MATH5315M: Applied probability & statistics

The aim of this module is to learn the most important concepts in probability theory and statistics, and to introduce and develop a number of methods in statistics usually used in applied financial analysis. On completion of the module, you will be expected to be able to communicate the theoretical and applied concepts of probability and statistics, and to carry out statistical tests and interpret the findings. The material can be split into three sections: fundamentals of probability and statistics, simple statistical analyses and time series methods.

The notes in this document are not exhaustive and should be supplemented with additional information from both the lectures and the recommended reading. Each chapter of the notes has been linked with chapters from the recommended reading to help with this. Throughout the notes there are examples that will be discussed in the lectures, and you will be expected to take notes covering the examples. Additional examples not in these notes will also be explained in the lectures.

In the notes, the mention of the implementation of the techniques in a computer-based statistical package has been kept to a minimum. However, you must bear in mind that we will be investigating how we can run these analyses on specific datasets. In the second half of the module, some of our focus will be on implementation in a statistical package called R. We use R because it

- is rich enough to handle all the techniques we cover in this module
- can be used to do more advanced analyses that you will be facing later in the programme
- is free and available on many operating systems, and
- is versatile enough for new methods to be easily created (which might be important in your future careers).

Martín López-García (September, 2019)

1 Data Summaries - (CT3 Unit 1)

Data exist in many different forms, and we need to understand the differences to perform useful statistical analyses. Often, simple statistical summaries and well-made graphs can remove the need for complicated (and sometimes expensive) analysis.

1.1 Types of Data

Batch data are a set of related observations.
Sample data are a set of observations that are chosen to be representative of some population.

Numerical (quantitative) - Discrete $\{0, 1, 2, 3, ...\}$ or $\{\frac{1}{2}, 1, 2, 4\}$
Continuous, e.g. height
Categorical (qualitative) - Attribute (dichotomous or binary, e.g., yes/no)
Nominal (e.g. colour) - “no order”
Ordinal (e.g. agree, disagree, don’t care...)

1.2 Data summaries

There many ways to display data: bar charts, 3D bar charts, pie charts, grouped frequency tables, histograms, stem and leaf diagrams, dot and line plots, line plots and box plots (to name some common types).

Example 1

Dataset: 18, 18, 19, 22, 22, 23, 25, 25, 25, 28, 30, 31, 33, 37, 41.
Stem and leaf diagram of ages

```
1|889
2|2235558
3|0137
4|1
```

Key: 3|2 denotes 32 years old.

Example 2

Dataset: 18, 18, 19, 22, 22, 23, 25, 25, 25, 28, 30, 31, 33, 37, 41.
Show bar chart using same intervals and talk about relative frequency.

Interval	Count	Relative frequency
15-	3	0.6
20-	3	0.6
25-	4	0.8
30-	4	0.4
40-50	1	0.1

Draw histogram.

Bimodality - *when data has 2 modes.*
Multimodality - *when data has more than 2 modes.*

We can also make numerical (rather than graphical) summaries of a dataset.

For a set of n observations, $x_1, x_2, ..., x_n$, the **mean** is defined by:

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

The **median** is the 'middle' observation in a data set which has been sorted. If n is odd, it is the $\frac{n+1}{2}^{\text{th}}$ observation. If n is even, it is the mid point between $\frac{n}{2}^{\text{th}}$ and $\frac{n+2}{2}^{\text{th}}$ observation.

The **mode** of a dataset is the value that occurs most frequently.

The mean, median and mode are all measures of *location.*

The **range** is the difference between the largest and smallest values in a data set:

$$\text{R} = \max_i(x_i) - \min_i(x_i).$$

The **interquartile range** is the difference between the upper and lower quartiles: IQR = UQ - LQ, where the lower quartile is the $25th$ percentile, and the upper quartile is the $75th$ precentile (see below).

A box plot is a useful visualisation tool that shows the LQ, IQR and UQ.

Percentiles extend the idea of quartiles. For example, the 5^{th} percentile is the value such that 5% of the data falls below it. The nearest rank method for calculating the rank at which the P^{th}-percentile will occur uses the formula:

$$n = \lceil\frac{P}{100} \times N\rceil$$

where N is the total number of observations. We equate the n^{th} observation with the P^{th}-percentile.

There are other methods like linear interpolation

Example 3

Given a dataset: 0, 2, 3, 8 and 7.
We have median value 3.
Including 3, we have LQ = 2 and UQ =7.
We have IQR = 5 - there is a lot of data in the range.
Produce a box plot.
What is the 30th percentile?
$n = \lceil \frac{30}{100} 5 \rceil = 2$. The second value in the ordered set is 2.

*Note, when using computer programs to show box plots - make sure outliers are plotted on the line. L(U)Q -(+) 1.5*IQR.*

For a dataset $x_1, ..., x_n$, with mean $\overline{x}$, we consider deviations from the mean:

$$\sum_{i=1}^{n} (x_i - \overline{x})^2 \qquad \text{[the sum of squared deviations from the mean].}$$

We define the **sample variance** to be:

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

and the **sample standard deviation** is:

$$s = \sqrt{s^2}.$$

Most of the data falls in $(\overline{x} - 3s, \overline{x} + 3s)$.

The range, IQR, variance and standard deviation are all measures of spread.

Example 4

Consider the following frequency table:

Result	Frequency	Cumulative Frequency
1	10	10
2	12	22
3	15	37
4	18	55
5	25	80
6	10	90
7	5	95

The range is (7-1)=6.
The median is at the 48th observation: 4.
The LQ is at the 24th obs: 3.
The UQ is at the 72nd obs: 5.
The IQR is (5-3) = 2. (Data concentrated around median because IQR < Range/2.)
The variance is 2.79 (to 2 d.p.) and the std. deviation is 1.67

2 Introduction to Probability - (CT3 Unit 2)

2.1 Definitions

The **sample space** $\mathcal{S}$ (for a well-defined experiment) is the set of all possible outcomes that might be observed.

An **event** A is defined as a subset of $\mathcal{S}$ ($\mathrm{A} \subseteq \mathcal{S}$).

A **complementary event** can be denoted by

$$\mathrm{A}^{\mathrm{c}} \text{ (or } \mathrm{A}', \overline{\mathrm{A}}, \neg\mathrm{A})$$

and is the subset of $\mathcal{S}$ with all possible events not included in A.

Intersection: Given two events A and B, their intersection $\mathrm{A}\cap\mathrm{B}$ denotes the set where both A and B occur.

Union: Given two events A and B, their union $\mathrm{A}\cup\mathrm{B}$ denotes the set where A and/or B occur.

Example 5

Imagine rolling a six-sided die.
$\mathcal{S} = \{1 \text{ is on top}, 2 \text{ is on top}, 3, 4, 5, 6\}$
The event that we have an even number on top is $\{2, 4, 6\}$, which is a subset of $\mathcal{S}$. The complementary event is $\{1, 3, 5\}$, which is also subset of $\mathcal{S}$.
The event that we have a number less than 1 on top is $\emptyset$, which is a subset of $\mathcal{S}$. The complementary event is $\mathcal{S}$, which is a rather boring subset of $\mathcal{S}$.

The null set (denoted $\emptyset$) is the empty set; that is, the set with no elements.

2.2 Basic probability axioms

I. $\mathrm{P}(\mathcal{S}) = 1$, the probability of something happening is 1.
II. $\mathrm{P}(\mathrm{A}) \geq 0 \ \forall \ \mathrm{A} \subset \mathcal{S}$, probability of something happening is non-negative.
III. If A and B are mutually exclusive events: $\mathrm{P}(\mathrm{A}\cup\mathrm{B}) = \mathrm{P}(\mathrm{A}) + \mathrm{P}\,(\mathrm{B})$.

Powerpoint: Unit2_Support.pptx

$\mathrm{P}(\mathrm{A}^{\mathrm{c}}) = 1 - \mathrm{P}(\mathrm{A})$

$\mathcal{S} = \mathrm{A}\cup\mathrm{A}^{\mathrm{c}} \qquad \mathrm{P}(\mathcal{S}) = \mathrm{P}(\mathrm{A}) + \mathrm{P}(\mathrm{A}^{\mathrm{c}})$

$\mathcal{S}^{\mathrm{c}} = \emptyset$, complement of $\mathcal{S}$ is empty, because there is no event.

$\mathrm{P}(\mathrm{A}\cup\mathrm{B}) = \mathrm{P}(\mathrm{A}) + \mathrm{P}(\mathrm{B})$ - $\mathrm{P}(\mathrm{A}\cap\mathrm{B})$

2.3 Conditional Probability

Consider two events, A and B. We might want to know the probability of event A occurring given that B has occurred. This is known as a conditional probability and is denoted by $P(A|B)$.

$$\mathrm{P}(\mathrm{A}|\mathrm{B}) = \frac{\mathrm{P}(\mathrm{A}\cap\mathrm{B})}{\mathrm{P}(\mathrm{A}\cap\mathrm{B}) + \mathrm{P}(\mathrm{A}^{c}\cap\mathrm{B})} = \frac{\mathrm{P}(\mathrm{A}\cap\mathrm{B})}{\mathrm{P}(\mathrm{B})}$$

$\mathrm{P}(\mathrm{A}\cap\mathrm{B}) = \mathrm{P}(\mathrm{A}|\mathrm{B})\mathrm{P}(\mathrm{B})$

For all of these, $\mathrm{P}(\mathrm{B}) > 0$, otherwise B cannot happen.

$$\mathrm{P}(\mathrm{A}) = \frac{\mathrm{P}(\mathrm{A}\cap\mathcal{S})}{\mathrm{P}(\mathcal{S})} = \mathrm{P}(\mathrm{A}|\mathcal{S})$$

Example 6

Imagine rolling a fair six-sided die again. What is the probability of getting a two given that you have got an even number?

$$P(2|\text{even}) = \frac{P(2 \text{ and even})}{P(\text{even})} = 1/3.$$

$$\begin{aligned} \text{A and B are independent} \quad &\iff \quad P(A|B) = P(A) \\ &\iff \quad P(A\cap B) = P(A)P(B) \\ &\iff \quad P(B|A) = P(B) \end{aligned}$$

Example 7

Is the event of rolling a 5 or a 6 (A) independent of the event of rolling an even number (B)?

$$P(A) = 1/3, P(B) = 1/2, P(A \cap B) = 1/6,$$
$$P(A|B) = 1/3 \text{ and } P(B|A) = 1/2.$$

What about the event of rolling a 3 or a 5 (C) and the event of rolling an odd number (D)?

2.4 Theorem of total probability

We have a space $\mathcal{S}$ divided into a partition of n mutually exclusive events. For example,

$$\mathcal{S} = \{1, 2, 3, 4, 5, 6\}, \text{with},$$

$$\mathrm{E}_1 = \{1, 2, 3\},\ \mathrm{E}_2 = \{4\} \text{ and } \mathrm{E}_3 = \{5, 6\}.$$

All outcomes are accounted for, but do not occur more than once:

$$\mathrm{E}_i \cap \mathrm{E}_j = \emptyset \ \text{ if } i \neq j$$
$$\mathrm{E}_1 \cup \mathrm{E}_2 \cup ... \cup \mathrm{E}_n = \mathcal{S}.$$

For any event, A $\subset \mathcal{S}$:

$$\mathrm{A} = (\mathrm{A} \cap \mathrm{E}_1) \cup (\mathrm{A} \cap \mathrm{E}_2) \cup ... \cup (\mathrm{A} \cap \mathrm{E}_n)$$

Therefore,

$$\mathrm{P}(\mathrm{A}) = \sum_{j=1}^{n} \mathrm{P}(\mathrm{A} \cap \mathrm{E}_j) = \sum_{j=1}^{n} \mathrm{P}(\mathrm{A}|\mathrm{E}_j)\mathrm{P}(\mathrm{E}_j).$$

Example 8

I am interested in the probability of Man Utd winning the Champions' League. After some thought, I propose

$$\begin{aligned}
P(\text{Win}|\text{no sig. injuries, no new signings}) &= 0.2,\\
P(\text{Win}|\text{sig. injuries, no new signings}) &= 0.05,\\
P(\text{Win}|\text{no sig. injuries, new signings}) &= 0.25,\\
P(\text{Win}|\text{sig. injuries, new signings}) &= 0.18.
\end{aligned}$$

What is my overall probability of Man Utd winning the Champions' League? This depends on what I think the chances are of each of the events that are being conditioned on. Suppose:

$$\begin{aligned}
P(\text{no sig. injuries, no new signings}) &= 0.4,\\
P(\text{sig. injuries, no new signings}) &= 0.05,\\
P(\text{no sig. injuries, new signings}) &= 0.2.
\end{aligned}$$

Note that the events here are mutually exclusive.
Now, by the theory of total probability,

$$\begin{aligned}
P(\text{Win}) &= 0.2 \times 0.4 + 0.05 \times 0.05 + 0.25 \times 0.2 + 0.18 \times 0.35\\
&= 0.1955.
\end{aligned}$$

2.5 Bayes Theorem

Draw rectangle to describe partition and relation to A.

Let E_1, E_2, ..., E_n be partitions of $\mathcal{S}$ and let $\mathrm{A} \subset \mathcal{S}$, then:

1) $\mathrm{P}(\mathrm{E}_i|\mathrm{A}) = \dfrac{\mathrm{P}(\mathrm{E}_i \cap \mathrm{A})}{\mathrm{P}(\mathrm{A})} \quad \forall\, i$

2) $\mathrm{P}(\mathrm{E}_i \cap \mathrm{A}) = \mathrm{P}(\mathrm{A} \cap \mathrm{E}_i) = \mathrm{P}(\mathrm{E}_j)\mathrm{P}(\mathrm{A}|\mathrm{E}_j)$

3) $\mathrm{P}(\mathrm{A}) = \displaystyle\sum_{j=1}^{n} \mathrm{P}(\mathrm{A} \cap \mathrm{E}_j)$

Thus,

$$\mathrm{P}(\mathrm{E}_i|\mathrm{A}) = \frac{\mathrm{P}(\mathrm{E}_i)\mathrm{P}(\mathrm{A}|\mathrm{E}_i)}{\displaystyle\sum_{j=1}^{n}\mathrm{P}(\mathrm{E}_j)\mathrm{P}(\mathrm{A}|\mathrm{E}_j)}$$

Example 9

A financial think-tank has speculated that the probability of a major bank collapse in the next ten years is 0.5 if there is no substantial government-imposed regulation and that the probability of a major bank collapse in the next ten years is 0.2 if there is substantial government-imposed regulation. Currently, it is thought that there is a 50% chance of further, substantial government-imposed regulation.
If we do not see a major bank collapse in the next ten years, what is the probability that there has been substantial government regulation?
First, let B be the event that a major bank collapses in the next ten years, and let G be the event that there is substantial government-imposed regulation.
From the above statements, we have

$$P(B|G^c) = 0.5, P(B|G) = 0.2 \text{ and } P(G) = P(G^c) = 0.5.$$

We are interested in

$$\begin{aligned} P(G|B^c) &= \frac{P(B^c|G)P(G)}{P(B^c|G)P(G) + P(B^c|G^c)P(G^c)} \text{ (using Bayes's theorem)} \\ &= \frac{0.8 \times 0.5}{0.8 \times 0.5 + 0.5 \times 0.5} = 0.62 \text{ (to 2 d.p.).} \end{aligned}$$

3 Random Variables (CT3 Unit 3)

3.1 Definition

Formally, a **random variable** is a rule for associating a number with each element in a sample space. If w is an element of $\mathcal{S}$ and we associate the number x with w, then $X(w) = x$.

For example, $\mathcal{S}$ is a countably infinite set:

$$\{0, 1, 2, 3, ...\} \quad \text{or,}$$

$$\{x : x = y^2, y = ... - 3, -2, -1, 0, 1, 2, 3, ...\}.$$

These both correspond to *discrete* random variables.

Probabilities are defined over the elements of $\mathcal{S}$, but shorthand is used in terms of the random variable such that $\mathrm{P}(X = x)$.

Example 10

Given 8 outcomes: $w_1, ..., w_8$. We are interested in three events:
w_1, w_2, w_3 are considered as a single event, w_4 and w_5 are considered as a single event and w_6, w_7, w_8 also.
Associate x_1 with event $\{w_1, w_2, w_3\}$ and x_2 with $\{w_4, w_5\}$,
by $\mathrm{P}(X = x_1)$, it is meant $\mathrm{P}(\mathrm{E}_1)$,
by $\mathrm{P}(X = x_2)$, it is meant $\mathrm{P}(\mathrm{E}_2)$, $\quad \mathrm{E}_2 = \{w_4, w_5\}$.

3.2 Probability Functions

The function $f_X(x) = \mathrm{P}(X = x)$, $\forall\ x$ in the range of the random variable, is called the probability function. These are requirements for this function:

$$\begin{aligned} f_X(x) &\geq 0 \quad \forall\ x, \\ \sum_x f_X(x) &= 1 \end{aligned}$$

The function $F_X(x) = \mathrm{P}(X \leq x)\ \ \forall\ x \in \mathbb{R}$. $F_X(x)$ is a monotonically increasing function, and it has maximum value of 1.

For discrete random variables with gaps of 1,

$$\mathrm{P}(X = x) = F_X(x) - F_X(x - 1).$$

The range of a continuous random variable is an interval (or collection of intervals) on $\mathbb{R}$. For continuous functions, there are infinitely many probabilities so we consider the density of the probability through a probability density function (pdf). The probability associated with an interval (a, b) is represented as $\mathrm{P}(a < X < b) = \mathrm{P}(a \leq X \leq b)$.

Probabilities are evaluated by integrating the pdf, $f_X(x)$.

$$\Longrightarrow \mathrm{P}(a < X < b) = \int_a^b f_X(x) \text{ dx}$$

The requirements for $f_X(x)$ to be a pdf are:

$$\begin{aligned} f_X(x) &\geq 0 \quad \forall \quad x \in \mathbb{R}, \\ \int_{-\infty}^{\infty} f_X(x)\, dx &= 1. \end{aligned}$$

The cumulative distribution function (cdf) is defined as $F_X(x) = \mathrm{P}(X \leq x)$. For a continuous random variable, this is a monotonically increasing function and is defined $\forall \; x \in \mathbb{R}$:

$$F_X(x) = \mathrm{P}(-\infty < X < x) = \int_{-\infty}^{x} f_X(t)\, dt.$$

Example 11

Suppose we have the following function:

$$f(x) = \begin{cases} 2x & \text{if } 0 \leq x < 0.25, \\ x^2 + 3/16 & \text{if } 0.25 \leq x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Is it a pdf?
First, $f(x) \geq 0 \, \forall x$.

$$\begin{aligned} \int_{-\infty}^{\infty} f(x) dx &= \int_0^{0.25} 2x dx + \int_{0.25}^{1} x^2 + 3/16 \, dx \\ &= \left[x^2\right]_0^{0.25} + \left[x^3/3 + (3/16)x\right]_{0.25}^{1} \\ &= 102/192 \neq 1. \end{aligned}$$

3.3 Expected Values

3.3.1 Mean

The mean is a measure of location for a random variable.

$\mathrm{E}[X]$ is the expectation of the random variable of X.

$\mathrm{E}[X]$ is calculated using sums if the RV is discrete, or integrals if continuous. For the discrete case,

$$\mathrm{E}[X] = \sum_x x f_X(x).$$

For the continuous case,

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx.$$

Example 12

If X is the random variable associated with a roll of a fair die, we have {1, 2, 3, 4, 5, 6}, $\quad \mathrm{P}(X = x) = \frac{1}{6} \quad \forall\, x, \quad \mathrm{P}(X = x) = f_X(x),$
thus, $\mathrm{E}[X] = \frac{21}{6} = 3.5.$

Example 13

Suppose we have random variable X with pdf:

$$f(x) = \begin{cases} 3x^2 & \text{if } 0 \leq x \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

What is $\mathrm{E}(X)$?

$$\begin{aligned} \mathrm{E}(X) &= \int_0^1 x \times 3x^2 dx = \int_0^1 3x^3 dx \\ &= \left[(3/4)x^4\right]_0^1 = 3/4. \end{aligned}$$

Perhaps draw pdf.

3.3.2 Functions of random variables

For functions of random variables,

$$\mathrm{E}[g(X)] = \sum_x g(x) f_X(x) \qquad \text{[discrete]},$$

$$\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx \qquad \text{[continuous]}.$$

Example 14

X	-1	0	1	2	3	4
$p(X)$	1/2	1/4	1/16	1/16	1/16	1/16

$E(X) = -\frac{1}{2} + \frac{10}{16} = \frac{1}{8} = 0.125.$

What is $E(X^2)$? (*Add X^2 to table*)

$E(X^2) = \frac{9}{16} + \frac{4}{16} + \frac{9}{16} + \frac{16}{16} = \frac{38}{16} = \frac{152}{64} = 2.375.$

Note that $E(X^2) \neq E(X)^2$:
$E(X)^2 = \left(\frac{1}{8}\right)^2 = \frac{1}{64} = 0.015625.$

3.4 Variance

Variance is a measure of uncertainty about a random variable.

$$\begin{aligned} \mathrm{Var}(X) &= \mathrm{E}[(X - \mathrm{E}(X))^2] \\ &= \mathrm{E}[X^2] - \{\mathrm{E}[X]\}^2 \end{aligned}$$

Demonstrate this - expand quadratic and use linear properties.

The standard deviation of X is then given by $\sqrt{\mathrm{Var}(X)}$.

Notation
Population mean: $\mathrm{E}[X] = \mu$
Population variance: $\mathrm{Var}(X) = \sigma^2$
Population standard deviation: σ

3.5 Linear Functions

Let the random variable $Y = aX + b, \quad a, b \in \mathbb{R}$ and $\mathrm{E}[X] = \mu$, then

$$\begin{aligned}\mathrm{E}[Y] &= \mathrm{E}[aX + b] \\ &= \mathrm{E}[aX] + E[b] \\ &= a\mathrm{E}[X] + b \\ &= a\mu + b\end{aligned}$$

and

$$\begin{aligned}\mathrm{Var}(Y) &= \mathrm{Var}[aX + b] \\ &= \mathrm{Var}[aX] + \mathrm{Var}[b] \\ &= a^2\mathrm{Var}[X] \\ &= a^2\sigma^2.\end{aligned}$$

Show why a and a^2.

Standardisation:

$$Z = \frac{(X - \mu)}{\sigma} \qquad (\mu \text{ and } \sigma \text{ are fixed for a random variable})$$

$$\begin{aligned}\text{then} \quad \mathrm{E}[Z] &= \mathrm{E}\left[\frac{X - \mu}{\sigma}\right] \\ &= \frac{1}{\sigma}(\mathrm{E}[X] - \mu) \;=\; 0. \\ \mathrm{Var}(Z) &= \frac{1}{\sigma^2}\mathrm{Var}(X) \;=\; 1.\end{aligned}$$

3.6 Taylor Series approximation (not in CT3)

For any real-valued function $g(x)$, we can write the function as an infinite series (expanding about $x = a$):

$$g(x) = g(a) + \frac{g'(a)(x - a)}{1!} + \frac{g''(a)(x - a)^2}{2!} + ... + \frac{g^{(n)}(x - a)^n}{n!} + ...$$

Let $Y = g(X),\ \mathrm{E}[X] = \mu,\ \mathrm{Var}(X) = \sigma^2$ and we can approximate the mean and variance of the transformed variable Y.

Use a Taylor series expansion of $g(.)$ about μ:

$$\begin{aligned}\mathrm{E}[Y] &\approx g(\mu)+\frac{1}{2}\sigma^2 g''(\mu),\\ \mathrm{Var}(Y) &\approx \sigma^2(g'(\mu))^2.\end{aligned}$$

Example 15

Let X be a random variable with $\mathrm{E}(X)=2$ and $\mathrm{Var}(X)=1$.
Consider the random variable $Y=\exp(-X^2)$.
What are the mean and variance of Y?

$$\begin{aligned}g(x) &= \exp(-x^2),\\ g'(x) &= \frac{d\exp(y)}{dy}\frac{-dx^2}{dx} \qquad \text{(using the chain rule, } y=-x^2\text{)}\\ &= -2x\exp(-x^2),\\ g''(x) &= -2x\frac{d\exp(-x^2)}{dx}+\frac{-d2x}{dx}\exp(-x^2) \qquad \text{(using the product rule)}\\ &= 4x^2\exp(-x^2)-2\exp(x^2).\end{aligned}$$

Now, using the TS approximation,

$$\begin{aligned}\mathrm{E}(Y) &\approx \exp(-4)+0.5\times 1\times(16\exp(-4)-2\exp(-4))\\ &= 8\exp(-4)=0.147 \text{ (to 3 d.p.)},\\ \mathrm{Var}(Y) &\approx 1\times(-4\exp(-4))^2=16\exp(-8)=0.0054 \text{ (to 4 d.p.)}.\end{aligned}$$

Note that $\exp(-4)\approx 0.0183$.

3.7 Moments (CT3 Units 1 and 3)

Moments are calculated from datasets or probability distributions for random variables. They give us a way of describing the key features of data or random variables. We have covered much of this already focussing on two of the moments and using a different notation. It is worth revisiting these looking from the perspective of moments because they give us an alternative way to describe both data and random variables.

3.7.1 Moments for a dataset (sample moments)

The mean and variance are special cases of the moments of a dataset. In general, the k^{th} order moment about a value α is defined by:

$$m_k(\alpha) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \alpha)^k$$

If $k = 1$, and $\alpha = 0$, then $m_k(\alpha) = \overline{x}$,

Therefore, $m_1(0) = \overline{x} = m'_1$ (1st raw moment — raw denotes moment about 0) and the variance $s^2 = \dfrac{n}{n-1} m_2(\overline{x})$.

Often, we write

$$m_2(\overline{x}) = m_2 \qquad (m'_1 = m_1)$$

So, m_k is the k^{th} moment about the sample mean and m'_k is the k^{th} raw moment (i.e. about zero).

The first four moments are related to the following:
m_1 - location,
m_2 - spread,
m_3 - skewness,
m_4 - kurtosis.

A measure for skewness is given by:

$$g_1 = \frac{m_3}{m_2^{3/2}} \qquad \text{(coefficient of skewness)}$$

The sign of g_1 gives the direction of the skew.

3.7.2 Moments for a random variable

Again, the mean and variance are special cases of moments for random variables. In general, the k^{th} order moment about the value of α is defined by:

$$\mu_k(\alpha) = \mathrm{E}[(X - \alpha)^k]$$

Therefore, $\mu_1(0) = \mathrm{E}[X] = \mu = \mu'_1 \;\; (= \mu_1)$

and $\qquad \mu_2(\mu) = \mathrm{E}[(X-\alpha)^2] = \sigma^2 = \mu_2$

These μ'_k are the raw moments (i.e. about zero) and μ_k are the centred moments (i.e. about the mean).

The coefficients of skewness for a random variable is given by $\gamma_1 = \dfrac{\mu_3}{\sigma^3}$. If the distribution of a random variable is symmetric, then $\gamma_1 = 0$.

3.8 Median and Mode for a random variable

An alternative to the mean for random variable X is the median:

$$\mathrm{P}(X < \text{median}) \leq 0.5$$
$$\mathrm{P}(X \leq \text{median}) \geq 0.5$$

If X is continuous, the median, lower and upper quartiles are defined respectively by:

$$\int_{-\infty}^{\text{median}} f_X(x)dx = 0.5 \qquad \int_{-\infty}^{LQ} f_X(x)dx = 0.25 \qquad \int_{-\infty}^{UQ} f_X(x)dx = 0.75$$

The mode is often easier to find because we use differentiation, The mode is the maximum of f_X(x). So, in many cases,

$$f'_X(\text{mode}) = 0 \;\text{ with }\; f''_X(\text{mode}) < 0.$$

This falls down in cases of bimodality and discontinuous distributions.

Example 16

(a) Let X be a random variable with pdf:

$$f_X(x) = \begin{cases} (1/263)(153 + 420x - 300x^2) & \text{if } 0 \leq x \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

$$\begin{aligned} \frac{df_X(x)}{dx} &= 420/263 - 600/263x, \\ \frac{df_X(\text{Mode})}{dx} &= 0 \\ \implies \text{Mode} &= 420/600 \ = \ 0.7. \end{aligned}$$

And the second derivative is negative. It is also worth checking the values at $x = 0$ and $x = 1$.

(b) Let X be a random variable with pdf:

$$f_X(x) = \begin{cases} 4 - 2x & \text{if } 1 \leq x \leq 2, \\ 0 & \text{otherwise.} \end{cases}$$

We have that

$$\begin{aligned} 0.5 &= \int_1^{\mathrm{m}} 4 - 2x \, dx, \\ \implies 0 &= -m^2 + 4m - 3.5 \\ \implies m &= 2 - \sqrt{2}/2. \end{aligned}$$

3.9 Functions of random variables (continuous)

If we have a random variable X and a function (with unique inverse) defined for all values of X, $u(.)$, then we can consider the properties of $Y = u(X)$ (we have already considered a transformation's impacts on expectation earlier).

$F_Y(y) = \mathrm{P}(Y \leq y)$ is what we are interested in and we can recover $f_Y(y)$ by differentiation of $F_Y(y)$.

We know $F_X(x)$, and $y = u(x)$ has a unique inverse:

$$x = w(y) = u^{-1}(y). \qquad (*)$$

If $u(x)$ is an increasing function, then

$$\begin{aligned} F_Y(y) &= \mathrm{P}(Y < y) = \mathrm{P}(u(X) < y) \\ &= \mathrm{P}(X < w(y)) = F_X(w(y)), \end{aligned}$$

and, hence,

$$\begin{aligned} f_Y(y) &= \frac{dF_X(w(y))}{dy} \\ &= \frac{dF_X(z)}{dz}\frac{dw(y)}{dy} \qquad \text{[using chain rule]} \\ &= f_X(w(y))\frac{dw(y)}{dy}. \end{aligned}$$

We know $F_X(x)$ and $y = u(x)$ has a unique inverse as given in $(*)$. If $u(x)$ is a decreasing function, then

$$\begin{aligned} F_Y(y) &= P(u(x) < y) = P(X > w(y)) \\ &= 1 - F_X(w(y)), \end{aligned}$$

and, hence,

$$\begin{aligned} f_Y(y) &= \frac{d(1 - F_X(w(y)))}{dy} \\ &= -f_X(w(y))\frac{dw(y)}{dy}. \end{aligned}$$

Both cases can be summed up in one result:

$$f_Y(y) = f_X(w(y))\left|\frac{d(w(y))}{dy}\right|$$

because the derivative of a increasing function will be positive and the derivative of a decreasing function will be negative.

Example 17

Let X be a random variable with pdf:

$$f_X(x) = \begin{cases} 4-2x & \text{if } 1 \leq x \leq 2, \\ 0 & \text{otherwise.} \end{cases}$$

Consider the random variable $Y = \sqrt{X}$. What is the pdf for Y?

$$\begin{aligned} y &= u(x) = \sqrt{x}. \\ x &= w(y) = y^2. \end{aligned}$$

Using the formula, we get

$$f_Y(y) = \begin{cases} 8y-4y^3 & \text{if } 1 \leq y \leq \sqrt{2}, \\ 0 & \text{otherwise.} \end{cases}$$

Maybe draw pdf and point out mode and approx median.

4 Discrete Distributions (CT3 Unit 4)

4.1 Uniform

Equal probability of each outcome in the sample space.

PF and CDF graphs

4.1.1 Probability Function

If we have k possible outcomes, we have
$\mathrm{P}(X = x) = \dfrac{1}{k}, \quad x = 1, 2, ..., k.$

4.1.2 Moments

Example 18

$$\mu = \mathrm{E}[X] = \frac{(1 + 2 + ... + k)}{k} = \frac{(k(k+1)/2)}{k} = \frac{(k+1)}{2},$$
$$\mu_2' = \mathrm{E}[X^2] = \frac{(1^2 + 2^2 + ... + k^2)}{k} = \frac{(k+1)(2k+1)}{6},$$
$$\sigma^2 = \mathrm{Var}(X) = \frac{(k^2 - 1)}{12}.$$

4.2 Bernoulli

$X \sim \mathrm{Bernoulli}(p)$,

Outcome of one trial with binary outcome, i.e. $\mathcal{S} = \{0, 1\}$

4.2.1 Probability Function

$\mathrm{P}(X = 1) = p, \quad \mathrm{P}(X = 0) = 1 - p.$

4.2.2 Moments

$\mathrm{E}[X] = p$.

Example 19

What is $\mathrm{Var}(X)$?
$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2$
$\mathrm{E}(X)^2 = p^2$
$\mathrm{E}(X^2) = (0^2) \times (1-p) + (1^2) \times (p) = p$
Therefore, $\mathrm{Var}(X) = p - p^2 = p(1-p)$
Note that, the closer that p is to 0.5, the more variability there is.

4.3 Binomial

$X \sim \mathrm{Bin}(n, p)$,

Number of successes in n Bernoulli trials (n independent Bernoulli trials with fixed p).

4.3.1 Probability Function

$\mathrm{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$

$\binom{n}{x}$ takes in order of successes into consideration. Note that
$\binom{n}{x} = {}^nC_x = {}_nC_x = \dfrac{n!}{x!(n-x)!}$ (counting up the combinations).

We could have:
00101, $x = 2$,
11000, $x = 2$, Same # of successes, but different order/combination
01010, $x = 2$.

4.3.2 Moments

$\mu = \mathrm{E}[X] = np$

$\sigma^2 = \mathrm{E}[(X-\mu)^2] = np(1-p)$

Example 20

From the population of accountants based in the UK working for private firms, it is believed that 80% of them got a pay-rise in real terms in the last financial year.
If we assume that pay rises are independent across accountants and we can take a random sample of 50 UK accountants working in private firms, what is the expected value of accountants with pay rises in the last year and what is the standard deviation?
Let X be the number of accountants with pay rises out of our random sample of 50.
We might say that $X \sim \mathrm{Bin}(50, 0.8)$, then

$$\begin{aligned}\mathrm{E}(X) &= np = 40,\\ \mathrm{Var}(X) &= np(1-p) = 8,\\ \text{Std Dev}(X) &= \sqrt{8} = 2.8 \text{ to 1 d.p..}\end{aligned}$$

4.4 Geometric

$X \sim \mathrm{Geo}(p)$,

Number of Bernoulli trials until the first success occurs.

4.4.1 Probability Function

$\mathrm{P}(X=x) = p(1-p)^{x-1}$

4.4.2 Moments

$\mu = \dfrac{1}{p}$
$\sigma^2 = \dfrac{(1-p)}{p^2}$

Example 21

If we were to pick accountants (as defined in the previous example) at random, how many would we expect to have to consider until we found one who had not had a pay rise?
Let A be the number of accountants we need to consider to find the first accountant who has not had a pay rise in real terms, then $A \sim \text{Geo}(0.2)$ and

$$\begin{aligned} \text{E}(A) &= 1/0.2 = 5, \\ \text{Var}(A) &= 0.8/0.04 = 20. \end{aligned}$$

Sometimes a Geometric random variable is defined in an alternative way: consider $Y = X - 1$, $Y \sim \text{Geo}(p)$ and thus,

$\text{P}(Y = y) = p(1-p)^y$

Now, $\mu = \dfrac{1}{p} - 1$ and $\sigma^2 = \dfrac{(1-p)}{p^2}$

Either including the success trial or not

4.5 Poisson

$X \sim \text{Po}(\lambda)$

Number of independent events in a specified time interval.

This distribution models the number of events in a specified time interval, when events occur one after another in a well-defined manner. This manner presumes that the events occur, singularly, at a constant rate and that the number of events that occur in two non-overlapping time intervals are independent.

4.5.1 Probability Function

$\text{P}(X = x) = \dfrac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, ..., \infty; \quad \lambda > 0,$

$$\mathrm{P}(X = x) = \frac{\lambda}{x}\mathrm{P}(X = x - 1).$$

4.5.2 Moments

$\mu = \mathrm{E}[X] = \lambda = \mathrm{Var}(X) = \sigma^2$

Example 22

Use top board so that R can be projected!
Assume that the number of insurance claims per hour (X, say) for some small insurance company is distributed as a Poisson with rate 1. The company is worried that they cannot cope with more than 4 claims per hour. What is the probability of getting more than 4 claims in an hour?
By hand, we could work out

$$P(X > 4) = 1 - P(X = 0) - \cdots - P(X = 4).$$

We can easily calculate this using R: `1-ppois(4,1)`: 0.0037 to 2 s.f..

The Poisson distribution provides an approximation to the binomial distribution when n is large and p is small. Typically, we will want $n > 100$ and $p < \frac{1}{20}$. Given $X \sim \mathrm{Bin}(n, p)$ with large n and small p, then $X \dot{\sim} \mathrm{Po}(np)$ can be used as a very good approximation.

Example 23

Let $X \sim Bin(1000, 0.01)$, then $X \dot{\sim} Po(10)$.
Show tables.

When events are described as occurring "as a Poisson process with rate lambda" or "randomly at a rate of lambda units per unit time", then the number of events which occur in a time period of length t follows the Poisson distribution with parameter λt.

Example 24

The rate of component failures at a factory is thought to be 0.02 per hour. If we accept that failures are happening independently and there are no "peak times", what is the probability of two failures in any 24 hour period? Let X be the number of failures in a 24 hour period, then $X \sim Po(24 \times 0.02) = Po(0.48)$.

$$P(X = 2) = \frac{0.48^2 \exp(-0.48)}{2!} = 0.07 \text{ (to 2 d.p.)}$$

The mean number of failures in a 24 hour period is 0.48.

5 Continuous Distributions (CT3 Unit 4)

5.1 Uniform

$X \sim \text{Uniform}(a, b)$

Equal probability density across an interval.

5.1.1 Probability density function

$f_X(x) = \dfrac{1}{b-a} \quad \text{for } x \in (a, b).$

Graph of density

5.1.2 Moments

$\mathrm{E}[X] = (a+b)/2 \quad \text{and} \quad \mathrm{Var}(X) = (b-a)^2/12.$

Example 25

$$\begin{aligned}
\mathrm{Var}(X) &= \mathrm{E}(X^2) - \mathrm{E}(X)^2, \\
\mathrm{E}(X)^2 &= \frac{(a+b)^2}{4}, \\
\mathrm{E}(X^2) &= \int_a^b \frac{x^2}{b-a} dx \\
&= \frac{1}{b-a}[x^3/3]_a^b \;=\; \frac{1}{3}\frac{b^3-a^3}{b-a}, \\
\mathrm{Var}(X) &= \frac{1}{12(b-a)}(4b^3 - 4a^3 - 3a^2b - 6ab^2 - 3b^3 + 3a^3 + 6a^2b + 3ab^2) \\
&= \frac{1}{12(b-a)}(b^3 - a^3 + 3a^2b - 3ab^2) \;=\; \frac{(b-a)^3}{12(b-a)}.
\end{aligned}$$

5.2 Beta

$x \sim \text{Be}(\alpha, \beta)$

$$f(x) \propto x^{\alpha-1}(1-x)^{\beta-1} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \qquad x \in [0,1]$$

A flexible distribution over an interval.

Graph of density

5.3 Gamma

$X \sim \text{Ga}(\alpha, \lambda)$ with $\alpha > 0$ and $\lambda > 0$.

A family of distributions that includes two special cases.
We define the gamma function $\Gamma(\alpha)$ as:

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy$$

$\Gamma(1) = 1, \quad \Gamma(\alpha) = (\alpha - 1)! \quad$ when $\alpha \in \mathbb{N}$.

5.3.1 Probability density function

$f_X(x) = \dfrac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}$ for $x > 0$.

5.3.2 Moments

$\text{E}[X] = \dfrac{\alpha}{\lambda}$ and $\text{Var}(X) = \dfrac{\alpha}{\lambda^2}$

5.4 Gamma special case - Exponential

Gamma with $\alpha = 1$, $X \sim \text{Exp}(\lambda)$ [$\sim$ Gamma(1, λ)].

5.4.1 Probability density function

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

5.4.2 Moments

$\mathrm{E}(X) = \dfrac{1}{\lambda}, \quad \mathrm{Var}(X) = \dfrac{1}{\lambda^2}.$

5.4.3 Cumulative distribution function

$$\begin{aligned} F_X(x) &= \int_0^x \lambda e^{\lambda t} \\ &= 1 - e^{-\lambda x}. \end{aligned}$$

The exponential distribution is used as a simple model of lifetimes. It also gives the distribution of waiting times between events regenerate by a Poisson process.

Diagram explaining waiting times and relation to Poisson

Example 26

Recall the insurance company from example 22. They had an hourly rate of 1 claim. They were worried about getting more than four claims in one hour, but, on reflection, they are actually worried about getting a claim less than five minutes after receiving the previous claim.
Let Y be the number of minutes between claims. The Poisson rate per minute is 1/60; therefore, $Y \sim Exp(1/60)$.
$\mathrm{E}(Y) = 60$ minutes and $\mathrm{Var}(Y) = 3600$ minutes^2 (which corresponds to a standard deviation of 60 minutes).
$P(Y < 5) = \int_0^5 \frac{\exp(-x/60)}{60} dx = 0.08$ (to 2 d.p. using `pexp(5,1/60)`.

5.5 Gamma special case - χ^2 -distribution

Gamma with $\alpha = \dfrac{\nu}{2}$, where ν is a positive integer and $\lambda = \frac{1}{2}$, then,

$$X \sim \chi^2_\nu \qquad \left[\sim \mathrm{Gamma}\left(\frac{\nu}{2}, \frac{1}{2}\right)\right]$$

5.5.1 Moments

$\mathrm{E}(X) = \nu, \quad \mathrm{Var}(X) = 2\nu.$

This distribution is important because it is widely used in Statistics.

Note: a χ^2 variable with ν=2 is an exponential with mean 2.

5.6 Normal Distribution (or Gaussian distribution)

$X \sim \mathrm{N}(\mu, \sigma^2)$

This distribution with its symmetrical "bell-shaped" density curve is of fundamental importance in both statistical theory and practice. Why?
(1) It often occurs in nature.
(2) It provides good approximations to other distributions - in particular it is a good approximation for the Binomial distribution.
(3) It provides a distribution widely used in Statistics.
(4) When we have large data sets, we often see normal behaviour.

5.6.1 Probability density function

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad x \in \mathbb{R}.$$

Its mean and variance are $\mathrm{E}(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$,

If $\mu = 0$ and $\sigma^2 = 1$, we have the standard normal distribution, $Z \sim N(0,1)$, for which we use the following notation:

$$f_Z(z) = \phi(Z), \quad F_Z(z) = \Phi(Z).$$

(Note that it is impossible to write out an analytic expression for $F_X(x)$.)

A linear function of a normal variable X is also normally distributed:

$$\begin{array}{rcl} \text{If } Y &=& aX + b, \\ \text{then } Y &\sim& N(a\mu + b, a^2\sigma^2). \end{array}$$

Example 27

Let $X \sim N(4, 4)$. What are $P(X < 3)$, $P(X \geq 6)$ and the 95th percentile of X's distribution?
We first standardise X: $Z = (X - 4)/\sqrt{4} = X/2 - 2$. Using properties of the normal distribution, $Z \sim N(0, 1)$.

$$P(X < 3) = P(Z < -0.5) = \Phi(-0.5) = 1 - \Phi(0.5) = 1 - 0.6915 = 0.31.$$

$$P(X \geq 6) = P(Z \geq 1) = 1 - \Phi(1) = 1 - 0.8413 = 0.16.$$

From tables,

$$P(Z < 1.6449) = 0.95 = P(X < 1.6449 \times 2 + 4),$$

which implies that the 95th percentile of X is 7.29 to 2 d.p..

5.7 Lognormal Distribution

The lognormal distribution is the distribution of a log-transformed normal random variable.

This distribution is important because it is used to model positive random variables. Many of which we see in finance.

If X represents some positive random variable and $Y = \log(X)$ with $Y \sim N(\mu, \sigma^2)$, then $X \sim LN(\mu, \sigma^2)$.

5.7.1 Probability density function

(See section 3.9.)

$$f_X(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{-\frac{(\log(x) - \mu)^2}{2\sigma^2}\right\} \quad \text{for } x > 0.$$

Note that $\mathrm{E}(X) \neq \mu, \mathrm{Var}(X) \neq \sigma^2$.

5.8 Student's t-distribution

$X \sim t_\nu$

Many sample statistics will have this distribution, and ν is the degrees of freedom parameter that has a large bearing on the kurtosis of the distribution.

Graph showing change in kurtosis

5.9 Random Number Generation

A basic simulation method starts with the generation of a uniformly distributed random variable on the interval (0, 1), call this U.

Suppose we want to generate a random value from the distribution of X that has $F_X(x)$.
For our random number u, we set $x = F_X^{-1}(u)$, and this is a random draw from the distribution of X.

Example 28

$$X \sim \text{Exp}(\lambda).$$

$$\begin{aligned}
F_X(x) &= 1 - e^{(-\lambda x)} \\
F_X(x) &= u = 1 - e^{(-\lambda x)} \\
&\implies e^{(-\lambda x)} = 1 - u \\
&\implies -\lambda x = \log(1 - u) \\
&\implies x = -\frac{1}{\lambda}\log(1 - u).
\end{aligned}$$

Sampling from the Exponential in this way is very important in Stochastic processes and simulation of Markov processes

6 Generating Functions (CT3 Unit 5)

6.1 Probability generating functions - Discrete

Let X be a counting variable which take values $0, 1, 2, 3, ...$ (i.e., $x \in \mathbb{N} \cup \{0\}$) with probabilities $p_0, p_1, p_2, p_3, ...$ respectively. We define the probability generating function (PGF) as

$$\begin{aligned} G_X(t) &= \mathrm{E}[t^X] \\ &= p_0 + p_1 t + p_2 t^2 + ... \end{aligned}$$

Thus, p_k equals the coefficient of t^k in $G_X(t)$. It is easy to see that $G_X(1) = 1$ and $G_X(0) = \mathrm{P}(X = 0)$. $G_X(t)$ exists at least for $|t| \leq 1$. This gives us an alternative mechanism for describing the distribution of a discrete random variable.

Also, if X and Y have the same probability generating function, then they have the same distribution.

6.2 Evaluating moments using PGFs

The PGF, $G_X(t)$, can be used relatively easily to find low-order moments.

If we differentiate $G_X(t)$ with respect to t and set $t = 1$, we end up with $\mathrm{E}[X]$:

$$\mathrm{E}[X] = \left. \frac{dG_X(t)}{dt} \right|_{t=1}.$$

Further differentiation yields higher order moments. Recall $\mathrm{Var}(X) = \mathrm{E}[X^2] - \{E[X]\}^2$, and, using PGFs, we have

$$\mathrm{E}[X^2] = \left. \frac{d^2 G_X(t)}{dt^2} \right|_{t=1} + \left. \frac{dG_X(t)}{dt} \right|_{t=1}.$$

Example 29

Show that the PGF for a uniform distribution over k outcomes, 1,...,k is

$$G_X(t) = \frac{t}{k}\frac{(1-t^k)}{(1-t)} \text{ for } t \neq 1.$$

$$\begin{aligned} G_X(t) &= \mathrm{E}(t^X) \;=\; \frac{1}{k}(t + t^2 + \cdots + t^k) \\ &= \frac{t}{k}\frac{(1-t^k)}{(1-t)} \end{aligned}$$

using the fact (sum of a geometric series) that

$$\sum_{k=0}^{n-1} ax^k = a\frac{(1-x^k)}{(1-x)} \text{ for } x \neq 1.$$

Example 30

Find the PGF of $X \sim \mathrm{Po}(\lambda)$.
$\mathrm{P}(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, 3, ...$

$$\begin{aligned} G_X(t) &= e^{-\lambda}\sum_{x=0}^{\infty}\frac{(\lambda t)^x}{x!} \\ &= e^{-\lambda}e^{\lambda t} \\ &= e^{\lambda(t-1)} \end{aligned}$$

Now, find the mean and variance of $X \sim \mathrm{Po}(\lambda)$.
$\mathrm{E}[X] = \left.\frac{dG_X(t)}{dt}\right|_{t=1} = \left.\frac{de^{\lambda(t-1)}}{dt}\right|_{t=1} = \left.\frac{de^{-\lambda}e^{\lambda t}}{dt}\right|_{t=1} = \left.\lambda e^{-\lambda}e^{\lambda t}\right|_{t=1}$
$= \lambda e^{-\lambda}e^{\lambda} = \lambda$
$\left.\frac{d^2G_X(t)}{dt^2}\right|_{t=1} = \lambda e^{-\lambda}\left.\frac{de^{\lambda t}}{dt}\right|_{t=1} = \left.\lambda^2 e^{-\lambda}e^{\lambda t}\right|_{t=1}$
$= \lambda^2$
Therefore, we can put all these into the variance equation,

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

6.3 Moment Generating Functions (Discrete and Continuous)

The moment generating function (MGF), $M_X(t)$, of a random variable X is given by:

$$M_X(t) = \mathrm{E}[\exp(tX)] \;\; \forall t \;\text{ for which this expectation exists.}$$

By expanding the exponential term, we get $\left[e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}\right]$

$$M_X(t) = 1 + t\mathrm{E}[X] + \frac{t^2}{2!}\mathrm{E}[X^2] + \frac{t^3}{3!}\mathrm{E}[X^3] + ...,$$

and, from this, it can be seen that the k^{th} raw moment is the coefficient of $\frac{t^k}{k!}$.

Another method to find the k^{th} raw moment is to differentiate $M_X(t)$ k times and set $t = 0$.

If an MGF can be found, then there is a **unique** distribution associated with it. So, if two random variables have the same MGF, then they have the same distribution. For a counting variable which has PGF $G_X(t)\left[= \mathrm{E}[t^X]\right]$, then its MGF can be found by substituting the e^t for t in $G_X(t)$:

$$M_X(t) = G_X(e^t).$$

Example 31

$X \sim \mathrm{Po}(\lambda)$,

$M_X(t) = \exp\{\lambda(e^t - 1)\} = \exp\{-\lambda\}\exp\{\lambda e^t\}$,

$\left.\frac{dM_X(t)}{dt}\right|_{t=0} = e^{-\lambda}\left.\frac{d\exp\{\lambda e^t\}}{dt}\right|_{t=0}$

$\left[\text{Letting } y = e^t, \frac{d\exp\{\lambda e^t\}}{dt} = \frac{d\exp\{\lambda e^t\}}{dy}\frac{dy}{dt} = \lambda\exp\{\lambda e^t\}e^t\right]$

Set $t = 0$, $\implies e^{-\lambda}\lambda e^{\lambda} = \lambda$

Example 32

$X \sim N(\mu, \sigma^2)$,

$\mathrm{M}_X(t) = \exp\{\mu t + \frac{1}{2}\sigma^2 t^2\}$

$\left[\text{expand using } \mathrm{e}^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}\right] = 1 + (\mu t + \frac{1}{2}\sigma^2 t^2) + \frac{1}{2}(\mu t + \frac{1}{2}\sigma^2 t^2)^2 + ...$

Reading off the coefficient of t we find $\mathrm{E}[X] = \mu$.

Reading off the coefficient of $\dfrac{t^2}{2!}$, we find $\mathrm{E}[X^2] = \sigma^2 + \mu^2$.

$\mathrm{Var}(X) = \mathrm{E}[X^2] - \{\mathrm{E}[X]\}^2 = \sigma^2$.

7 Joint Distribution (CT3 Unit 6)

Defining several random variables on a sample space gives rise to a multivariate distribution. If you have two random variables, then you have a bivariate distribution.

Example 33

Bivariate - We have random variables X and Y whose joint probability distribution is defined by the following table:

		x		
		1	2	3
	1	0.10	0.10	0.05
y	2	0.15	0.10	0.05
	3	0.20	0.05	0.00
	4	0.15	0.05	0.00

These probabilities add up to 1.
For example, from the table, we can see that $\mathrm{P}(X = 3, Y = 1) = 0.05$.

7.1 Notation and properties

The function $f_{X,Y}(x, y) = \mathrm{P}(X = x, Y = y)$, defined for all possible values of x and y, is the joint probability function.

To have a joint probability function for two (discrete) random variables, we require

$$\begin{aligned} f_{X,Y}(x,y) &\geq 0 \quad \text{(every probability is positive),} \\ \sum_x \sum_y f_{X,Y}(x,y) &= 1 \quad \text{(all probabilities add up to 1).} \end{aligned}$$

In the case of a pair of continuous variables, we have a function $f_{X,Y}(x, y)$ defined over the x, y - plane, called the joint probability density function.

The probability of X, Y taking values in some region is given by:

$$\mathrm{P}(x_1 < X < x_2, y_1 < Y < y_2) = \int_{y_1}^{y_2} \int_{x_1}^{x_2} f(x,y)dxdy$$

We are integrating the area under a surface defined by $z = f(x, y)$ so we need to look at double integration.

The joint distribution function is defined by

$$F(x,y) = \mathrm{P}(X < x, Y < y),$$

and it is related to the density function through

$$\frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y} = f_{X,Y}(x,y).$$

To qualify as a joint probability density function, we must have that

$$\begin{aligned} f_{X,Y}(x,y) \geq 0, \quad \forall\ x,y, \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)dxdy = 1. \end{aligned}$$

Example 34

We have

$$f_{X,Y}(x,y) = \begin{cases} axy & 0 \leq x \leq 2, 1 \leq y \leq 2, \\ 0 & \text{otherwise.} \end{cases}$$

Find a such that $f_{X,Y}(x,y)$ is a pdf.

$$\begin{aligned} \int_1^2 \int_0^2 axydxdy &= 1 = \int_1^2 [ax^2y/2]_{x=0}^{x=2}dy \\ &= \int_1^2 2aydy = [ay^2]_1^2 \\ &= 4a - a = 3a = 1 \implies a = 1/3. \end{aligned}$$

Of course, for this choice of a, $f(x,y) \geq 0\ \forall x,y$.

7.2 Marginal Probability (Density) Functions

The marginal probability function for a discrete random variable X is defined as

$$f_X(x) = \sum_y f_{X,Y}(x,y).$$

Example 35

$\mathrm{P}(X = 3) = 0.1$ (from earlier table, add up column for $X = 3$)
$\mathrm{P}(Y = 1) = 0.25$

In the continuous case, f_X(x) is obtained by integrating over the possible values of y:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy.$$

Example 36

From example 35, we have

$$f_{X,Y}(x,y) = \begin{cases} \frac{xy}{3} & 0 \leq x \leq 2, 1 \leq y \leq 2, \\ 0 & \text{otherwise.} \end{cases}$$

What are the marginal densities for X and Y?

$$f_X(x) = \int_1^2 \frac{xy}{3} dy = \frac{1}{6}[xy^2]_1^2 = x/2.$$

$$f_Y(y) = \int_0^2 \frac{xy}{3} dx = 2y/3.$$

Perhaps sketch these to show they are pdfs.

7.3 Conditional Probability (Density) Function

The distribution of X for a particular value of Y is called the conditional distribution of X given $Y = y$.

For a discrete variable,

$$\mathrm{P}_{X|Y=y}(x|y) = \mathrm{P}(X = x|Y = y) = \frac{\mathrm{P}_{X,Y}(x,y)}{\mathrm{P}_Y(y)}.$$

Example 37

From the earlier table:

$\mathrm{P}(X = 2|Y = 2) = \dfrac{\mathrm{P}(X = 2, Y = 2)}{P(Y = 2)} = \dfrac{0.1}{0.3} = \dfrac{1}{3}$

$\mathrm{P}(Y = 4|X = 1) = \dfrac{0.15}{0.6} = \dfrac{1}{4}$

$\mathrm{P}(Y = 2|X = 2) = \dfrac{0.1}{0.3} = \dfrac{1}{3}$

$\mathrm{P}(Y = 3|X = 3) = 0$

$\mathrm{P}(X = 3|Y = 3) = 0$

$\mathrm{P}(X = 2|Y = 3) = \dfrac{0.05}{0.25} = \dfrac{1}{5} \qquad (*)$

$\mathrm{P}(Y = 3|X = 2) = \dfrac{0.05}{0.3} = \dfrac{1}{6}$

Example 38

Bayes's theorem:

$\mathrm{P}(X = 2|Y = 3) = \dfrac{\mathrm{P}(Y = 3|X = 2)\mathrm{P}(X = 2)}{\mathrm{P}(Y = 3)} = \dfrac{1}{6}\dfrac{3/10}{1/4} = \dfrac{12}{60} = \dfrac{1}{5}$, which is the same as $(*)$.

For continuous variables, $f_{X|Y=y}(x|y)$ is the conditional density, and we get the conditional probabilities:

$$\begin{aligned}\mathrm{P}(x_1 < X < x_2|Y = y) &= \int_{x_1}^{x_2} f_{X|Y=y}(x|y)dx \\ &= \int_{x_1}^{x_2} \frac{f_{X,Y}(x, y)}{f_Y(y)}dx.\end{aligned}$$

Example 39

From example 35, we have

$$f_{X,Y}(x, y) = \begin{cases} \frac{xy}{3} & 0 \leq x \leq 2, 1 \leq y \leq 2, \\ 0 & \text{otherwise.} \end{cases}$$

What is $P(X < 1|Y = 2)$?

$$f_{X|Y=2}(x|y = 2) = \frac{f_{X,Y}(x, 2)}{f_Y(2)} = \frac{2x/3}{f_Y(2)}.$$

From example 37, we have $f_Y(y) = 2y/3$; therefore, $f_Y(2) = 4/3$.

$$P(X < 1|Y = 2) = \int_0^1 x/2dx = 1/4.$$

7.4 Independence of Random Variables

Consider a pair of random variables (X, Y) and suppose that Y does not actually depend on X at all. It follows that the conditional distribution of $Y|X$, $f(y|x)$, is simply the marginal distribution of Y, $f_Y(y)$. So,

$$\begin{aligned} f_Y(y) \quad &= \quad f(y|x) \;=\; \frac{f_{X,Y}(x,y)}{f_X(x)} \\ &\implies \quad f_{X,Y}(x,y) \;=\; f_X(x)f_Y(y). \end{aligned}$$

X and Y are independent if, and only if, the joint probability density function for X and Y is the product of the two marginals,

$$f_{X,Y}(x,y) = f_X(x)f_Y(y).$$

Example 40

(a) From example 35, we have

$$f_{X,Y}(x,y) = \begin{cases} \frac{xy}{3} & 0 \le x \le 2, 1 \le y \le 2, \\ 0 & \text{otherwise.} \end{cases}$$

From example 37, we have $f_X(x) = x/2$ and $f_Y(y) = 2y/3$. Therefore, $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ and X and Y are independent.

(b) Consider the joint density for X and Y:

$$f_{X,Y}(x,y) = \begin{cases} x/5 + y/5 & 0 \le x \le 2, 1 \le y \le 2, \\ 0 & \text{otherwise.} \end{cases}$$

It is easy to show that $f_X(x) = x/5 + 3/10$ and $f_Y(y) = 2(y+1)/5$. Therefore, X and Y are not independent.

7.5 Expectation of Functions of two variables

The basic rule for expectation is: $\sum_{\text{values}}$ value × probability of seeing that value.

For discrete variables, $\mathrm{E}[g(x,y)] = \sum_x \sum_y g(x,y)\mathrm{P}_{X,Y}(x,y)$,

and, for continuous variables, $\mathrm{E}[g(x,y)] = \int_x \int_y g(x,y) f_{X,Y}(x,y) dx dy$.
Let $g(X,Y) = X + Y$ then $\mathrm{E}[X+Y] = \mathrm{E}[X] + \mathrm{E}[Y]$.
Let $g(X,Y) = ag_1(X) + bg_2(Y)$ then $\mathrm{E}[g(X,Y)] = a\mathrm{E}[g_1(X)] + b\mathrm{E}[g_2(X)]$ (by the linearity of expectation)

Consider a product of two functions: $h(X,Y) = h_1(X)h_2(Y)$. In general,

$$\mathrm{E}[h(X,Y)] \neq \mathrm{E}[h_1(X)]\mathrm{E}[h_2(Y)],$$

but, if X and Y are **independent**, we can write

$$\mathrm{E}[h(X,Y)] = \mathrm{E}[h_1(X)]\mathrm{E}[h_2(Y)].$$

Example 41

From example 35, we have

$$f_{X,Y}(x,y) = \begin{cases} \frac{xy}{3} & 0 \leq x \leq 2, 1 \leq y \leq 2, \\ 0 & \text{otherwise,} \end{cases}$$

and, from example 37, we have $f_X(x) = x/2$ and $f_Y(y) = 2y/3$. What is $\mathrm{E}(XY)$?

$$\mathrm{E}(XY) = \int_0^2 \int_1^2 \frac{x^2y^2}{3} dy dx = \int_0^2 (7/9)x^2 dx = 56/27.$$

Because X and Y are independent, we can also use $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$. We can easily find that $\mathrm{E}(X) = 4/3$ and $\mathrm{E}(Y) = 14/9$. Multiplying these together gives 56/27.

7.6 Covariance and Correlation Coefficient

The covariance between X and Y is defined as:

$$\begin{aligned} \mathrm{Cov}(X,Y) &= \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])] \\ &= \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y]. \\ \text{Note that } \mathrm{Cov}(X,X) &= \mathrm{E}[(X - \mathrm{E}[X])(X - \mathrm{E}[X])] \\ &= \mathrm{E}[(X - \mathrm{E}[X])^2] \\ &= \mathrm{Var}(X). \end{aligned}$$

$\rho_{X,Y}$ is the correlation coefficient such that $\rho_{X,Y} = \dfrac{\mathrm{Cov}(X,Y)}{\sqrt{(\mathrm{Var}(X)\mathrm{Var}(Y))}}$.

This is a quantity between -1 and 1. The closer $|\rho_{X,Y}|$ is to 1 the stronger the *linear* relationship. If $\rho_{X,Y} = 0$, we say that X and Y are uncorrelated, and they have no linear relationship.

If X and Y are independent, then,

$$\rho_{X,Y} = 0.$$

If $\rho_{X,Y} = 0$, however, X and Y may still be dependent.

Slides about correlation strength

Example 42

Consider the joint density for X and Y:

$$f_{X,Y}(x, y) = \begin{cases} x/5 + y/5 & 0 \leq x \leq 2, 1 \leq y \leq 2, \\ 0 & \text{otherwise.} \end{cases}$$

It is easy to show that $f_X(x) = x/5 + 3/10$ and $f_Y(y) = 2(y + 1)/5$.
From the marginals, we need $\mathrm{E}(X) = 17/15$ and $\mathrm{E}(Y) = 23/15$. We can also get $\mathrm{E}(XY) = 26/15$.
Therefore, $\mathrm{Cov}(X, Y) = 26/15 - (17 \times 23)/(15 \times 15) = -0.004$ to 3 decimal places; hence, we know that there is negative correlation between X and Y.

Some Additional Facts about Covariances

$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y)$

$\mathrm{Cov}(aX+bY, cS+dT) = ac\mathrm{Cov}(X, S)+ad\mathrm{Cov}(X, T)+bc\mathrm{Cov}(Y, S)+bd\mathrm{Cov}(Y, T)$

Now, if $a = b = c = d = 1$ and $X = S, Y = T$, then
$\mathrm{Cov}(X + Y, X + Y) = \mathrm{Var}(X + Y)$.

Also, note that $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$.

7.7 Distributions of linear combinations of r.v.s via generating functions

7.7.1 PGFs

Suppose that we have X and Y independent counting variables with PGFs $\mathrm{G}_X(t)$ and $\mathrm{G}_Y(t)$ respectively. Let

$$S = c_1X + c_2Y$$

Then,

$$\begin{aligned} G_S(t) &= \mathrm{E}[t^{(c_1X+c_2Y)}] \\ &= \mathrm{E}[t^{c_1X}t^{c_2Y}] = \mathrm{E}[t^{c_1X}]\mathrm{E}[t^{c_2Y}] \\ &= G_X(t^{c_1})G_Y(t^{c_2}) \end{aligned}$$

In the simple case of $Z = X + Y$,

$$G_Z(t) = G_X(t)G_Y(t).$$

Example 43

$X \sim \mathrm{Po}(\lambda)$ and $Y \sim \mathrm{Po}(\gamma)$, with X and Y independent.
X has PGF, $G_X(t) = \exp\{\lambda(\mathrm{t}$ - $1)\}$ and Y has PGF, $G_Y(t) = \exp\{\gamma(\mathrm{t}$ - $1)\}$.
So, $Z = X + Y$ has PGF,

$$\begin{aligned} G_Z(t) &= \exp\{\lambda(t-1)\}\exp\{\gamma(t-1)\} \\ &= \exp\{(\lambda+\gamma)(t-1)\}. \end{aligned}$$

We get $Z \sim \mathrm{Po}(\lambda + \gamma)$ by uniqueness of PGFs.

7.7.2 Moment generating functions

Suppose X and Y are independent random variables with MGFs, $M_X(t)$ and $M_Y(t)$. Let

$$S = c_1X + c_2Y,$$

then,

$$\begin{aligned} M_S(t) &= \mathrm{E}[e^{(c_1X+c_2Y)t}] \\ &= \mathrm{E}[e^{c_1Xt}e^{c_2Yt}] \\ &= \mathrm{E}[e^{c_1Xt}]\mathrm{E}[e^{c_2Yt}] \\ &= M_X(c_1t)M_Y(c_2t). \end{aligned}$$

Example 44

$X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$, with X and Y independent.

$$M_X(t) = \exp\left(\mu_X t + \frac{1}{2}\sigma_X^2 t^2\right)$$

$Z = X + Y$ has MGF,

$$\begin{aligned} M_Z(t) &= \exp\left(\mu_X t + \frac{1}{2}\sigma_X^2 t^2\right)\exp\left(\mu_Y t + \frac{1}{2}\sigma_Y^2 t^2\right) \\ &= \exp\left\{(\mu_X + \mu_Y)t + \frac{1}{2}(\sigma_X^2 + \sigma_Y^2)t^2\right\}. \end{aligned}$$

Therefore, $Z \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$ through the uniqueness of MGFs.

8 The Central Limit Theorem (CT3 Unit 7)

If $X_1, X_2, ...X_n$ is a sequence of independent and identically distributed (iid) random variables with mean μ and variance σ^2 (both finite), then the distribution of

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

approaches the standard normal distribution, $N(0, 1)$, as $n \to \infty$. Therefore, both $\dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ and $\dfrac{(\sum X_i) - n\mu}{\sqrt{n\sigma^2}}$, are approximately distributed as $N(0, 1)$ for "large" n.

Alternatively, the unstandardised form can be used:

$$\overline{X} \mathrel{\dot{\sim}} N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{and} \quad \sum_{i=1}^{n} X_i \mathrel{\dot{\sim}} N(n\mu, n\sigma^2),$$

Note that by "large" n, we mean roughly over 30.

Demonstration in R?

8.1 Normal Approximations

When summing a number of independent random variables, we often use the CLT as a basis for an approximation

Example 45

Let X_i be i.i.d. Bernoulli random variables such that

$$\begin{aligned} \mathrm{P}(X_i = 1) &= \theta, \\ \mathrm{P}(X_i = 0) &= 1 - \theta. \end{aligned}$$

Consider a sequence $X_1, X_2, ..., X_n$ of such random variables, and set

$$X = \sum_{i=1}^{n} X_i, X \sim \mathrm{Bin}(n, \theta), \quad \text{also note that} \quad \frac{X}{n} = \overline{X}$$

By the Central Limit Theorem, for large n,

$$\overline{X} \overset{\cdot}{\sim} N(\mu, \frac{\sigma^2}{n}) \quad \text{or} \quad \sum_{i=1}^{n} X_i \overset{\cdot}{\sim} N(n\mu, n\sigma^2).$$

For the Bernoulli distribution,

$$\begin{aligned} \mu &= \mathrm{E}[X_i] = \theta, \\ \sigma^2 &= \mathrm{Var}(X_i) = \theta(1-\theta). \end{aligned}$$

So, if $X \sim \mathrm{Bin}(n, \theta)$, then we can say, $X \overset{\cdot}{\sim} N(n\theta, n\theta(1-\theta))$.

Let $Y \sim N(n\theta, n\theta(1-\theta))$. Therefore, $X \overset{\cdot}{\sim} Y$. Example: $P(X \leq 5) \approx P(Y < 5.5)$.

Example 46

Consider a sequence $U_1, U_2, ..., U_n$ of i.i.d. random variables where $U_i \sim \mathrm{Uniform}(0,1) \forall i$ and n is large. What is the distribution of $T = \pi \sum_{i=1}^{n} U_i$? We know that $\mathrm{E}(U_i) = 1/2$ and $\mathrm{Var}(U_i) = 1/12$. Therefore, using the central limit theorem, we have

$$\begin{aligned} \sum_{i=1}^{n} U_i &\overset{\cdot}{\sim} N(n/2, n/12), \\ \pi \sum_{i=1}^{n} U_i &\overset{\cdot}{\sim} N(\pi n/2, \pi^2 n/12). \end{aligned}$$

End of 1st third of course

9 Sampling & Inference (CT3 Unit 8)

When a sample is taken from a population, the sample information can be used to infer certain things about the population. A random sample is made up of i.i.d. random variables, X_i. We use the notation $\underline{X} = \{X_1, X_2, ..., X_n\}$ for a random sample. An observed sample is denoted by $\underline{x} = \{x_1, x_2, ..., x_n\}$. The population density is denoted by $\mathrm{f}(x;\theta)$ where θ represents any distribution parameters.

A statistic is a function of $\underline{X}$ alone and does not involve any component of θ.

Example 47

Thus, $\overline{X} = \displaystyle\sum_{i=1}^{n} \frac{X_i}{n}$ and $s^2 = \displaystyle\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ are statistics whereas $\displaystyle\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ is not a statistic unless μ is known.

A statistic is denoted generally by $\mathrm{g}(\underline{X})$, and, because it is a function of something random, it is random itself.

9.1 Moments of the sample mean and variance

Suppose that i.i.d. random variables X_i have mean μ and variance σ^2. Now, the sample mean is

$$\begin{aligned}
\overline{X} &= \sum_{i=1}^{n} \frac{X_i}{n}. \\
\mathrm{E}\left(\sum_{i=1}^{n} X_i\right) &= \sum_{i=1}^{n} \mathrm{E}[X_i] = \sum_{i=1}^{n} \mu = n\mu, \\
\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) &= \sum_{i=1}^{n} \mathrm{Var}(X_i) = n\sigma^2.
\end{aligned}$$

So,

$$\mathrm{E}[\overline{X}] = \frac{\mathrm{E}[\sum X_i]}{n} = \mu,$$

$$\mathrm{Var}(\overline{X}) = \frac{\mathrm{Var}\,(\sum X_i)}{n^2} = \frac{\sigma^2}{n},$$

and

$$\text{Std. Dev.}(\overline{X}) = \sqrt{\text{Var}(\overline{X})} = \frac{\sigma}{\sqrt{n}},$$

which is the standard error of the mean.

The sample variance is $S^2 = \frac{1}{n-1}\sum(X_i - \overline{X})^2$.

So we have

$$\text{E}(S^2) \;=\; \frac{1}{n-1}\sum \text{E}\left[(X_i - \overline{X})^2\right],$$

and,

$$\begin{aligned} \text{E}\left[(X_i - \overline{X})^2\right] &= \text{E}\left\{\left[(X_i - \mu) - (\overline{X} - \mu)\right]^2\right\} \\ &= \text{E}\left[(X_i - \mu)^2\right] - 2\text{E}\left[(X_i - \mu)(\overline{X} - \mu)\right] + \text{E}\left[(\overline{X} - \mu)^2\right]. \end{aligned}$$

We can rewrite this expression using:

$$\begin{aligned} \text{E}\left[(X_i - \mu)^2\right] &= \sigma^2, \\ \text{E}\left[(\overline{X} - \mu)^2\right] &= \frac{\sigma^2}{n}, \\ \text{E}\left[(X_i - \mu)(\overline{X} - \mu)\right] &= \frac{\sigma^2}{n}. \end{aligned}$$

Therefore,

$$\begin{aligned} \text{E}(S^2) &= \frac{1}{n-1}\sum\left(\sigma^2 - \frac{\sigma^2}{n}\right) \\ &= \sigma^2. \end{aligned}$$

Do the extended version on the board and mention issues with deriving Var(S^2).

9.2 Sampling distributions when data are normal

The Central Limit Theorem provides a distribution for $\overline{X}$ without any assumptions about the distribution of the X$_i$ if n is large enough.

Recall, for large n,

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

If X_i are i.i.d. normal, then

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

We can rearrange this to get the "Z result".

The sampling distribution for S^2 when sampling from $N(\mu, \sigma^2)$ is

$$\frac{(n-1)}{\sigma^2} S^2 \sim \chi^2_{n-1}.$$

[
Informal sketch of Proof — not examinable!
Firstly we note that the moment generating function of a χ^2_k random variable is $M_X(t) = (1-2t)^{-k/2}$. Using this fact, it is then easy to verify that, if $X_i \sim \chi^2_{k_i}, \quad i = 1, 2, \ldots, M$, and independent, then $\sum X_i \sim \chi^2_K$ where $K = \sum k_i$.

Now we use an inductive argument. Let $\overline{X}_j$ and S^2_j denote the sample mean and variance of the first j observations (the order is immaterial). Then we note

$$(n-1)S^2_n = (n-2)S^2_{n-1} + \left(\frac{n-1}{n}\right)(X_n - \overline{X}_{n-1})^2.$$

Now consider $n = 2$. Defining $0 \times S^2_1 = 0$, we have that

$$S^2_2 = \frac{1}{2}(X_2 - X_1)^2.$$

Since the distribution of $(X_2 - X_1)/\sqrt{2}$ is $N(0,1)$ then $S^2_2 \sim \chi^2_1$. Now proceed by induction, in which we need to check that $(j/(j+1))(X_{j+1} - \overline{X}_j)^2 \sim \chi^2_1$ and is also independent of S^2_j.
]

The χ^2_{n-1} distribution can be derived from $\displaystyle\sum_{i=1}^{n} Z_i^2$ where $Z_i \sim N(0,1)$, because $\displaystyle\sum_{i=1}^{n} Z_i^2 \sim \chi^2_n$ by definition. (In our case we have $n-1$ degrees of freedom because $\overline{X}$ is used in place of μ.)

$$\left[\frac{n-1}{\sigma^2}S^2 = \frac{n-1}{n-1}\sum_{i=1}^{n}\frac{(X_i-\overline{X})^2}{\sigma^2}, \text{dividing by } \sigma^2 \text{ gives Var} = 1.\right]$$

Powerpoint slides: Sampling.pptx

We know that S^2 is related to the χ^2_ν distribution when sampling from a normal distribution. We know that the mean and variance of a χ^2_ν random variable are ν and 2ν respectively.

$$\begin{aligned}
\mathrm{E}\left[\frac{(n-1)S^2}{\sigma^2}\right] &= n-1 \implies \mathrm{E}[S^2] = \sigma^2. \\
\mathrm{Var}\left(\frac{(n-1)S^2}{\sigma^2}\right) &= 2(n-1) \\
\implies \frac{(n-1)^2}{\sigma^4}\mathrm{Var}(S^2) &= 2(n-1) \implies \mathrm{Var}(S^2) = \frac{2\sigma^4}{(n-1)}.
\end{aligned}$$

9.3 The sampling distribution for $\overline{X}$ when σ^2 is unknown

When σ^2 is unknown, we replace σ^2 with a suitable estimate, and, if we choose the usual estimate, we will replace the normal sampling distribution with a t-distribution.

Plot comparing t with normal

A t-distribution has "heavier" tails than the normal distribution: it has more probability density at extreme values.

The t-result is similar to the previous result for S^2 (the "z" result) with σ^2 replaced by S^2 and N(0, 1) replaced by t_{n-1}. So, we have

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

The general t_k variable can be defined as $\mathrm{t}_k = \dfrac{N(0,1)}{\sqrt{\chi^2_k/k}}$, where the N(0, 1) part and χ^2_k part are independent.

From before, $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ and $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$, and, because these are independent, we get

$$\begin{aligned}\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \times \frac{\sigma\sqrt{n-1}}{\sqrt{n-1}S} &= \frac{(\overline{X}-\mu)(\sqrt{n-1})}{S/\sqrt{n}(\sqrt{n-1})} \\ &= \left(\frac{\overline{X}-\mu}{S/\sqrt{n}}\right) \sim t_{n-1}.\end{aligned}$$

Now, for a t$_k$ variable, we have the mean is 0, and variance is $\frac{k}{k-2}$. (This makes it difficult for small n).

$$\begin{aligned}\mathrm{E}\left[\frac{\overline{X}-\mu}{S/\sqrt{n}}\right] &= 0 \implies \mathrm{E}[\overline{X}] = \mu. \\ \mathrm{Var}\left(\frac{\overline{X}-\mu}{S/\sqrt{n}}\right) &= \frac{n-1}{n-3}.\end{aligned}$$

9.4 The F-result for variance ratios

The F-distribution is defined by $\dfrac{U/\nu_1}{V/\nu_2}$ where U and V are independent χ^2 random variables with ν_1 and ν_2 degrees of freedom respectively.

If independent random samples of size n_1 and n_2 are taken from normal populations with variance σ_1^2 and σ_2^2 respectively, then

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim \mathcal{F}_{n_1-1,n_2-1}$$

or

$$\frac{S_2^2/\sigma_2^2}{S_1^2/\sigma_1^2} \sim \mathcal{F}_{n_2-1,n_1-1}$$

We have that $\dfrac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi^2_{n_1-1}$ and $\dfrac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi^2_{n_2-1}$ (both independent).

So,

$$\frac{(n_1-1)S_1^2\sigma_2^2(n_2-1)}{\sigma_1^2(n_2-1)S_2^2(n_1-1)} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim \mathcal{F}_{n_1-1,n_2-1} \qquad \left[\implies \frac{\chi^2_{n_1-1}}{\chi^2_{n_2-1}}\right].$$

We note that

$$\begin{aligned}
\mathrm{E}\left(\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}\right) &= \mathrm{E}\left(\frac{\chi^2_{n_1-1}/(n_1-1)}{\chi^2_{n_2-1}/(n_2-1)}\right)\\
&= \mathrm{E}\left(\frac{\chi^2_{n_1-1}}{n_1-1}\right)\mathrm{E}\left(\frac{n_2-1}{\chi^2_{n_2-1}}\right) \qquad \text{(since independent)}\\
&= \left(\frac{n_1-1}{n_1-1}\right)\left(\frac{n_2-1}{n_2-3}\right) \qquad (*)\\
&= \frac{n_2-1}{n_2-3},
\end{aligned}$$

and so for reasonably large n_2 we have

$$\frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \approx \frac{n_2-1}{n_2-3} \approx 1.$$

The statement (*) can be checked using integration by parts. Note also,

- If $X \sim F_{p,q}$, then $1/X \sim F_{q,p}$
- If $X \sim t_q$, then $X^2 \sim F_{1,q}$
- If $X \sim F_{p,q}$, then $(p/q)X/(1+(p/q)X) \sim \mathrm{Be}(p/1, q/2)$.

10 Point Estimation (CT3 Unit 9)

10.1 The method of moments

To get estimates of parameters using the method of moments, we set sample moments to be equal to expressions for population moments and solve the resulting equations for the distribution parameters.

10.1.1 One parameter case

The simplest case is where we equate the population mean with the sample mean and solve for the unknown parameter. That is, set

$$\mathrm{E}[X] = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Example 48

$X \sim N(\mu, \sigma^2)$ with known σ^2,
then using the method of moments,

$$\mathrm{E}[X] = \mu \overset{\text{set}}{=} \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}.$$

We have our estimator for μ is $\overline{x}$.

Example 49

$X \sim$ Uni(0, θ),

$$\mathrm{E}[X] = \frac{\theta}{2} \overset{\text{set}}{=} \overline{x}.$$

Our estimator for θ is $2\overline{x}$.
What if $\overline{x} = 1.5$?
$X \sim$ Uni(0, 3)

Example 50

$X \sim \mathrm{N}(0, \sigma^2)$

$$\mathrm{E}[X] = 0$$

This is not helpful because there is no σ^2 in this equation,

$$\mathrm{Var}(X) = \sigma^2 \overset{\text{set}}{=} S^2$$

10.1.2 The two parameter case

This involves equating the 1st and 2nd moments of the population with the sample mean and variance respectively.

Example 51

$X \sim \mathrm{N}(\mu, \sigma^2)$ with both parameters unknown.

$$\mathrm{E}[X] = \mu = \overline{x}.$$

and,

$$\mathrm{Var}(X) = \sigma^2 = s^2.$$

Example 52

$X \sim \mathrm{Be}(\alpha, \beta)$

$$\mathrm{E}[X] = \frac{\alpha}{\alpha+\beta} \text{ and } \mathrm{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

$$\overline{x} = \frac{\alpha}{\alpha+\beta}$$
$$\implies \beta = \frac{\alpha}{\overline{x}} - \alpha$$
$$s^2 = \frac{\alpha^2(\frac{1}{\overline{x}}-1)}{(\alpha-\alpha+\frac{\alpha}{\overline{x}})^2(\frac{\alpha}{\overline{x}}-1)} = \frac{(\frac{1}{\overline{x}}-1)\overline{x}^2}{\frac{\alpha}{\overline{x}}-1}$$
$$\left(\frac{\alpha}{\overline{x}}-1\right) = \frac{\overline{x}-\overline{x}^2}{s^2}$$
$$\alpha = \frac{\overline{x}^2-\overline{x}^3}{s^2} + \overline{x} = \overline{x}\left(\frac{\overline{x}(1-\overline{x})}{s^2}-1\right)$$
$$\beta = \frac{\alpha}{\overline{x}} - \alpha \implies \beta = (1-\overline{x})\left(\frac{\overline{x}(1-\overline{x})}{s^2}-1\right)$$

10.2 Maximum Likelihood

The method of moments does not take into account the full distribution: it only accounts for one or two moments and we need infinitely many to describe the distribution.

Powerpoint slides: Sampling.pptx

10.2.1 The one-parameter case

The first step is to write down a likelihood,

$$l(\theta; \underline{x}) \propto \prod_{i=1}^{n} f_{X_i}(x_i|\theta)$$

for a random sample $x_1, x_2, x_3, ...x_n$ from a population with density $f(x|\theta)$.

Example 53

$X \sim \text{Bin}(20, p)$
We have observed $x = 5$, then

$$\begin{aligned} l(p; x = 5) &\propto \text{P}(X = 5) \\ &\propto (p)^5(1-p)^{20-5} \\ &= (p)^5(1-p)^{15}, \qquad \text{for } \ p \in (0,1). \end{aligned}$$

Might need to explain why taking logs is a good idea and has no effect: show graphs in Sampling.pptx

Differentiating the likelihood (or log-likelihood) and solving for θ after equating with 0, we find the maximum of the likelihood and we call this the maximum likelihood estimator (MLE), $\hat{\theta}$. (Of course, we should also check that the 2^{nd} derivative is negative at that value of θ.)

MLEs has the invariance property, which means that, if $\hat{\theta}$ is the MLE of θ, then the MLE of g(θ) is g$(\hat{\theta})$.

Example 54

We have a random sample of size n from $X_i \sim \text{Exp}(\lambda)$, with

$$f_X(x) = \lambda e^{-\lambda x}, \text{ for } \ x \in (0, \infty)$$

We want to estimate λ given a sample,

$$\begin{aligned} l(\lambda; \underline{x}) &\propto \prod_{i=1}^{n} \lambda e^{-\lambda x_i} \\ &= \lambda^n e^{-\lambda \sum x_i} \end{aligned}$$

$$\begin{aligned} L(\lambda; x) &= \log(l(\lambda; \underline{x})) \\ &= \log(\lambda^n) + \log(e^{-\lambda \sum x_i}) + c \\ &= n\log\lambda - \lambda \sum_{i=1}^{n} x_i + c. \\ \frac{\partial L(\lambda; x)}{\partial \lambda} &= \frac{n}{\lambda} - \sum_{i=1}^{n} x_i \end{aligned}$$

Equate this to zero,

$$\implies \tfrac{n}{\lambda} = \sum_{i=1}^{n} x_i \implies \frac{1}{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x}$$

$$\implies \hat{\lambda} = \tfrac{1}{\overline{x}} \text{ and } \frac{\partial^2 L(\lambda; \overline{x})}{\partial \lambda^2} = \frac{-n}{\lambda^2}$$

For $\hat{\lambda} = \dfrac{1}{\overline{x}}$, this is negative and we have a max at $\hat{\lambda}$. So $\hat{\lambda}$ is a MLE.

10.3 Unbiasedness

If we have a random sample $\underline{\text{X}} = \{X_1, ...X_n\}$ from a distribution with unknown parameter θ and $g(\underline{X})$ is an estimator for θ, we want

$$\text{E}[g(\underline{X})] = \theta. \quad \text{(This is unbiasedness)}$$

If an estimator is biased the bias is measured by $\text{E}[g(\underline{X})] - \theta$. This property is not preserved under non-linear transformations of the estimator and parameter.

Example 55

An example of a biased estimator is $\frac{\sum_i (X_i - \overline{X})^2}{n}$ for variance.
An unbiased estimator for variance is

$$\begin{aligned} &\frac{\sum_i (X_i - \overline{X})^2}{n-1}, \\ \mathrm{E}\left[\frac{\sum_i (X_i - \overline{X})^2}{n-1}\right] &= \sigma^2; \\ \text{whereas, } \mathrm{E}\left[\frac{\sum_i (X_i - \overline{X})^2}{n}\right] &= \frac{\sigma^2(n-1)}{n}. \end{aligned}$$

And the bias is,

$$\frac{(n-1)\sigma^2}{n} - \sigma^2 = \frac{-\sigma^2}{n}.$$

10.4 Mean Squared Error (MSE)

We are also interested in the accuracy of the statistic. One way of measuring this is through the mean squared error of the statistic. The MSE of an estimator g($\underline{X}$) for θ is defined by,

$$\mathrm{MSE}(g(\underline{X})) = \mathrm{E}[(g(\underline{X}) - \theta)^2].$$

This is not always the variance of g($\underline{X}$)

$$\mathrm{Var}(g(\underline{X})) = \mathrm{E}[\{g(\underline{X}) - \mathrm{E}[g(\underline{X})]\}^2].$$

The MSE is only equal to this variance when

$$\mathrm{E}[g(\underline{X})] = \theta \quad \text{(i.e., when g}(\underline{\mathrm{X}})\text{ is unbiased).}$$

$$\mathrm{MSE} = \mathrm{Variance} + \mathrm{Bias}^2.$$

This expression can be proved as follows:

$$
\begin{aligned}
MSE[g(\underline{X})] &= E[(g(\underline{X}) - \theta)^2] \\
&= E[\{(g(\underline{X}) - E[g(\underline{X})]) + (E[g(\underline{X})] - \theta)\}^2] \\
&= E[(g(\underline{X}) - E[g(\underline{X})])^2] + 2(E[g(\underline{X})] - \theta)E[g(\underline{X}) - E[g(\underline{X})]] \\
&\quad + (E[g(\underline{X})] - \theta)^2 \\
&= Var[g(\underline{X})] + bias(g(\underline{X}))^2
\end{aligned}
$$

Example 56

Let X_i $(i = 1, \ldots, n)$ be i.i.d. normal with mean μ and variance σ^2. We know that

$$\mathrm{E}(S^2) = \sigma^2 \text{ and } \mathrm{Var}(S^2) = \frac{2\sigma^4}{(n-1)} = \mathrm{MSE}(S^2).$$

Also, consider

$$\mathrm{E}\left[\frac{\sum_i (X_i - \overline{X})^2}{n}\right] = \frac{\sigma^2(n-1)}{n} \text{ and } \mathrm{Var}\left[\frac{\sum_i (X_i - \overline{X})^2}{n}\right] = \frac{2(n-1)\sigma^4}{n^2}.$$

Therefore,

$$
\begin{aligned}
\mathrm{MSE}\left[\frac{\sum_i (X_i - \overline{X})^2}{n}\right] = \frac{(2n-1)\sigma^4}{n^2} &< \frac{(2n-1)\sigma^4}{n(n-1)} \\
&< \frac{2n\sigma^4}{n(n-1)} < \mathrm{MSE}(S^2).
\end{aligned}
$$

Show final two slides of Sampling.pptx

We can be biased and more accurate at the same time.

11 Confidence Intervals (CT3 Unit 10)

A **confidence interval** provides an 'interval estimate' of an unknown parameter (instead of just a 'point estimate'). Therefore, we get some appreciation of the accuracy of the estimate.

A $100\,(1-\alpha)\,\%$ confidence interval for θ is defined by specifying random variables $\hat{\theta}_1\left(\underline{\mathrm{X}}\right)$ and $\hat{\theta}_2\left(\underline{\mathrm{X}}\right)$ such that:

$$\Pr\left(\hat{\theta}_1\left(\underline{\mathrm{X}}\right) < \theta < \hat{\theta}_2\left(\underline{\mathrm{X}}\right)\right) = 1-\alpha,$$

which gives the confidence interval:

$$\left(\hat{\theta}_1\left(\underline{\mathrm{X}}\right), \hat{\theta}_2\left(\underline{\mathrm{X}}\right)\right).$$

There are infinitely many choices for the random variables given α, so confidence intervals are not unique. The usual choice for α is $\alpha = 0.05$ so we will have a 95% confidence interval. In the long run, if we repeated the experiment that yielded $\underline{\mathrm{X}}$ and we calculate a 95% interval in the same manner, 95% of such intervals will contain θ.

11.1 Confidence Intervals for Normally Distributed Random Variables

If we are sampling from a $N\left(\mu, \sigma^2\right)$ with σ^2 known, we know that

$$\frac{\overline{X}-\mu}{\sqrt{\sigma^2/n}} \sim N\left(0,1\right)$$

From tables, we know that

$$\begin{aligned}
& \Pr\left(-1.96 < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.95 \\
\Rightarrow\quad & \Pr\left(-1.96\frac{\sigma}{\sqrt{n}} < \overline{X}-\mu < 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95 \\
\Rightarrow\quad & \Pr\left(-\overline{X}-1.96\frac{\sigma}{\sqrt{n}} < -\mu < -\overline{X}+1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95 \\
\Rightarrow\quad & \Pr\left(\overline{X}-1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X}+1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95
\end{aligned}$$

Figure of a normal curve, confidence interval shaded

This can also be written as $\overline{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$.

If we look at the width of the interval, the larger n is (i.e., the more data we have), the more confidence we have in the estimate.

Powerpoint slides: CIs.pptx

Example 57

Let $X_i \overset{i.i.d.}{\sim} N(\mu, 3)$ and $\underline{\mathrm{X}} = \{X_1, X_2, X_3\}$, so that $\overline{X} \sim N(\mu, 1)$.
Suppose $\underline{\mathrm{x}} = \{-1, 1, 2\}$ and $\overline{x} = \frac{2}{3}$
We can construct a 95% confidence interval using $\overline{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$ so, approximately, we have $(-1.3, 2.6)$. In this example, we know that μ is really 0, so

$$\overline{X} \sim N(0, 1),$$

and we can calculate

$$\begin{aligned}\Pr\left(-1.3 < \overline{X} < 2.6\right) &= \Pr\left(\overline{X} < 2.6\right) - \Pr\left(\overline{X} < -1.3\right) \\ &= 0.9953 - 0.0968 \\ &= 0.8985.\end{aligned}$$

Remember the following:

- A 95% confidence interval is **not** 95% probability that the statistic is inside that interval,
- It is **not** a 95% probability that the true parameter is in that interval.
- It **is** such that, if we repeated the same experiment many times and we constructed the confidence interval in the same way, 95% of such intervals would contain θ.

We can also get asymmetric confidence intervals:

$$
\begin{aligned}
& \Pr\left(-\infty < \frac{\overline{X}-\mu}{\sqrt{\sigma^2/n}} < 1.6449\right) = 0.95, \\
\Rightarrow \quad & \Pr\left(\overline{X} - 1.6449\frac{\sigma}{\sqrt{n}} < \mu < \infty\right) = 0.95, \\
\textbf{or} \quad & \Pr\left(\overline{X} - 1.8808\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 2.0537\frac{\sigma}{\sqrt{n}}\right) = 0.95.
\end{aligned}
$$

Example 58

We wish to estimate the average percentage change in workforce over the past year for FTSE listed companies. From years of study, we know that they are normally distributed with variance of 3. We have access to this percentage change for just 10 companies. We have calculated $\overline{x} = 0.14\%$. Find a 99% confidence interval for μ.

We know $\Pr(Z < 2.5758) = 0.995$ for $Z \sim N(0,1)$. Hence,

$$
\Pr\left(\overline{X} - 2.5758\sqrt{\frac{3}{10}} < \mu < \overline{X} + 2.5758\sqrt{\frac{3}{10}}\right) = 0.99
$$

So, for our samples, a 99% confidence interval for μ is $(-1.27\%, 1.55\%)$, *i.e.* the width of the confidence interval is 2.82%.

We decide that this confidence interval is not accurate enough. How many companies' figures do we need to get a 99% confidence interval width of at most 1%?

We have that width $= 2 \times 2.5758\sqrt{\frac{3}{n}}$,

$$
\begin{aligned}
\Rightarrow \quad & 1 = 2 \times 2.5758 \times \frac{\sqrt{3}}{\sqrt{n}} \\
\Rightarrow \quad & \sqrt{n} = 2 \times 2.5758 \times \sqrt{3} \\
\Rightarrow \quad & n = 79.617
\end{aligned}
$$

Therefore, if we want to find a 99% CI of at most width 1, we need to get info from at least 80 companies.

If we have the sampling distribution for a statistic, we can derive a confidence interval. Previously, we had

$$
\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1} \text{ for } X_i \overset{i.i.d.}{\sim} N\left(\mu, \sigma^2\right), i = 1, ..., n, \text{ and } \mu, \sigma^2 \text{ unknown.}
$$

Here, a 95% confidence interval for μ is given by

$$\overline{X} \pm t_{0.025,n-1} \frac{S}{\sqrt{n}}.$$

We also had

$$\frac{(n-1)\, S^2}{\sigma^2} \sim \chi^2_{n-1} \text{ for same } X_i.$$

Therefore, a 95% confidence interval for σ^2 is

$$\left(\frac{(n-1)\, S^2}{\chi^2_{0.975,n-1}}, \frac{(n-1)\, S^2}{\chi^2_{0.025,n-1}} \right).$$

Example 59

(Workforce-change example continued)
Now, let's say we do not know the variance, but we still have data for 10 companies, $\overline{x} = 0.14\%, s^2 = 3$. Find a 99% confidence interval for μ.
If $T \sim t_9$, Then $\Pr(T < 3.2498) = 0.995$, so a 99% confidence interval for μ is given by

$$\overline{x} \pm 3.2498 \frac{\sqrt{3}}{\sqrt{10}} \Rightarrow (-1.64\%, 1.92\%), \qquad \text{(which is a width of 3.56\%).}$$

Find a 99% confidence interval for σ^2,

$$\begin{aligned} \chi^2_{0.005,9} &= 1.7359 \\ \chi^2_{0.995,9} &= 23.5894 \\ &\Rightarrow (1.14, 15.56). \end{aligned}$$

11.2 Confidence intervals for two sample problems

A comparison of parameters for two populations can be considered by taking **independent** random samples from each population. When samples are independent,

$$\text{Var}\left(\overline{X_1} - \overline{X_2}\right) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

If samples are not independent,

$$\text{Var}\left(\overline{X_1} - \overline{X_2}\right) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} - 2\text{Cov}\left(\overline{X_1}, \overline{X_2}\right).$$

So, in this case, the covariance could have a massive impact on the distribution of $\overline{X_1} - \overline{X_2}$ and the resulting confidence intervals.

11.2.1 Normal means

If $\overline{x_1}$ and $\overline{x_2}$ are means from independent random samples of size n_1 and n_2 (and we know that they follow normal distributions with variances σ_1^2 and σ_2^2), then a $100\,(1-\alpha)\,\%$ confidence interval for the difference in population means $\mu_1 - \mu_2$ is:

$$(\overline{x_1} - \overline{x_2}) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

where $z_{\alpha/2}$ is given by $\Pr\big(Z < z_{\alpha/2}\big) = \big(1 - \frac{\alpha}{2}\big)$ and $Z \sim N\,(0,1)$.
If σ_1^2 and σ_2^2 are unknown, but is it believed that $\sigma_1^2 = \sigma_2^2$, and we have sample statistics s_1^2 and s_2^2, then a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is:

$$(\overline{x}_1 - \overline{x}_2) \pm t_{\alpha/2, n_1+n_2-2}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

where $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$ is the pooled variance.

Example 60

We want to estimate the difference between the average salaries in $(\pounds k)$ at two companies. We believe that the salaries are normally distributed and each population has the same variance, but we don't know the variance:

$$\begin{aligned} n_1 &= 20 & n_2 &= 12 \\ \overline{x}_1 &= 31.5 & \overline{x}_2 &= 30.1 \\ s_1^2 &= 4.1 & s_2^2 &= 4.8 \end{aligned}$$

Then $\overline{x}_1 - \overline{x}_2 = 1.4$, and $s_p^2 = \frac{19 \times 4.1 + 11 \times 4.8}{20+12-2} = 4.356$. This is closer to s_1^2 because $n_1 > n_2$. Also, $t_{0.025,30} = 2.0423$.
Hence, the 95% confidence interval for $\mu_1 - \mu_2$ is:

$$1.4 \pm 2.0423\sqrt{4.356}\sqrt{\frac{1}{20} + \frac{1}{12}}, \Rightarrow (0.16, 2.96)$$

Roughly speaking, this confidence interval covers 0, so we cannot rule out $\mu_1 - \mu_2 = 0$ or, equivalently, $\mu_1 = \mu_2$.

11.2.2 Two sample normal variances

For two population variances, we consider $\frac{\sigma_1^2}{\sigma_2^2}$ rather than $\sigma_1^2 - \sigma_2^2$.
We have that:

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim \mathcal{F}_{n_1-1,n_2-1},$$

which leads to a 95% confidence interval of:

$$\frac{S_1^2}{S_2^2} \times \frac{1}{\mathcal{F}_{0.975,n_1-1,n_2-1}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2} \times \mathcal{F}_{0.975,n_2-1,n_1-1}.$$

Let $\mathcal{F}_{P/100,n_1-1,n_2-1}$ be the Pth percentile of the F-distribution, then

$$\mathcal{F}_{P/100,n_1-1,n_2-1} = \frac{1}{\mathcal{F}_{1-P/100,n_2-1,n_1-1}}.$$

We will be looking for confidence intervals that include 1, that is because $\sigma_1^2/\sigma_2^2 = 1 \Rightarrow \sigma_1^2 = \sigma_2^2$.

Example 61

Assume we have two independent samples that come from normal distributions: $n_1 = 4$, $n_2 = 5$, $s_1^2 = 3.33$ and $s_2^2 = 1.5$. From tables (or R: `qf`), we know that $\mathcal{F}_{0.975,3,4} = 9.98$ and $\mathcal{F}_{0.975,4,3} = 15.1$. Therefore, a 95% CI for σ_1^2/σ_2^2 is given by (0.22,34) to 2 s.f..

12 Hypothesis Testing (CT3 Unit 11)

A **hypothesis** is a proposed possible explanation for a set of circumstances. For example, we might hypothesise that the mean value for some population is some specified value.

The basic hypothesis being tested is the **null** hypothesis (H_0). In a hypothesis test, we contrast this with the **alternative** hypothesis (H_1).
For example, $H_0 : \mu = 0$ vs. $H_1 : \mu \neq 0$.

A **test** is a rule that divides the space of possible experimental results into two regions. In one region, the results are consistent with H_0. In the other, they are not (this second region is the **critical region**).

The test answers the question 'Do the data provide sufficient evidence to justify rejecting H_0?'. For the test, we need to find a suitable test statistic for which we can define a distribution under the null hypothesis.

The **level of significance** of the test, α, is the probability of committing a **type I error**; that is, rejecting H_0 when H_0 is true.

The probability of committing a **type II erro**r is β. This is the probability of not rejecting H_0 when it is false. $(1 - \beta)$ is called the **power** of the test.

An ideal test would minimise both α and β. In practice, this is difficult so we often fix α and pick a test that minimises β.

12.1 Single Sample Tests for Means

We have a sample from $N(\mu, \sigma^2)$, and we would like to know if it is reasonable to say $\mu = \mu_0$.
So we have:

$$\begin{array}{lrl} H_0 : \mu = \mu_0 \text{ vs. } H_1 : & \mu \neq \mu_0 & \text{two-sided} \\ & \mu > \mu_0,\ \mu < \mu_0 & \text{one-sided} \end{array}$$

If σ^2 is known, our test statistic is:

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1) \text{ under } H_0, \text{ or } X \sim N\left(\mu_0, \sigma^2\right).$$

If σ^2 is unknown, our test statistic is:

$$T = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}, \text{ under } H_0.$$

Example 62

We have $X_i \overset{i.i.d.}{\sim} N(\mu, 3)$ and $\underline{\mathrm{x}} = \{-1, 1, 2\}$. Our test is:

$$H_0 : \mu = 0 \text{ vs. } H_1 : \mu \neq 0.$$

Under H_0, $\frac{\overline{X}}{\sqrt{3}/\sqrt{3}} \sim N(0, 1)$.
We want to test at a 5% level of significance. The critical region is outside $\mu_0 \pm 1.96 = \pm 1.96$.
[**Diagram here showing critical region**]
Our test statistic $z = (\overline{x} - \mu_0) / \frac{\sigma}{\sqrt{n}} = \overline{x} = \frac{2}{3}$. In this case, $z = \overline{x}$: this is **not** always true. This is outside the critical region. Therefore, we do not have enough evidence to reject H_0 at a 5% level.

In general, the critical region is the complement of:

$$\begin{array}{cl} \mu_0 \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} & \text{for a 2-sided test and known } \sigma^2 \\ \text{or} & \\ \mu_0 \pm t_{\alpha/2, n-1} \frac{s}{\sqrt{n}} & \text{for a 2-sided test and unknown } \sigma^2. \end{array}$$

Example 63

Let $X_i \overset{i.i.d.}{\sim} N(\mu, 3)$ and $\underline{\mathrm{x}} = \{-1, 1, 2\}$. Our test is:

$$H_0 : \mu = -1 \text{ vs. } H_1 : \mu > -1.$$

Under H_0, $\frac{\left(\overline{X} - (-1)\right)}{\sqrt{3}/\sqrt{3}} \sim N(0, 1)$.
We test at a 5% level of significance, and our critical region is $\overline{X} + 1 > 1.6449$. With our data, $\overline{x} + 1 = \frac{5}{3} = 1.\dot{6}$, which is in the critical region. So we have enough evidence to reject H_0 at the 5% level.

Example 64

Let $X_i \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$ and $\underline{\text{x}} = \{-1, 1, 2\}$. The sample variance is 7/3. Again, our test is:

$$H_0 : \mu = -1 \text{ vs. } H_1 : \mu > -1.$$

Under H_0, $\frac{(\overline{X}-(-1))}{\sqrt{7/3}/\sqrt{3}} \sim t_2$.
We test at a 5% level of significance, and our critical region is $(\overline{X} + 1)/\sqrt{7/9} > 2.92$.
With our data, $(\overline{x}+1)/\sqrt{7/9} = 1.89$ (to 2 d.p.), which is not in the critical region. So we do not have enough evidence to reject H_0 at the 5% level.

12.2 P-Value

We can also assess the strength of the rejection of H_0 using the p-value.
The **p-value** is the probability of seeing a test statistic as extreme as the one in our experiment under the assumptions of H_0.

12.2.1 What the p-value is not

- It is **not** the probability that H_0 is true (or false).
- It is **not** the probability of getting the value of the test statistic.

Example 65

Let $X_i \overset{i.i.d.}{\sim} N(\mu, 3)$ and $\underline{\mathrm{x}} = \{-1, 1, 2\}$, so $\overline{x} = \frac{2}{3}$.

(a) $H_0 : \mu = 0$ vs. $H_1 : \mu \neq 0$
Then to find the p-value:

$$\begin{aligned} p\text{-value} &= \Pr\left(\left|\overline{X}\right| > \frac{2}{3} \middle| H_0 \right), \\ &= \Pr\left(\overline{X} < -\frac{2}{3} \middle| H_0 \right) + \Pr\left(\overline{X} > \frac{2}{3} \middle| H_0 \right), \\ &= 0.2525 + (1 - 0.7475), \\ &= 0.5050. \end{aligned}$$

So, under the null hypothesis, a value this extreme for the test statistic is not that unlikely.

(b) $H_0 : \mu = -1$ vs. $H_1 : \mu > -1$
Then to find the p-value:

$$\begin{aligned} p\text{-value} &= \Pr\left(\overline{X} > \frac{2}{3} \middle| H_0 \right), \\ &= 0.0478. \end{aligned}$$

Hence, the p-value< 0.05.
So we can reject H_0 at the 5% level, but, as p-value> 0.01, we cannot reject H_0 at the 1% level.

12.3 Single Sample Tests for Variance

We have a sample of size n from $N(0, 1)$. Our hypothesis will be:

$$H_0 : \sigma^2 = \sigma_0^2 \text{ vs. } H_1 : \sigma^2 \neq \sigma_0^2.$$

The test statistic is:

$$\frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}, \text{ under } H_0.$$

Example 66

Let $X_i \overset{i.i.d.}{\sim} N(\mu, \sigma^2), i = 1, ..., 15$ with $s^2 = 23.2$. We test:

$$H_0 : \sigma^2 = 10 \text{ vs. } H_1 : \sigma^2 > 10.$$

We might want to test at a 1% level of significance. Now,

$$\frac{14s^2}{10} > \chi^2_{0.99,14} = 29.14$$

As $s^2 = 23.2, \frac{14}{10} \times 23.2 = 32.48 > 29.14$. Hence, there is enough evidence at the 1% level to reject H_0. In this case,

$$p\text{-value} = \Pr\left(\frac{14s^2}{10} > 32.48 \middle| H_0 \right) = 0.0038 < 1\%.$$

12.4 Two-sample tests for population means and variances

We have independent samples of size n_1 and n_2 from $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ respectively, and we will consider:

$$\begin{aligned} H_0 : &\quad \mu_1 - \mu_2 = \delta \\ \text{vs.} & \\ H_1 : &\quad \mu_1 - \mu_2 \neq \delta \end{aligned}$$

If σ_1^2 and σ_2^2 are known, the test statistic under H_0 is

$$Z = \frac{\overline{X_1} - \overline{X_2} - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1).$$

If σ_1^2 and σ_2^2 are unknown, but we are assuming that $\sigma_1^2 = \sigma_2^2$, then under H_0,

$$T = \frac{\overline{X_1} - \overline{X_2} - \delta}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2},$$

where the pooled variance is

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$

This is the two-sample t-test.

For variances, a suitable test might be:

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ vs. } H_1 : \sigma_1^2 \neq \sigma_2^2$$

Here, under H_0, the test statistic is simply:

$$\frac{S_1^2}{S_2^2} \sim \mathcal{F}_{n_1-1,n_2-1}.$$

This is a formal test for the equal variances needed for the two-sampled t-test.

In what we have looked at so far, an important consideration is whether the two samples from each population are independent. However, this is not always the case.

Example 67

Consider a random sample of companies, which we have recorded profits for 2010 and 2011:

Company	2010 (£k)	2011 (£k)
A	33	29
B	32	28
C	27	41
⋮	⋮	⋮

Note that the 2011 sample depends upon the sample shown for 2010. We might be interested in whether the mean profit is the same across the two years:

$$H_0 : \mu_{2010} = \mu_{2011} \text{ vs. } H_1 : \mu_{2010} \neq \mu_{2011}$$

When we have paired data, we consider differences instead: $\delta = \mu_{2011} - \mu_{2010}$ to give the new test:

$$H_0 : \delta = 0 \text{ vs. } H_1 : \delta \neq 0$$

When data is paired, you take the difference between the two years, and treat as one sample test. We might say that $\delta \sim N(0, \sigma^2)$ under the null hypothesis.

12.5 χ^2 tests

This type of test is used to test the hypothesis that a variable follows some specified distribution. It depends on us having count data for the random variables of interest and a way of calculating the expected count given the specified distribution.

Example 68

We are testing if a six-sided die is fair. A suitable model is:

$$P\left(X=x\right)=\frac{1}{6}, \qquad x=1,2,...,6.$$

We have the test:

$$\begin{aligned} H_0 :&\quad X \text{ has this uniform distribution,}\\ \text{vs.}&\\ H_1 :&\quad X \text{ does } \textbf{not} \text{ have this distribution,}\end{aligned}$$

We are going to assume that we have a simple random sample (independent). I throw the die 300 times to get:

x :	1	2	3	4	5	6
f_i:	43	56	54	47	41	59
e_i :	50	50	50	50	50	50
$f_i - e_i$:	-7	6	4	-3	-9	9
$\left(f_i - e_i\right)^2$	49	36	16	9	81	81

Our test statistic is:

$$\sum\left[\frac{\left(f_i-e_i\right)^2}{e_i}\right] \mathrel{\dot{\sim}} \chi^2_{k-1},$$

where k is the number of possible values of x ($k=6$). In this case,

$$\sum_{i=1}^{6}\left[\frac{\left(f_i-e_i\right)^2}{e_i}\right]=5.44.$$

Under H_0: $\sum\left[\frac{(f_i-e_i)^2}{e_i}\right]\sim\chi^2_5$, and from tables, if $Y\sim\chi^2_5$, $P\left(Y>11.070\right)=0.05$.
So 5.44 is well outside the critical region (at the 5% level). Therefore, we do not have enough evidence to reject that the die is fair.

12.6 Contingency Tables

Draw a 2-way contingency table (maybe use example 70)

A contingency table is a two-way table of counts obtained when sample observations are classified according to two categorical variables. In this case, the null hypothesis is:

$$\begin{aligned} H_0: &\quad \text{the two classification criteria are independent,} \\ &\qquad \text{vs.} \\ H_1: &\quad \text{the two classification criteria have some dependence.} \end{aligned}$$

The test statistic is essentially a comparison between observed values in the table and values that we would expect to get if the criteria were independent. Expected value of cell in table $=\frac{\text{(row total} \times \text{column total)}}{\text{table total}}$.

Example 69

Consider results from a job survey.

	North	South	Total
Agriculture	*6*	*4*	10
Industry	*18*	*12*	30
Office	*36*	*24*	60
Total	60	40	100

Talk about maintaining proportions.

Now, our test statistic is given by

$$\sum_{\text{cells of the table}} \left[\frac{(f_i - e_i)^2}{e_i} \right] \dot{\sim} \chi^2_k.$$

where the degrees of freedom
$k =$ (Number of rows $-$ 1) $\times$ (Number of columns $-$ 1).

Example 70

A job survey was conducted for 200 individuals, with the following results:

	Job Category				
f_i	Unskilled	Skilled	Professional	Unemployed	Total
School	34	22	8	8	72
College	18	18	33	5	74
University	3	13	31	7	54
Total	55	53	72	20	200

We have:

$$\begin{aligned} H_0 &: \text{ Education is independent of job category} \\ &\text{vs.} \\ H_1 &: \text{ Education is not independent of job category} \end{aligned}$$

As expected frequency=$\frac{\text{(row total} \times \text{column total)}}{\text{table total}}$, we get:

	Job Category				
e_i	Unskilled	Skilled	Professional	Unemployed	Total
School	19.80	19.08	25.92	7.20	72
College	20.35	19.61	26.64	7.40	74
University	14.85	14.31	19.44	5.4	54

Note that: Degrees of freedom = 6.

In this case, $\sum \left[\frac{(f_i - e_i)^2}{e_i}\right] = 42.73$, then, if $Y \sim \chi^2_6, \mathrm{P}\,(Y > 42.73) = 0.0000001$.

Therefore, we can reject H_0 at a 5% (or 1%, or even 0.0001%) level.

Note: The χ^2 distribution is an approximation here. If any e_i is small, then it is a poor approximation. By small, we should avoid $e_i < 5$. If you get $e_i < 5$, you may need to combine categories or drop them all together.

13 Correlation and Regression (CT3 Unit 12)

We use **correlation** to measure the strength of linear relationships between two variables. This is closely related to the problem of fitting a straight line through some data points. The process of fitting a straight line is called **linear regression**.
In linear regression, we use a straight line to describe the relationship between y, the **response** (or **dependent**) variable, and x, the **explanatory** (or **independent**) variable.

Example 71

This example is taken from the CT3 book (unit 12, page 3).
A sample of ten claims and corresponding payments on settlement for household policies is taken from the business of an insurance company.
The amounts are in units of £100. We denote the claim by x and the corresponding payment by y.
Now, move to the slides: Correlation and regression.pptx.

13.1 Correlation

We are going to use the following samples statistics throughout this section:

$$\begin{aligned} S_{xx} &= \sum_{i=1}^{n}(x_i-\overline{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = \sum x_i^2 - n\overline{x}^2, \\ S_{yy} &= \sum_{i=1}^{n}(y_i-\overline{y})^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{n} = \sum y_i^2 - n\overline{y}^2, \\ (S_{yx} =)\quad S_{xy} &= \sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}) = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} \\ &= \sum x_i y_i - n\overline{x}\,\overline{y}, \end{aligned}$$

where we have n observations of X_i, Y_i pairs.

Derive the relationship for S_{xy} and relate to variance and covariance.

Correlation is defined as

$$r = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}.$$

Correlation, r, must be between $[-1, 1]$. If r is close to 1, there is a strong positive correlation.

Powerpoint slides: Correlation.pptx

We have $|S_{xy}| \leq \max\{S_{xx}, S_{yy}\}$, because $|r| \leq 1$.

Example 72

Continuation of example 72: see slides.

13.2 The Simple Linear Model

Given a set of n pairs of data $(x_i, y_i), i = 1, ..., n$; the y_i are regarded as a response random variable Y_i given x_i. For the purpose of this type of analysis, the x_i, the values of the explanatory variable, are regarded as constants.
The **simple linear model** is the following relationship:

$$Y_i = \alpha + \beta x_i + e_i, \quad i = 1, ..., n,$$

where the e_i are uncorrelated with mean 0 and have common variance σ^2; i.e., $\mathrm{E}(e_i) = 0$, $\mathrm{Var}(e_i) = \sigma^2$ and $\mathrm{Cov}(e_i, e_j) = 0, \forall\ i \neq j$. Note that:

$$\begin{aligned} \mathrm{E}\left[Y_i | x_i\right] &= \mathrm{E}\left[\alpha + \beta x_i + e_i\right] \\ &= \mathrm{E}[\alpha] + \mathrm{E}[\beta x_i] + \mathrm{E}[e_i] \\ &= \alpha + \beta x_i. \end{aligned}$$

Here, β is the slope parameter and α is the intercept. To fit the model, we must estimate α, β and σ^2. The fitted regression line (or fitted model) that we can use to estimate Y given x is given by

$$\hat{y} = \hat{\alpha} + \hat{\beta} x,$$

where

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}}$$

and

$$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}.$$

These estimates come from minimising the sum of squared errors for the regression model (exercise).

We want to find α and β such that $\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$ is minimised. Let's set

$$\begin{aligned}
h(\alpha,\beta) &= \sum_{i=1}^{n}[(\alpha+\beta x_i)-y_i]^2, \\
\frac{\partial h(\alpha,\beta)}{\partial\alpha} &= \sum_{i=1}^{n} 2(\alpha+\beta x_i - y_i), \\
\textit{set this to zero: } \sum_{i=1}^{n} y_i &= n\hat{\alpha} + \hat{\beta}\sum_{i=1}^{n} x_i \\
\implies \hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x}. \\
\frac{\partial h(\alpha,\beta)}{\partial\beta} &= \sum_{i=1}^{n} 2x_i(\alpha+\beta x_i - y_i), \\
\textit{set to zero \& sub in } \hat{\alpha}\textit{: } \sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i &= \hat{\beta}\left(\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i\right) \\
\implies \hat{\beta} &= S_{xy}/S_{xx}.
\end{aligned}$$

We could also perform the second partial derivative test to check it is a minimum.

Powerpoint slides: Correlation and regression.pptx

The estimate for the error variance is:

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n-2}\sum r_i^2,$$

where $r_i = (y_i - \hat{y}_i)$ are the residuals from the fitted model.

Warning: Lookout for outliers as they can give regression line leverage and skew data.

We also have that $\hat{\beta}$ is the observed value of a statistic $\hat{B}$, whose sampling distribution has the properties:

$$\begin{aligned}
\mathrm{E}\left(\hat{B}\right) &= \beta, \\
\mathrm{Var}\left(\hat{B}\right) &= \frac{\sigma^2}{S_{xx}}.
\end{aligned}$$

To help understand the 'goodness of fit' of our model to the data, we investigate the total variation in the response, which is $S_{yy} = \sum (y_i - \overline{y})^2$.

In particular, we want to know how much of the variation can be explained by our model as opposed to random variables. We look at departures from predicted responses:

$$\begin{aligned} y_i - \overline{y} &= (y_i - \hat{y}_i) + (\hat{y}_i - \overline{y}), \\ & \text{which gives,} \\ (y_i - \overline{y})^2 &= \left[(y_i - \hat{y}_i) + (\hat{y}_i - \overline{y})\right]^2, \end{aligned}$$

and, using $\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0$, we can get:

$$S_{yy} = \sum (y_i - \overline{y})^2 = \quad \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \overline{y})^2.$$

i.e.,

Total sum of squares = Residual sum of squares + Regression sum of squares

Note: This is sometimes written as SST=SSR+SSE, or TSS=RSS+ERR.

SST = sum of squares in total, SSR = sum of squares explained by the regression, SSE = sum of squares for the errors.
TSS = total sum of squares, RSS = regression sum of squares, ERR = error sum of squares

We will use SSTOT=SSRES+SSREG. For computation,

$$\begin{aligned} \text{SSTOT} = &\ S_{yy} \\ \text{SSREG} = &\ \sum_{i=1}^{n} \left[\left(\hat{\alpha} + \hat{\beta} x_i \right) - \left(\hat{\alpha} + \hat{\beta} \overline{x} \right) \right]^2 = \frac{S_{xy}^2}{S_{xx}} \\ \text{SSRES} = &\ S_{yy} - \frac{S_{xy}^2}{S_{xx}} \end{aligned}$$

For this model, the level of correlation between X and Y is highly related to the goodness of fit. The proportion of the total variability in y is explained by the linear model and is called the **coefficient of determination**,

$$R^2 = \frac{\text{SSREG}}{\text{SSTOT}} = \frac{S_{xy}^2}{S_{xx} S_{yy}}.$$

Note that:

- If $R = 1$, the model provides a perfect fit,
- If $R = 0$, the model is doing nothing,
- For a linear model, R^2 is the square of the correlation coefficient.

Powerpoint slides: Correlation and regression (part 2).pptx

Example 73

Continuing the earlier example, we have SSTOT = 7.159, SSREG = 6.573 and SSRES = 0.586.
And the coefficient of determination is 91.8%.

13.3 The Normal Linear Model

We have already seen that $\hat{\beta}$ is a sample statistic for some statistic $\hat{B}$, which has $\mathrm{E}\left[\hat{B}\right] = \beta$ and $\mathrm{Var}\left(\hat{B}\right) = \frac{\sigma^2}{S_{xx}}$.
We might want to test if $\hat{B}$ is significantly different from zero, and, to do this, we need to make assumptions about the distributions of the errors.
Our model is:

$$Y_i = \alpha + \beta x_i + e_i,$$

and, now, because we have $e_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$, it follows that given x_i, the Y_i are normally distributed random variables with

$$\begin{aligned} \mathrm{E}\left[Y_i\right] &= \alpha + \beta x_i, \\ \mathrm{Var}(Y_i) &= \sigma^2. \end{aligned}$$

It can be shown that $\hat{B} \sim N\left(\beta, \frac{\sigma^2}{S_{xx}}\right)$, $\hat{B}$ and σ^2 are independent, and $\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$.
Now, if σ^2 is known,

$$\frac{\hat{B} - \beta}{(\sigma^2/S_{xx})^{\frac{1}{2}}} \sim N(0, 1).$$

Of course, this is rarely the case and, because $\hat{B}$ and $\hat{\sigma}^2$ are independent,

$$\frac{\frac{\hat{B}-\beta}{(\sigma^2/S_{xx})^{\frac{1}{2}}}}{\left[\left(\frac{(n-2)\hat{\sigma}^2}{\sigma^2}\right)/(n-2)\right]^{\frac{1}{2}}} \sim t_{n-2}$$

$$\Rightarrow \qquad \frac{\hat{B} - \beta}{(\hat{\sigma}^2/S_{xx})^{\frac{1}{2}}} \sim t_{n-2}.$$

Using this distribution, we can calculate confidence intervals for β and test hypotheses like $H_0 : \beta = 0$ vs. $H_1 : \beta \neq 0$.

Example 74

A 95% CI for β is given by

$$\hat{\beta} \pm t_{0.025,n-2} \left(\frac{\hat{\sigma}^2}{S_{xx}}\right)^{1/2}.$$

In our example, $\hat{\beta} = 0.8823, \hat{\sigma}^2 = 0.0732, S_{xx} = 8.444$ and $n = 10$. From tables, $t_{0.025,8} = -2.306$. Therefore, a 95% CI for β is

$$0.8823 \pm 2.306(0.0732/8.444)^{1/2} \;=\; 0.8823 \pm 0.2147,$$

or (0.6676, 1.0970).

Consider the following hypothesis test at the 1% level of significance: $H_0 : \beta = 0$ vs. $H_1 : \beta \neq 0$.

Under H_0,

$$\frac{\hat{B}}{(\hat{\sigma}^2/S_{xx})^{\frac{1}{2}}} \sim t_8.$$

Using a two-sided test, we will reject the null hypothesis if $\left|\frac{\hat{\beta}}{(\hat{\sigma}^2/S_{xx})^{1/2}}\right| >$ 3.355 (which is the 99.5% point of a t_8 distribution). Our test statistic is 9.476 to 3 d.p.. Therefore, we have enough evidence to reject the hypothesis that $\beta = 0$. In terms of our problem, this means that our data give support to a (positive) linear relationship between claim value and the amount paid out by the insurer.

13.4 Using our model to estimate mean responses and to predict individual responses

Let μ^* be the expected response for a value x^*:

$$\mu^* = \mathrm{E}\left[Y|x^*\right] = \alpha + \beta x^*,$$

μ^* is estimated by $\hat{\mu}^* = \hat{\alpha} + \hat{\beta}x^*$, and the variance of this estimator is

$$\mathrm{Var}(\hat{\mu}^*) = \left\{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}\right\}\hat{\sigma}^2.$$

We can find:

$$\frac{(\hat{\mu}^* - \mu^*)}{\text{s.e.}(\hat{\mu}^*)} \sim t_{n-2},$$

where the standard error of $\hat{\mu}^*$ is:

$$\text{s.e.}\,(\hat{\mu}^*) = \left[\left\{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}\right\}\hat{\sigma}^2\right]^{\frac{1}{2}}.$$

This can be used to produce confidence intervals for μ^*.
Instead of considering the mean response, we sometimes need to look at predicted responses for some x^*, which we denote y^*.
The estimate is:

$$\hat{y}^* = \hat{\alpha}^* + \hat{\beta}x^*.$$

However, the uncertainty associated with this estimate is greater. In this case,

$$\frac{(\hat{y}^* - y^*)}{\text{s.e.}(\hat{y}^*)} \sim t_{n-2},$$

where

$$\text{s.e.}(\hat{y}^*) = \left[\left\{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}\right\}\hat{\sigma}^2\right]^{\frac{1}{2}}.$$

Example 75

We are interested in estimating what we pay on average and variability there could be in payouts when people are claiming £350.

$$\begin{aligned}
\mu^* &= \hat{\alpha} + \hat{\beta}x^* \\
&= 0.1636 + 0.8823 \times 3.5 \;=\; \pounds 325.17 \;=\; \hat{y}^*. \\
\text{s.e.}(\hat{\mu}^*) &= \left[\left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right)\hat{\sigma}^2\right]^{1/2} \;=\; 0.086. \\
\text{s.e.}(\hat{y}^*) &= \left[\left(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right)\hat{\sigma}^2\right]^{1/2} \;=\; 0.284.
\end{aligned}$$

The standard error is three times bigger for the possible value of prediction rather than the mean.

13.5 Multiple Linear Regression

An extension of the simple linear model is made by allowing for multiple explanatory variables:

$$Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + e_i, \qquad i = 1, ..., n.$$

with $e_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$.
We end up with very similar sampling distributions that enable us to test if $\beta_i = 0$ or not. We can also use the same sum of squares partition to get the population of variance explained by the model.

14 ANalysis Of VAriance (ANOVA) (CT3 Unit 13)

A 'one-way analysis of variance' (or one-way ANOVA) is used to compare k 'treatments' when an experiment provides n_i responses for treatment i, $i = 1, ..., k$. The data are then $n\,(= \sum_i n_i)$ responses y_{ij}, where y_{ij} is the j^{th} observation for the i^{th} treatment. 'Treatment' for us could mean company, country, employment status, etc..

Essentially, we want to know if there is a difference in the mean response across the differential treatments. The model is:

$$Y_{ij} = \mu + \tau_i + e_{ij}, \quad i = 1, ..., k, j = 1, ..., n, \quad e_{ij} \overset{i.i.d.}{\sim} N(0, \sigma^2).$$

We define the overall mean to be:

$$\mu = \frac{1}{n} \sum_i \sum_j \mathrm{E}\left[Y_{ij}\right]$$

and τ_i is the difference made on average by treatment i.

Estimation of the parameters follows a least-squares maximisation procedure:

$$\begin{aligned}
\hat{\mu} &= \overline{Y}_{\bullet\bullet} \text{ where } \overline{Y}_{\bullet\bullet} = \frac{1}{n} \sum_i \sum_j Y_{ij} \\
\hat{\tau}_i &= \overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet} \text{ where } \overline{Y}_{i\bullet} = \frac{1}{n_i} \sum_{j=1}^{n} Y_{ij}. \\
\text{Also,} \qquad \hat{\sigma}^2 &= \frac{1}{n-k} \sum_i \sum_j \left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 \\
\text{and} \qquad & \frac{1}{\sigma^2} \sum_i \sum_j \left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 \sim \chi^2_{n-k}.
\end{aligned}$$

Again, we partition the variability:

$$\begin{aligned}
\sum_i \sum_j \left(Y_{ij} - \overline{Y}_{\bullet\bullet}\right)^2 &= \sum_i \sum_j \left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 + \sum_i n_i \left(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}\right)^2. \\
(\text{SSTOT} \quad &= \quad \text{SSRES} \quad + \quad \text{SSTRE})
\end{aligned}$$

A typical hypothesis test for this type of model is:

$$\begin{aligned}
&H_0 : \tau_i = 0, \quad i = 1, ..., k, \\
&\qquad \text{vs.} \\
&H_1 : \tau_i \neq 0, \quad \text{for at least one } i.
\end{aligned}$$

It can be shown that, under H_0,

$$\begin{aligned} F &= \left[\frac{\text{SSTRE}}{k-1}\right] \Big/ \left[\frac{\text{SSRES}}{n-k}\right] \\ &= \frac{\text{Between treatment mean squared error}}{\text{Residual mean squared error}} \end{aligned}$$

and $F \sim \mathcal{F}_{k-1,n-k}$. We reject H_0 if F is large.
We usually set out the components of this test in an ANOVA table:

Source of variation	Degrees of freedom	Sum of squares	Mean Squares	F-statistic
Between treatments	$k-1$	SSTRE	$\frac{\text{SSTRE}}{k-1}$	$\frac{\text{SSTRE}\times(n-k)}{\text{SSRES}\times(k-1)}$
Residual	$n-k$	SSRES	$\frac{\text{SSRES}}{n-k}$	
Total	$n-1$	SSTOT		

Example 76

We have 5 companies and we want to know if accountants within each company have significantly different average salaries:

$$\begin{gathered} H_0 : \tau_i = 0, \quad i = 1, ..., 5 \\ \text{vs.} \\ H_1 : \tau_i \neq 0, \quad \text{for at least one } i \end{gathered}$$

The ANOVA table gives:

Source of variation	Degrees of freedom	Sum of squares	Mean Squares	F-statistic
Between treatments	4	128.9	32.23	1.85
Residual	45	782.2	17.38	
Total	49	911.1		

From tables, $\mathcal{F}_{4,45}(95\%) = 2.58$. $1.85 < 2.58$ so we don't have enough evidence at the 5% level to reject H_0.

15 Assessing Model Fit Numerically

When we have several competing models, there are many metrics available to help us decide which model is 'best'. If we are simply interested in how well a model fits data, we may consider the **root mean squared-errors** (RMSE):

$$\begin{aligned} \text{MSE} &= \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2, \\ \text{RMSE} &= \sqrt{\text{MSE}}. \end{aligned}$$

We might be interested in finding the model with smallest RMSE. In general, the more variables/parameters that we add to our model, the lower the RMSE, *regardless* of whether the variables are related or if the extra parameters explain the variation (this is called overfitting).

Diagram demonstrating overfitting

15.1 Information Criteria

The Akaike information criterion (AIC) provides a way of judging models that takes into account the number of parameters being used as well as the fit. We use the following formula:

$$\text{AIC} = 2k - 2\ln(l^*),$$

where k is the number of parameters and l^* is the maximised likelihood for the fitted model.
Again, we want to minimise AIC. AIC is smaller for simpler models (small k) and for models with larger maximum likelihood.
There is another metric that penalises for the number of parameters even more: the **Baysian Information Criterion** (BIC):

$$\text{BIC} = \ln(n)k - 2\ln(l^*).$$

where n is the number of observations. (If $n > 8$, BIC will be larger than AIC.) We try to find the model with the smallest BIC.

Example 77

Fit three different models to 50 observations.

Model	No. of parameters	l^*	RMSE	AIC	BIC
1	3	8.28	0.4821	1.77	9.58
2	6	14.16	0.4511	6.70	18.7
3	100	18.31	0.0026	194.19	385.39

From this table, we can see we penalise the model where we use 100 parameters. In this example, there are 50 observations and 100 parameters: we should not use this model.

15.2 Cross-validation

An alternative approach to avoiding over-fitting— and to make comparisons between models — which does not inherently favour a specific model assumption, is to use *cross-validation.*

The main idea is to split the n observations into two complementary sets: a training or learning set, say $L \subset \{1, \ldots, n\}$ and a test set, say $T = \{1, \ldots, n\} \cap L^c$. Any model or procedure is then fit to the data in L, but then models are compared using the data in T. To give an example, a regression model (which could be nonparametric) could be fit to observations $(x_i, y_i), i \in L$ and then the model could be used to predict y for each $x_i, i \in T$, giving $\hat{y}_i, i \in T$. Then comparisons can be made by considering

$$\sum_{\{i, i \in T\}} (y_i - \hat{y}_i)^2.$$

Such an approach can be used for selecting smoothing parameters, by successively choosing $T = \{1\}, T = \{2\}, \ldots, T = \{n\}$ and this is known as "leave-one-out cross-validation".

For the purposes of model or method comparisons it would be common to choose the size of L to be approximately $n/2$, with observations selected at random from the original set. When forecasting is the objective (as in the next chapter) it would then be more natural to take the test set T to be the latest observations available.

16 Time Series (Brooks Ch 5)

A **time series** is a set of variables (or data) that are indexed by some notion of time.

Examples:

- Daily close price for some stock,
- Monthly interest rates for some national bank (time refers to month),
- Stock price after trades (time is ordering or trades).

A **time series model** is a model that uses past time series observations to provide a mechanism to predict future values; for example, using historical stock data to predict future prices. The time series models we will be looking at will not contain any external explanatory variables. Our models are just based on past observations (this is called an *empirical* approach to modelling).

16.1 Notation and Important Concepts

Our time series are sequences of random variables:

$$\ldots, X_0, X_1, X_2, ... \text{ or } \ldots, Y_0, Y_1, Y_2, ...,$$

or, once they have been observed, sequences of data values:

$$x_0, x_1, x_2, ..., x_n \text{ or } y_0, y_1, y_2, ...y_n.$$

We also have the concept of **lag**. There is a lag of 1 between x_1 and x_2, a lag of 2 between x_5 and x_7, etc.. In general, there is a lag of s between x_k and x_{k+s}.
A **strictly stationary process** is a time series where for any time indices and integer k,

$$F_{Y_{t_1},Y_{t_2},...Y_{t_r}}(y_1, ..., y_r) = F_{Y_{t_{1+k}},Y_{t_{2+k}},...Y_{t_{r+k}}}(y_1, ..., y_r)$$

That is, the joint distribution of the variables making up the time series does not change over time.
A **weakly stationary process** is a time series such that:

$$\begin{aligned} \mathrm{E}\left[Y_t\right] &= \mu, & \forall\, t, \\ \mathrm{Var}\left(Y_t\right) &= \sigma^2 < \infty, & \forall\, t, \\ \mathrm{Cov}\left(Y_{t_1}, Y_{t_2}\right) &= \gamma_{t_2-t_1}, & \forall\, t_1, t_2. \end{aligned}$$

The **autocovariance** between two time points $t_2 > t_1$ is

$$\gamma_{t_2 - t_1} = \mathrm{Cov}\left(Y_{t_1}, Y_{t_2}\right).$$

The **autocorrelation** at lag s is given by

$$\tau_s = \gamma_s / \gamma_0$$

if we have a weakly stationary process.

"Auto-" comes from the Greek for self.

The partial autocorrelation at lag s, denoted by α_s, is the autocorrelation between Y_t and Y_{t+s} with the linear dependence of Y_{t+1} to Y_{t+s-1} removed. It can also be thought of as the autocorrelation between Y_t and Y_{t+s} that is not accounted for by the autocorrelation at lags 1 to $s-1$.

Think about a perfect linear relationship between two neighbouring variables in a time series. What does this mean for autocorrelation and partial autocorrelation?

There are three main plots that we will consider when analysing time series:

i) Data value vs. time

ii) Autocorrelation vs. lag

iii) Partial autocorrelation vs. lag

All of these can be used to help select a time series model.

Diagram showing plot types: Univariate time series.pptx

16.2 White Noise Process

The model is:

$$X_t = \varepsilon_t, \quad \varepsilon_t \overset{i.i.d.}{\sim} G_\varepsilon\left(0, \sigma_\varepsilon^2\right)$$

where G_ε is any distribution $(t, \mathcal{F}, ...)$ that satisfies some properties that are given below. Note that σ_ε^2 is the only parameter. We have **Gaussian white noise** if $G_\varepsilon\left(0, \sigma_\varepsilon^2\right) = N\left(0, \sigma_\varepsilon^2\right)$.
White noise processes have the following properties:

i) $\mathrm{E}\left(X_t\right)=0, \qquad \forall\, t,$

ii) $\mathrm{Var}\left(X_t\right)=\sigma_\varepsilon^2, \qquad \forall\, t,$

iii) $\mathrm{Cov}\left(X_t, X_{t+k}\right)=0,\ \forall\, t, k \neq 0.$

Explain consequences of each

16.3 Autoregressive Process

An **autoregressive process model** of order p, AR(p), is:

$$X_t=c+\sum_{i=1}^{p}\phi_i X_{t-i}+\varepsilon_t, \quad \varepsilon_t \overset{i.i.d.}{\sim} G_\varepsilon\left(0,\sigma_\varepsilon^2\right),$$

where $\phi_1, ..., \phi_p, c, \sigma_\varepsilon^2$ are model parameters.

Consider an AR(1) model such that $X_t=c+\phi_1 X_{t-1}+\varepsilon_t$, then

i) $\mathrm{E}\left[X_t\right]=c+\phi_1 \mathrm{E}\left[X_{t-1}\right]$,
but $\mathrm{E}\left[X_{t-1}\right]=c+\phi_1 \mathrm{E}\left[X_{t-2}\right]$,
so,

$$\begin{aligned}\mathrm{E}\left[X_t\right] &= c+\phi_1\left(c+\phi_1 \mathrm{E}\left[X_{t-2}\right]\right)\\ &= c(1+\phi_1+\phi_1^2+...).\end{aligned}$$

Now, if $|\phi_1|<1$, this is a geometrically decreasing sum and

$$\mathrm{E}\left[X_t\right]=\frac{c}{1-\phi_1}.$$

ii) For variance,

$$\begin{aligned}\mathrm{Var}\left(X_t\right) &= \phi_1^2 \mathrm{Var}\left(X_{t-1}\right)+\sigma_\varepsilon^2\\ &= \frac{\sigma_\varepsilon^2}{(1-\phi_1^2)}\end{aligned}$$

if and only if $|\phi_1|<1$.

iii) Similarly, for autocovariances at lag s,

$$\mathrm{Cov}\left(X_t, X_{t+s}\right)=\gamma_s=\frac{\phi_1^s \sigma_\varepsilon^2}{(1-\phi_1^2)},$$

and the autocorrelation at lag s is

$$\begin{aligned}\rho_s &= \frac{\gamma_s}{\gamma_0} = \frac{\text{Cov}\left(X_t, X_{t+s}\right)}{\sqrt{\text{Var}(X_t)\text{Var}(X_{t+s})}} \\ &= \phi_1^s.\end{aligned}$$

Example 78

Consider the AR(1) model. What is the autocovariance at lag 1?

$$\begin{aligned}\text{Cov}\left(X_t, X_{t+1}\right) &= \text{Cov}\left(X_t, c + \phi_1 X_t + \varepsilon_{t+1}\right), \\ &= \text{Cov}\left(X_t, c\right) + \text{Cov}\left(X_t, \phi_1 X_t\right) + \text{Cov}\left(X_t, \varepsilon_{t+1}\right) \\ &= \phi_1 \text{Cov}\left(X_t, X_t\right), \\ &= \phi_1 \text{Var}\left(X_t\right), \\ &= \frac{\phi_1 \sigma_\varepsilon^2}{(1-\phi_1^2)}.\end{aligned}$$

Powerpoint slides: Univariate time series.pptx

More generally, we now consider the acf and pacf of the AR(p) model. If it is stationary then $\text{E}[X_t] = \mu$ for all t and so

$$\begin{aligned}\mu = \text{E}[X_t] &= c + \phi_1 \text{E}[X_{t-1}] + \cdots + \phi_p \text{E}[X_{t-p}] + \text{E}[\varepsilon_t] \\ &= c + \sum_{i=1}^{p} \phi_i \mu\end{aligned}$$

and so

$$\mu(1 - \phi_1 - \cdots - \phi_p) = c \tag{1}$$

and

$$\mu = \text{E}[X_t] = \frac{c}{1 - \phi_1 - \cdots - \phi_p}.$$

Note that the condition for an AR(p) process to be stationary is that the roots of $\Phi(B) = 0$ lie outside the unit circle, where $\Phi(B)$ is defined by

$$\Phi(B)X_t = (1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)X_t = X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p}$$

and B is known as the “backshift operator” (see Section 16.9).

To obtain the covariances (and then the autocorrelations) we first note that — using (1) —

$$X_t - \mu = \sum_{i=1}^{p} \phi_i B^i (X_t - \mu) + \varepsilon_t$$

and so, without loss of generality, we can set $\mu = 0$ and replace $X_t - \mu$ by X_t. Then

$$\begin{aligned} \gamma_k = \mathrm{E}[X_t X_{t-k}] &= \sum_{i=1}^{p} \phi_i \mathrm{E}[X_t X_{t-k}] + \mathrm{E}[X_{y-k}\varepsilon_t] \\ &= \sum_{i=1}^{p} \phi_i \gamma_{k-i} \qquad \text{for} \quad k > 0 \end{aligned}$$

where we note that $\gamma_k = \gamma_{-k}$.

Dividing by $\gamma_0 = \mathrm{Var}[X_t]$ gives a recursive relationship

$$\rho_k = \phi_1 \rho_{k-1} + \cdots \phi_p \rho_{k-p} \qquad (k > 0).$$

The **partial autocorrelation** can be derived as follows. Consider the regression model where X_{t+k} is regressed on $X_{t+k-1}, \ldots, X_t$, i.e.

$$X_{t+k} = \phi_{k1} X_{t+k-1} + \cdots + \phi_{kk} X_t + \varepsilon_{t+k}$$

where the subscripts on ϕ show both the number of regressors (k) and the coefficient label $(1, \ldots, k)$. Mutiplying through by X_t, taking expectations, and proceeding as above, we obtain

$$\rho_j = \phi_{k1} \rho_{j-1} + \cdots + \phi_{kk} \rho_{j-k}.$$

Thus for $k = 1, 2, \ldots$ gives a system of equations:

$$\begin{array}{ccccccc} \rho_1 & = & \phi_{k1}\rho_0 & + & \cdots & + & \phi_{kk}\rho_{k-1} \\ \vdots & = & \vdots & & & & \vdots \\ \rho_k & = & \phi_{k1}\rho_{k-1} & + & \cdots & + & \phi_{kk}\rho_0 \end{array}$$

(where $\rho_0 = 1$).

Using Cramer's rule, we can solve (for $k = 1, 2, \ldots$) to obtain

$$\begin{aligned}
\phi_{11} &= \rho_1 \\
\phi_{22} &= \frac{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & \rho_2 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{vmatrix}} = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2} \\
\vdots \quad & \quad \vdots \\
\phi_{kk} &= \frac{\begin{vmatrix} 1 & \rho_1 & \cdots & \rho_{k-2} & \rho_1 \\ \vdots & \vdots & & \vdots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \cdots & \rho_1 & \rho_k \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 & \cdots & \rho_{k-2} & \rho_{k-1} \\ \vdots & \vdots & & \vdots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \cdots & \rho_1 & 1 \end{vmatrix}}
\end{aligned}$$

We can obtain the pacf for the AR(p) model we note that, when $k > p$ the last column in the numerator of ϕ_{kk} can be written as a linear combination of previous columns, and so the determinant will be zero, and so we see that, for an AR(p) process

$$\phi_{kk} = 0 \qquad \text{for} \quad k > p.$$

16.4 Moving average process

A **moving average model** of order q, MA(q), is

$$X_t = c + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}, \quad \varepsilon_t \overset{i.i.d.}{\sim} G_\varepsilon \left(0, \sigma_\varepsilon^2\right),$$

where all error terms are independent and $c, \theta_1, ..., \theta_q, \sigma_\varepsilon^2$ are model parameters. Consider MA(1),

$$\begin{aligned}
X_t &= c + \varepsilon_t + \theta_1 \varepsilon_{t-1}, \\
\mathrm{E}\left[X_t\right] &= c, \\
\mathrm{Var}\left(X_t\right) &= \sigma_\varepsilon^2 + \theta_1^2 \sigma_\varepsilon^2 = \sigma_\varepsilon^2 \left(1 + \theta_1^2\right).
\end{aligned}$$

We also have

$$\begin{aligned}
\mathrm{Cov}\left(X_t, X_{t+1}\right) &= \gamma_1 \\
&= \mathrm{Cov}\left(c + \varepsilon_t + \theta_1 \varepsilon_{t-1}, c + \varepsilon_{t+1} + \theta_1 \varepsilon_t\right) \\
&= \theta_1 \mathrm{Cov}(\varepsilon_t, \varepsilon_t) \\
&= \theta_1 \sigma_\varepsilon^2,
\end{aligned}$$

because $\text{Cov}(\varepsilon_t, \varepsilon_t) = \sigma_\varepsilon^2$ and $\text{Cov}(c, c) = \text{Cov}(\varepsilon_t, \varepsilon_{t-1}) = 0$.
Further,

$$\rho_1 = \frac{\theta_1}{1+\theta_1^2}, \quad \rho_s = 0 \qquad \forall s > 1.$$

More generally, we consider the MA(q) process. Using similar calculations as for the AR(p) process we see that, if

$$X_t - \mu = (1 + \sum_{i=1}^{q} \theta_i B^i)\varepsilon_t$$

then

$$\sigma_X^2 = \text{Var}[X_t] = \gamma_0 = \sigma_\varepsilon^2 \sum_{i=0}^{q} \theta_i^2 \qquad \text{with} \quad \theta_0 = 1$$

and

$$\gamma_k = \begin{cases} \sigma_\varepsilon^2(\theta_k + \theta_1\theta_{k+1} + \cdots + \theta_{q-k}\theta_q) & k = 1, \ldots, q \\ 0 & k > q \end{cases}$$

and so

$$\rho_k = \begin{cases} \frac{\theta_k + \theta_1\theta_{k+1} + \cdots + \theta_{q-k}\theta_q}{1+\theta_1^2+\cdots+\theta_q^2} & k = 1, \ldots, q \\ 0 & k > q \end{cases}$$

16.5 Autoregressive Moving Average Process

An **autoregressive moving average model** with integer parameters p and q, ARMA(p, q) is

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}, \quad \varepsilon_t \overset{i.i.d.}{\sim} G_\varepsilon\left(0, \sigma_\varepsilon^2\right),$$

where $c, \theta_1, ..., \theta_q, \sigma_\varepsilon^2, \phi_1, ..., \phi_p$ are model parameters. Consider ARMA(1,2):

$$\begin{aligned} X_t &= c + \varepsilon_t + \phi_1 X_{t-1} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2}, \\ \text{E}\left[X_t\right] &= c + \phi_1 \text{E}\left[X_{t-1}\right], \\ &= \frac{c}{1-\phi_1}, \\ \text{Var}\left(X_t\right) &= \sigma_\varepsilon^2 + \text{Var}\left(\phi_1 X_{t-1} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2}\right). \end{aligned}$$

Because $X_{t-1}, \varepsilon_{t-1}$ and ε_{t-2} are related, it is not simple to derive this.

16.6 Identifying which model to use

We could use the `acf(.)` and `pacf(.)` plots to help identify models to fit. For an AR(p) model:

acf: magnitude decreases geometrically,
pacf: no significant partial autocorrelation after lag p.

For a MA(q) model:

acf: no significant autocorrelation after lag q,
pacf: magnitude decreases geometrically.

For a ARMA(p, q):

acf: like acf for AR model after first few lags,
pacf: like pacf for MA model after first few lags.

We can also use model selection criteria, AIC and BIC, to help select our model.

16.7 Model Checking

Choosing a good model for a set of data is not always easy, so it is important to compare models (for example using AIC), and then to check that the residuals from a fitted model form a white noise process. This can be checked by plotting the residuals, plotting an acf of the residuals, and using the Box-Pierce or Ljung-Box "portmanteau" tests, for which the test statistics are given by

$$Q = n \sum_{k=1}^{K} \hat{\rho}_k^2$$

and

$$Q = n(n+2) \sum_{k=1}^{K} (n-k)^{-1} \hat{\rho}_k^2$$

respectively. Here, K is a user-chosen number to reflect the number of lags which are to be considered, and the statistic is compared to a χ^2 distribution with $K - p - q$ degrees of freedom.

16.8 Forecasting with Time Series Models

Often, we will want to ask: what value do we expect our time series to take several steps ahead? One way to think about this is to work out what the expectation and variance for the time series are several steps ahead. If we do it this way, then we have a point estimate and some appreciation of its accuracy.

Example 79

Consider the AR(2) model: $X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \varepsilon_t$.
We know that $c = 0, \phi_1 = 0.2, \phi_2 = -0.5$ and $\sigma^2_\varepsilon = 1$ (we will denote these by $\mathcal{P}$).
We observe $x_0 = 0.5, x_1 = 0.9$. What are x_2, x_3, x_4?

$$\begin{aligned}
\mathrm{E}\left[X_2|x_0, x_1, \mathcal{P}\right] &= (0.2)(0.9) - (0.5)(0.5), \\
&= -0.07, \\
\mathrm{Var}\left(X_2|x_0, x_1, \mathcal{P}\right) &= \sigma^2_\varepsilon = 1, \\
\\
\mathrm{E}\left[X_3|x_0, x_1, \mathcal{P}\right] &= \phi \mathrm{E}\left[X_2|x_0, x_1, \mathcal{P}\right] + \phi_2 x_1, \\
&= (0.2)(-0.07) + (-0.5)(0.9), \\
&= -0.464, \\
\mathrm{Var}\left(X_3|x_0, x_1, \mathcal{P}\right) &= \phi_1^2 \mathrm{Var}\left(X_2|x_0, x_1, \mathcal{P}\right) + \sigma^2_\varepsilon, \\
&= 0.04 + 1, \\
&= 1.04.
\end{aligned}$$

The further away we go away from 0, the more uncertain we are that σ^2 will increase:

$$\begin{aligned}
\mathrm{E}\left[X_4|x_0, x_1, \mathcal{P}\right] &= (0.2)(-0.464) + (-0.5)(-0.07), \\
&= -0.0578
\end{aligned}$$

$\mathrm{Var}\left(X_4|x_0, x_1, \mathcal{P}\right)$ is more complicated due to the covariance between X_2 and X_3.

$$\begin{aligned}
\mathrm{Var}\left(X_4|x_0, x_1, \mathcal{P}\right) &= \mathrm{Var}(\phi_1 X_3 + \phi_2 X_2|x_0, x_1, \mathcal{P}) + \sigma^2_\varepsilon, \\
\mathrm{Var}(\phi_1 X_3 + \phi_2 X_2|x_0, x_1, \mathcal{P}) &= \phi_1^2 \mathrm{Var}(X_3) + \phi_2^2 \mathrm{Var}(X_2) \\
&\qquad + 2\phi_1\phi_2 \mathrm{Cov}(X_2, X_3), \\
\mathrm{Cov}(X_2, X_3) &= \mathrm{Cov}(X_2, c + \phi_1 X_2 + \phi_2 x_1 + \varepsilon_3) \\
&= \phi_1 \mathrm{Var}(X_2), \\
\mathrm{Var}\left(X_4|x_0, x_1, \mathcal{P}\right) &= \sigma^2_\varepsilon + \phi_1^2 \mathrm{Var}(X_3) + (\phi_2^2 + \phi_1)\mathrm{Var}(X_2) \\
&= 1 + (0.2)^2 1.04 + (0.5^2 + 0.2)1 \ = \ 1.4916.
\end{aligned}$$

As t increases, $\mathrm{E}\left[X_t|x_0, x_1, \mathcal{P}\right]$ will tend to the mean of the process and $\mathrm{Var}\left(X_t|x_0, x_1, \mathcal{P}\right)$ will grow until we lose memory of the given points.

Example 80

Consider the MA(1) model, $X_t = c + \theta_1\epsilon_{t-1} + \epsilon_t$. We are given that $\{c = 3, \theta_1 = 0.5, \sigma_\varepsilon^2 = 4\}$.
If we know $x_0 = 3.5, x_1 = 3.9, \varepsilon_0 = 0.2, \varepsilon_1 = 0.8$, then:

$$\begin{aligned}
\mathrm{E}\left[X_2|x_0, x_1, \mathcal{P}\right] &= 3 + (0.5)(0.8), \\
&= 3.4, \\
\mathrm{Var}\left(X_2|x_0, x_1, \mathcal{P}\right) &= \sigma_\varepsilon^2, \\
&= 4,
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}\left[X_3|x_0, x_1, \mathcal{P}\right] &= 3(=c), \\
\mathrm{Var}\left(X_3|x_0, x_1, \mathcal{P}\right) &= (1 + \theta_1^2)\sigma_\varepsilon^2. \\
&= 5
\end{aligned}$$

As t increases, these remain the same.
For an ARMA model, we can do similar things. Fortunately, it is easy to do this in R.

16.9 Lag Operator

The **i-step lag operator** (sometimes called the "backshift" operator), B^i, is defined by $B^iX_t = X_{t-i}$.
We can define our AR(p) model in terms of this operator:

$$X_t = c + \left(\sum_{i=1}^{p} \phi_i B^i\right) X_t + \varepsilon_t,$$
$$\left(1 - \sum_{i=1}^{p} \phi_i B^i\right) X_t = c + \varepsilon_t.$$

This helps us define the next type of model using first differences:

$$X_1 - X_0, X_2 - X_1, X_3 - X_2, \ldots$$
$$(1 - B)X_t = X_t - X_{t-1}.$$

This gives us another way of writing previous models. For example, the AR(1) model can be written as

$$(1 - \phi B)X_t = c + \varepsilon_t.$$

16.10 Autoregressive Integrated Moving Average Process (ARIMA)

An ARIMA model with integer parameters, p, d, q, ARIMA(p, d, q), is

$$\left(1-\sum_{i=1}^{p}\phi_i B^i\right)(1-B)^d X_t \;=\; c+\left(1+\sum_{i=1}^{q}\theta_i B^i\right)\varepsilon_t, \qquad \text{where } \varepsilon_t \overset{i.i.d.}{\sim} G_\varepsilon\left(0,\sigma_\varepsilon^2\right),$$

where $c, \theta_1, ..., \theta_q, \sigma_\varepsilon^2, \phi_1, ..., \phi_p$ are model parameters.
All of our models so far are contained in this model:

$$\begin{aligned} \text{AR}(p) &\Leftrightarrow \text{ARIMA}(p,0,0),\\ \text{MA}(q) &\Leftrightarrow \text{ARIMA}(0,0,q),\\ \text{ARMA}(p,q) &\Leftrightarrow \text{ARIMA}(p,0,q). \end{aligned}$$

"Integrated" comes from the the definition that a time series is integrated of order 0 if it admits a moving average representation with

$$\sum_{k=0}^{\infty} \mid {\theta_k}^2 \mid < \infty,$$

where θ is the possibly infinite vector of moving average weights. This implies that the autocovariance is decaying to 0 sufficiently quickly. This is a necessary, but not sufficient condition for a stationary process. Therefore, all stationary processes are $I(0)$, but not all $I(0)$ processes are stationary.

Example 81

Demonstrate fitting in R: arima.r.

16.11 Vector Autoregressive Models

A **vector autoregressive model** with parameter p, VAR(p) is:

$$\begin{aligned} X_{1,t} &= c_1+\sum_{i=1}^{p}\phi_{1,1,i}X_{1,t-i}+...+\sum_{i=1}^{p}\phi_{1,n,i}X_{n,t-i}+\varepsilon_{1,t},\\ &\vdots\\ X_{n,t} &= c_n+\sum_{i=1}^{p}\phi_{n,1,i}X_{1,t-i}+...+\sum_{i=1}^{p}\phi_{n,n,i}X_{n,t-i}+\varepsilon_{n,t}. \end{aligned}$$

where all $\varepsilon_{i,j} \overset{i.i.d.}{\sim} G_\varepsilon\left(0, \sigma_\varepsilon^2\right)$.

Actually, this error structure is a simplified version of the full VAR model which has a covariance matrix to model dependence in the errors across the time series.

Suppose we have n time series:

$$\begin{aligned} 1^{st} \text{ time series} : &\quad ..., X_{1,0}, X_{1,1}, X_{1,2}, ... \\ 2^{nd} \text{ time series} : &\quad ..., X_{2,0}, X_{2,1}, X_{2,2}, ... \\ \vdots \qquad & \\ n^{th} \text{ time series} : &\quad ..., X_{n,0}, X_{n,1}, X_{n,2}, ... \end{aligned}$$

We could model $\{X_{1,t}\}$ using an AR(p) process:

$$X_{1,t} = c_1 + \sum_{i=1}^{p} \phi_{1,1,i} X_{1,t-i} + \varepsilon_{1,t}$$

We could model $\{X_{r,t}\}$ using a separate AR(p) process:

$$X_{r,t} = c_r + \sum_{i=1}^{p} \phi_{r,r,i} X_{r,t-i} + \varepsilon_{r,t}$$

What if we thought that $X_{1,t}$ depended on both its previous values and previous values from other time series? We could write:

$$X_{1,t} = c_1 + \sum_{i=1}^{p} \phi_{1,1,i} X_{1,t-i} + \sum_{i=1}^{p} \phi_{1,2,i} X_{2,t-i} + ... + \sum_{i=1}^{p} \phi_{1,n,i} X_{n,t-i} + \varepsilon_{1,t}.$$

We could then do this for all n time series simultaneously, and this is a VAR(p) model. Each time series is a constant plus a linear combination of its p previous values, and the p previous values from the other time series, plus error.

Remember: For $\phi_{r,s,i}$, r is the index for the time series that you are modelling, s is the index of the time series that you are considering past values for and i is the number of ‘steps back’ that we are considering.

17 Modelling Time Series with Changing Variability (Brooks Ch 8)

The models we have looked at so far have been like:

$$X_t = \text{constant} + (\text{coefficients} \times \text{random variables}) + \text{white noise}.$$

These do not help if the variance is changing over time.

17.1 Checking time series assumptions

Consider a simple AR(1) model: $X_t = c + \phi_1 X_{t-1} + \varepsilon_t$.
Apart from checking the acf and pacf are behaving correctly, we need to consider if $\varepsilon_t \overset{i.i.d.}{\sim} G_\varepsilon\left(0, \sigma_\varepsilon^2\right)$ is appropriate.

Make link to checking the assumptions of a linear model.

Let's say we have fitted a model in R using `TSmodel = arima(data, c(1, 0, 0))`.
The error terms (or residuals, innovations), that are the difference between the fitted model and the observed data, are automatically stored in `TSmodel$residuals`.
To assess if $\mathrm{E}\left(\varepsilon_t\right)$ is appropriate, you might look at `mean`(...) or `hist`(...). To assess independence, we could plot the residuals against time, `plot`(...), or the autocorrelation in the residuals, `acf`(...). These plot can also tell us about variability of data over time.

Changing mean: try ARIMA (looks at the differences),
Changing variance: try ARCH (fits model to the variance).

17.2 Returns and (Crude) Volatility

Note that there are several definitions for both of these

We have an observed time series:

$$x_0, x_1, ..., x_n.$$

We can create another time series by looking at differences at lag 1 (in R, use `diff`(...)).

$$x_1 - x_0, x_2 - x_1, ..., x_n - x_{n-1}$$

In finance, we are often interested in a rescaled version of this given by the series of **arithmetic returns**:

$$r_1, r_2, ..., r_n, \quad \text{where} \quad r_t = \frac{x_t - x_{t-1}}{x_{t-1}}.$$

Example 82

The price at close of BP stock was 417.0 and 423.1 over two consecutive days (6.1 gain). For G4S, the prices at the same time were 241.7 and 246.2 (4.5 gain).

$$\begin{aligned} \text{Return for BP}: &\quad \tfrac{423.1-417.0}{417.0} &= 0.0146 \\ \text{Return for G4S}: &\quad \tfrac{246.2-241.7}{241.7} &= 0.0186 \end{aligned}$$

Because the arithmetic return can change by orders of magnitude from day to day, the logarithmic returns are usually reported.

$$lr_1, lr_2, ..., lr_n, \text{ where } lr_t = \ln(x_t) - \ln(x_{t-1}) = \ln\left(\frac{x_t}{x_{t-1}}\right).$$

One measure of volatility of a time series over n time steps is:

$$\sum_{i=1}^{n} \frac{\left(lr_i - \overline{lr}\right)^2}{(n-1)} \quad \text{(the sample variance for log returns).}$$

If the variance of our time series is fixed over time, this volatility measure should be fixed over time. In a time series based on stock prices, we often see changes in volatility over time, and high volatility occurs in bursts. Further, low volatility over a period will (on average) lead to more low volatility, and high volatility will (on average) lead to more high volatility. This suggests correlation in volatility over time.

If variability is unchanged over time, we say that the time series model is **homoscedastic**. If variability changes over time, we say the time series model is **heteroscedastic**.

17.3 Autoregressive Conditional Heteroscedastic (ARCH) Model

An ARCH model is essentially two models: one for the mean behaviour and one for the variance. The typical set-up is to have an ARMA(p, q) for the

mean behaviour and a separate model for the variance based on variances from previous time steps.
Recall that ARMA(p,q) is

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

with independent $\varepsilon_t | \sigma_t^2 \sim G_\varepsilon\left(0, \sigma_t^2\right)$. Now,

$$\begin{aligned} \text{ARCH}(q^*): \quad \sigma_t^2 = & \quad \alpha_0 + \sum_{i=1}^{q^*} \alpha_i \nu_{t-i}^2, \\ \nu_t = & \quad \sigma_t z_t, \\ \text{where} \quad z_t \overset{i.i.d.}{\sim} & \quad G_z\left(0, 1\right). \end{aligned}$$

The parameters for this model are the ARMA parameters, plus $\alpha_0, \alpha_1, ..., \alpha_{q^*}$ and σ_z^2.
α_0 is the expected variance over the time series;
$\alpha_1, ..., \alpha_{q^*}$ tell us how related the current variance is to variances at previous timesteps;
σ_z^2 tells us how far σ_t^2 could move in time.
Essentially, the bigger the α_i, the more volatile the time series, and the bigger σ_z^2, the more heteroscedastic the time series.

Example 83

The ARCH(1) model is given by

$$\begin{aligned} \sigma_t^2 &= \alpha_0 + \alpha_1 \nu_{t-1}^2 \\ \nu_t^2 &= \sigma_t^2 z_t^2 = (\alpha_0 + \alpha_1 \nu_{t-1}^2) z_t^2 \end{aligned}$$

with $\text{Var}[z_t] = 1$ and $\text{E}[z_t] = 0$. This is a bit like ar AR(1) process but in ν_t^2 rather than ν_t, and with multiplicative noise with a mean of 1, rather than additive noise with a mean of zero.
We have

$$\begin{aligned} \text{E}[\nu_t^2 \mid \nu_t, \ldots] &= \text{E}\{(\alpha_0 + \alpha_1 \nu_{t-1}^2) z_t^2 \mid \nu_{t-1}, \nu_{t-2}, \ldots\} \\ &= \alpha_0 + \alpha_1 \nu_{t-1}^2 \text{E}(z_t^2 \mid \nu_{t-1}, \nu_{t-2}, \ldots) \\ &= \alpha_0 + \alpha_1 \nu_{t-1}^2 \end{aligned}$$

If ν_{t-1} has a large (absolute) value, the σ_t is larger than usual, and so ν_t is also expected to have an unusually large magnitude. This volatility propagates since when ν_t has large deviation, then σ_{t+1}^2 is large and so on. Similarly for small values.
So, unusually large volatility in ν_t tends to persist, though not forever.

Say something about model fitting: point to tomorrow's practical? And mention fGarch for fitting in R (fGARCH stands for family GARCH which encompasses most of the GARCH variants).

17.4 Generalised Autoregressive Conditional Heteroscedastic (GARCH) Model

This is the same as ARCH, but with a MA process as well as the AR process for σ_t^2.

The GARCH(p^*, q^*) model is thus defined by

$$\begin{aligned} \nu_t &= \sigma_t z_t, \\ \sigma_t^2 &= \omega + \sum_{i=1}^{p^*} \alpha_i \nu_{t-1}^2 + \sum_{j=1}^{q^*} \beta_j \sigma_{k-j}^2, \end{aligned}$$

where $\omega > 0, \alpha_i \geq 0, \beta_j \geq 0$, and the innovation sequence z_t is i.i.d. with $\mathrm{E}[z_t] = 0$ and $\mathrm{Var}[z_t] = 1$.