 3\theta^2/4$.
Hence $\text{Var}(\hat{\theta}) = 4\text{Var}(\bar{X}) = 3\theta^2/n$ and $SE(\hat{\theta}) = \theta\sqrt{3/n}$.

94. The $100(1-\alpha)\%$ symmetric confidence interval is

$$\left[\bar{x} - z_\gamma \times \frac{4}{10}, \bar{x} + z_\gamma \times \frac{4}{10} \right] \; (\gamma = 1 - \frac{\alpha}{2})$$

where z_γ is the 100γ percentile of the standard normal distribution. The width of the CI is $0.8z_\gamma$. The width of the quoted confidence interval is 1.316. Therefore, assuming that the quoted interval is symmetric,

$$0.8z_\gamma = 1.316 \Rightarrow z_\gamma = 1.645 \Rightarrow \gamma = 0.95 \text{ (`pnorm(1.645)` in R)}.$$

This implies that $\alpha = 0.1$ and hence $100(1-\alpha) = 90$, i.e. the confidence level is 90%.

95. Assuming that although the 972 homeowners are all insured within the same company they constitute a random sample from the population of all homeowners in the city, the 95% interval is given approximately by

$$\left[\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right],$$

where $n = 972$ and $\hat{p} = 357/972$. The interval is therefore $[0.337, 0.398]$.

96. The $100(1-\alpha)\%$ confidence interval is, in the usual notation,

$$\left[\bar{x} \pm \text{ critical value } \frac{s}{\sqrt{n}}\right],$$

where the critical value is the $100(1-\alpha/2)th$ percentile of the t-distribution with $n-1$ degrees of freedom. Here $\bar{x} = 121$ and $s = 5$.

(a) For the 95% CI, critical value = 1.753 (`qt(0.95,df=15)` in `R`) and the interval is

$$[121 - 1.753 \times 5/4, 121 + 1.753 \times 5/4],$$

i.e. $[118.81, 123.19]$.

(b) For the 99% CI, critical value = 2.947 (`qt(0.995,df=15)` in `R`) and the interval is $[117.32, 124.68]$.

(c) We obtain the following table

n	`R` command	$t_{0.975}(n-1)$	Confidence Limits	
16	`qt(0.975,df=15)`	2.131	118.34	123.66
25	`qt(0.975,df=24)`	2.064	118.94	123.06
100	`qt(0.975,df=99)`	1.984	120.01	121.99
225	`qt(0.975,df=224)`	1.971	120.34	121.66
400	`qt(0.975,df=399)`	1.966	120.51	121.49

97. For the 95% the critical value is As 2.306 (`qt(0.975,df=8)` in `R`), the interval is

$$\left[\bar{x} - 2.306 \times \frac{s}{3}, \bar{x} + 2.306 \times \frac{s}{3}\right].$$

The midpoint is $\bar{x}$ and therefore

$$\bar{x} = \frac{22.4 + 25.6}{2} = 24.0.$$

This is an unbiased estimate of μ.
Also

$$\frac{2.306s}{3} = \frac{25.6 - 22.4}{2} = 1.6.$$

Hence $s = 2.0815$ so that $s^2 = 4.333$. This is an unbiased estimate of σ^2.

98. In the usual notation,

$$\sum x_i = 13.676, \ \sum x_i^2 = 18.703628.$$

These lead to

$$\bar{x} = 1.3676, \ s = 0.00606.$$

Also, for the 95% CI, critical value = 2.262 (`qt(0.975,df=9)` in `R`).

Thus a 95% confidence interval for the mean is

$$\left[1.3676 - 2.262 \times \frac{0.00606}{\sqrt{10}}, 1.3676 + 2.262 \times \frac{0.00606}{\sqrt{10}}\right],$$

i.e. $[1.3633, 1.3719]$.

A 95% confidence interval for the mean error is obtained by subtracting the true wavelength of 1.372 from each endpoint. This gives $[-0.0087, -0.0001]$. As this contains negative values only, we conclude that the device tends to underestimate the true value.

99. Using x to refer to the early attempts and y to refer to the later ones, we find from the data that

$$\sum x_i = 723.7, \ \sum x_i^2 = 104896.71,$$
$$\sum y_i = 672.6, \ \sum y_i^2 = 90684.38.$$

This gives

$$\bar{x} = 144.74, \ s_x^2 = 37.093,$$
$$\bar{y} = 134.52, \ s_y^2 = 51.557.$$

Confidence limits for the change in mean time are

$$\bar{y} - \bar{x} \pm \text{ critical value } \sqrt{\frac{4s_x^2 + 4s_y^2}{8}\left(\frac{1}{5} + \frac{1}{5}\right)},$$

leading to the interval $[-18.05, -2.39]$, as critical value $= 1.860$ (`qt(0.95,df=8)` in `R`).

As it contains only negative values, this suggests that there is a real decrease in the mean time taken to complete the cube. We have assumed that the two samples are independent random samples from normal distributions of equal variance.

100. Let $d_1, d_2, \ldots, d_{10}$ denote the differences in levels before and after treatment. Their values are

$$0.32, 0.16, -0.18, 0.11, 0.22, 0.00, 0.47, 0.31, 0.49, -0.12.$$

Then $\sum_{i=1}^{10} d_i = 1.78$ and $\sum_{i=1}^{10} d_i^2 = 0.7924$ so that $\bar{d} = 0.178$, $s_d = 0.2299$.

A 95% confidence interval for the mean difference $\mu_1 - \mu_2$ is

$$\left[\bar{d} \pm \text{ critical value } \frac{s_d}{\sqrt{10}}\right],$$

i.e. $\left[0.178 - 2.262 \times \frac{0.2299}{\sqrt{10}}, 0.178 + 2.262 \times \frac{0.2299}{\sqrt{10}}\right]$ or $[0.014, 0.342]$, as critical value $= 2.262$ (`qt(0.975,df=9)` in `R`).

Note that the two samples are not independent. Thus the standard method of finding a confidence interval for $\mu_1 - \mu_2$, as used in Question 9 for example, would be inappropriate.

101. The mean of the heights is 1.624 and the standard deviation is 0.04326. A 99% confidence interval for the mean height is therefore

$$\left[1.624 - 3.250 \times \frac{0.04326}{\sqrt{10}}, 1.624 + 3.250 \times \frac{0.04326}{\sqrt{10}}\right],$$

i.e. $[1.580, 1.668]$, as critical value $= 3.250$ (`qt(0.995,df=9)` in `R`).

If we were told that the true mean height was 1.67 m then, discounting the possibility that this information is false, we would conclude that our sample is not a random sample from the population of all fifteen-year-old boys or that we have such a sample but an event with probability 0.01 has occurred, namely that the 99% confidence interval does not contain the true mean height.

Appendix C

Homework Sheets

Homework Sheet 1

1. [**3 marks**] Suppose we have the data: $x_1 = 1,\ x_2 = 2, \ldots, x_n = n$. Find the mean and the variance. For variance use the divisor n instead of $n-1$.

   For questions 2, 3, and 4, use the summation symbol, $\sum$. For example, for any sequence of numbers $z_1,\ z_2,\ \ldots,\ z_n$,

   $$\sum_{i=1}^{n} z_i = z_1 + z_2 + \cdots + z_n.$$

   For example, when $n = 4$ and $z_1 = 1, z_2 = 4$ and $z_3 = 9$, and $z_4 = 16$, we have $\sum_{i=1}^{n} z_i = 1 + 4 + 9 + 16 = 30$. Do not write the $\sum$ symbol on its own; always write the limits, e.g. $\sum_{i=1}^{n}$.

2. [**3 marks**] Suppose $y_i = ax_i + b$ for $i = 1, \ldots, n$ where a and b are real numbers. Show that:

   (a) $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i = a\bar{x} + b$ and

   (b) $\mathrm{Var}(y) = a^2\mathrm{Var}(x)$ where $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

   and for variance you may either use the divisor n, i.e. $\mathrm{Var}(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$ or $n-1$, i.e. $\mathrm{Var}(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$. The divisor does not matter as the results hold regardless. **Hint:** For part (b), start with the left hand side, $\mathrm{Var}(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$ and substitute y_i and $\bar{y}$ in terms of x_i and $\bar{x}$.

3. [**1 mark**] Suppose $a \le x_i \le b$ for $i = 1, \ldots, n$. Show that $a \le \bar{x} \le b$.

4. [**3 marks**] Prove that for any set of numbers $x_1,\ x_2, \ldots, x_n$,

   $$\left(x_1^2 + x_2^2 + \cdots + x_n^2\right) \ge \frac{(x_1 + x_2 + \ldots x_n)^2}{n},$$

i.e. sum of squares of n numbers is greater than equal to the square of the sum divided by n. **Hint:** You may start by assuming $\sum_{i=1}^{n}(x_i - \overline{x})^2 \geq 0$ and then expand the square within the summation symbol. This is a version of the famous and very important Cauchy-Schwartz inequality in Mathematics.

5. Read the computer failure data by issuing the command `cfail <- scan("compfail.txt")`. [Before issuing the command, make sure that you have set the correct working directory so that `R` can find this data file.]

   (a) [**2 marks**] Describe the data using the commands `summary`, `var` and `table`. Also obtain an histogram (`hist`) plot and a boxplot (`boxplot`) of the data and write a brief comment on the shape of the distribution.

   (b) [**2 marks**] Use `plot(cfail)` to produce a plot of the failure numbers (y-axis) against the order in which the data were collected (x-axis). Do you detect any trend? **Hint:** You may modify the command to `plot(cfail, type="l")` to perhaps get a better looking plot. The argument is lower case letter "l" not the number 1.

   (c) [**2 marks**] The object `cfail` is a vector of length 104. The first 52 values are for the weeks in the first year and the last 52 values are for the second year. Find the two annual means and variances.

   Hint: To do this we need to create an index corresponding to each of the 104 values. The index should take the value 1 for the first 52 values (in the first year) and 2 for the remaining values in the second year. Use the `rep` (for replicate) command to generate this. Issue `year <- rep(c(1, 2), each=52).` This will replicate 1, 52 times followed by 2, 52 times and put the resulting vector in year. Type `cbind(year, cfail)` and see what happens. Now you can calculate the year-wise means by issuing the command: `tapply(X=cfail, INDEX=year, FUN=mean)`. The function `tapply` applies the supplied FUN argument to the X argument separately for each unique index supplied by the INDEX argument. That is, the command returns the mean value for each group of observations corresponding to distinct values of the INDEX variable `year`.

6. Fortune magazine publishes a list of the world's billionaires each year. The 1992 list includes 225 individuals. Their wealth, age, and geographic location (**A**sia, **E**urope, **M**iddle East, **U**nited States, and **O**ther) are reported. Variables are: wealth: Wealth of family or individual in billions of dollars; age: Age in years (for families it is the maximum age of family members); region: Region of the World (Asia, Europe, Middle East, United States and Other). The head and tail values of the data set are given below.

wealth	age	region
37.0	50	M
24.0	88	U
⋮	⋮	⋮
1	9	M
1	59	E

Read the data by issuing the following command.

`bill <- read.table("billionaires.txt", head=T)`.

(a) [**2 marks**] Obtain the summary of the data set and comment on each of the three columns of data.

(b) [**2 marks**] Obtain the mean and variance of the wealth for each of the 5 regions of the world and comment.

Hint: You can use the `tapply(X=bill$wealth, INDEX=bill$region, FUN=mean)` command to do this.

(c) [**2 marks**] Produce a set of boxplots to demonstrate the difference in distribution of wealth according to different parts of the world and comment on it.

Hint: You may issue the command: `boxplot(wealth` $\sim$ `region, data =bill)`

(d) [**3 marks**] Produce a scatter plot of the wealth against age and comment on the relationship between them. Do you see any outlying observations? Do you think a linear relationship between age and wealth will be sensible?

Try and ensure all your plots have informative titles and axis labelling – The function `title` and the function arguments `xlab` and `ylab`, for example `ylab="wealth in billions of US dollars"` are helpful for this.

Homework Sheet 2

1. [**1 mark**] I select two cards from a pack of 52 cards and observe the colour of each. Which of the following is an appropriate sample space S for the possible outcomes?

   (a) S = {red, black}

   (b) S = {(red, red), (red, black), (black, red), (black, black)}

   (c) S = {0, 1, 2}

   (d) None of the above

2. [**1 mark**] A builder requires the services of both a plumber and an electrician. If there are 12 plumbers and 9 electricians available in the area, the number of choices of the pair of contractors is

   (a) 96

   (b) 88

   (c) 108

   (d) none of the above

3. [**1 mark**] In a particular game, a fair die is rolled. If the number of spots showing is either 4 or 5 you win £1, if the number of spots showing is 6 you win £4, and if the number of spots showing is 1, 2 or 3 you win nothing. If it costs you £1 to play the game, the probability that you win more than the cost of playing is

   (a) 0

   (b) 1/6

   (c) 1/3

   (d) 2/3

4. A rental car company has 10 Japanese and 15 European cars waiting to be serviced on a particular Saturday morning. Because there are so few mechanics available, only 6 cars can be serviced. Calculate

   (a) [**1 mark**] the number of outcomes in the sample space (i.e. the number of possible selections of 6 cars).

   (b) [**1 mark**] the number of outcomes in the event "3 of the selected cars are Japanese and the other 3 are European".

   (c) [**1 mark**] the probability (to 3 decimal places) that the event in (b) occurs, when the 6 cars are chosen at random.

5. Shortly after being put into service, some buses manufactured by a certain company have developed cracks on the underside of the main frame. Suppose a particular city has 25 of these buses, and cracks have actually appeared in 8 of them.

   (a) [**1 mark**] How many ways are there to select a sample of 5 buses from the 25 for a thorough inspection?

   (b) [**1 mark**] How many of these samples of 5 buses contain exactly 4 with visible cracks?

   (c) [**1 mark**] If a sample of 5 buses is chosen at random, what is the probability that exactly 4 of the 5 will have visible cracks (to 3 dp)?

   (d) [**3 marks**] If buses are selected as in part (c), what is the probability that at least 4 of those selected will have visible cracks (to 3 dp)?

6. [**4 marks**] The probability that a man reads The Sun is 0.6, while the probability that he reads both The Sun and The Guardian is 0.1 and the probability that he reads neither paper is 0.2. What is the probability that he reads The Guardian? Draw a Venn diagram to illustrate your answer.

7. [**5 + 4 marks**] Write down all 27 ways in which three distinguishable balls A, B and C can be distributed among three boxes. If each ball were to be placed into a box selected at random, what probabilities would you assign to each way?

   Show that, if the balls are made indistinguishable, the number of distinct ways reduces to 10. If again each ball were to be placed into a box selected at random, what probabilities would you assign to the 10 ways?

   Hint: For the second part, replace both B and C by A in the long table of 27 allocations. Then find out which rows of the table are distinct and for each distinct row count the identical rows.

Homework Sheet 3

1. [**1 mark**] Event A occurs with probability 0.8. The conditional probability that event B occurs given that A occurs is 0.5. The probability that both A and B occur is

   (a) 0.3,

   (b) 0.4,

   (c) 0.8,

   (d) cannot be determined from the information given

2. [**1 mark**] An event A occurs with probability 0.5. An event B occurs with probability 0.6. The probability that both A and B occurs is 0.1. The conditional probability of A given B is

   (a) 0.3,

   (b) 0.2,

   (c) 1/6,

   (d) cannot be determined from the information given

   You only need to give the letters corresponding to the correct answers in questions 1 and 2.

3. [**3 marks**] In a multiple choice examination paper, n answers are given for each question ($n = 4$ in the above two questions). Suppose a candidate knows the answer to a question with probability p. If he does not know it he guesses, choosing one of the answers at random. Use the Bayes Theorem to show that the probability that he knew the answer given that his answer is correct, is $\frac{np}{1+(n-1)p}$.

4. [**4 marks**] Two events A and B are such that the probability of B occurring given that A has occurred is 4 times the probability of A, and the probability that A occurs given that B has occurred is 9 times the probability that B occurs. If the probability that at least one of the events occurs is 7/48, find the probability of event A occurring.

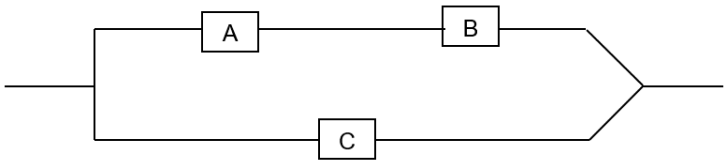
5. Recall the definition of the independence of two events A and B.

   (a) [**2 marks**] Show that if $P(A) = 1$, then $P(A \cap B) = P(B)$.

   (b) [**2 marks**] Prove that any event A with $P(A) = 0$ or $P(A) = 1$ is independent of every event B.

   (c) [**3 marks**] Prove that if the events A and B are independent, then every pair of events A' and B; A and B'; A' and B' are independent.

6. [**3 marks**] A water supply system has three pumps A, B and C arranged as below, where the pumps operate independently.



   The system functions if either A and B operate or C operates, or all three operate. If A and B have reliabilities 90% and C has reliability 95%, what is the reliability of the water supply system?

7. (a) [**1 mark**] A twin-engine aircraft can fly with at least one engine working. If each engine has a probability of failure during flight of q and the engines operate independently, what is the probability of a successful flight?

   (b) [**2 marks**] A four-engine aircraft can fly with at least two of its engines working. If all the engines operate independently and each has a probability of failure of q, what is the probability that a flight will be successful?

   (c) [**2 marks**] For what range of values of q would you prefer to fly in a twin-engine aircraft rather than a four-engine aircraft?

   (d) [**1 mark**] Discuss briefly (in no more than two sentences) whether the assumption of independent operation of the engines is reasonable.

Homework Sheet 4

1. [**2 marks**] The random variable X has probability function

$$p_x = \begin{cases} \frac{1}{14}(1+x) & \text{if } x = 1, 2, 3, 4 \\ 0 & \text{otherwise.} \end{cases}$$

Find the mean and variance of X.

Hint: This is a discrete random variable taking only four values. Hence use summation rather than integration to find expectation and variance.

2. [**4 marks**] Suppose that X is a random variable and that $E(X) = \mu$. Show that, for any constant a,

$$E[(X-a)^2] = Var(X) + (\mu - a)^2.$$

What value of a will minimise $E[(X-a)^2]$?

Hint: Start with the left hand side, write $X - a = X - \mu + \mu - a$ and then expand the square and take expectations. Use properties of expectations discussed in Lecture 12.

3. In a certain district of a city, the proportion of people who vote Conservative is 0.3. Let a random sample of 100 voters be selected from the district.

   (a) [**2 marks**] Calculate the mean, μ and variance, σ^2 of the number of people in the sample who vote Conservative.

   (b) [**2 marks**] If X denotes the number in the sample who vote Conservative and μ and σ are the mean and standard deviation respectively as calculated in (a), use the `pbinom` or the `dbinom` in `R` to evaluate

      i. $P(\mu - \sigma \leq X \leq \mu + \sigma)$,

      ii. $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma)$.

4. Suppose that over a period of several years the average annual number of deaths from a certain non-contagious disease has been 10. If the number of deaths follows a Poisson model, what is the probability that

   (a) [**1 mark**] seven people will die from this disease this year,

   (b) [**1 mark**] 10 or more people will die from this disease this year,

   (c) [**1 mark**] there will be no deaths from this disease this year?

5. [**2 + 2 marks**] The performance of a computer is studied over a period of time, during which it makes 10^{15} computations. If the probability that a particular computation is incorrect equals 10^{-14}, independent of other computations, write down an expression for the probability that fewer than 10 errors occur and calculate an approximation to its value using `R`.

   Hint: For the second part, use the Poisson approximation to the binomial distribution.

6. The random variable X has probability density function

$$f(x) = \begin{cases} kx^2(1-x) & \text{if } 0 \leq x \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

   (a) [**2 marks**] Find the value of k.

   (b) [**2 marks**] Find the probability that X lies in the range $(0, \frac{1}{2})$.

   (c) [**4 marks**] In 100 independent observations of X, how many on average will fall in the range $(0, \frac{1}{2})$?

Homework Sheet 5

1. If X has the distribution $N(4, 16)$, find

   (a) [**2 marks**] the number a such that $P(|X - 4| \leq a) = 0.95$,

   (b) [**1 marks**] the median,

   (c) [**1 mark**] the 90th percentile,

   (d) [**2 marks**] the interquartile range.

2. [**4 marks**] The fractional parts of 100 numbers are distributed uniformly between 0 and 1. The numbers are first rounded to the nearest integer and then added. Using the CLT, find an approximation for the probability that the error in the sum due to rounding lies between –0.5 and 0.5. **Hint:** First argue that the individual errors are uniformly distributed and find the mean and variance of the uniform distribution.

3. A man buys a new die and throws it 600 times. He does not yet know if the die is fair!

   (a) [**2 marks**] Show that the probability that he obtains between 90 and 100 'sixes' if the die is fair is 0.3968 approximately.

   (b) [**3 marks**] From the first part we see that $P(90 \leq X \leq 100) \approx 0.3968$ where X denotes the number of 'sixes' from 600 throws. If the die is fair we can expect about 100 'sixes' from 600 throws. Between what two limits symmetrically placed about 100 would the number sixes obtained lie with probability 0.95 if the die is fair? That is, find N such that $P(100 - N \leq X \leq 100 + N) = 0.95$. You may use the continuity correction of $\frac{1}{2}$ on each of the limits inside the probability statement.

   (c) [**2 marks**] What might he conclude if he obtained 120 sixes?

4. The random variables X, Y have the following joint probability table.

		Y		
		1	2	3
X	1	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{16}$
	2	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$
	3	$\frac{1}{16}$	$\frac{1}{4}$	$\frac{1}{16}$

   (a) [**1 mark**] Find the marginal probability functions of X and Y.

   (b) [**1 mark**] Show that $P(X = 1, Y = 2) = P(X = 1)P(Y = 2)$. Can it be concluded that X and Y are independent?

5. Two random variables X, Y each take only the values 0 and 1, and their joint probability table is as follows.

		Y	
		0	1
X	0	a	b
	1	c	d

   (a) [**2 marks**] Show that $Cov(X, Y) = ad - bc$.

   (b) [**2 marks**] Use the definition of independence to verify that $ad = bc$ when X and Y are independent.

   (c) [**2 marks**] Show that X and Y are independent when $ad = bc$.

Homework Sheet 6

1. [**6 marks**] Show that if $X_1, X_2, \ldots X_n$ is a random sample from a distribution having mean μ and variance σ^2, then

$$Y = \sum_{i=1}^{n} a_i X_i = a_1 X_1 + a_2 X_2 + \cdots + a_n X_n$$

is an unbiased estimator for μ provided $\sum_{i=1}^{n} a_i = 1$, and find an expression for $Var(Y)$.

Hint: See notes for Lecture 19.

[**4 marks**] In the case $n = 4$, determine which of the following estimators are unbiased for μ:

(a) $Y_1 = (X_1 + X_2 + X_3 + X_4)/4$,
(b) $Y_2 = X_1 - X_2 + X_3 - X_4$,
(c) $Y_3 = (2X_1 + 3X_4)/5$,
(d) $Y_4 = (X_1 + X_3)/2$.

Hint: Find the expectation of each Y_i and see which one equals μ.

[**2 marks**] Which is the most efficient of those estimators which are unbiased?

Hint: The most efficient estimator is the one with the least variance.

2. [**3 marks**] In an experiment, 100 observations were taken from a normal distribution with variance 16. The experimenter quoted $[1.545, 2.861]$ as the confidence interval for μ. What level of confidence was used?

Hint: The variance is known and the distribution is normal. Hence you will have to use the normal distribution to obtain confidence intervals. See Lecture 24.

3. The heights of n randomly selected seven-year-old children were measured. The sample mean and standard deviation were found to be 121 cm and 5 cm respectively. Assuming that height is normally distributed, calculate the following confidence intervals for the mean height of seven-year-old children:

   (a) [**2 marks**] 90% with $n = 16$,

   (b) [**2 marks**] 99% with $n = 16$,

   (c) [**2 marks**] 95% with $n = 16, 25, 100, 225, 400$.

   Hint: You will have to use the t-distribution based intervals; see notes for Lecture 26.

4. [**4 marks**] The random variable X has a normal distribution with standard deviation 3.5 but unknown mean μ. The hypothesis $H_0 : \mu = 3$ is to be tested against $H_1 : \mu = 4$ by taking a random sample of size 50 from the distribution of X and rejecting H_0 if the sample mean exceeds 3.4. Calculate the Type I and Type II error probabilities of this procedure.

   Hint: See notes for Lecture 27.

Homework Sheet 7

1. [**5 marks**] A wholesale greengrocer decides to buy a field of winter cabbages if he can convince himself that their mean weight exceeds 1.2 kg. Accordingly he cuts 12 cabbages at random and weighs them with the following results (in kg):

   $$1.26, 1.19, 1.17, 1.24, 1.23, 1.25, 1.20, 1.18, 1.23, 1.21, 1.19, 1.17.$$

   Should the greengrocer buy the cabbages? Use a 10% level of significance.

   Hint: Let μ denote the true mean of the cabbages. Test the hypotheses $H_0 : \mu \leq 1.2; H_1 : \mu > 1.2$. See the fast food waiting time example in Lecture 28.

2. [**7 marks**] A market gardener, wishing to compare the effectiveness of two fertilizers, used one fertilizer throughout the growing season on half of his plants and used the second one on the other half. The yields of the surviving plants are shown below, measured in kilograms.

Fertilizer A	6.72,	9.17,	7.43,	6.53,	10.20,	6.88,	6.28,	6.73,
	7.38,	6.95,	7.03,	8.05,	7.57,	6.90,	7.63.	
Fertilizer B	7.73,	10.34,	8.59,	5.71,	5.42,	7.94,	8.29,	
	7.63,	9.41,	9.35,	9.62,	8.94.			

   Stating all assumptions made, test at a 10% significance level the hypothesis that the fertilizers are equally effective against the alternative that they are not.

   Hint: Two sample t-test. See notes for Lecture 29.

3. [**7 marks**] Eight young English county cricket batsmen were awarded scholarships which enabled them to spend the winter in Australia playing club cricket. Their first-class batting averages in the preceding and following seasons were as follows.

Batsman	1	2	3	4	5	6	7	8
Average before	29.43	21.21	31.23	36.27	22.28	30.06	27.60	43.19
Average after	31.26	24.95	29.74	33.43	28.50	30.35	29.16	47.24

   Is there a significant improvement in their batting averages between seasons? Could any change be attributed to the winter practice?

   Hint: Paired t-test, see Lecture 29. The alternative hypothesis is one-sided.

4. [**6 marks**] In an experiment to study the effect of a certain concentration of insulin on blood glucose levels in rats, each member of a random sample of 10 rats was treated with insulin. The blood glucose level of each rat was measured both before and after treatment. The results, in suitable units, were as follows.

Rat	1	2	3	4	5	6	7	8	9	10
Level before	2.30	2.01	1.92	1.89	2.15	1.93	2.32	1.98	2.21	1.78
Level after	1.98	1.85	2.10	1.78	1.93	1.93	1.85	1.67	1.72	1.90

   Find a 95% confidence interval for the mean difference stating all the assumptions.

Prof Sujit Sahu studied statistics and mathematics at the University of Calcutta and the Indian Statistical Institute and went on to obtain his PhD at the University of Connecticut (USA) in 1994. He joined the University of Southampton in 1999. His research area is Bayesian statistical modelling and computation.



Acknowledgements

These notes are taken largely from the MATH1024 notes developed over the years by many Southampton colleagues, most recently by Prof Dankmar Böhning and Dr Vesna Perisic.

Miss Joanne Ellison helped enormously in typesetting and proofreading these notes.

Many worked examples were taken from a series of books written by F.D.J. Dunstan, A.B.J. Nix, J.F. Reynolds and R.J. Rowlands, published by R.N.D. Publications.



Mathematical
Sciences Lecture Note Series