Maths for Materials and Design

Probability and Statistics

Contents

1 Lecture 1: Probability

1.1 Probability of an Event

We can categorise the probability of an event happening into two possibilities. Ones where we know the probability exactly (a-priori) and ones where we use past knowledge (statistics) to make an estimate (a-posteriori).

A-Priori: This is where we can do a calculation based on what we know about the system. For example, a fair die. We know the probability of rolling a six is 1/6.

A-Posteriori: The probability is estimated from a collection of empirical data from similar events in the past. For example, of 126 matches in the FA cup between premier league and non-league sides, premier league teams won 102 times. So, the probability of a premier league side beating a non-league side next is 102/126 based on past events. But, it is an estimate with too many factors involved to make an exact calculation.

1.1.1 A-Priori Examples

Example 1.1. *If we have a specific outcome in mind, let's call it A. For example, the probability of rolling a 6 when we roll a die. We can find the probability of outcome A (denoted P(A)) by considering the number of possible outcomes where A occurs as a fraction of* all possible events*:*

Solution:

$$P(A) = \frac{\textit{No. of outcomes where A occurs}}{\textit{No. of possible events}}$$

Example 1.2. *If a card is drawn at random from a standard deck of 52 playing cards, what is the probability that it is a face card (i.e. a jack, queen or king)?*

Solution:

Let's call the event of picking a face card, A. How many cards do we have?
4 kings
4 queens
4 jacks
So in total we have 12 out of a possible 52. Thus:

$$P(A) = \frac{12}{52} = \frac{3}{13}$$

1.2 Venn diagrams

Venn diagrams are used to visualise sets of outcomes. Let's think again about the 52 cards. If we pick one card, then we have 52 possibilities in total - we call this the **universal set**.

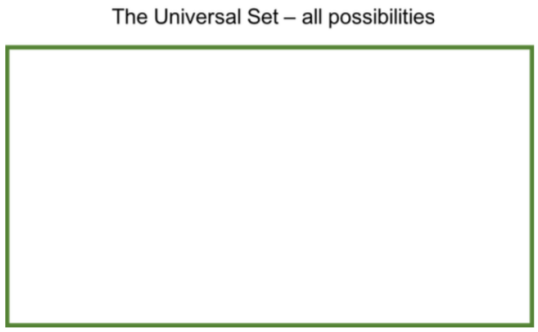


Figure 1: The universal set

We then called the event of drawing a picture card A, with probability $P(A)$. This is a subset of the universal set:

Figure 2: $P(A)$

What about the possibility of getting a spade? Call this S with probability, $P(S)$.

There are 13 spades, thus:

$$P(S) = \frac{13}{52} = \frac{1}{4}$$

1.2.1 Venn Diagram Examples

Example 1.3. *Draw a Venn diagram showing the probability of drawing a spade (S)*

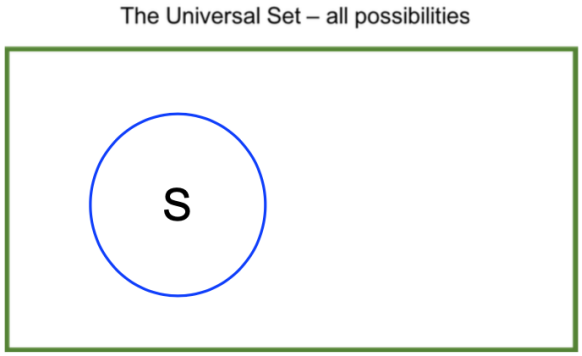


Figure 3: $P(S)$

Example 1.4. *Draw a Venn diagram representing the probability of getting a spade (S) or a diamond (D).*

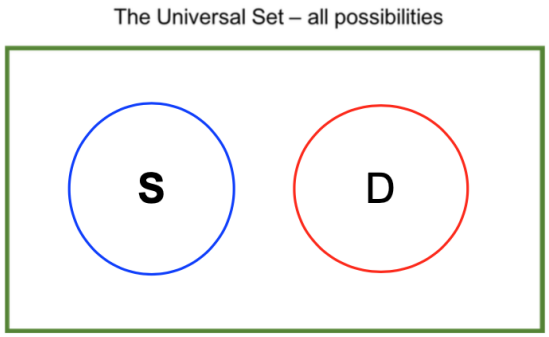


Figure 4: $P(S \text{ or } D)$

$$\begin{aligned} P(S \text{ or } D) &= P(S) + P(D) \\ &= \frac{13}{52} + \frac{13}{52} \\ &= \frac{26}{52} = \frac{1}{2} \end{aligned}$$

Example 1.5. *From the full pack of cards (the universal set), draw a Venn diagram showing the probability of drawing either a face card (Z) or a spade (S).*

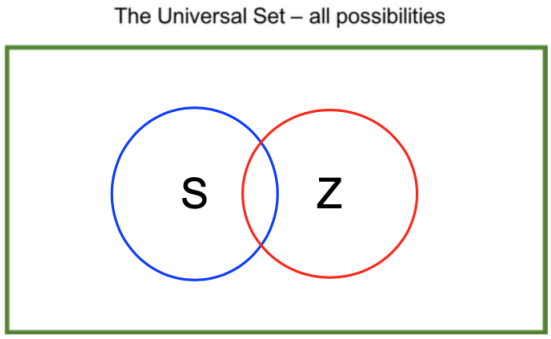


Figure 5: $P(Z \text{ or } S)$

$$
\begin{aligned}
P(Z \text{ or } S) &= P(Z) + P(S) - P(Z \text{ and } S) \\
&= \frac{3}{13} + \frac{13}{52} - \frac{3}{52} \\
&= \frac{12}{52} + \frac{13}{52} - \frac{3}{52} \\
&= \frac{12 + 13 - 3}{52} \\
&= \frac{22}{52} = \frac{11}{26}
\end{aligned}
$$

In this case, the card can be ***both*** *a spade and a face card, so we have to subtract the probability that it is both, as this is counted twice. In the Venn diagram, this is interpreted as an overlap as they are occurring at the same time, so to work out the area we have to subtract the overlapping part once.*

In general, for two events A and B, the probability that **either** A or B (or both) occur is given by:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

Example 1.6. *Now, imagine we again pick a card. Draw a Venn diagram representing the probability that we get a face card (Z) or a 5 (call it P(F))?*

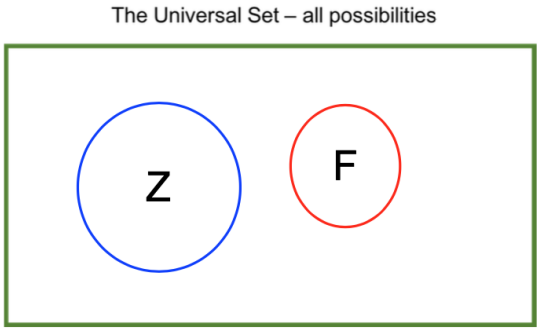


Figure 6: $P(Z \text{ or } F)$

$$P(Z) = \frac{3}{13} \qquad P(F) = \frac{4}{52}$$

Then the probability of getting a face card (Z) or a five (F) is:

$$\begin{aligned} P(Z \text{ or } F) &= P(Z) + P(F) - P(Z \text{ and } F) \\ &= \frac{3}{13} + \frac{4}{52} - P(Z \text{ and } F) \end{aligned}$$

But what is the probability of getting both a face and a 5? Because there exists no one card that is both, then $P(Z \text{ and } F) = 0$. These two events are ***mutually exclusive****, meaning that they cannot happen at the same time.*

$$\begin{aligned} P(Z \text{ or } F) &= \frac{3}{13} + \frac{4}{52} - 0 \\ &= \frac{12}{52} + \frac{4}{52} \\ &= \frac{16}{52} = \frac{4}{13} \end{aligned}$$

If two events A and B are **mutually exclusive** we can write:

$$P(A \text{ or } B) = P(A) + P(B)$$

1.2.2 Complements

For a given event A, the complement of A (denoted $\bar{A}$) represents all the possible outcomes in which A does **not** occur.

$$P(A) + P(\bar{A}) = 1$$

Using our example of the probability of a face card being drawn:

$$P(A) = \frac{3}{13}$$

Then the probability that a card is *not* a face card is:

$$P(\bar{A}) = 1 - \frac{3}{13} = \frac{10}{13}$$

1.3 Conditional Probability

Two events A and B are **dependent** if the occurrence of A affects the probability of B occuring. When this happens we can talk about the probability of B, **given that** A **has already occurred**, which we denote:

$$P(B|A)$$

The probability that both B and A occur is written as:

$$P(A \text{ and } B) = P(A)P(B|A)$$

If the occurrence of A does not affect B we say that they are **independent** and then:

$$P(A \text{ and } B) = P(A)P(B)$$

as $P(B) = P(B|A)$, since the probability of B occuring is not affected by the fact that A is known to have occured.

However, how can we calculate a conditional probability if the events *are* depedent?

In the case of $P(A|B)$, we can think of the "pool" of outcomes that we are interested in as being reduced to only those where B has already occured. So for what fraction of *those* events did A occur?

This can be expressed as a formula:

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

which of course is just a re-arrangment of the first formula on this page!

1.4 Summary: Decision Tree

We can decide which equation to use by using the following decision tree.

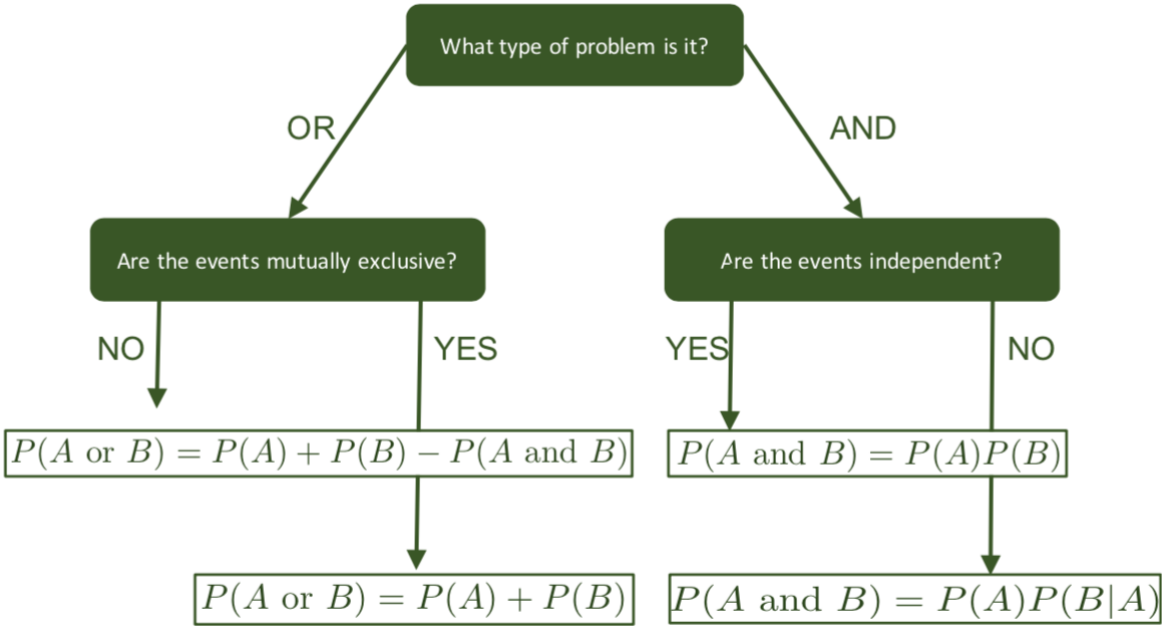


Figure 7: Decision tree

1.5 Probability Tree

Example 1.7. *A box contains 14 light bulbs, 12 of which work and 2 are broken. 2 bulbs are chosen at random without replacement. Let's call the outcome that the first bulb works A and that the second bulb works B.*

Construct a probability tree to account for the probabilities of all possible options.

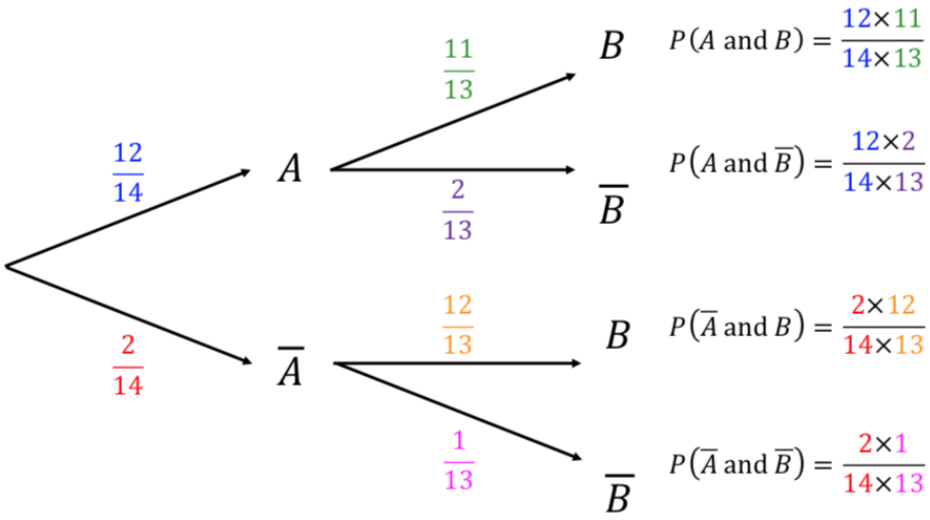


Figure 8: Example probability tree

What is the probability that ***both*** *bulbs work?*

$$P(A \text{ and } B) = \frac{12}{14} \times \frac{11}{13} = \frac{132}{182}$$

What is the probability that only ***one*** *bulb works?*

$$\begin{aligned} P(A \text{ and } \bar{B}) + P(\bar{A} \text{ and } B) &= \frac{12}{14} \times \frac{2}{13} + \frac{2}{14} \times \frac{12}{13} \\ &= \frac{24}{182} + \frac{24}{182} \\ &= \frac{48}{182} = \frac{24}{91} \end{aligned}$$

1.6 Probability distributions

In the next lecture we will be looking at probability distributions. A probability distribution is basically a graph of the probability of an event occurring as a function of the event.

Example 1.8. *Sketch a probability distribution graph of the possible outcomes of rolling a fair die.*

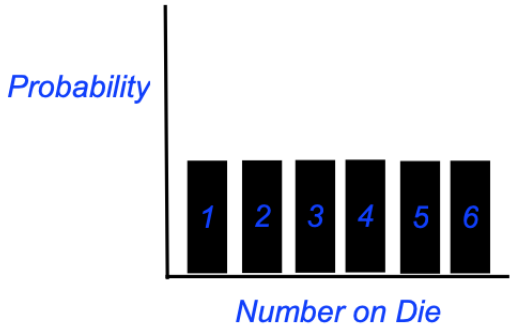


Figure 9: Example probability distribution graph

*This is a **uniform** probability distribution because each outcome is equally as likely.*

Example 1.9 (Bonus example: complete this at home if we don't have time in class!).

(a) *A six-sided die is rolled. What is the probability that the result is either a multiple of 2 or a multiple of 3?*

Solution:

Event A: it is a multiple of 2. In this case the options are 2, 4 or 6.

$$P(A) = \frac{3}{6} = \frac{1}{2}$$

Event B: it is a multiple of 3. In this case the options are 3 and 6.

$$P(B) = \frac{2}{6} = \frac{1}{3}$$

But 6 is both multiple of 2 and 3, so both A and B can occur at the same time. Thus they are not mutually exclusive.

$$\begin{aligned} P(A \text{ or } B) &= P(A) + P(B) - P(A \text{ and } B) \\ &= \frac{3}{6} + \frac{2}{6} - \frac{1}{6} \\ &= \frac{4}{6} = \frac{2}{3} \end{aligned}$$

(b) *The die is rolled a second time. What is the probability that the result is less than 3 or greater than 4?*

Solution:

Event A: a number less than 3. Our options are 1 and 2.

$$P(A) = \frac{2}{6} = \frac{1}{3}$$

Event B: it is a number greater than 4. In this case the options are 5 and 6.

$$P(B) = \frac{2}{6} = \frac{1}{3}$$

But the number cannot be both less than 3 and greater than 4 at the same time, so the results are mutually exclusive. Hence:

$$\begin{aligned} P(A \text{ or } B) &= P(A) + P(B) \\ &= \frac{1}{3} + \frac{1}{3} = \frac{2}{3} \end{aligned}$$

2 Lecture 2: Data analysis and Statistics

2.1 Introduction to Data

We can think of data as a collection of measured quantities. For example, the number of people who have visited a park or the temperature of the Arctic.

We call any quantity that can take a number of different values a **variable**. These quantities (or variables) can be one of two kinds:

- **Discrete** – a variable that can be counted or that has a fixed set of values. For example, the number of visitors to a park (you can't have half or 0.2 of a person).
- **Continuous** – a variable that can be measured on a continuous scale (depending on the precision of your instrument/measurement). For example, temperature or height.

2.2 How can we analyse data?

2.2.1 Order Data

We could begin by simply writing the numbers in order from smallest to largest. This allows us to quickly look at the data and see several things, such as the smallest and largest numbers in the set. This is fine for small data sets but not much use for large ones. It also doesn't tell us much about the data.

2.2.2 Group Data

If the range (the range is the largest number minus the smallest number in the set) is large it is often helpful to to group values into regular groups, or classes.

Example 1

The number of hours worked per week by employees at a factory are:

45	31	46	25	57
40	59	11	38	34
57	37	43	51	33
44	47	42	46	50
38	22	33	39	66

Create a tally table

Overtime Hours	Tally Marks	Frequency
10-19	|	1
20-29	||	2
30-39	𝍸 |||	8
40-49	𝍸 |||	8
50-59	𝍸	5
60-69	|	1

If data is continuous (to a given precision) then you have to be more careful about grouping data. For example, if you had the same groups as the table above and had continuous data and one of your points was 19.5, which group would it enter? We would have to modify our group range to reflect the precision of our measurements/data.

Overtime Hours	Tally Marks	Frequency
10.0-19.9		
20.0-29.9		
30.0-39.9		
40.0-49.9		
50.0-59.9		
60.0-69.9		

2.2.3 Histograms

Once data is tallied (made into a frequency distribution) we can construct a histogram. This is essentially a graphical representation of a frequency distribution, where we draw vertical rectangular blocks so that:

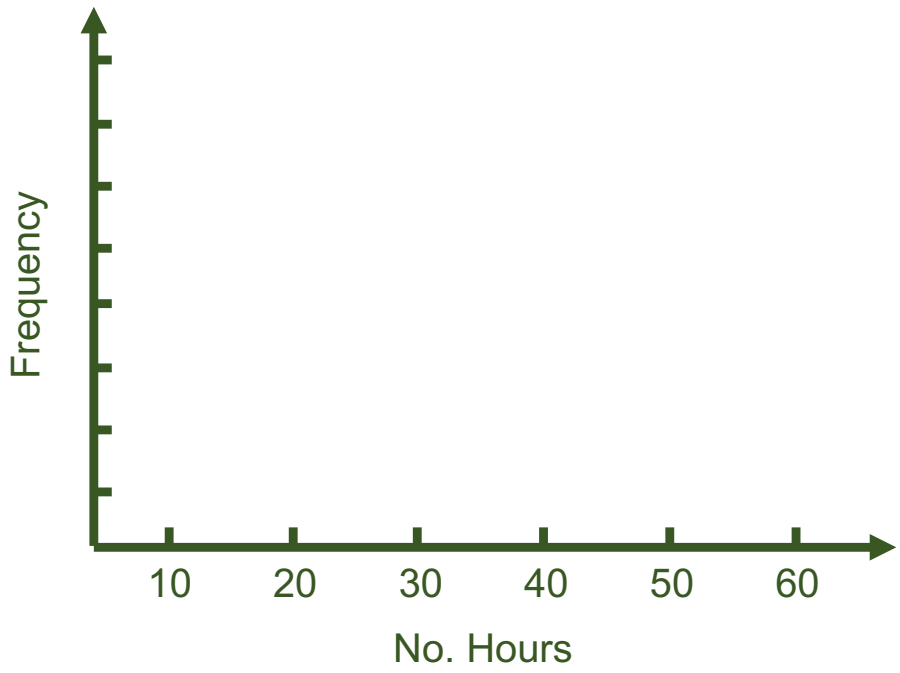
- The centre of the base indicates the central value of the class.
- The area represents the class frequency.

In general, we can say that the:

$$\boxed{\text{frequency distribution} = \frac{\text{frequency}}{\text{width}}}$$

If each class width is **regular** (they all have the same width) then the frequency is often denoted by the height. This is the most common form of a histogram.

Complete the following histogram using the data from *Example 1*:



2.3 Measures of centrality

We can calculate some properties to give us an indication of where "most" or the middle of the data lies.

Arithmetic Mean

The arithmetic mean is defined as the sum of the data $\{x_i\}$ divided by the number N of data points. This is written mathematically as:

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

In Excel this can be calculated using the formula `=AVERAGE( ... )`

Geometric Mean

The geometric mean is only defined for a set of N data points, $\{x_i\}$, with **positive** values of x_i. It is defined as:

$$\bar{x}_g = \sqrt[N]{x_1 x_2 x_3 \dots x_N}$$

This can also be written as:

$$\bar{x}_g = \left(\prod_{i=1}^{N} x_i \right)^{1/N}$$

We will be primarily concerned with the arithmetic mean.

Median

The median is the value that separates an ordered list into two. For an odd number of values this value is the middle number, for example:

$$\boxed{1, 4, \underbrace{5}_{\text{median}}, 8, 10}$$

If there are an even number of values, the median is the halfway point between the middle two numbers, for example:

$$\boxed{1, 4, \underbrace{5, 7}_{\text{middle numbers}}, 8, 10}$$

The median in this case is then

$$\frac{5+7}{2} = \frac{12}{6} = 6$$

The median is more useful as a measure of the average if there are very large number or very small numbers that are not typical of the set. Such values would tend to skew the mean, making it less useful.

Example 2

Calculate the arithmetic mean and median of this data set:

29 32 26 40 12 20 35 190

To calculate the median we first order the data:

12 20 26 29 32 35 40 190

The median is then:

$$\frac{29+32}{2} = 30.5$$

The arithmetic mean is:

$$\frac{12+20+26+29+32+35+40+190}{8} = 48$$

This may be an extreme example, though it shows the point that the arithmetic mean is not always the best value, it depends on the dataset.

Mean for grouped data

When using grouped data, we can calculate the mean by multiplying the midpoints and frequencies of each group, sum them, then divide by the total size of the sample.

Example:

Class	Frequency f	Midpoint x_i	$f \cdot x_i$
15-20	2	17.5	35
20-25	1	22.5	22.5
25-30	3	27.5	82.5
30-35	2	32.5	65
35-40	1	37.5	37.5
40-45	1	42.5	42.5

The mean is then:

$$\frac{\sum f \cdot x_i}{\sum f} = \frac{1}{10} \cdot 285 = 28.5$$

Median for grouped data

When using grouped data, we can calculate the **cumulative frequency** to help us locate the median.

Example:

18.3 20.2 24.0 26.9 27.1 28.0 32.4 34.0 39.3 41.7

Class	Frequency	Cumulative freq.
15-20	2	2
20-25	1	3
25-30	3	6
30-35	2	8
35-40	1	9
40-45	1	10

As there are 10 values in total, the median is the $\frac{10+1}{2} = 5.5^{th}$ value. Looking for the first class where the cumulative frequency reaches this value or greater, we find that it is the group $25 - 30$. Taking the midpoint of this group, the median is therefore 27.5.

Mode

The mode is the value of the sample that occurs the most. This is fine if you have, say, a discrete set of integer data, e.g. the age in years of people working in an office:

18 19 19 22 27 33 40

The mode is then 19.

However, if you have a continuous variable it is possible that none of them repeat. For example, if you measure something to 2 decimal places and collect 100 data points with a range of 10 there are 1000 possible values that the measurement can take. Sometimes even discrete data does not have a mode (no value repeats). What you can then do is group the data using a tally chart and find the **modal group**.

Example 3

Group the following data set into groups of 16-20, 21-25, 26-30 etc. to find the model group.

18 20 24 26 27 28 32 35 40 41 50

Class	Frequency
16-20	2
21-25	1
26-30	3
31-35	2
36-40	1
41-45	1
46-50	1

The modal group is then 26-30.

2.4 Dispersion or Spread of Data

We often want to know how spread a given data set is so that we know how far values stray from the mean. This is useful when taking repeated experimental measurements. The spread of data allows us to estimate the error in our measurement. The spread can be taken in many different ways. Here we will review and use a few of them.

Range and Quartiles

The range is the difference between the largest and smallest data values.

$$\text{Range} = \text{maximum value} - \text{minimum value}$$

As well as the range, often data can be characterised by the *so-called* upper and lower quartiles.

The lower quartile (L_{25}) is the median of the lower half of the data.

The upper quartile (U_{25}) is the median of the upper half of the data.

The difference between the upper and lower quartiles is the **interquartile range** (IQR):

$$IQR = U_{25} - L_{25}$$

Example 4

Calculate the range and IQR of the following data:

1 1 2 3 3 3 4 5 5 7 7 8 10 19

The range is the largest value minus the lowest value:

$$19 - 1 = 18$$

There are 14 data points, so the median of the first 7 is the 4th value. Thus,

$$L_{25} = 3$$

The median of points 8-14 is the 11th value. Thus,

$$U_{25} = 7$$

And so we have:

$$IQR = U_{25} - L_{25} = 7 - 3 = 4$$

Mean Squared Deviation or Variance

The mean squared deviation (σ^2, or sometimes written $\overline{\Delta x^2}$) is also known as the mean squared displacement or variance is the average of the square of the deviation of each data point from the arithmetic mean.

$$\sigma^2 \;=\; \overline{\Delta x^2} \;=\; \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2$$

You may wonder why don't we just calculate the average deviation from the mean? Well, the deviation from the mean of any data point, x_i is given by $x_i - \bar{x}$. If we try to calculate the mean of this quantity we get:

$$\begin{aligned}
\overline{x_i - \bar{x}} \;=\; \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x}) \;&=\; \underbrace{\frac{1}{N}\sum_{i=1}^{N} x_i}_{\text{This is just } \bar{x}} \;-\; \frac{1}{N}\underbrace{\sum_{i=1}^{N}\bar{x}}_{\text{This is just N lots of } \bar{x}} \\
&=\; \bar{x} - \frac{N\bar{x}}{N} \\
&=\; \bar{x} - \bar{x} = 0
\end{aligned}$$

The mean displacement would always just be **zero**.

Standard Deviation

The standard deviation is just the square root of the variance:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\bar{x} - x_i)^2}$$

σ has the same units as x_i or $\bar{x}$.

In particular, this is the "population" standard deviation, which we shall generally use on this module rather than the "sample" standard deviation. When working in EXCEL, we can call this using `=STDEV.P( ... )` where the dots are replaced by the range of cells that we want to take the standard deviation of.

If we were using the sample standard deviation with an unbiased estimator, this would be equivalent to the root mean squared error from the mean.

Standard Error

The final quantity that we need to know about is the standard error. This is most often used when discussing experimental data where repeated measurements are taken.

It is defined as:

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$$

where N is the number of measurements taken.

This is essentially an estimate of the confidence of how accurately we have measured the mean.

Difference between the standard deviation and the standard error

Let's say we are measuring the temperature of a bar. Even though the *average* temperature in the room remains constant, there are constant random fluctuations that lead to small changes in the temperature.

Even if we take measurements until the end of time, the standard deviation will remain the same, because the size of the fluctuations is (on average) the same.

However, our estimate of the mean becomes more and more accurate and our "confidence" in our measurement of the temperature of the bar improves, so the *error* is reduced.

3 Lecture 3: Probability distributions

3.1 Overview

- Histograms illustrate the frequency distribution of a data set for variable x in discrete bins.
- If we place our data in smaller and smaller bins, this discrete distribution approaches a continuous curve, described by a frequency distribution function $g(x)$.
- If we then "normalise" the data, by dividing by the total number of data points, we instead obtain a probability distribution $p(x)$ of our variable x
- In such a distribution, the area between $a \leq x \leq b$ gives the probability that x takes a value in this range.
- Key properties:

$$\int_{-\infty}^{\infty} p(x)\mathrm{d}x = 1 \qquad \bar{x} = \int_{-\infty}^{\infty} xp(x)\mathrm{d}x \qquad \sigma^2 = \int_{-\infty}^{\infty} (x-\bar{x})^2 p(x)\mathrm{d}x$$

- A common distribution for lengths, weights, heights, etc. is the normal distribution.
- This is described by the normal probability distribution function:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, \mathrm{e}^{-(x-\mu)^2/2\sigma^2}$$

  where μ is the mean, and σ is the standard deviation.
- If X is a normally-distributed variable with mean μ and standard deviation σ, this is written as:

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

- The "standard" normal distribution is given by:

$$X \sim N(0, 1)$$

  and we can use it (as described by the normal distribution table) to determine the probability that a normally-distributed variable lies within a given range.
- Log-normal distribution:
  - This is similar to the normal distribution, but asymmetric with a bias towards smaller values.
  - An example is found in the distribution of grain sizes in polycrystalline materials.

- The probability density function is:

$$P(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}$$

- The mean is $exp(\mu + \frac{1}{2}\sigma^2)$
- The median is $exp(\mu)$
- The mode is $exp(\mu - \sigma^2)$
- The variance is $(exp(\sigma^2) - 1)exp(2\mu + \sigma^2)$

- Lorentzian/Cauchy distribution:
  - This can be obtained from the amplitude of certain damped oscillations of masses on springs.
  - Can appear similar to the normal distribution, but with more mass at the extremes (does not fall away to zero so quickly).
  - Because of this, the integrals for the mean and variance are technically undefined.

3.2 Discrete and continuous distribution functions

In a research study, the size of garden ants was collected over a 6 month period using a sensitive detector that automatically measured the length to an accuracy of 1μm (10^{-6}m). In total 10,000 ants were measured. A histogram was then made by tallying the sizes into equal sized classes.

With 10 classes, the frequency distribution is plotted below:

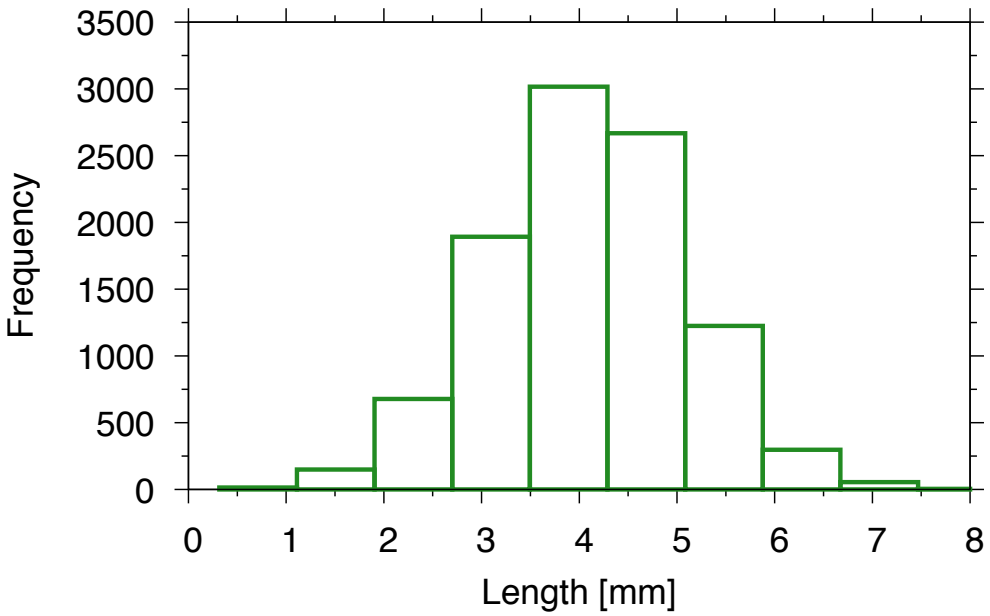


Figure 10: Frequency Distribution (10 classes)

Each class has a width of 0.8mm

What is the modal size?

Just by inspection, the modal size is between 3.5mm and 4.3mm.

Probability again

Look again at the data with 10 classes.

What is the probability that if we measured the length of another ant it would have a length between 3.5 and 4.3mm?

Recall from lecture 1:

$$P(A) = \frac{\text{No. of outcomes where } A \text{ occurs}}{\text{No. of possible eventts}}$$

We can see that there were around 300 (precisely 3016) ants in this range and we measured 10,000 in total. Therefore, our probability is:

$$P(\text{Ant between 3.5 and 4.3mm}) = \frac{3016}{10000} = 0.3016$$

It is often convenient to divide the frequency in each class by the total number in the sample as this gives us the probability distribution function.

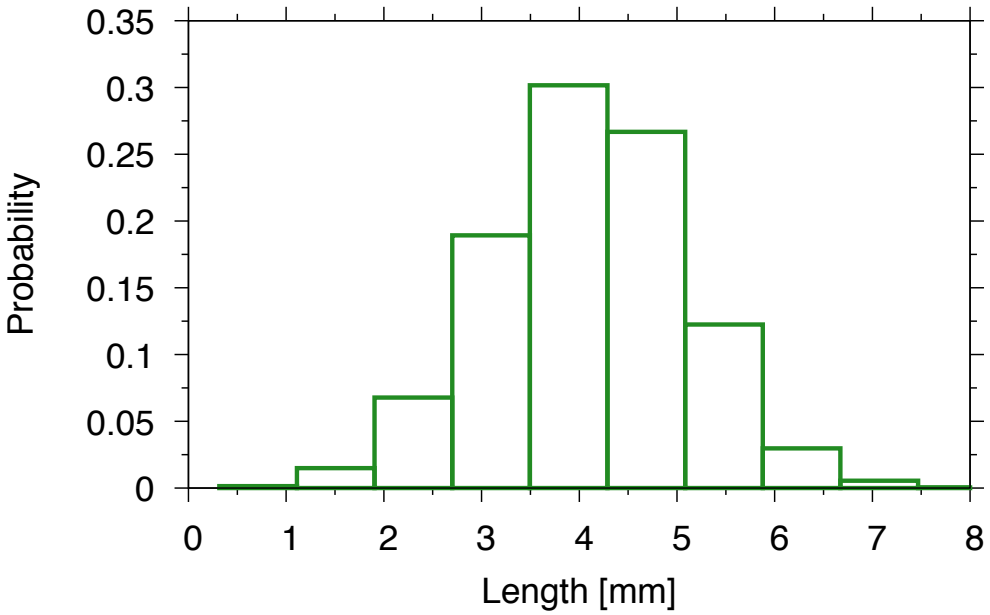


Figure 11: Probability Distribution

Continuous Distributions

If our tool for measuring the length is sufficiently accurate, we can split the histogram into a greater number of classes.

Let's see what happens if we have 30 classes:

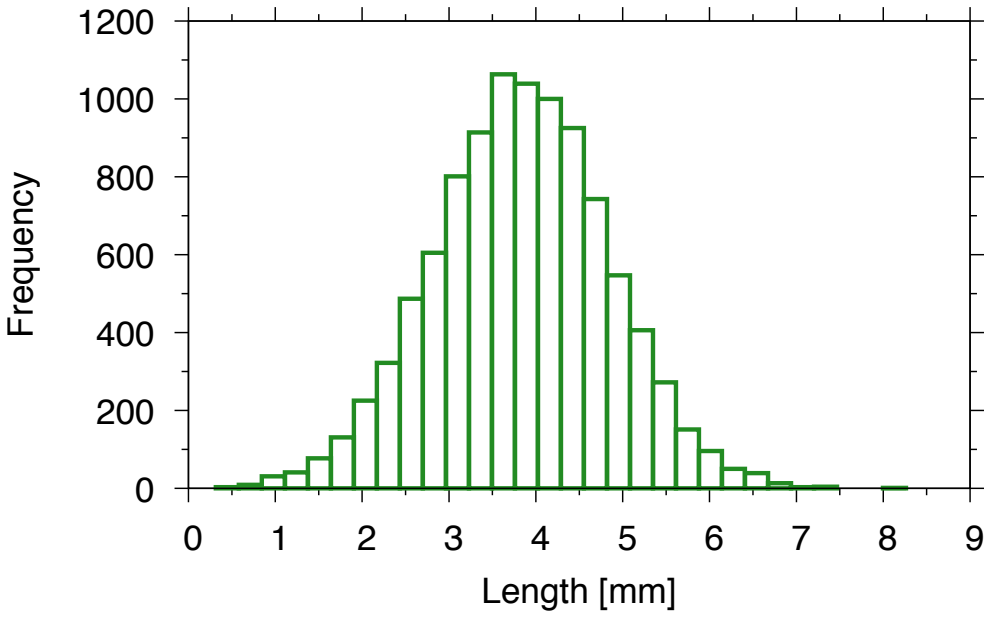


Figure 12: Frequency Distribution (30 classes)

Now we can see that the mode is more specifically somewhere between 3.5mm and 3.75mm.

The second thing to note is that we have a **smaller frequency in each class**.

Now let's look at the data with 100 classes.

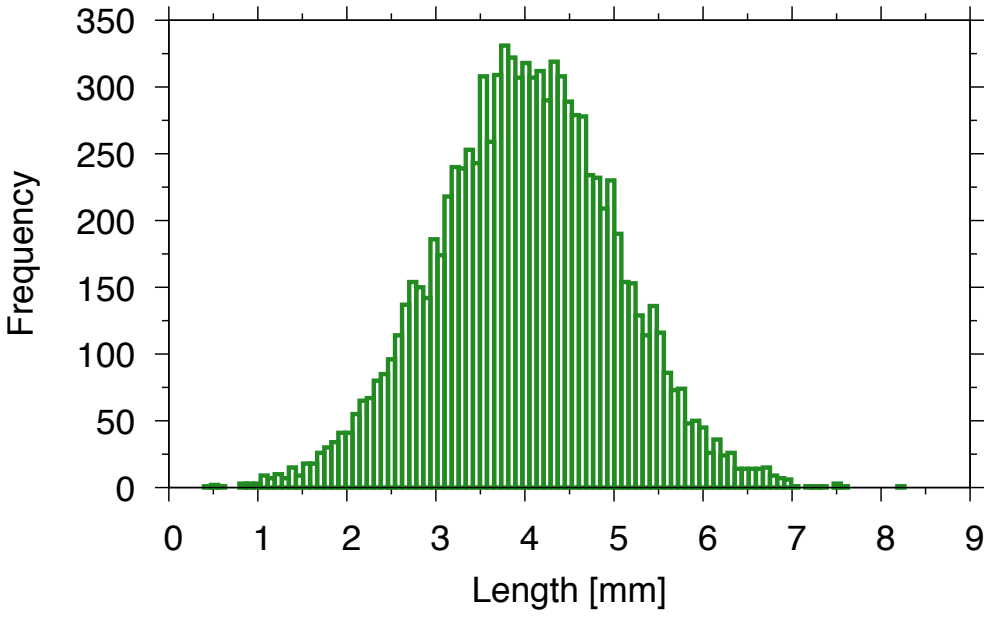


Figure 13: Frequency Distribution (100 classes)

The class width has again become smaller and the frequency in each class also goes down.

In cases like this (measuring lengths) where we have a lot of continuous data, we can approximate the distribution as having a **continuous** form. In many cases, we can even approximate a relatively large number of samples as having a continuous form. Distributions of real data tend to have certain forms.

If we have a form, $g(x)$:

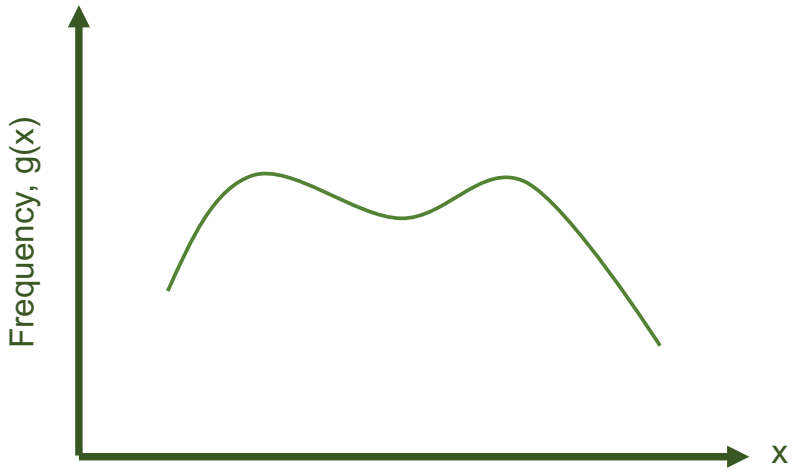


Figure 14: Continuous frequency distribution

Then, just like the discrete frequency distribution, the number of data points is the sum of the areas (remember the area of a histogram represents the number of samples in that class).

If we have a continuous function we can then calculate this using integration:

$$N = \int_{-\infty}^{\infty} g(x)dx$$

Here we should use the limits of $\pm\infty$ because in general our data can be anything. In most cases we can just take limits between two sensible points (where there is no more data).

Continuous Probability Distribution

Just as in the discrete case, we can divide our continuous frequency distribution $g(x)$ by the total number of points (the area under the curve).

We can then define our probability distribution $p(x)$ as:

$$\boxed{p(x) = \frac{1}{N} g(x)}$$

But since N is just the area under $g(x)$ (the sum of all frequencies):

$$\boxed{p(x) = \frac{g(x)}{\int_{-\infty}^{\infty} g(x)\mathrm{d}x}}$$

If we then add up all of our probabilities $p(x)$, they must equal 1 (you can't have the chance of something being larger than 1). This means that the total area under the entire probability distribution curve must be equal to 1.

Writing this as an integral:

$$\boxed{\int_{-\infty}^{\infty} p(x)\mathrm{d}x = 1}$$

As the probability distribution curve is the same as the frequency distribution curve, but **normalised** (i.e. it is divided by the number of samples), it has the same properties for the mean and standard deviation:

Mean and variance for a variable x with probability distribution $p(x)$:

$$\bar{x} = \int_{-\infty}^{\infty} x p(x)\mathrm{d}x$$

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \bar{x})^2 p(x)\mathrm{d}x$$

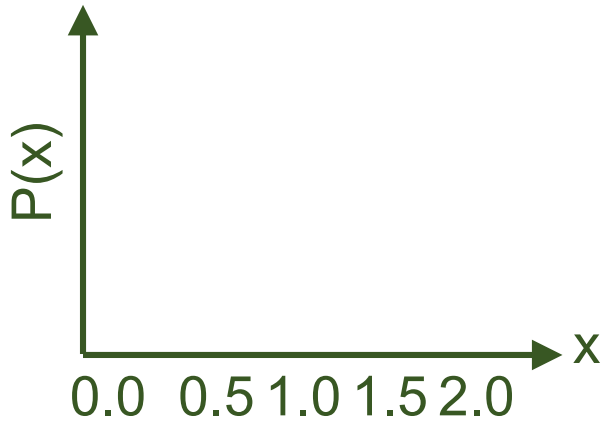
Let's look at an example of calculating the mean and variance of a continuous distribution.

Assume we have a uniform distribution between zero and 1, calculate the mean and standard deviation.

A uniform distribution has probability density function:

$$p(x) = \frac{1}{\text{upper limit} - \text{lower limit}} = \frac{1}{1-0} = 1 \quad \text{between the limits,}$$

and $p(x) = 0$ everywhere else.



The average is given by:

$$\bar{x} = \int_{-\infty}^{\infty} xp(x)\mathrm{d}x = \int_0^1 x \times 1\mathrm{d}x = \left[\frac{x^2}{2}\right]_0^1 = \frac{1^2}{2} - \frac{0^2}{2} = \frac{1}{2}$$

The variance is given by:

$$\begin{aligned}
\sigma^2 &= \int_{-\infty}^{\infty} (x-\bar{x})^2 p(x)\mathrm{d}x \\
&= \int_0^1 (x-0.5)^2 \times 1\mathrm{d}x \\
&= \left[\frac{(x-0.5)^3}{3}\right]_0^1 \\
&= \frac{(1-0.5)^3}{3} - \frac{(0-0.5)^3}{3} \\
&= \frac{0.5^3}{3} + \frac{0.5^3}{3} \\
&= 0.083333\ldots
\end{aligned}$$

3.3 Normal Distribution

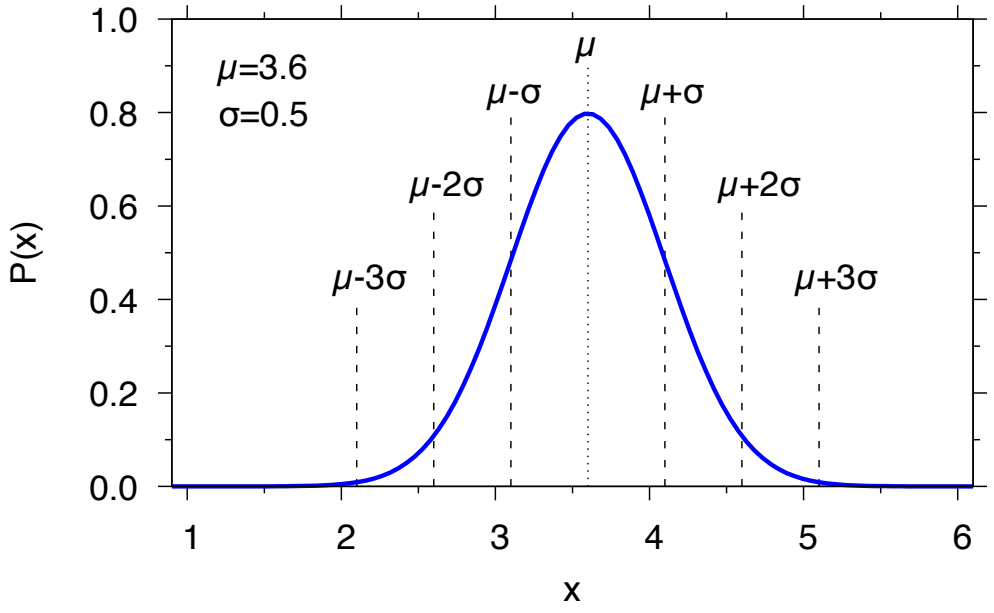
The form of the ant data curve follows a **normal** distribution. With the curve in an analytical form it becomes much easier to extract things like the mean or standard deviation. You may have come across this curve before, it is also known as a Gaussian distribution or a Bell curve. It is often useful to be able to fit a normal distribution to a data set so that the standard deviation and mean (and other quantities) can be quickly determined.

If we **normalise** the normal distribution by dividing by the total area beneath the curve, we obtain the probability. We call this the **normal probability distribution function**, and it takes the form:

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Here, μ is the **mean** and σ is the **standard deviation**.

Let's look at a graph of the normal probability distribution:



The dotted-lines drawn on the graph represent 1, 2 and 3 standard deviations above and below the mean. Around 68% of all of the points on a normal distribution curve will be within 1 standard deviation (from $\mu - \sigma$ to $\mu + \sigma$). Around 95% within 2 standard deviations and 99% within 3 standard deviations.

Remember that in general if we have a probability distribution function, $P(x)$, then we can find its mean by evaluating:

$$\bar{x} = \int_{-\infty}^{\infty} xP(x)\mathrm{d}x$$

This will also hold for this particular probability distribution function:

$$\begin{aligned}\bar{x} &= \int_{-\infty}^{\infty} xP(x)\mathrm{d}x \\ &= \int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\mathrm{d}x \\ &= \mu\end{aligned}$$

Our variance (σ^2) was given by $\sigma^2 = \int_{-\infty}^{\infty}(x-\bar{x})^2 g(x)\mathrm{d}x$, which in terms of our normal probability distribution function is:

$$\sigma^2 = \int_{-\infty}^{\infty}(x-\bar{x})^2 P(x)\mathrm{d}x$$

We will verify these two expressions in the tutorials.

Notation:

If X is a normally-distributed variable with a mean, μ and standard deviation, σ, then we write:

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

In words: X follows a normal distribution with mean, μ, and variance σ^2.

Standard Normal Distribution

The standard normal distribution is a special case of the normal probability distribution, where:

$$\mu = 0 \quad \text{and} \quad \sigma = 1$$

This curve is useful because it allows us to use a table to quickly work out the integrals. Let's look at the table on the next page.

The diagram shows the area that we are interested in from the centre of the curve (zero for the standard normal) to some value z. In the table, z is measured in the number of standard deviations. The number of standard deviations to 1 decimal place is located on the left hand side and the columns of the table (0-9) correspond to the second decimal place.

So, to find the probability that variable $X \sim N(\mu, \sigma^2)$ lies in the range $\mu < X < x$, we convert to the standard distribution using:

$$z = \frac{|x - \mu|}{\sigma}$$

This tells us that our value x lies z standard deviations away from the mean.

Then look up z in the standard normal distribution table to obtain the probability/area under $0 < Z < z$ in the standard distribution, which corresponds to the probability $P(\mu < X < x)$ in our original distribution.

To find other ranges, such as $x < X < +\infty$, we may need to use some extra steps using the fact that exactly half of the distribution is above and below the mean.

Notation:

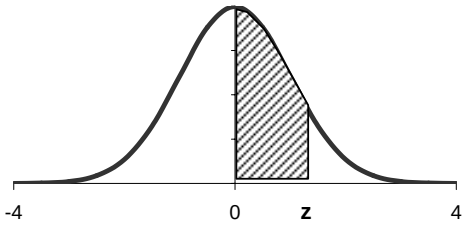
The standard normal distribution is often denoted, Z, and we would write:

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

In words: Z follows a normal distribution with mean, 0, and variance 1.

Standard Normal Distribution

Areas under the Standard Normal Curve from 0 to z



z	0	1	2	3	4	5	6	7	8	9
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998
3.6	.4998	.4998	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.7	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.8	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.9	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000

Example:

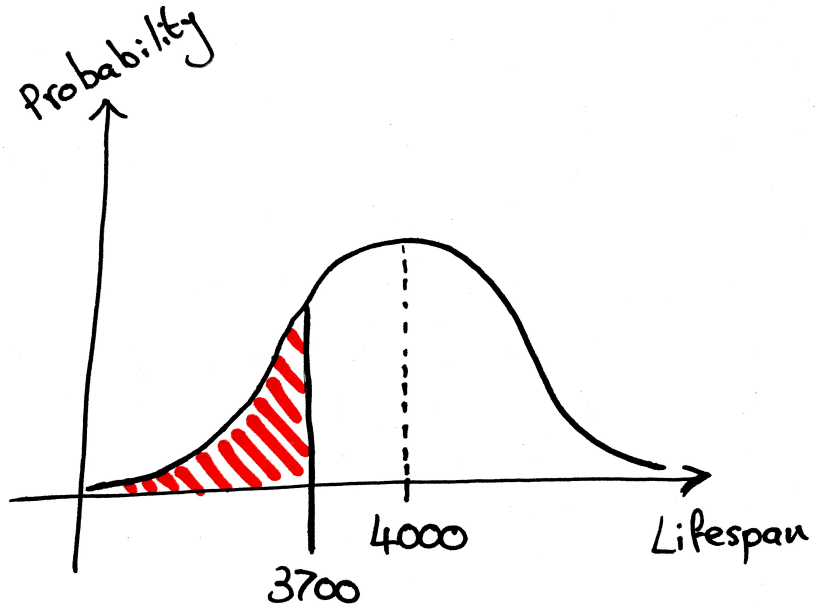
A company produces a particular microprocessor which is used to control industrial robots. Existing data indicates that the life span of the microprocessors is described by a Normal distribution with mean $\mu = 4000$ hours and standard deviation $\sigma = 200$ hours.

Determine the probability that the life span of such a microprocessor is:

1. Less than 3700 hours.
2. Between 3700 hours and 4250 hours.
3. More than 4250 hours.

In each case, it is helpful to draw the curve first.

(i) We want the area below 3700:



Let's call the lifespan variable X, then we want $P(X < 3700)$.

This can be found by calculating area from 3700 to 4000 (which is the same as that from 4000 to 4300) and getting $1/2$ minus that value.

Calculate the number of standard deviations from the average:

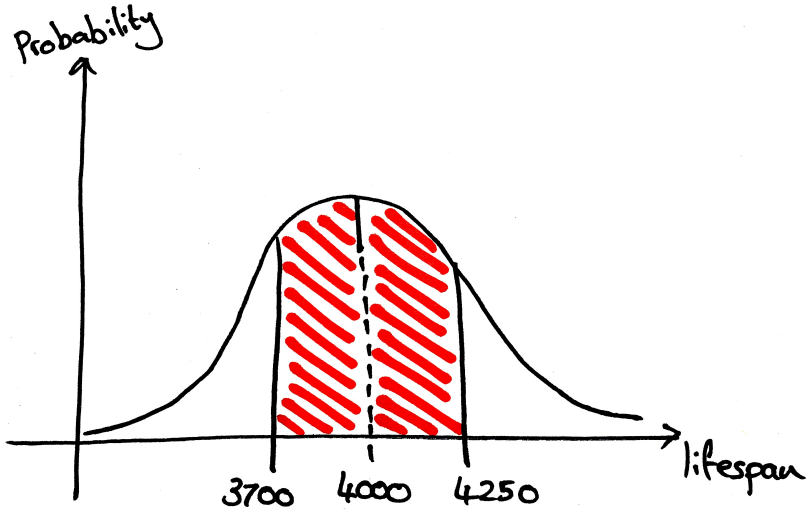
$$\text{No. Standard Deviations} = \frac{4000 - 3700}{200} = 1.5$$

From looking at the table, this value is 0.4332.

Thus, our area is:

$$\begin{aligned} P(X < 3700) &= 0.5 - P(3700 < X < 4000) \\ &= 0.5 - 0.4332 \\ &= 0.0668 \end{aligned}$$

(ii) We want the area between 3700 and 4250:



We already have the area from 3700 to 4000 (0.4332), now we need the area from 4000 to 4250.

Calculate the number of standard deviations from the average:

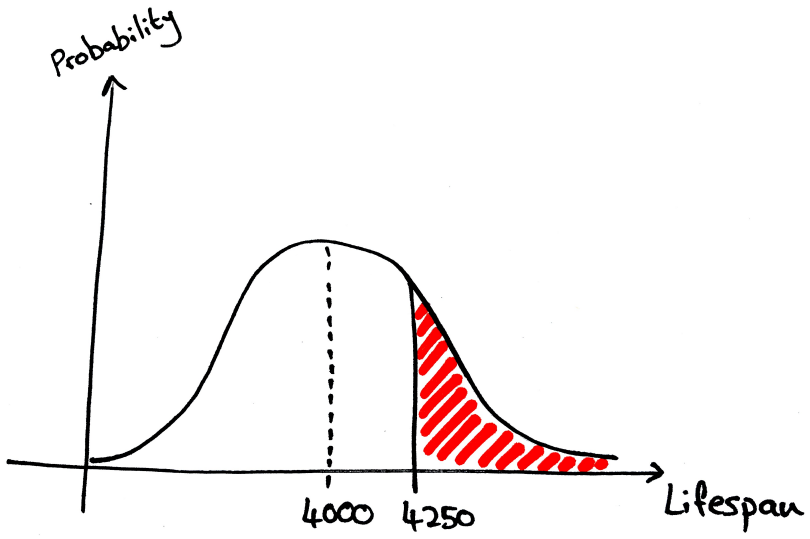
$$\text{No. Standard Deviations} = \frac{4250 - 4000}{200} = 1.25$$

From looking at the table, this value is 0.3944.

Thus, our total area is:

$$\begin{aligned} P(3700 < X < 4250) &= P(3700 < X < 4000) + P(4000 < X < 4250) \\ &= 0.4332 + 0.3944 \\ &= 0.8276 \end{aligned}$$

(iii) We want the area above 4250:



We already have the area from 4000 to 4250 (0.3944).

So all we need to do is subtract this from 0.5:

$$\begin{aligned} P(X > 4250) &= 0.5 - P(4000 < X < 4250) \\ &= 0.5 - 0.3944 \\ &= 0.1056 \end{aligned}$$

3.3.1 Cumulative Frequency Distribution

The cumulative frequency distribution (CFD) is the sum of the area under normal distribution curve, i.e. the integral of the normal probability distribution curve. As discussed above, the area under the normal probability distribution curve goes to 1, so as the CDF tends to infinity, its value should also tend to 1. The integral is not possible to solve exactly, instead we write the function in terms of what is known as the error function.

The error function is by definition:

$$erf(x) = \frac{1}{\sqrt{\pi}} \int_{-x}^{x} e^{-t^2} dt$$

In our case (for the normal probability distribution) we have μ and σ appearing and the equation takes the form (we use the symbol Φ for the CDF):

$$\Phi(x) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$$

Below we can see what this curve looks like for a range of values of μ and σ:

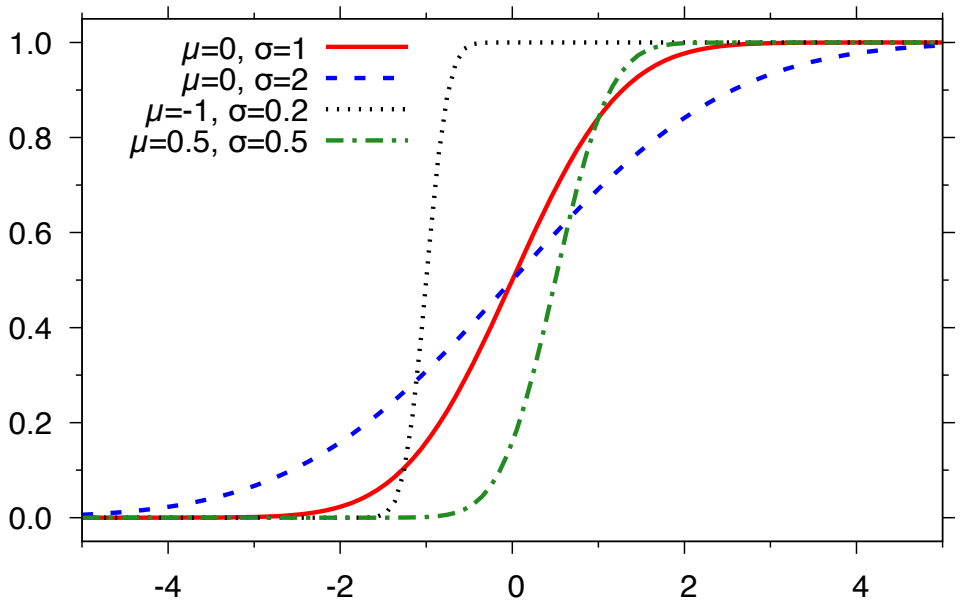


Figure 15: Cumulative frequency distributions for normal distributions

3.3.2 Summary: Normal Distribution

To summarise the important properties of a normal distribution $X \sim N(\mu, \sigma^2)$:

Mean	μ
Median	μ
Mode	μ
Variance	σ^2
Cumulative distribution function	$\frac{1}{2} + \frac{1}{2} \operatorname{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)$

Table 1: Statistical properties of the normal distribution

3.4 Other examples of continuous probability distributions

3.4.1 The Log-Normal Distribution

We have already come across the Gaussian distribution which crops up in many different places. One of the other common distributions is the log-normal distribution.

An example of where one might find a log-normal distribution of a property is when looking at grain sizes in polycrystalline materials. Below is an atomic force microscope image of a surface of three different semiconductor devices made of $Cu(In,Ga)Se_2$.

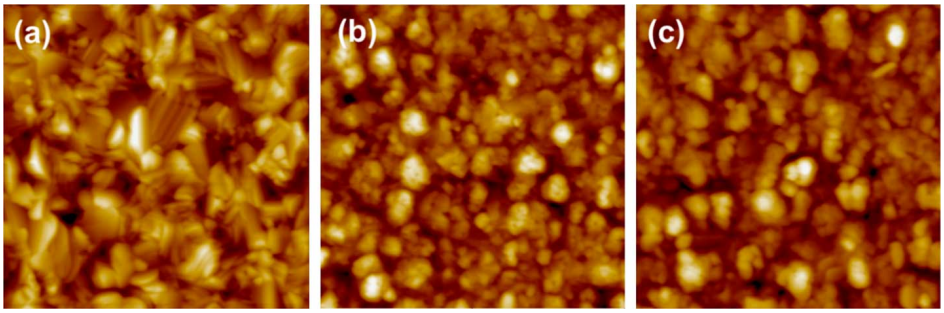


Figure 16: AFM images of three devices taken from Microscopy Today, Volume 26, Issue 3 pages 32-39 (2018).

Using computer software it is possible to analyse these images to find a distribution of grain sizes as shown below.

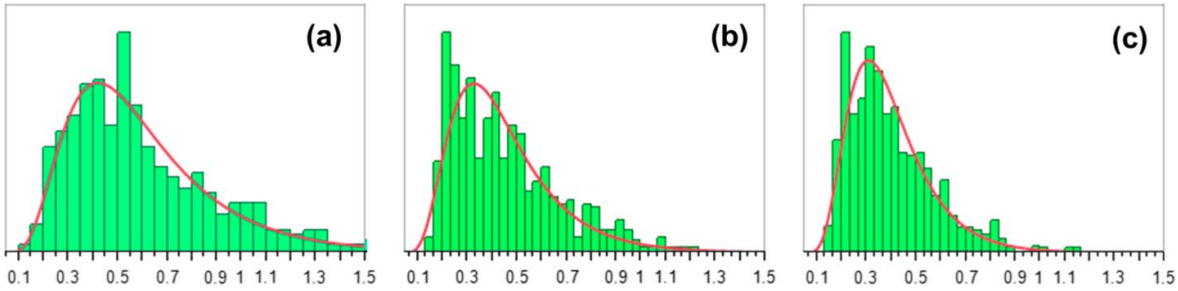


Figure 17: Grain size distributions of three devices taken from Microscopy Today, Volume 26, Issue 3 pages 32-39 (2018).

A log-normal distribution has the following analytical form for a variable, x:

$$P(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}$$

The formula looks similar in form to the normal distribution, but with an extra $1/x$ at the front and $\ln(x)$ in the exponential. Unlike a normal distribution, whose mean is μ (the middle of the curve) and the variance just given by σ^2, the log-normal distribution has the following properties:

Mean	$\exp\left(\mu + \frac{\sigma^2}{2}\right)$
Median	$\exp(\mu)$
Mode	$\exp(\mu - \sigma^2)$
Variance	$[\exp(\sigma^2) - 1]\exp(2\mu + \sigma^2)$
Cumulative distribution function	$\frac{1}{2} + \frac{1}{2}\,\mathrm{erf}\left(\frac{\ln(x)-\mu}{\sqrt{2}\sigma}\right)$

Table 2: Statistical properties of the log-normal distribution

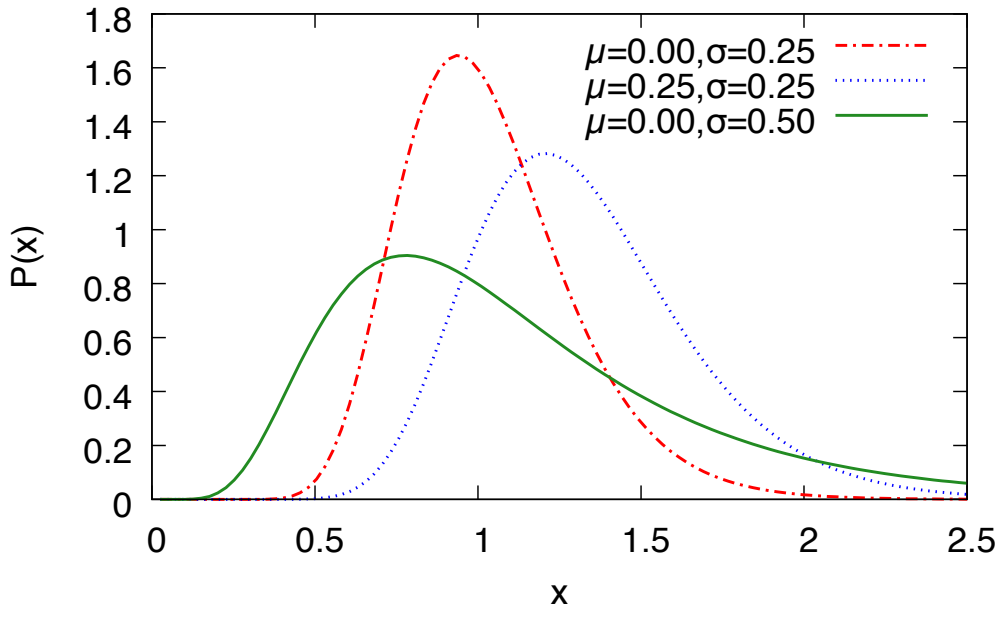


Figure 18: Examples of log-normal distributions

3.4.2 The Lorentzian Distribution

A different type of distribution that is unrelated to the normal distribution is one known as the Lorentzian distribution or the Cauchy distribution.

This distribution can arise in the case where we have a driven and damped resonator. For example, a mass on a spring will naturally damp its oscillations if given an initial push. However, if you drive the mass with a periodic force like $F_0 \cos(\omega t)$, the mass will eventually reach a steady state.

This can be described by solving the following ordinary differential equation (ODE):

$$\frac{d^2x(t)}{dt^2} + \frac{\gamma dx(t)}{dt} + \omega_0^2 x(t) = F_0 \cos(\omega t) \tag{1}$$

For example with $\omega_0 = 2$ and $\omega = 0.8$:

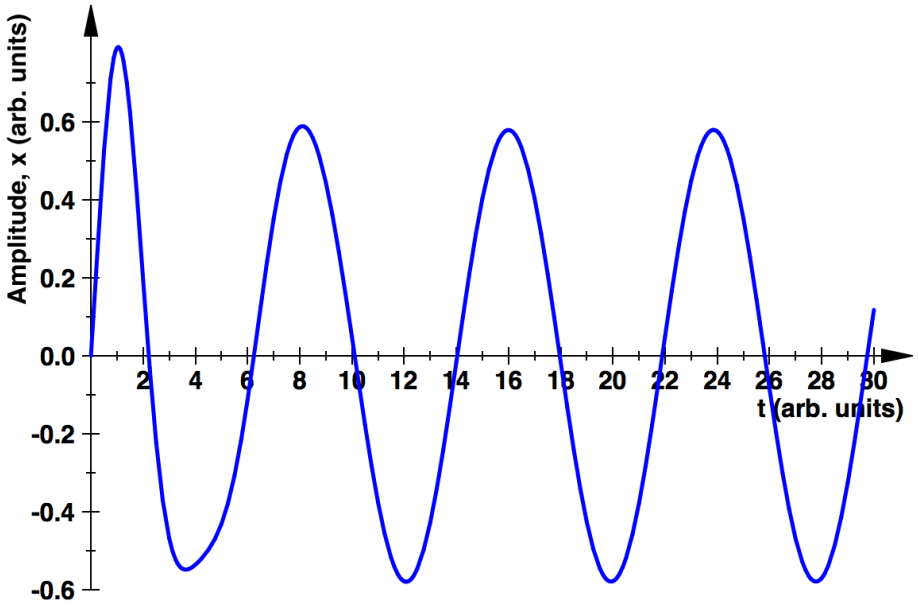


Figure 19: Solution to this ODE with $\omega_0 = 2$ and $\omega = 0.8$.

If we now solve the equation with $\omega_0 = 2$ and $\omega_0 = 1.2$, we see that the amplitude is larger:

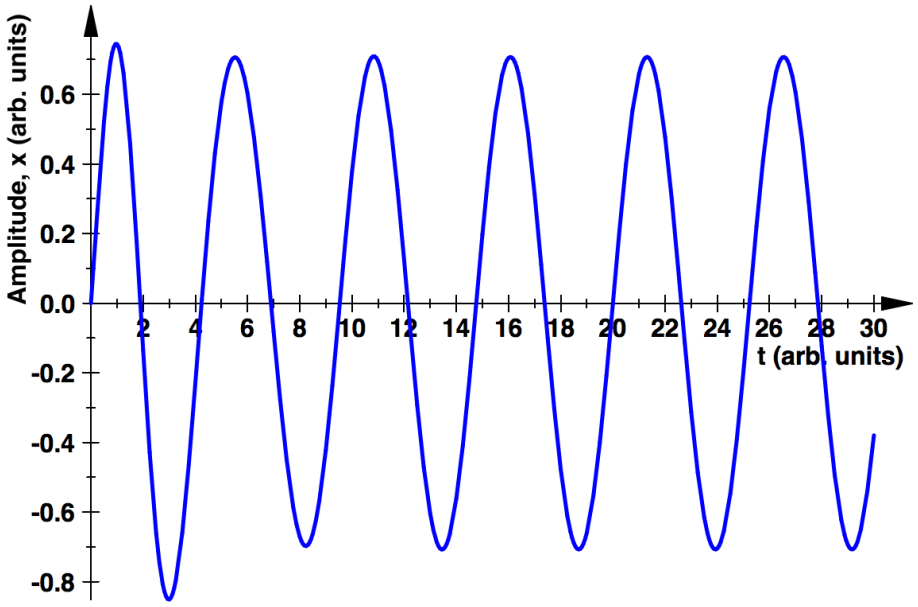


Figure 20: Solution to Eq. **??** with $\omega_0 = 2$ and $\omega = 1.2$.

As we drive the system to resonance the amplitude increases and as we go beyond resonance the amplitude decreases once more.

If we plot the amplitude as a function of the frequency we obtain a Lorentzian distribution, which has the following (normalised) form:

$$L(x) = \frac{1}{\pi} \frac{\frac{1}{2}\Gamma}{(x - x_0)^2 + \left(\frac{1}{2}\Gamma\right)^2}$$

Examples of this distribution for various parameter choices are shown in Figure 21.

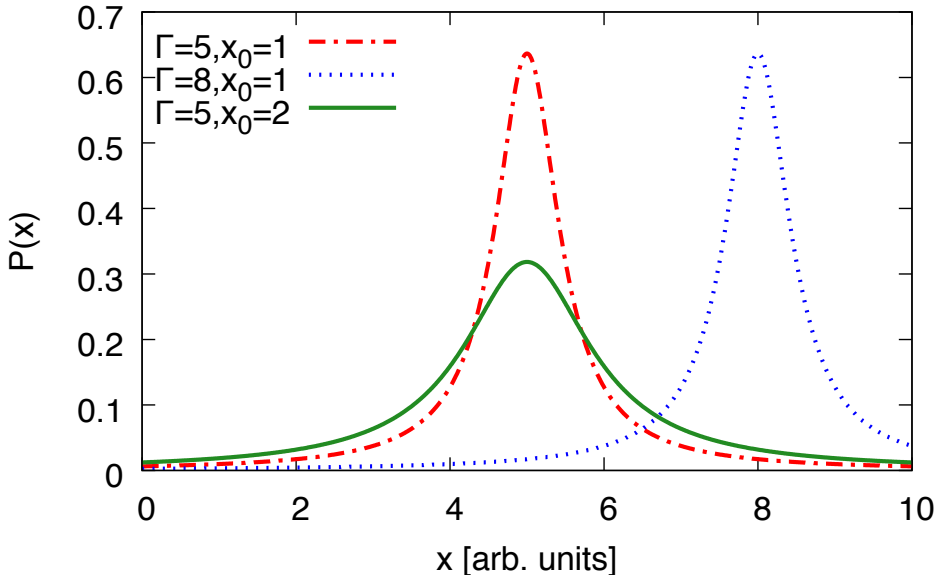


Figure 21: Examples of Lorentzian functions for different values of Γ and x_0.

Unlike the normal and log-normal functions, however, the Lorentzian technically does **not** have a mean and a variance as defined by:

$$\bar{x} = \int_{-\infty}^{\infty} xL(x)\mathrm{d}x \qquad \text{and} \qquad \sigma^2 = \int_{-\infty}^{\infty} (x-\bar{x})^2 L(x)\mathrm{d}x$$

because the function does not converge to zero at infinity. However, we can approximate the mean and variance using the integrals above by taking them from some large negative value to some large positive value. The Lorentzian does, however, have a median and a mode.

These properties are summarised in the table:

Mean	undefined
Median	x_0
Mode	x_0
Variance	undefined
Cumulative distribution function	$\frac{1}{\pi}\,\mathrm{atan}\left(\frac{x-x_0}{\Gamma}\right)+\frac{1}{2}$

Table 3: Statistical properties of the Lorentzian distribution

4 Lecture 4: Combining Normal Distributions, and Chauvenet's Criterion

4.1 Overview

- Previously, we extended histograms to probability distributions, and performed calculations using the normal distribution in particular.
- This week we will combine distributions.
- Key theory: If In general, if X and Y are independent random variables, then:

  $$Mean(X \pm Y) = Mean(X) \pm Mean(Y)$$

  and

  $$Var(X \pm Y) = Var(X) + Var(Y)$$

- Applying this to Normally-distributed variables in particular, adding and subtracting them will give a new variable that also obeys a normal distribution. If:

  $$X \sim N(\mu_x, \sigma_x^2) \quad \text{and} \quad Y \sim N(\mu_y, \sigma_y^2)$$

  Then:

  $$X \pm Y \sim N(\mu_x \pm \mu_y, \sigma_x^2 + \sigma_y^2)$$

Chauvenet's Criterion (3)

- When collecting data, some points may seem extreme and unrepresentative: outliers
- How can we deal with these in a systematic way?
- We can use Chauvenet's criterion to classify outliers in a sample if:

  it is from a normally-distributed Population

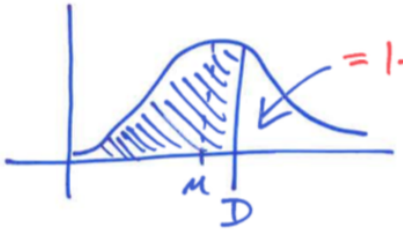
Procedure for applying Chauvenet's Criterion:

④

For each data point x_i in the sample:

① $D = \dfrac{|x_i - \mu|}{\sigma}$ how many s.d. from the mean is x_i?

② What is the probability of a point being at least that far above the mean in a standard normal distribution?



$= 1 - NORMSDIST(D)$

$P(x > D)$

③ How many points in a sample of size N would I expect to be that far away?

$$N_E = N \times P(x > D)$$

④ If $N_E < 0.5$ Reject!

4.2 Combining Data Sets

Let's say we have two sets of data, for example the size of a bored hole for a bolt and the width of the bolt. We may know something about each set, such as the mean and standard deviation of the hole and of the bolts, but can we work out how many pairs are likely to fit?

Consider the following example:

Assume we have the lengths of two sets each of four metal bars.

The first set all lie between 70 and 90:

Say, 72, 80, 88, 77.

Call the lengths of these bars $\{a_i\}$, the set A.

What is their mean and variance?

$$\begin{aligned}\bar{a} &= \frac{1}{4}\sum_{i=1}^{4} a_i \\ &= \frac{72+80+88+77}{4} \\ &= 79.25\end{aligned}$$

$$\begin{aligned}\sigma_a^2 &= \frac{1}{4}\sum_{i=1}^{4}(a_i-\bar{a})^2 \\ &= \frac{(72-79.25)^2+(80-79.25)^2+(88-79.25)^2+(77-79.25)^2}{4} \\ &= 33.6875\end{aligned}$$

The second set of 4 lengths lie between 20 and 30:

Say, 29, 21, 25, 26

Call the lengths of these bars $\{b_i\}$, the set B.

What is their mean and variance?

$$\begin{aligned}\bar{b} &= \frac{1}{4}\sum_{i=1}^{4} b_i \\ &= \frac{29+21+25+26}{4} \\ &= 25.25\end{aligned}$$

$$\begin{aligned}\sigma_b^2 &= \frac{1}{4}\sum_{i=1}^{4}(b_i-\bar{b})^2 \\ &= \frac{(29-25.25)^2+(21-25.25)^2+(25-25.25)^2+(26-25.25)^2}{4} \\ &= 8.1875\end{aligned}$$

One bar from set A and one bar from set B are chosen at random.
What is the mean and variance of the total length of two bars combined?

We have to combine the two lengths taking into account all the possible combinations:

Let's call the lengths, $\{l_i\}$, set L.

Write down all of the possible combinations. There are $4 \times 4 = 16$ possible pairings:

Select bar 1, which has a length of 72.
This can then be combined with each of the bars in B, giving:

$72 + 29 = 101$
$72 + 21 = 93$
$72 + 25 = 97$
$72 + 26 = 98$

Select bar 2 . . .

$80 + 29 = 109$
$80 + 21 = 101$
$80 + 25 = 105$
$80 + 26 = 106$

Select bar 3 . . .

$88 + 29 = 117$
$88 + 21 = 109$
$88 + 25 = 113$
$88 + 26 = 114$

Select bar 4 . . .

$77 + 29 = 106$
$77 + 21 = 98$
$77 + 25 = 102$
$77 + 26 = 103$

What is the mean and variance of this new set, L?

Add up these 16 possible combined lengths:

$$101 + 93 + 97 + 98 + 109 + 101 + 105 + 106 + 117$$
$$+109 + 113 + 114 + 106 + 98 + 102 + 103$$
$$= 1672$$

Find the mean:

$$\bar{l} \;=\; \frac{1692}{16} \;=\; 104.5$$

$$\bar{a} + \bar{b} \;=\; 79.25 + 25.25 \;=\; 104.5$$

Add up the square deviation:

$$(101 - 104.5)^2 + (93 - 104.5)^2 + (97 - 104.5)^2 + (98 - 104.5)^2$$
$$+(109 - 104.5)^2 + (101 - 104.5)^2 + (105 - 104.5)^2 + (106 - 104.5)^2$$
$$+(117 - 104.5)^2 + (109 - 104.5)^2 + (113 - 104.5)^2 + (114 - 104.5)^2$$
$$+(106 - 104.5)^2 + (98 - 104.5)^2 + (102 - 104.5)^2 + (103 - 104.5)^2$$
$$= 670$$

Find the variance:

$$\sigma_l^2 \;=\; \frac{670}{16} \;=\; 41.875$$

$$\sigma_a^2 + \sigma_b^2 \;=\; 33.6875 + 8.1875 \;=\; 41.875$$

So we have found that the mean is the sum of the two means: $\bar{l} = \bar{a} + \bar{b}$.

The variance is the sum of the two variances: $\sigma_l^2 = \sigma_a^2 + \sigma_b^2$.

Adding and Subtracting Distributions

In general, if X and Y are independent random variables, then:

$$\boxed{\text{Mean}(X \pm Y) = \text{Mean}(X) \pm \text{Mean}(Y)}$$

and

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

Note that the variances are never subtracted.

Applying this to Normally-distributed variables in particular, adding and subtracting them will give a new variable that also obeys a normal distribution.

If:

$$X \sim N(\mu_x, \sigma_x^2) \quad \text{and} \quad Y \sim N(\mu_y, \sigma_y^2)$$

And if:

$$S = X + Y \quad \text{and} \quad D = X - Y$$

Then:

$$S \sim N(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2) \quad \text{and} \quad D \sim N(\mu_x - \mu_y, \sigma_x^2 + \sigma_y^2)$$

Example 1

Two companies produce ball bearings whose masses obey normal distributions. Company A produces ball bearings whose mass in grams is given by $A \sim N(6.1, 0.5^2)$, while Company B's products have mass $B \sim N(5.8, 0.5^2)$.

What is the probability distribution of the difference in the masses of ball bearings made between the two companies?

Let variable D be the difference, in terms of how much larger a ball bearing made by Company A is than one made by company B:

$$D = A - B$$

Hence subtract the means but add the variances:

$$D \sim N(6.1 - 5.8, 0.5^2 + 0.5^2)$$

which can be simplied to either:

$$D \sim N(0.3, 0.5) \qquad \text{or} \qquad D \sim N(0.3, 0.707^2)$$

Example 2

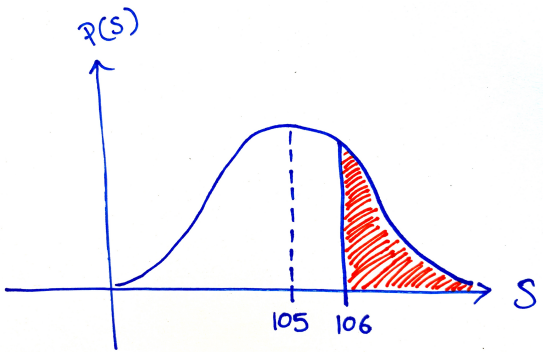
A company assembles swords that consist of two components: the blade and the hilt (i.e. the handle). The steel blades B (excluding the tang that overlaps with the hilt) are manufactured with lengths (in cm) given by a normal distribution $B \sim N(85, 4)$, and the hilts H are made of wood and have length obeying $H \sim N(20, 1)$.

What is the probability that a given sword exceeds 106cm in length?

Let variable S be the length of a sword given by the combination of a blade B and hilt H.

$$B \sim N(85, 4) \quad \text{and} \quad H \sim N(20, 1)$$

$$S = B + H \quad \Longrightarrow \quad S \sim N(85 + 20, 4 + 1) \quad \Longrightarrow \quad S \sim N(105, 5)$$



$$d = \frac{|h - \mu|}{\sigma} = \frac{106 - 105}{\sqrt{5}} = \frac{1}{\sqrt{5}} = 0.4472$$

So 106 is 0.45 standard deviations above the mean to 2 d.p.

From the tables, we find that this corresponds to an area of 0.1736. Hence:

$$P(105 < S < 106) = 0.1736$$

and so

$$\begin{aligned} P(S > 106) &= P(S > 105) - P(105 < S < 106) \\ &= 0.5 - 0.1736 \\ &= 0.3264 \end{aligned}$$

4.3 Chauvenet's Criterion

Sometimes when collecting data, we find that some values lie very far outside of the "expected" range, which we call **outliers**. This could be due to an error in the measurement, or simply a random fluctuation. When this occurs, there are statistical procedures that can help us decide whether a data point should be rejected from the data set or not. This can be controversial as some scientists believe that one should never discard any measurements. Chauvenet's theorem provide one systematic method of doing so.

> Chauvenet's theorem is a method to determine *so-called* outliers. It **only** applies to data samples that are known to have been taken from a **normally-distributed population**.

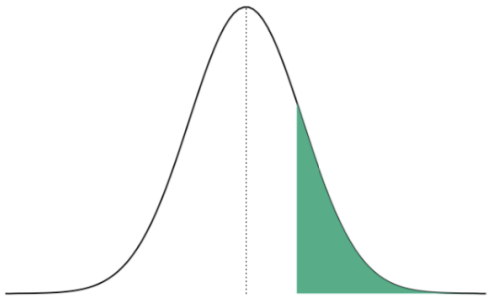
The process involves checking whether a data point lies beyond a certain distance (more than so many standard deviations) away from the mean. It is way of saying "is this sample *representative* of the population from which is was drawn"?

The number of standard deviations σ a value x is from the mean μ is given by:

$$D = \frac{|x - \mu|}{\sigma}$$

For a given data point, if the number of points in a sample data set of this size that would be expected to be lying beyond that many standard deviations above the mean is less than a half, then we reject that point.

To calculate the number of expected points we first need to know the probability ($P_{x>D}$) of finding a point beyond D in the *standard* normal distribution. This is drawn schematically below. We can use an Excel function called NORMSDIST to find this, which returns one minus this area (i.e. the size of the white area).



The number N_E of points, from our sample of size N, that are expected to lie farther away than D above the mean is given by:

$$N_E = N \times P_{x>D}$$

Chauvenet's Criterion:

If N_E is less than 0.5, the data point is rejected.

If $N_E < 0.5$, that would mean that (when we consider the probability of also being at least D standard deviations *below* the mean) that overall we expect *less than one point in a sample this size* to be that far from the mean, and so we discount it as not being representative of the population.

Procedure: Chauvenet's Theorem

The procedure is thus, given a set of N data points $\{x_i\}$:

1. Determine the mean:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

2. For each data point x_i, determine the number of standard deviations D from the mean it is, using:

$$D = \frac{|x_i - \mu|}{\sigma}$$

3. Calculate the probability $P_{x<D}$ in the standard normal distribution using the Excel function `NORMSDIST(D)`

4. Calculate:

$$P_{x>D} = 1 - P_{x<D}$$

   So this gives us the probability that a point is at least D standard deviations *above* the mean.

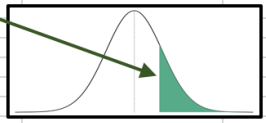
5. Calculate the number of data points in a sample of size N that we expect to be beyond this using:

$$N_E = N \times P_{x>D}$$

6. If this value is less than 0.5, reject data point x_i as an outlier.

Let's look at an example in the Excel spreadsheet:
`MMaDLectureExampleChauvenetsTheorem.xlsx`.

	A	B	C	D	E	F	G	H
1	Data	Deviation from Mean	Number of SD's from Mean	Area below	Area of outliers	No. expected values outside	Accept/Reject	Data, outliers removed
2	6.01	=A2-$B$18	=ABS(A2-$B$18)/$B$19	=NORMSDIST(C2)	=1-D2	=E2*12	=IF(F2<0.5,"Reject","Accept")	=IF(G2="Accept",A2,"")
3	0.03	=A3-$B$18	=ABS(A3-$B$18)/$B$19	=NORMSDIST(C3)	=1-D3	=E3*12	=IF(F3<0.5,"Reject","Accept")	=IF(G3="Accept",A3,"")
4	6.77	=A4-$B$18	=ABS(A4-$B$18)/$B$19	=NORMSDIST(C4)	=1-D4	=E4*12	=IF(F4<0.5,"Reject","Accept")	=IF(G4="Accept",A4,"")
5	7.93	=A5-$B$18	=ABS(A5-$B$18)/$B$19	=NORMSDIST(C5)	=1-D5	=E5*12	=IF(F5<0.5,"Reject","Accept")	=IF(G5="Accept",A5,"")
6	7.48	=A6-$B$18	=ABS(A6-$B$18)/$B$19	=NORMSDIST(C6)	=1-D6	=E6*12	=IF(F6<0.5,"Reject","Accept")	=IF(G6="Accept",A6,"")
7	6.34	=A7-$B$18	=ABS(A7-$B$18)/$B$19	=NORMSDIST(C7)	=1-D7	=E7*12	=IF(F7<0.5,"Reject","Accept")	=IF(G7="Accept",A7,"")
8	6.72	=A8-$B$18	=ABS(A8-$B$18)/$B$19	=NORMSDIST(C8)	=1-D8	=E8*12	=IF(F8<0.5,"Reject","Accept")	=IF(G8="Accept",A8,"")
9	7.45	=A9-$B$18	=ABS(A9-$B$18)/$B$19	=NORMSDIST(C9)	=1-D9	=E9*12	=IF(F9<0.5,"Reject","Accept")	=IF(G9="Accept",A9,"")
10	7.03	=A10-$B$18	=ABS(A10-$B$18)/$B$19	=NORMSDIST(C10)	=1-D10	=E10*12	=IF(F10<0.5,"Reject","Accept")	=IF(G10="Accept",A10,"")
11	13.01	=A11-$B$18	=ABS(A11-$B$18)/$B$19	=NORMSDIST(C11)	=1-D11	=E11*12	=IF(F11<0.5,"Reject","Accept")	=IF(G11="Accept",A11,"")
12	6.63	=A12-$B$18	=ABS(A12-$B$18)/$B$19	=NORMSDIST(C12)	=1-D12	=E12*12	=IF(F12<0.5,"Reject","Accept")	=IF(G12="Accept",A12,"")
13	7.16	=A13-$B$18	=ABS(A13-$B$18)/$B$19	=NORMSDIST(C13)	=1-D13	=E13*12	=IF(F13<0.5,"Reject","Accept")	=IF(G13="Accept",A13,"")
14								
15								
16	Original	Data		Data with outliers	removed			
17	Number of points	12		Mean	=AVERAGE(H2:H13)			
18	Mean	=AVERAGE(A2:A13)		Std Dev	=STDEV.P(H2:H13)			
19	Std Dev	=STDEV.P(A2:A13)						
20								
21								
22								



NORMSDIST(x) - where x is the number of standard deviations from the mean:

This command gives the area under the standard normal curve in the range from $-\infty$ to x.

5 Lecture 5:
Correlation and Curve Fitting

5.1 Summary

How can we quantify the "goodness of fit" of a model $y = f(x)$ to a data set (x_i, y_i)? (2)

$y = f(x)$: $y = mx + c$ (straight line)

The "coefficient of determination" R^2:

$$R^2 = 1 - \frac{SS_{Res}}{SS_{tot}}$$

$$SS_{Res} = \sum_{i=1}^{N} (y_i - f(x_i))^2$$

(y_i: actual value; $f(x_i)$: predicted by model)

(SS_{Res}: sum of square residuals)

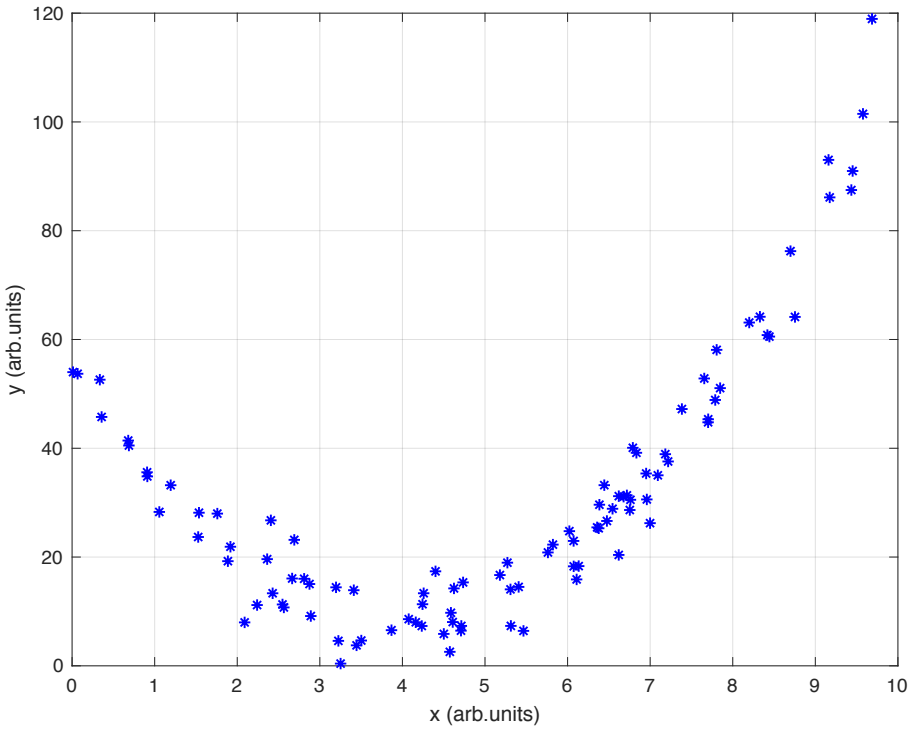
$$SS_{tot} = \sum_{i=1}^{N} (y_i - \bar{y})^2$$

5.2 Curve Fitting

Often we can obtain a set of experimental data, and hypothesise that the relationship between the independent variable (what we control) and the dependent variable (what we measure) is described by some function. If we could determine the exact nature of the relationship, this could provide us with some insight into the physical processes that generated the data, and would give us a means to make further predictions by extrapolating or interpolating the fitted curve.

Once we have decided on a general form of the relationship between our variables (e.g. linear, quadratic, expontial, power law), curve fitting is the process of finding the **set of parameter values** that best fits the set of data.

Suppose we have the set of data shown below:



In this case, you might look at that data and think that a quadratic function $y = ax^2+bx+c$ might fit through the data. But what should the values of a, b and c be that would give the best fit?

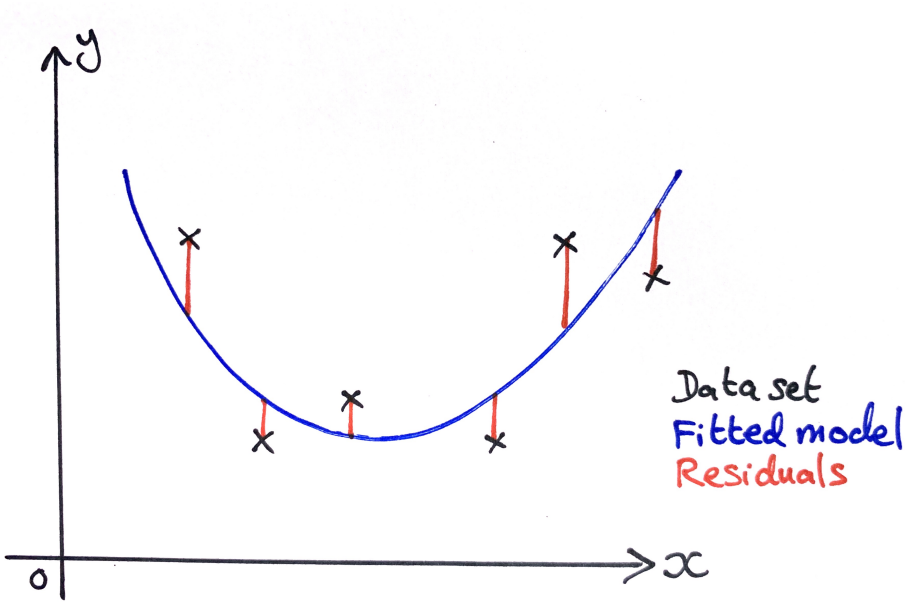
In other cases it may not be be obvious what function one should choose to fit a set of data, and so trial and error sometimes needs to be applied.

The procedure for curve fitting, broadly speaking, is:

1. Select an appropriate function that you think might match the data.
2. Choose some starting parameters (e.g. values for a, b and c in the function above).
3. Use some software to adjust your parameters to find "good" values of the parameters a, b and c.

How do we decide what a "good" fit is? Usually this is done by:

Minimising the sum of the squared differences ("residuals") between the actually-observed y-values and those predicted by our model with the current choice of parameters. The software will iterate through many choices of parameter values, and thus draw lots of different quadratic curves (blue), until the smallest total sum of the square of the residuals (the red distances between the model and the actual black data points) is obtained.



We can use various pieces of software to do this. First, let's consider a MATLAB procedure.

Curve fitting in MATLAB

The main file `fit2.m` that you need to execute is shown on the next page. This will generate some noisy data, and attempt to fit a curve using the function `lsqcurvefit`. This will vary the parameters in an attempt to find the least total difference in the squares of the actual and the model's predicted values at each of the 100 data points.

You also need another file `f2.m` in the **same** directory. This file contains the form of the function that you are trying to fit (as you can see, it contains a generic quadratic equation) and has the same name as the function in the first argument of `lsqcurvefit` in the main program. This is telling MATLAB to fit the function found in that secondary file to the data.

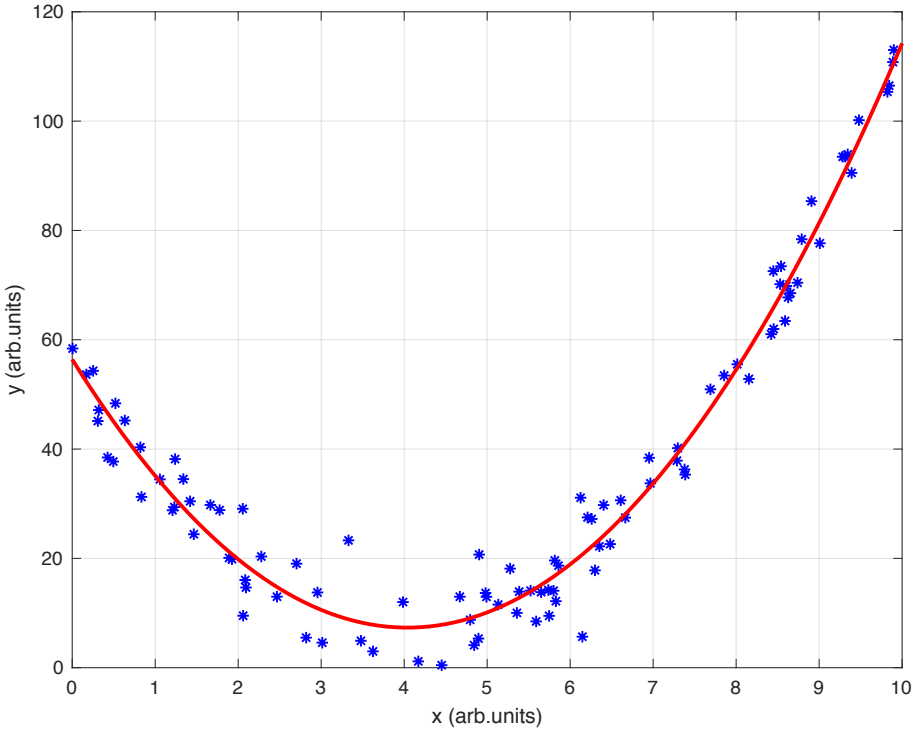
File: f2.m

```
1 function y=f(constant,x)
2     y=constant(1)*(x-constant(2)).^2+constant(3);
3 end
```

File: fit2.m

```
% Generate some random numbers to add some noise
x = rand(1,100).*10;
noise = randn(1,100);

% Make the data follow an x^2 curve but add in the noise
y = 3*(x-4).^2 + 8 + 5*noise;

% Plot the initial graph of randomly generated data
figure
plot(x,y,'b*')
xlabel('x (arb.units)')
ylabel('y (arb.units)')
grid on

% Execute the curve fit using function f2 with the initial
% guess of a, b and c of zero against the data x and y
constant = lsqcurvefit(@f2, [0;0;0], x, y);

% Copy the constants found from fitting to three variables:
% a, b and c
a=constant(1)
b=constant(2)
c=constant(3)

% Create a "space" for the x fitted data
xfit=0:0.1:10;

% Find the y-values using the fitted function
yfit=f2(constant,xfit);

% Plot both the data and the fit
figure

plot(x,y,'b*')
hold on
plot(xfit,yfit,'r','linewidth',2)
xlabel('x (arb.units)')
ylabel('y (arb.units)')
grid on
```

If you run this code you will see the figure below:



There are different algorithms that Matlab can use to fit the constants (a, b and c in this case). The algorithm we have chosen using the function `lsqcurvefit`) is a least squares fit, which is the most standard method. This is basically working by adjusting a, b and c until the **sum of the square distances between the data points and the function is minimised**.

In our case this means for our function f_2 and set of y data, $\{y_i\}$, we want to find the choice of three constants in constructing function f_2 that minimises:

$$min\left(\sum_{i=1}^{N}\left(f_2(x_i)-y_i\right)^2\right)$$

Procedure: Curve Fitting in Excel

When we use Excel, we have to do this a bit more manually using the SOLVER function. This may need to be added first. On PC, go to File>Options>Add-ins, and select Solver. For OSX, go to Tools>Excel Add-ins and again select Solver. It will now be available in the Data tab.

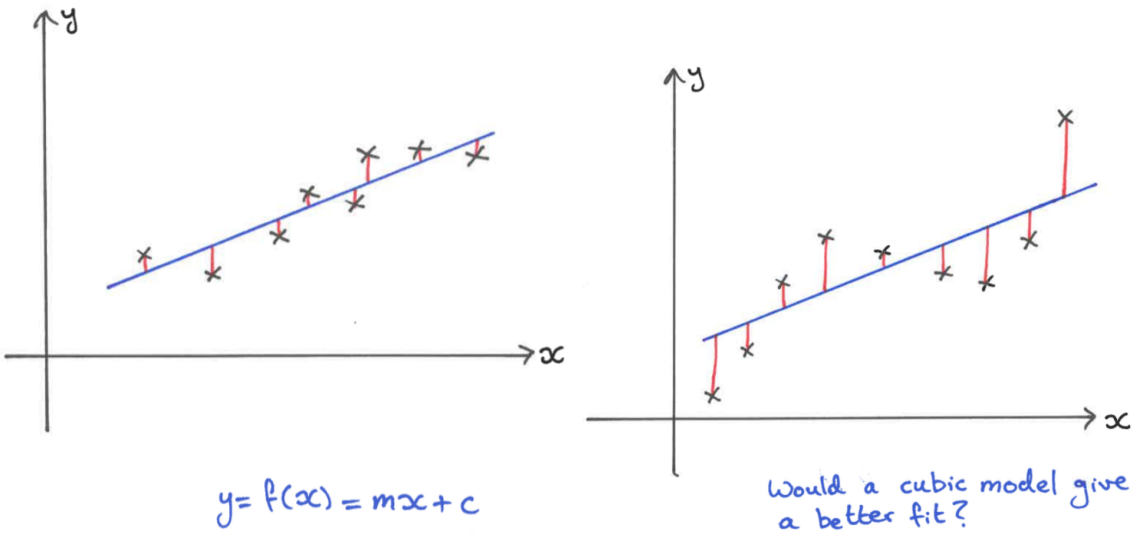
1. You should begin with columns containing the observed (actual) set of data, one column each with x-values and one with y-values.
2. Input some estimate parameter values in cells away from the columns of data.
3. Create a column of y-values predicted by the model, that reference the parameter value cells.
4. Create a column of the squares of the differences between the actual y-values and those estimated by the model. These are the "square residuals".
5. Sum all of these in a single cell.
6. Open SOLVER in the Data tab.
7. We want to **minimise** the sum of the square residuals.
8. By changing the variable cells containing the values of the parameters.
9. Ensure that "Make unconstrained variables non-negative" is NOT checked.
10. Click "Solve"!

In the lecture, we will now demonstrate this procedure by fitting a linear relationship (a straight line) to a set of data points gathered while investigating heat expansion in metals.

We will use the Excel workbook `ExcelCurveFittingExample.xlsx`

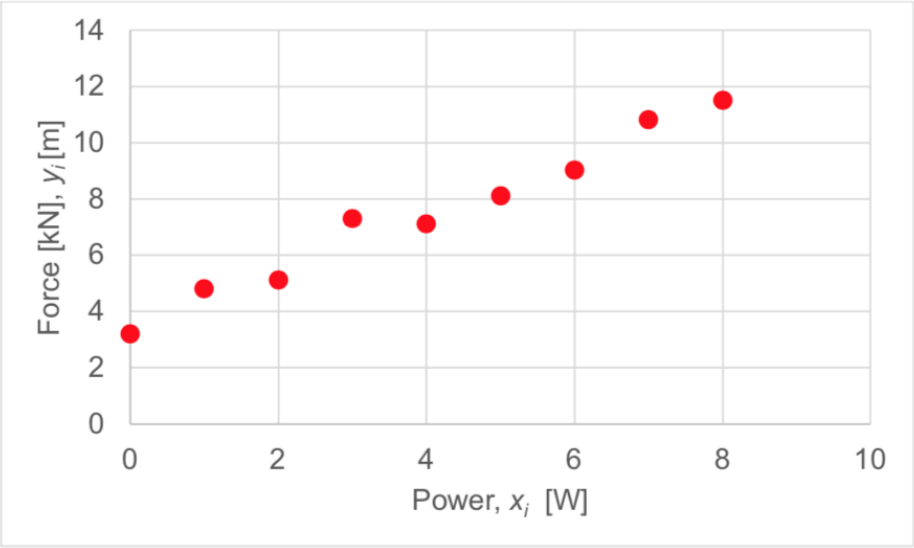
5.3 Correlation

Now that we have looked at curve fitting, whereby we had a set of data and we wanted to choose the best parameter values that fit a function to the data set. But we could choose several different models and fit the best parameter choices in each case. How would we know which described the data best by giving the closest fit?



There are several ways of quantifying the quality of a fit. We have already discussed what we are trying to do when we fit a curve: minimise the (total squared) difference between the curve and the data points. So the sum of the square differences between the data points and the fitted function is one measure of how good the fit is. However, there are sometimes problems with it, as it can be minimised yet the fitted curve still doesn't look very close to the data.

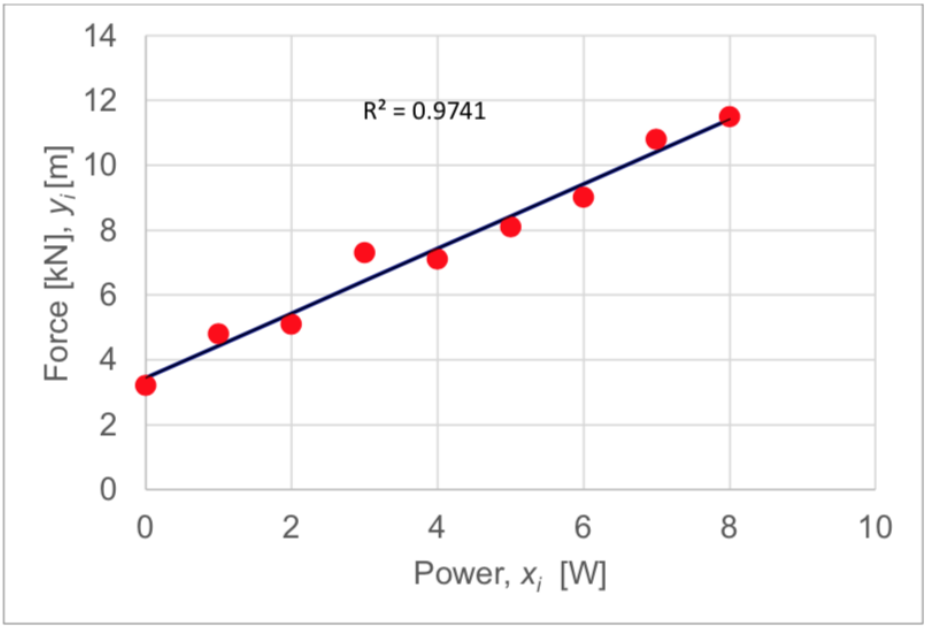
A value called R^2 is the standard approach to quantifying how good the fit is. It is also known as the coefficient of determination. If y_i is a set of data that we want to fit, for example, see the graph below of the force exerted, y_i, by a press as a function of its operating power, x_i.



Using SOLVER, we can fit a function to this curve. These data look to be linear, so we can try to fit the function:

$$f(x) = mx + c$$

We can easily get Excel to add a trendline to these data. Excel can (for the built in functions) calculate R^2, which is displayed on the chart.



For a set of N data points (x_i, y_i), to which a model is fitted given by $y = f(x)$,

R^2 is calculated according to:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

where

$$SS_{res} = \sum_{i=1}^{N} \big(y_i - f(x_i)\big)^2$$

and

$$SS_{tot} = \sum_{i=1}^{N} \big(y_i - \bar{y}\big)^2$$

A value of R^2 equal to 1 means that the curve fits the data perfectly. A smaller value means a poorer fit.

The fraction $\frac{SS_{res}}{SS_{tot}}$ is needed to scale the relative size of the residuals appropriately. Essentially, it means that the greater the variance in the y-values of the data set (i.e. the larger the value of SS_{tot}, the larger the squared residuals (the error between the data and the model) can be whilst still being considered a "good" fit.

So when we are presented with a data set, we might have an idea from the context of what kind of model to fit (for example, in the tutorial questions for this week we might know that the data is taken from a probability distribution that might be roughly normal or log-normal).

Otherwise, we could create a scatter plot of the data, and make an educated guess from the plot of what model(s) might be suitable. Then we would undertake a curve-fitting procedure in each case to find the *best version of each model*, and finally compare the resulting R^2-value to evaluate the best version of each model and determine which one produced the overall best result.

Here we show the formulae used in the Excel file "RSquaredExample". In this workbook, we fit the straight line relationship between force and power of a press discussed in the previous pages.

	A	B	C	D	E	F
1	Power, x_i [W]	Force, y_i [kN]	Model [kN]	Square Displacement [kN^2]		y_i-mean(y_i)
2	0	3.2	=$F$13*A2+$F$14	=(C2-B2)^2		=(B2-$B$11)^2
3	1	4.8	=$F$13*A3+$F$14	=(C3-B3)^2		=(B3-$B$11)^2
4	2	5.1	=$F$13*A4+$F$14	=(C4-B4)^2		=(B4-$B$11)^2
5	3	7.3	=$F$13*A5+$F$14	=(C5-B5)^2		=(B5-$B$11)^2
6	4	7.1	=$F$13*A6+$F$14	=(C6-B6)^2		=(B6-$B$11)^2
7	5	8.1	=$F$13*A7+$F$14	=(C7-B7)^2		=(B7-$B$11)^2
8	6	9.0123	=$F$13*A8+$F$14	=(C8-B8)^2		=(B8-$B$11)^2
9	7	10.8	=$F$13*A9+$F$14	=(C9-B9)^2		=(B9-$B$11)^2
10	8	11.5	=$F$13*A10+$F$14	=(C10-B10)^2		=(B10-$B$11)^2
11	Mean of Data:	=AVERAGE(B2:B10)	Sum of Square Residuals:	=SUM(D2:D10)	Sum of Squares:	=SUM(F2:F10)
12					The parameters to optimise:	
13					m	0.997076564970542
14					c	3.44639300594564
15			Sum of Square Residuals (SSres)	=D11		
16			Sum of Squares (SStot)	=F11		
17			R^2	=1-D15/D16		

1. First, we create a third column containing the y-values predicted by the model (in this case $y = mx + c$).

2. As before, use the SOLVER tool to obtain the values of the parameters m and c that minimise the sum of the square residuals SS_{res}.

3. Then we simply need to also calculate the sum of the squares SS_{tot}.

4. Finally calculate the R^2 value using the formula above.

6 Lecture 6:
Discrete distributions and Poisson Processes

6.1 Discrete probability distributions

The normal/Gaussian, log-normal and Lorentzian distributions that we have looked at are all **continuous probability distributions** as they describe the probable values of a continuous random variable.

There are also discrete probability distributions that show the probability of a discrete random variable taking a particular value. This could be useful if we need the probability of whether an event occurs or not (using the Binomial distribution), or if we wanted to know *how many times* a discrete event is likely to occur.

6.1.1 Discrete versus continuous probability distributions

A continuous random variable, such as the height h of a student, can take any value in some interval. Thus we have been interested in the probability that it falls within a particular range $P(a < h < b)$ using a continuous probability distribution.

A discrete random variable, such as the number of students X in a class, can only take values from some discrete set. A discrete probability distribution would therefore tell us the likelihood of X taking specific discrete values:

$$P(X = 0), \quad P(X = 1), \quad P(X = 2), \quad \ldots$$

6.2 Binomial distribution

Consider an experiment with n independent trials, where there are exactly two outcomes of each trial: success with probability p, or failure with probability $1 - p$.

For example, tossing a (potentially unfair) coin n times, with probability p of getting a head each time.

- If you conducted $n = 100$ coin tosses, what is the probability of getting exactly 17 heads?
- What is the mean number of heads that you can expect?
- What is the variance of the distribution of the number of heads?

Because there can only be an integer number of heads, the variable H "number of heads" has a discrete distribution. In particular, for an experiment with n independent trials with two outcomes, the probability of the number X of "successes" is given by the **Binomial distribution**:

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

where

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

is the binomial coefficient and ! is the factorial symbol.

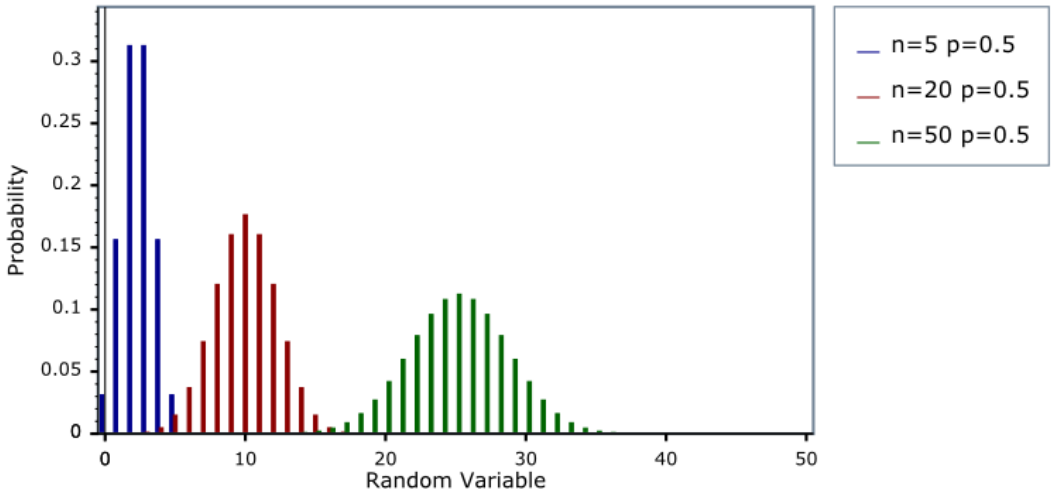
A binomial distribution with n trials and p probability of success has mean:

$$E(X) = np$$

and variance

$$Var(X) = np(1-p)$$

The shape of the distribution depends on n and p:



Heads and tails:

Returning to the coin toss problem then:

- The probability of 17 heads is:

$$P(H = 17) = \binom{100}{17} p^{17}(1-p)^{83}$$

$$= \frac{100!}{17!\,83!} p^{17}(1-p)^{83}$$

The fraction is the number of different combinations of choosing 17 heads from 100 coins: about 1 with 268 zeros written after it!

If the coin is *fair* ($p = 0.5$), this becomes:

$$P(H = 17) = \frac{100!}{17!\,83!}(0.5)^{17}(0.5)^{83} = 5.246 \times 10^{-12}$$

- The expected number of heads is:

$$E(H) = np = 100p$$

For a fair coin:

$$E(H) = 100 \times 0.5 = 50$$

as we would probably have guessed!

- The variance in the number of heads is:

$$Var(H) = np(1-p) = 100p(1-p)$$

For a fair coin:

$$Var(H) = 100 \times 0.5(1-0.5) = 25$$

Example:

A component supplier claims that 95% of its catalogue items are in stock at any time.

A particular order for 20 different components is returned with three items missing as being out of stock.

Is this a likely outcome, given the supplier's claim?

Solution:

Each item is either in stock or not, and the probability of each item being out of stock is 5%, so the binomial distribution applies:

$$P(k \text{ items out of stock}) = \binom{20}{k} 0.05^k 0.95^{20-k}$$

So the probability that *at least three* items are out of stock is:

$$\begin{aligned} P(3+ \text{ items out of stock}) &= P(3) + P(4) + \cdots + P(20) \\ &= 1 - P(0) - P(1) - P(2) \\ &= 1 - 0.3585 - 0.3774 - 0.1887 \\ &= 0.0755 \end{aligned}$$

So this is unlikely to occur.

6.3 Poisson processes and the Poisson distribution

A Poisson process describes a situation where discrete events occur from time to time and may be characterised either in terms of the probability of an event happening at each moment in time, or by the distribution of the amount of time between two subsequent events.

The specific criteria are:

- Events are independent of each other. The occurrence of one event does not affect the probability another event will occur.
- The **average** rate of events (the number of events per a given time period) is constant.
- Two events cannot occur at the exact same time.

Typical examples include:

- the arrival of customers to a queue
- the arrival of jobs to a printer
- the arrival of telephone calls to an exchange
- breakdowns of a machine
- the arrival of claims to an insurance company
- the arrival of vehicles at a traffic intersection.

For engineers, the main applications are in statistical quality control as you try to determine the likelihood that a product (e.g. the components of an aircraft jet engine) will fail in a given time, and in the design of computer and internet networks as they have to be designed to handle search requests that may occur according to a Poisson process.

For such a process, consider a particular interval of time. If λ is the rate at which the event can be expected to occur over that interval, then the probability that the *actual* number of events which occur X in that time period is equal to k is given by:

$$P(X = k) = \frac{\lambda^k \, \mathrm{e}^{-\lambda}}{k!} \qquad k = 0, 1, 2, 3, \ldots$$

and the distribution of the different values (i.e. the probable number of events that occur in that time interval) is given by the Possion distribution.

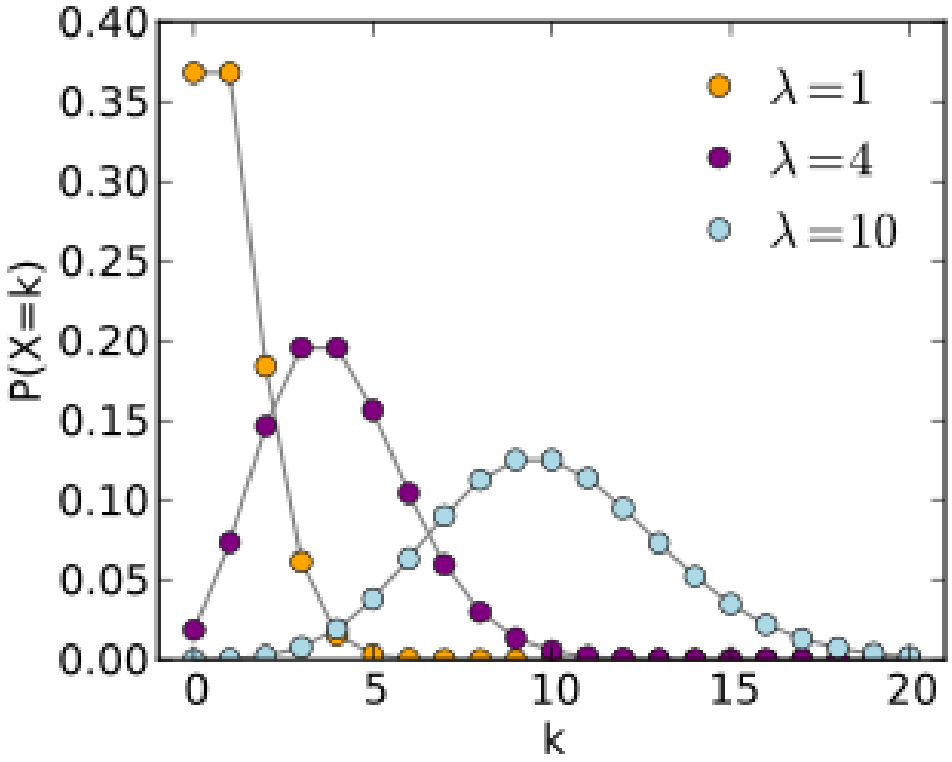


Figure 22: Poisson distribution with three choices of λ

Both the mean and the variance of the Poisson Distribution are equal to λ.

Example 6.1. *If you knew that (on average) my internet connection disconnects three times in a 30-minute game of* Rainbow Six: Siege *and that this is a Poisson process, then the probability that it will drop* exactly twice *in the next match is:*

$$P(X = 2) \;=\; \frac{3^2\,\mathrm{e}^{-3}}{2!} \;=\; \frac{9}{2}\,\mathrm{e}^{-3} \;\approx\; 0.224$$

where the variable X is defined as the number of internet disconnects during a 30-minute match.

Example 6.2. *On average, 12 vehicles arrive at an intersection every hour. What is the probability that:*

(a) Exactly 8 cars arrive in a given hour?

(b) Exactly 6 cars arrive in a given 20 minute period?

(c) No more than 2 cars arrive in a given hour?

Solution:

(a) Let X be the number of cars that arrive in one hour.

In this case, $\lambda = 12$, and we want $P(X = 8)$

$$P(X = 8) = \frac{12^8 \, e^{-12}}{8!} \approx 0.0655$$

(b) Let Y be the number of cars that arrive in a 20 minute interval.

This time, $\lambda = 12/3 = 4$ and we want $P(Y = 6)$

$$P(Y = 6) = \frac{4^6 \, e^{-4}}{6!} \approx 0.1042$$

(c) Back to $\lambda = 12$ but we want to find:

$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$.

$$P(X = 0) = \frac{12^0 \, e^{-12}}{0!} \approx 0.00000614$$

$$P(X = 1) = \frac{12^1 \, e^{-12}}{1!} \approx 0.00007373$$

$$P(X = 2) = \frac{12^2 \, e^{-12}}{2!} \approx 0.00044238$$

$$\begin{aligned} P(X \leq 2) &= P(X = 0) + P(X = 1) + P(X = 2) \\ &= 0.00000614 + 0.00007373 + 0.00044238 \\ &= 0.000522 \end{aligned}$$

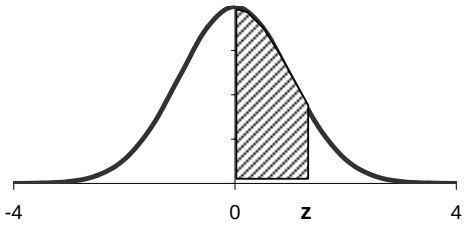
7 Revision

So, what should I be able to do at this point in the course?

- Calculate basic probabilities using discrete random variables (e.g. dice rolls, picking cards from a deck). Understand how this these calculations are affected by conditional probabilities, independence, and mutual exclusivity.
- Draw and interpret Venn diagrams.
- Calculate the mean, median, mode, standard deviation, range, IQR, standard error and variance of a data set.
- Draw the probability tree of a series of events with multiple discrete outcomes.
- Explain the difference between discrete and continuous variables.
- Draw histograms for grouped data.
- Understand what a probability distribution is, and how to interpret one.
- Calculate probabilities from normal distributions.
- Understand what a cumulative frequency distribution is, and how to interpret one.
- Add and subtract normal distributions.
- Understand what sort of processes would be give rise to a log-normal, Lorentzian, or Poisson distribution.
- Calculate the probabilities of events governed by Poisson processes.
- Fit a curve (given the likely form of the model) to a data set in MATLAB or in Excel.
- Determine the R^2-value to describe how well the curve fits the data.
- Apply Chauvenet's criterion to identify and remove outliers from a given data set.

Standard Normal Distribution

Areas under the Standard Normal Curve from 0 to z



z	0	1	2	3	4	5	6	7	8	9
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998
3.6	.4998	.4998	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.7	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.8	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.9	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000