A Level Statistics

Histograms and Cumulative Frequency

This is really a reminder from GCSE.

Histograms are similar to bar charts apart from the consideration of areas. In a bar chart, all of the bars are the same width and the only thing that matters is the height of the bar. In a histogram, the area is the important thing.

Example

Draw a histogram for the following information.

<table>
<thead>
<tr>
<th>Height (feet)</th>
<th>Frequency</th>
<th>Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 - 2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2 - 4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4 - 5</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>5 - 6</td>
<td>8</td>
<td>16</td>
</tr>
<tr>
<td>6 - 8</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

(Ignore relative frequency for now). It is difficult to draw a bar chart for this information, because the class divisions for the height are not the same.

When drawing a histogram, the y-axis is labelled ‘frequency density’ or "relative frequency". You must work out the relative frequency before you can draw a histogram. To do this, first decide upon a standard width for the groups. Some of the heights are grouped into 2s (0-2, 2-4, 6-8) and some into 1s (4-5, 5-6). Most are 2s, so we shall call
the standard width 2. To make the areas match, we must double the values for frequency which have a class division of 1 (since 1 is half of 2). Therefore the figures in the 4-5 and the 5-6 columns must be doubled. If any of the class divisions were 4 (for example if there was a 8-12 group), these figures would be halved. This is because the area of this "bar" will be twice the standard width of 2 unless we half the frequency.

- Area of bar = frequency x standard width

Cumulative Frequency

The cumulative frequency is the running total of the frequencies. On a graph, it can be represented by a cumulative frequency polygon, where straight lines join up the points, or a cumulative frequency curve.

Example

<table>
<thead>
<tr>
<th>Height (cm)</th>
<th>Frequency</th>
<th>Cumulative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 - 100</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>100 - 120</td>
<td>6</td>
<td>10 (= 4 + 6)</td>
</tr>
<tr>
<td>Class Interval</td>
<td>Frequency</td>
<td>Cumulative Frequency</td>
</tr>
<tr>
<td>----------------</td>
<td>-----------</td>
<td>----------------------</td>
</tr>
<tr>
<td>120 - 140</td>
<td>3</td>
<td>13 (= 4 + 6 + 3)</td>
</tr>
<tr>
<td>140 - 160</td>
<td>2</td>
<td>15 (= 4 + 6 + 3 + 2)</td>
</tr>
<tr>
<td>160 - 180</td>
<td>6</td>
<td>21</td>
</tr>
<tr>
<td>180 - 220</td>
<td>4</td>
<td>25</td>
</tr>
</tbody>
</table>

These data are used to draw a cumulative frequency polygon by plotting the cumulative frequencies against the upper class boundaries.

## Averages

### The Median Value

The median of a group of numbers is the number in the middle, when the numbers are in order of magnitude. For example, if the set of numbers is 4, 1, 6, 2, 6, 7, 8, the median is 6:

1, 2, 4, 6, 6, 7, 8  (6 is the middle value when the numbers are in order)

If you have n numbers in a group, the median is the (n + 1)/2 th value. For example, there are 7 numbers in the example above, so replace n by 7 and the median is the (7 + 1)/2 th value = 4th value. The 4th value is 6.

On a histogram, the median value occurs where the whole histogram is divided into two equal parts. An estimate of the median can be found using algebraic methods. However, an easier method would be to use the data to draw a cumulative frequency polygon and estimate the median using that.

### Mean

There are four types of average: mean, mode, median and range. The mean is what most people mean when they say "average". It is found by adding up all of the numbers you have to find the mean of, and dividing by the number of numbers. So the mean of 3, 5, 7, 3 and 5 is 23/5 = 4.6 .

When you are given data which has been grouped, the mean is $\Sigma fx / \Sigma f$, where f is the frequency and x is the midpoint of the group ($\Sigma$ means "the sum of").
**Example**

Work out an estimate for the mean height.

<table>
<thead>
<tr>
<th>Height (cm)</th>
<th>Number of People (f)</th>
<th>Midpoint (x)</th>
<th>fx (f multiplied by x)</th>
</tr>
</thead>
<tbody>
<tr>
<td>101-120</td>
<td>1</td>
<td>110.5</td>
<td>110.5</td>
</tr>
<tr>
<td>121-130</td>
<td>3</td>
<td>125.5</td>
<td>376.5</td>
</tr>
<tr>
<td>131-140</td>
<td>5</td>
<td>135.5</td>
<td>677.5</td>
</tr>
<tr>
<td>141-150</td>
<td>7</td>
<td>145.5</td>
<td>1018.5</td>
</tr>
<tr>
<td>151-160</td>
<td>4</td>
<td>155.5</td>
<td>622</td>
</tr>
<tr>
<td>161-170</td>
<td>2</td>
<td>165.5</td>
<td>331</td>
</tr>
<tr>
<td>171-190</td>
<td>1</td>
<td>180.5</td>
<td>180.5</td>
</tr>
</tbody>
</table>

\[ \Sigma fx = 3316.5 \]
\[ \Sigma f = 23 \]
\[ \text{mean} = \frac{3316.5}{23} = 144 \text{cm} \ (3 \text{s.f.}) \]

**Mode**

The mode is the number in a set of numbers which occurs the most. So the modal value of 5, 6, 3, 4, 5, 2, 5 and 3 is 5, because there are more 5s than any other number. On a histogram, the modal class is the class with the largest frequency density.

**Range**

The range is the largest number in a set minus the smallest number. So the range of 5, 7, 9 and 14 is \( (14 - 5) = 9 \).
Measures of Dispersion

Measures of dispersion measure how spread out a set of data is.

Variance and Standard Deviation

The formulae for the variance and standard deviation are given below. \( \mu \) means the mean of the data.

Variance: \[ \sigma^2 = \frac{\sum (x_r - \mu)^2}{n} \]

The standard deviation, \( \sigma \), is the square root of the variance.

What the formula means:
1. \( x_r - \mu \) means take each value in turn and subtract the mean from each value.
2. \( (x_r - \mu)^2 \) means square each of the results obtained from step (1). This is to get rid of any minus signs.
3. \( \sum (x_r - \mu)^2 \) means add up all of the results obtained from step (2).
4. Divide step (3) by \( n \), which is the number of numbers
5. For the standard deviation, square root the answer to step (4).

Example

Find the variance and standard deviation of the following numbers: 1, 3, 5, 5, 6, 7, 9, 10.

The mean = \( \frac{46}{8} = 5.75 \)

(Step 1): (1 - 5.75), (3 - 5.75), (5 - 5.75), (5 - 5.75), (6 - 5.75), (7 - 5.75), (9 - 5.75), (10 - 5.75)

= -4.75, -2.75, -0.75, -0.75, 0.25, 1.25, 3.25, 4.25

(Step 2): 22.563, 7.563, 0.563, 0.563, 0.063, 1.563, 10.563, 18.063

(Step 3): 22.563 + 7.563 + 0.563 + 0.563 + 0.063 + 1.563 + 10.563 + 18.063

= 61.504

(Step 4): \( n = 8 \), therefore variance = \( \frac{61.504}{8} = 7.69 \) (3sf)

(Step 5): standard deviation = \( 2.77 \) (3sf)
**Adding or Multiplying Data by a Constant**

If a constant, k, is added to each number in a set of data, the mean will be increased by k and the standard deviation will be unaltered (since the spread of the data will be unchanged).

If the data is multiplied by the constant k, the mean and standard deviation will both be multiplied by k.

**Grouped Data**

There are many ways of writing the formula for the standard deviation. The one above is for a basic list of numbers. The formula for the variance when the data is grouped is as follows. The standard deviation can be found by taking the square root of this value.

\[
\text{Mean} = \frac{\Sigma fx}{\Sigma f}
\]

\[
\text{Variance}, \sigma^2 = \frac{\Sigma fx^2}{\Sigma f} - \left( \frac{\Sigma fx}{\Sigma f} \right)^2
\]

**Example**

The table shows marks (out of 10) obtained by 20 people in a test

<table>
<thead>
<tr>
<th>Mark (x)</th>
<th>Frequency (f)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
</tr>
</tbody>
</table>
Work out the variance of this data. In such questions, it is often easiest to set your working out in a table:

<table>
<thead>
<tr>
<th>fx</th>
<th>fx²</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>12</td>
<td>48</td>
</tr>
<tr>
<td>10</td>
<td>50</td>
</tr>
<tr>
<td>30</td>
<td>180</td>
</tr>
<tr>
<td>35</td>
<td>245</td>
</tr>
<tr>
<td>16</td>
<td>128</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>100</td>
</tr>
</tbody>
</table>

Σf = 20  
Σfx = 118  
Σfx² = 764

\[
\text{variance} = \frac{\Sigma fx^2 - (\Sigma fx)^2}{\Sigma f} = \frac{764 - (118)^2}{20} = \frac{38.2 - 34.81}{20} = 3.39
\]
Quartiles

If we divide a cumulative frequency curve into quarters, the value at the lower quarter is referred to as the lower quartile, the value at the middle gives the median and the value at the upper quarter is the upper quartile.

A set of numbers may be as follows: 8, 14, 15, 16, 17, 18, 19, 50. The mean of these numbers is 19.625. However, the extremes in this set (8 and 50) distort the range. The inter-quartile range is a method of measuring the spread of the numbers by finding the middle 50% of the values. It is useful since it ignore the extreme values. It is a method of measuring the spread of the data.

The lower quartile is \(\frac{(n+1)}{4}\) th value (n is the cumulative frequency, i.e. 157 in this case) and the upper quartile is the \(\frac{3(n+1)}{4}\) th value. The difference between these two is the inter-quartile range (IQR).

In the above example, the upper quartile is the 118.5th value and the lower quartile is the 39.5th value. If we draw a cumulative frequency curve, we see that the lower quartile, therefore, is about 17 and the upper quartile is about 37. Therefore the IQR is 20 (bear in mind that this is a rough sketch- if you plot the values on graph paper you will get a more accurate value).
Box and Whisker Diagrams

Given some data, we can draw a box and whisker diagram (or box plot) to show the spread of the data. The diagram shows the quartiles of the data, using these as an indication of the spread.

The diagram is made up of a "box", which lies between the upper and lower quartiles. The median can also be indicated by dividing the box into two.

The "whiskers" are straight line extending from the ends of the box to the maximum and minimum values.

Outliers

When collecting data, often a result is collected which seems "wrong". In other words, it is much higher or much lower than all of the other values. Such points are known as "outliers". On a box and whisker diagram, outliers should be excluded from the whisker portion of the diagram. Instead, plot them individually, labelling them as outliers.

Skewness

See: Skewness

If the whisker to the right of the box is longer than the one to the left, there is more extreme values towards the positive end and so the distribution is positively skewed.

Similarly, if the whisker to the left is longer, the distribution is negatively skewed.
**Probability**

*See also the probability section in the GCSE section for topics such as probability trees.*

The probability of an event occurring is the chance or likelihood of it occurring. The probability of an event A, written P(A), can be between zero and one, with P(A) = 1 indicating that the event will certainly happen and with P(A) = 0 indicating that event A will certainly not happen.

Probability = \[\frac{\text{the number of successful outcomes of an experiment}}{\text{the number of possible outcomes}}\]

So, for example, if a coin were tossed, the probability of obtaining a head = \(\frac{1}{2}\), since there are 2 possible outcomes (heads or tails) and 1 of these is the ‘successful’ outcome.

**Using Set Notation**

Probability can be studied in conjunction with set theory, with Venn Diagrams being particularly useful in analysis.

The probability of a certain event occurring, for example, can be represented by P(A). The probability of a different event occurring can be written P(B). Clearly, therefore, for two events A and B,

\[
P(A) + P(B) - P(A \cap B) = P(A \cup B)
\]

P(A \cap B) represents the probability of A AND B occurring. P(A \cup B) represents the probability of A OR B occurring.

This can be shown on a Venn diagram. The rectangle represents the sample space, which is all of the possible outcomes of the experiment. The circle labelled as A represents event A. In other words, all of the points within A represent possible ways of achieving the outcome of A. Similarly for B.
So in the diagram, \( P(A) + P(B) \) is the whole of A (the whole circle) + the whole of B (so we have counted the middle bit twice).

A" is the complement of A and means everything not in A. So \( P(A") \) is the probability that A does not occur. Note that the probability that A occurs + the probability that A does not occur = 1 (one or the other must happen). So \( P(A) + P(A") = 1 \). Thus:

- \( P(A") = 1 - P(A) \)

**Mutual Exclusive Events**

Events A and B are mutually exclusive if they have no events in common. In other words, if A occurs B cannot occur and vice-versa. On a Venn Diagram, this would mean that the circles representing events A and B would not overlap.

If, for example, we are asked to pick a card from a pack of 52, the probability that the card is red is \( \frac{1}{2} \). The probability that the card is a club is \( \frac{1}{4} \). However, if the card is red it can"t be a club. These events are therefore mutually exclusive.

If two events are mutually exclusive, \( P(A \cap B) = 0 \), so

- \( P(A) + P(B) = P(A \cup B) \)

**Independent Events**

Two events are independent if the first one does not influence the second. For example, if a bag contains 2 blue balls and 2 red balls and two balls are selected randomly, the events are:
- a) independent if the first ball is replaced after being selected
- b) not independent if the first ball is removed without being replaced. In this instance, there are only three balls remaining in the bag so the probabilities of selecting the various colours have changed.
Two events are independent if (and only if):

- \( P(A \cap B) = P(A)P(B) \)

This is known as the multiplication law.

**Conditional Probability**

*Conditional probability* is the probability of an event occurring, given that another event has occurred. For example, the probability of John doing mathematics at A-Level, given that he is doing physics may be quite high. P(A|B) means the probability of A occurring, given that B has occurred. For two events A and B,

- \( P(A \cap B) = P(A|B)P(B) \)

and similarly \( P(A \cap B) = P(B|A)P(A) \).

If two events are mutually exclusive, then \( P(A|B) = 0 \).

**Independence**

Using the above condition for independence, we deduce that if two events are independent, then:

\[
P(A)P(B) = P(A|B)P(B) = P(B|A)P(A), \text{ or:}
\]

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

**Example**

A six-sided die is thrown. What is the probability that the number thrown is prime, given that it is odd.

The probability of obtaining an odd number is \( 3/6 = \frac{1}{2} \). Of these odd numbers, 2 of them are prime (3 and 5).

\[
P(\text{prime | odd}) = \frac{P(\text{prime and odd})}{P(\text{odd})} = \frac{2/6}{3/6} = \frac{2}{3}
\]
Skewness

A normal distribution is a bell-shaped distribution of data where the mean, median and mode all coincide. A frequency curve showing a normal distribution would look like this:

In a normal distribution, approximately 68% of the values lie within one standard deviation of the mean and approximately 95% of the data lies within two standard deviations of the mean.

If there are extreme values towards the positive end of a distribution, the distribution is said to be positively skewed. In a positively skewed distribution, the mean is greater than the mode. For example:
A negatively skewed distribution, on the other hand, has a mean which is less than the mode because of the presence of extreme values at the negative end of the distribution.

There are a number of ways of measuring skewness:

Pearson’s coefficient of skewness = \[
\frac{\text{mean} - \text{mode}}{\text{Standard deviation}} = \frac{3(\text{mean} - \text{median})}{\text{Standard deviation}}
\]

Quartile measure of skewness = \[
\frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1}
\]

**Linear Regression**

**Scatter Diagrams**

We often wish to look at the relationship between two things (e.g. between a person's height and weight) by comparing data for each of these things. A good way of doing this is by drawing a scatter diagram.

"Regression" is the process of finding the function satisfied by the points on the scatter diagram. Of course, the points might not fit the function exactly but the aim is to get as close as possible. "Linear" means that the function we are looking for is a straight line (so our function \( f \) will be of the form \( f(x) = mx + c \) for constants \( m \) and \( c \)).

Here is a scatter diagram with a regression line drawn in:
Correlation

Correlation is a term used to describe how strong the relationship between the two variables appears to be.

We say that there is a positive linear correlation if y increases as x increases and we say there is a negative linear correlation if y decreases as x increases. There is no correlation if x and y do not appear to be related.

Explanatory and Response Variables

In many experiments, one of the variables is fixed or controlled and the point of the experiment is to determine how the other variable varies with the first. The fixed/controlled variable is known as the explanatory or independent variable and the other variable is known as the response or dependent variable.

I shall use "x" for my explanatory variable and "y" for my response variable, but I could have used any letters.
Regression Lines

By Eye

If there is very little scatter (we say there is a strong correlation between the variables), a regression line can be drawn "by eye". You should make sure that your line passes through the mean point (the point (x,y) where x is mean of the data collected for the explanatory variable and y is the mean of the data collected for the response variable).

Two Regression Lines

When there is a reasonable amount of scatter, we can draw two different regression lines depending upon which variable we consider to be the most accurate. The first is a line of regression of y on x, which can be used to estimate y given x. The other is a line of regression of x on y, used to estimate x given y.

If there is a perfect correlation between the data (in other words, if all the points lie on a straight line), then the two regression lines will be the same.

Least Squares Regression Lines

This is a method of finding a regression line without estimating where the line should go by eye.

If the equation of the regression line is \( y = ax + b \), we need to find what a and b are. We find these by solving the "normal equations".

Normal Equations

The "normal equations" for the line of regression of y on x are:

\[
\Sigma y = na + b \Sigma x
\]

\[
\Sigma xy = a \Sigma x + b \Sigma x^2
\]

The values of a and b are found by solving these equations simultaneously.

For the line of regression of x on y, the "normal equations" are the same but with x and y swapped.
The Product Moment Correlation Coefficient

The product moment correlation coefficient is a measurement of the degree of scatter. It is usually denoted by $r$ and $r$ can be any value between $-1$ and $1$. It is defined as follows:

$$r = \frac{s_{xy}}{s_x s_y}$$

where $s_{xy}$ is the covariance of $x$ and $y$, $s_x = \frac{1}{n} \sum (x - \bar{x})$.

**Correlation**

The product moment correlation coefficient (pmcc) can be used to tell us how strong the correlation between two variables is.

A positive value indicates a positive correlation and the higher the value, the stronger the correlation. Similarly, a negative value indicates a negative correlation and the lower the value the stronger the correlation.

If there is a perfect positive correlation (in other words the points all lie on a straight line that goes up from left to right), then $r = 1$.

If there is a perfect negative correlation, then $r = -1$.

If there is no correlation, then $r = 0$. $r$ would also be equal to zero if the variables were related in a non-linear way (they might lie on a quadratic curve rather than a straight line, for example).
Discrete Random Variables

A probability distribution is a table of values showing the probabilities of various outcomes of an experiment.

For example, if a coin is tossed three times, the number of heads obtained can be 0, 1, 2 or 3. The probabilities of each of these possibilities can be tabulated as shown:

<table>
<thead>
<tr>
<th>Number of Heads</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>1/8</td>
<td>3/8</td>
<td>3/8</td>
<td>1/8</td>
</tr>
</tbody>
</table>

A discrete variable is a variable which can only take a countable number of values. In this example, the number of heads can only take 4 values (0, 1, 2, 3) and so the variable is discrete. The variable is said to be random if the sum of the probabilities is one.

Probability Density Function

The probability density function (p.d.f.) of X (or probability mass function) is a function which allocates probabilities. Put simply, it is a function which tells you the probability of certain events occurring. The usual notation that is used is P(X = x) = something. The random variable (r.v.) X is the event that we are considering. So in the above example, X represents the number of heads that we throw. So P(X = 0) means "the probability that no heads are thrown". Here, P(X = 0) = 1/8 (the probability that we throw no heads is 1/8).

In the above example, we could therefore have written:

<table>
<thead>
<tr>
<th>x</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(X = x)</td>
<td>1/8</td>
<td>3/8</td>
<td>3/8</td>
<td>1/8</td>
</tr>
</tbody>
</table>

Quite often, the probability density function will be given to you in terms of x. In the above example, P(X = x) = \( \frac{3C_3}{2^3} \) (see permutations and combinations for the meaning of \( 3C_3 \)).

Example

A die is thrown repeatedly until a 6 is obtained. Find the probability density function for the number times we throw the die.
Let $X$ be the random variable representing the number of times we throw the die.

$P(X = 1) = \frac{1}{6}$ (if we only throw the die once, we get a 6 on our first throw. The probability of this is $\frac{1}{6}$).

$P(X = 2) = (\frac{5}{6}) \times (\frac{1}{6})$ (if we throw the die twice before getting a 6, we must throw something that isn't a 6 with our first throw, the probability of which is $\frac{5}{6}$ and we must throw a 6 on our second throw, the probability of which is $\frac{1}{6}$)

etc

In general, $P(X = x) = (\frac{5}{6})^{(x-1)} \times (\frac{1}{6})$

**Cumulative Distribution Function**

The cumulative distribution function (c.d.f.) of a discrete random variable $X$ is the function $F(t)$ which tells you the probability that $X$ is less than or equal to $t$. So if $X$ has p.d.f. $P(X = x)$, we have:

$F(t) = P(X \leq t) = \sum P(X = x)$.

In other words, for each value that $X$ can be which is less than or equal to $t$, work out the probability that $X$ is that value and add up all such results.

**Example**

In the above example where the die is thrown repeatedly, let's work out $P(X \leq t)$ for some values of $t$.

$P(X \leq 1)$ is the probability that the number of throws until we get a 6 is less than or equal to 1. So it is either 0 or 1.

$P(X = 0) = 0$ and $P(X = 1) = \frac{1}{6}$. Hence $P(X \leq 1) = \frac{1}{6}$

Similarly, $P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) = 0 + \frac{1}{6} + \frac{5}{36} = \frac{11}{36}$
Expectation

The **expected value** (or mean) of X, where X is a discrete random variable, is a weighted average of the possible values that X can take, each value being weighted according to the probability of that event occurring. The expected value of X is usually written as E(X) or m.

- \( E(X) = \sum x \times P(X = x) \)

So the expected value is the sum of: [(each of the possible outcomes) × (the probability of the outcome occurring)].

In more concrete terms, the expectation is what you would expect the outcome of an experiment to be on average.

**Example**

What is the expected value when we roll a fair die?

There are six possible outcomes: 1, 2, 3, 4, 5, 6. Each of these has a probability of 1/6 of occurring. Let X represent the outcome of the experiment.

Therefore \( P(X = 1) = \frac{1}{6} \) (this means that the probability that the outcome of the experiment is 1 is \( \frac{1}{6} \))
\( P(X = 2) = \frac{1}{6} \) (the probability that you throw a 2 is \( \frac{1}{6} \))
\( P(X = 3) = \frac{1}{6} \) (the probability that you throw a 3 is \( \frac{1}{6} \))
\( P(X = 4) = \frac{1}{6} \) (the probability that you throw a 4 is \( \frac{1}{6} \))
\( P(X = 5) = \frac{1}{6} \) (the probability that you throw a 5 is \( \frac{1}{6} \))
\( P(X = 6) = \frac{1}{6} \) (the probability that you throw a 6 is \( \frac{1}{6} \))

\( E(X) = 1 \times P(X = 1) + 2 \times P(X = 2) + 3 \times P(X = 3) + 4 \times P(X = 4) + 5 \times P(X = 5) + 6 \times P(X = 6) \)

Therefore \( E(X) = \frac{1}{6} + \frac{2}{6} + \frac{3}{6} + \frac{4}{6} + \frac{5}{6} + \frac{6}{6} = \frac{7}{2} \)

So the expectation is 3.5. If you think about it, 3.5 is halfway between the possible values the die can take and so this is what you should have expected.

**Expected Value of a Function of X**

To find \( E[f(X)] \), where \( f(X) \) is a function of X, use the following formula:
•  \( E[ f(X) ] = \sum f(x)P(X = x) \)

**Example**

For the above experiment (with the die), calculate \( E(X^2) \)

Using our notation above, \( f(x) = x^2 \)

\[ f(1) = 1, \; f(2) = 4, \; f(3) = 9, \; f(4) = 16, \; f(5) = 25, \; f(6) = 36 \]

\[ P(X = 1) = \frac{1}{6}, \; P(X = 2) = \frac{1}{6}, \; \text{etc} \]

So \( E(X^2) = \frac{1}{6} + \frac{4}{6} + \frac{9}{6} + \frac{16}{6} + \frac{25}{6} + \frac{36}{6} = \frac{91}{6} = 15.167 \)

The expected value of a constant is just the constant, so for example \( E(1) = 1 \).

Multiplying a random variable by a constant multiplies the expected value by that constant, so \( E[2X] = 2E[X] \).

A useful formula, where \( a \) and \( b \) are constants, is:

•  \( E[aX + b] = aE[X] + b \)

[This says that expectation is a linear operator].

**Variance**

The variance of a random variable tells us something about the spread of the possible values of the variable. For a discrete random variable \( X \), the variance of \( X \) is written as \( \text{Var}(X) \).

•  \( \text{Var}(X) = E[ (X - m)^2 ] \) \( m \) is the expected value \( E(X) \)

This can also be written as:

•  \( \text{Var}(X) = E(X^2) - m^2 \)

The *standard deviation* of \( X \) is the square root of \( \text{Var}(X) \).

Note that the variance does not behave in the same way as expectation when we multiply and add constants to random variables. In fact:

•  \( \text{Var}[aX + b] = a^2\text{Var}(X) \)
You is because: \( \text{Var}[aX + b] = E[(aX + b)^2] - (E[aX + b])^2 \).

\[
= E[a^2X^2 + 2abX + b^2] - (aE(X) + b)^2 \\
= a^2E(X^2) + 2abE(X) + b^2 - a^2E^2(X) - 2abE(X) - b^2 \\
= a^2E(X^2) - a^2E^2(X) = a^2 \text{Var}(X)
\]

**The Discrete Uniform Distribution**

There are a number of important types of discrete random variables. The simplest is the uniform distribution.

A random variable with p.d.f. given by:

\[
P(X = x) = \frac{1}{k+1} \text{ for all values of } x = 0, \ldots, k \\
P(X = x) = 0 \text{ for other values of } x
\]

where \(k\) is a constant, is said to be follow a uniform distribution.

**Example**

Suppose we throw a die. Let \(X\) be the random variable denoting what number is thrown.

\[
P(X = 1) = \frac{1}{6} \\
P(X = 2) = \frac{1}{6} \text{ etc}
\]

In fact, \(P(X = x) = \frac{1}{6}\) for all \(x\) between 1 and 6. Hence we have a uniform distribution.

**Expectation and Variance**

We can find the expectation and variance of the discrete uniform distribution:

Suppose \(P(X = x) = \frac{1}{k+1}\) for all values of \(x = 0, \ldots, k\).

Then \(E(X) = 1.P(X = 1) + 2.P(X = 2) + \ldots + k.P(X = k)\)

\[
= \frac{1}{k+1} + 2\frac{1}{k+1} + 3\frac{1}{k+1} + \ldots + k\frac{1}{k+1} \\
= \left(\frac{1}{k+1}\right)(1 + 2 + \ldots + k) \\
= \left(\frac{1}{k+1}\right) \times \frac{1}{2}k \left[2 + (k - 1)\right] \quad \text{(summing the arithmetic progression)} \\
= \frac{1}{2}k
\]
It turns out that the variance is:
\[
\frac{k(k+2)}{12}
\]

---

**The Normal Distribution**

A continuous random variable \(X\) follows a **normal distribution** if it has the following probability density function (p.d.f.):

\[
f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x - \mu)^2}{2\sigma^2}}
\]

The parameters of the distribution are \(\mu\) and \(\sigma^2\), where \(\mu\) is the mean (expectation) of the distribution and \(\sigma^2\) is the variance. We write \(X \sim N(\mu, \sigma^2)\) to mean that the random variable \(X\) has a normal distribution with parameters \(\mu\) and \(\sigma^2\).

The normal distribution is symmetrical about its mean:

---

**The Standard Normal Distribution**

If \(Z \sim N(0, 1)\), then \(Z\) is said to follow a **standard normal** distribution.

\(P(Z < z)\) is known as the cumulative distribution function of the random variable \(Z\). For the standard normal distribution, this is usually denoted by \(\Phi(z)\).
Normally, you would work out the c.d.f. by doing some integration. However, it is impossible to do this for the normal distribution and so results have to be looked up in statistical tables.

**Example**

Find $P(-1.96 < Z < 1.96)$, where $Z \sim N(0, 1)$.

This is equal to $P(Z < 1.96) - P(Z < -1.96)$

$= \Phi(1.96) - \Phi(-1.96) = \Phi(1.96) - (1 - \Phi(1.96)) = 2 \Phi(1.96) - 1$

Now we need to look in a table to find out what $\Phi(1.96)$ is.

If you look in statistical tables for the standard normal distribution, you should be able to find a line:

<table>
<thead>
<tr>
<th>$z$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.9</td>
<td>.9713</td>
<td>.9719</td>
<td>.9726</td>
<td>.9732</td>
<td>.9738</td>
<td>.9744</td>
<td>.9750</td>
<td>.9756</td>
<td>.9761</td>
<td>.9767</td>
</tr>
</tbody>
</table>

To find $\Phi(1.96)$, read across the 1.9 line until you get to 6 (for 1.96). So $\Phi(1.96) = 0.975$.

Hence $P(-1.96 < Z < 1.96) = 2 \times 0.975 - 1 = 0.95$.

This result says that the central 95% of the distribution lies between -1.96 and 1.96.

**Some Identities**

- $\Phi(-z) = P(Z < -z) = 1 - \Phi(z)$
- $P(Z > z) = 1 - \Phi(z)$
- $P(a < Z < b) = \Phi(b) - \Phi(a)$

**Standardising**

Now, the mean and variance of the normal distribution can be any value and so clearly there can't be a statistical table for each one. Instead, we convert to the standard normal distribution - we can also use statistical tables for the standard normal distribution to find the c.d.f. of any normal distribution.

We use the following trick:
If \( X \sim N(\mu, \sigma^2) \), then put:

\[
Z = \frac{X - \mu}{\sigma}
\]

It turns out that \( Z \sim N(0, 1) \). Note that it is \( \sigma \) and not \( \sigma^2 \) on the denominator!

**Example**

If \( X \sim N(4, 9) \), find \( P(X < 12) \).

First we "standardise". The mean is 4 and \( \sigma \) is 3 (the square root of 9).

\[
P(X < 12) = P\left[\frac{X - 4}{3} < \frac{12 - 4}{3}\right]
\]

(subtracting 4 and dividing each side of the inequality by 3)

\[
= P(Z < 8/3)
\]

\[
= \Phi(8/3) = \Phi(2.67) = 0.9962
\]

---

**The Binomial Distribution**

If a discrete random variable \( X \) has the following probability density function (p.d.f.), it is said to have a **binomial distribution**:

\[
P(X = x) = \binom{n}{x} p^x q^{(n-x)}, \text{ where } q = 1 - p
\]

\( p \) can be considered as the probability of a success, and \( q \) the probability of a failure.

Note: \( \binom{n}{r} \) ("n choose r") is more commonly written \( \binom{n}{r} \), but I shall use the former because it is easier to write on a computer. It means the number of ways of choosing \( r \) objects from a collection of \( n \) objects (see permutations and combinations).

If a random variable \( X \) has a binomial distribution, we write \( X \sim B(n, p) \) (~ means ‘has distribution…’).
n and p are known as the **parameters** of the distribution (n can be any integer greater than 0 and p can be any number between 0 and 1). All random variables with a binomial distribution have the above p.d.f., but may have different parameters (different values for n and p).

**Example**

A coin is thrown 10 times. Find the probability density function for X, where X is the random variable representing the number of heads obtained.

The probability of throwing a head is \( \frac{1}{2} \) and the probability of throwing a tail is \( \frac{1}{2} \). Therefore, the probability of throwing 8 tails is \( \left( \frac{1}{2} \right)^8 \)

If we throw 2 heads and 8 tails, we could have thrown them HTTTTTHTTT, or TTHTHTTTTT, or in a number of other ways. In fact, the total number of ways of throwing 2 heads and 8 tails is \( \binom{10}{2} \) (see the permutations and combinations section).

Hence the probability of throwing 2 heads and 8 tails is \( \binom{10}{2} \times \left( \frac{1}{2} \right)^2 \times \left( \frac{1}{2} \right)^8 \). As you can see this has a Binomial distribution, where \( n = 10, \ p = \frac{1}{2} \).

You can see, therefore, that the p.d.f. is going to be: \( P(X = x) = \binom{10}{x} \left( \frac{1}{2} \right)^{10-x} \left( \frac{1}{2} \right)^x \).

From this, we can work out the probability of throwing, for example, 3 heads (put \( x = 3 \)).

**Expectation and Variance**

If \( X \sim B(n,p) \), then the expectation and variance is given by:

- \( E(X) = np \)
- \( \text{Var}(X) = npq \)

**Example**

In the above example, what is the expected number of heads thrown?

\( E(X) = np \)

Now in the above example, \( p = \) probability of throwing a head = \( \frac{1}{2} \) . \( n = \) number of throws = 10

Hence expected number of heads = 5.
This is what you would expect: if you throw a coin 10 times you would expect 5 heads and 5 tails on average.

### The Poisson Distribution

A discrete random variable $X$ with a probability distribution function (p.d.f.) of the form:

$$P(X = x) = \frac{(e^{-\lambda}) \lambda^x}{x!}$$

is said to be a Poisson random variable with parameter $\lambda$. We write $X \sim \text{Po}(\lambda)$

#### Expectation and Variance

If $X \sim \text{Po}(\lambda)$, then:

- $\mathbb{E}(X) = \lambda$
- $\text{Var}(X) = \lambda$

#### Sums of Poissons

Suppose $X$ and $Y$ are independent Poisson random variables with parameters $\lambda$ and $\mu$ respectively. Then $X + Y$ has a Poisson distribution with parameter $\lambda + \mu$. In other words:

- If $X \sim \text{Po}(\lambda)$ and $Y \sim \text{Po}(\mu)$, then $X + Y \sim \text{Po}(\lambda + \mu)$

#### Random Events

The Poisson distribution is useful because many random events follow it.

If a random event has a mean number of occurrences $\lambda$ in a given time period, then the number of occurrences within that time period will follow a Poisson distribution.

For example, the occurrence of earthquakes could be considered to be a random event. If there are 5 major earthquakes each year, then the number of earthquakes in any given year will have a Poisson distribution with parameter 5.
Example

There are 50 misprints in a book which has 250 pages. Find the probability that page 100 has no misprints.

The average number of misprints on a page is 50/250 = 0.2. Therefore, if we let X be the random variable denoting the number of misprints on a page, X will follow a Poisson distribution with parameter 0.2.

Since the average number of misprints on a page is 0.2, the parameter, \( \lambda \) of the distribution is equal to 0.2.

\[
P(X = 0) = \frac{(e^{-0.2})(0.2^0)}{0!} = 0.819 \quad (3sf)
\]

**Binomial Approximation**

The Poisson distribution can be used as an approximation to the binomial distribution.

- A Binomial distribution with parameters n and p can be approximated by a Poisson distribution with parameter np.

**Continuous Random Variables**

A **continuous random variable** is a random variable where the data can take infinitely many values. For example, a random variable measuring the time taken for something to be done is continuous since there are an infinite number of possible times that can be taken.

For any continuous random variable with probability density function \( f(x) \), we have that:

\[
\int_{\text{all } x} f(x) \, dx = 1
\]
This is a useful fact.

**Example**

X is a continuous random variable with probability density function given by \( f(x) = cx \) for \( 0 \leq x \leq 1 \), where \( c \) is a constant. Find \( c \).

If we integrate \( f(x) \) between 0 and 1 we get \( c/2 \). Hence \( c/2 = 1 \) (from the useful fact above!), giving \( c = 2 \).

**Cumulative Distribution Function (c.d.f.)**

If \( X \) is a continuous random variable with p.d.f. \( f(x) \) defined on \( a \leq x \leq b \), then the **cumulative distribution function** (c.d.f.), written \( F(t) \) is given by:

\[
F(t) = P(X \leq t) = \int_a^t f(x) \, dx
\]

So the c.d.f. is found by integrating the p.d.f. between the minimum value of \( X \) and \( t \).

Similarly, the probability density function of a continuous random variable can be obtained by differentiating the cumulative distribution.

The c.d.f. can be used to find out the probability of a random variable being between two values:

\( P(s \leq X \leq t) = \) the probability that \( X \) is between \( s \) and \( t \). But this is equal to the probability that \( X \leq t \) minus the probability that \( X \leq s \).

[We want the probability that \( X \) is in the red area:]
Hence:

- \( P(s \leq X \leq t) = P(X \leq t) - P(X \leq s) = F(t) - F(s) \)

**Expectation and Variance**

With discrete random variables, we had that the expectation was \( \Sigma x \ P(X = x) \), where \( P(X = x) \) was the p.d.f.. It may come as no surprise that to find the expectation of a continuous random variable, we integrate rather than sum, i.e.:

\[
E(X) = \int \ x \ f(x) \ dx
\]

As with discrete random variables, \( \text{Var}(X) = E(X^2) - [E(X)]^2 \)
The Uniform Distribution

A continuous random variable X which has probability density function given by:

\[ f(x) = \frac{1}{b - a} \quad \text{for} \quad a \leq x \leq b \]

(and \( f(x) = 0 \) if \( x \) is not between \( a \) and \( b \)) follows a uniform distribution with parameters \( a \) and \( b \). We write \( X \sim U(a,b) \)

Remember that the area under the graph of the random variable must be equal to 1 (see continuous random variables).

Expectation and Variance

If \( X \sim U(a,b) \), then:

- \( E(X) = \frac{1}{2} (a + b) \)
- \( \text{Var}(X) = \frac{1}{12} (b - a)^2 \)

Proof of Expectation

\[
\int_{a}^{b} \frac{x}{b - a} \, dx = \frac{1}{b - a} \left[ \frac{x^2}{2} \right]_{a}^{b} = \frac{b^2 - a^2}{2(b - a)} = \frac{b + a}{2}
\]

Cumulative Distribution Function

The cumulative distribution function can be found by integrating the p.d.f between 0 and \( t \):
Normal Approximations

Binomial Approximation

The normal distribution can be used as an approximation to the binomial distribution, under certain circumstances, namely:

- If $X \sim B(n, p)$ and if $n$ is large and/or $p$ is close to $\frac{1}{2}$, then $X$ is approximately $N(np, npq)$

(where $q = 1 - p$).

In some cases, working out a problem using the Normal distribution may be easier than using a Binomial.

Poisson Approximation

The normal distribution can also be used to approximate the Poisson distribution for large values of $\lambda$ (the mean of the Poisson distribution).

- If $X \sim Po(\lambda)$ then for large values of $\lambda$, $X \sim N(\lambda, \lambda)$ approximately.

Continuity Correction

The binomial and Poisson distributions are discrete random variables, whereas the normal distribution is continuous. We need to take this into account when we are using the normal distribution to approximate a binomial or Poisson using a continuity correction.
In the discrete distribution, each probability is represented by a rectangle (right hand diagram):

![Image]

When working out probabilities, we want to include whole rectangles, which is what continuity correction is all about.

**Example**

Suppose we toss a fair coin 20 times. What is the probability of getting between 9 and 11 heads?

Let X be the random variable representing the number of heads thrown.

\[ X \sim \text{Bin}(20, \frac{1}{2}) \]

Since \( p \) is close to \( \frac{1}{2} \) (it equals \( \frac{1}{2} \)), we can use the normal approximation to the binomial.

\[ X \sim N(20 \times \frac{1}{2}, 20 \times \frac{1}{2} \times \frac{1}{2}) \text{ so } X \sim N(10, 5). \]

In this diagram, the rectangles represent the binomial distribution and the curve is the normal distribution:

![Image]

We want \( P(9 \leq X \leq 11) \), which is the red shaded area. Notice that the first rectangle starts at 8.5 and the last rectangle ends at 11.5. Using a continuity correction, therefore, our probability becomes \( P(8.5 < X < 11.5) \) in the normal distribution.
Standardising:

\[
\Pr\left(\frac{9 - 10}{\sqrt{5}} < \frac{X - 10}{\sqrt{5}} < \frac{11 - 10}{\sqrt{5}}\right) = \Pr(-0.447 < Z < 0.447) = 2 \times 0.67 - 1 \text{ (using tables)} = 0.34
\]

**Sampling 1**

A **population** is a group that we want to find information about. It might be a group of people or it could be simply a group of numbers.

A **census** is when information about *every* member of the population is collected. The disadvantage of this method is that if the population is large, it can be difficult to collect and process so much information.

A **sample survey** is when information is collected from a small representation of the population. For example, if we were trying to find out what Britain's favourite TV show is, it would be impractical to ask everybody in the country (as a census would do). Instead, 1000 people might be chosen and asked.

A **sampling unit** is a person/object to be sampled. The sampling units must be defined in such a way that any one member of a population is not sampled more than once. For example, if we were trying to find out how many cars the average family owns, the sampling unit would be a household rather than an individual, because we wouldn't want to ask two people from the same household.

A **sampling frame** is the collection of all of the sampling units. Ideally, this should cover the whole population.
Hypothesis Testing

The parameters of a distribution are those quantities that you need to specify when describing the distribution. For example, a normal distribution has parameters \( \mu \) and \( \sigma^2 \) and a Poisson distribution has parameter \( \lambda \).

If we know that some data comes from a certain distribution, but the parameter is unknown, we might try to predict what the parameter is. Hypothesis testing is about working out how likely our predictions are.

Hypotheses

The null hypothesis, denoted by \( H_0 \), is a prediction about a parameter (so if we are dealing with a normal distribution, we might predict the mean or the variance of the distribution).

We also have an alternative hypothesis, denoted by \( H_1 \). We then perform a test to decide whether or not we should reject the null hypothesis in favour of the alternative.

Testing

Suppose we are given a value and told that it comes from a certain distribution, but we don"t know what the parameter of that distribution is.

Suppose we make a null hypothesis about the parameter. We test how likely it is that the value we were given could have come from the distribution with this predicted parameter.

For example, suppose we are told that the value of 3 has come from a Poisson distribution. We might want to test the null hypothesis that the parameter (which is the mean) of the Poisson distribution is 9. So we work out how likely it is that the value of 3 could have come from a Poisson distribution with parameter 9. If it"s not very likely, we reject the null hypothesis in favour of the alternative.
Critical Region

But what exactly is "not very likely"?

We choose a region known as the critical region. If the result of our test lies in this region, then we reject the null hypothesis in favour of the alternative.

One and Two Tailed Tests

Suppose we have a null hypothesis $H_0$ and an alternative hypothesis $H_1$. We consider the distribution given by the null hypothesis and perform a test to determine whether or not the null hypothesis should be rejected in favour of the alternative hypothesis.

There are two different types of tests that can be performed. A one-tailed test looks for an increase or decrease in the parameter whereas a two-tailed test looks for any change in the parameter (which can be any change - increase or decrease).

We can perform the test at any level (usually 1%, 5% or 10%). For example, performing the test at a 5% level means that there is a 5% chance of wrongly rejecting $H_0$.

If we perform the test at the 5% level and decide to reject the null hypothesis, we say "there is significant evidence at the 5% level to suggest the hypothesis is false".

One-Tailed Test

We choose a critical region. In a one-tailed test, the critical region will have just one part (the red area below). If our sample value lies in this region, we reject the null hypothesis in favour of the alternative.

Suppose we are looking for a definite decrease. Then the critical region will be to the left. Note, however, that in the one-tailed test the value of the parameter can be as high as you like.
Example

Suppose we are given that $X$ has a Poisson distribution and we want to carry out a hypothesis test on the mean, $\lambda$, based upon a sample observation of 3.

Suppose the hypotheses are:
- $H_0$: $\lambda = 9$
- $H_1$: $\lambda < 9$

We want to test if it is "reasonable" for the observed value of 3 to have come from a Poisson distribution with parameter 9. So what is the probability that a value as low as 3 has come from a Poisson distribution with parameter 9?

\[ P(X \leq 3) = 0.0212 \text{ (this has come from a Poisson table)} \]

The probability is less than 0.05, so there is less than a 5% chance that the value has come from a Poisson distribution with parameter 9. We therefore reject the null hypothesis in favour of the alternative at the 5% level.

However, the probability is greater than 0.01, so we would not reject the null hypothesis in favour of the alternative at the 1% level.

Two-Tailed Test

In a two-tailed test, we are looking for either an increase or a decrease. So, for example, $H_0$ might be that the mean is equal to 9 (as before). This time, however, $H_1$ would be that the mean is not equal to 9. In this case, therefore, the critical region has two parts:
Example

Let's test the parameter $p$ of a Binomial distribution at the 10% level.

Suppose a coin is tossed 10 times and we get 7 heads. We want to test whether or not the coin is fair. If the coin is fair, $p = 0.5$. Put this as the null hypothesis:

$H_0: p = 0.5$
$H_1: p \neq 0.5$

Now, because the test is 2-tailed, the critical region has two parts. Half of the critical region is to the right and half is to the left. So the critical region contains both the top 5% of the distribution and the bottom 5% of the distribution (since we are testing at the 10% level).

If $H_0$ is true, $X \sim \text{Bin}(10, 0.5)$.

If the null hypothesis is true, what is the probability that $X$ is 7 or above?
$P(X \geq 7) = 1 - P(X < 7) = 1 - P(X \leq 6) = 1 - 0.8281 = 0.1719$

Is this in the critical region? No- because the probability that $X$ is at least 7 is not less than 0.05 (5%), which is what we need it to be.

So there is not significant evidence at the 10% level to reject the null hypothesis.
Sampling 2

Random Sampling

The idea of random sampling is that each member of the sample frame has an equal chance of being selected.

One way of doing this is to assign each member of the sample frame a number. Random numbers are then generated (using a computer or from a table) and those members of the sample frame whose numbers come out are sampled.

Systematic Sampling

Instead of choosing the members to be sampled using random numbers (which might be difficult and time consuming for large populations), systematic sampling uses a simple rule to choose people. For example, every 10th member of the sample frame could be selected.

Stratified Sampling

Stratified sampling can be used when the population in question is split up into groups who are likely to behave differently. For example, if we were trying to find the nation’s favourite television programme, most children would probably like different programmes to most adults.

Each group is sampled separately and the results are put together.

In the television example, if children make up 20% of the population, we would make sure that children make up 20% of the total sample.

Quota Sampling

Quota sampling involves splitting the population into groups and sampling a given number of people from each group.

This method is easy to implement when carrying out market research. For example, if someone is interviewing people at a shopping centre, they may have been told to interview 50 men and 50 women. It doesn’t matter how they choose the 50, as long as they interview that many.
If there is no sampling frame (list of sampling units), the above sampling methods can't really be implemented. Quota sampling might be the only real possibility.

Random Samples

A random sample is a collection of independent random variables X₁, X₂, ..., Xₙ, all with the same probability distribution.

The Sample Mean

The sample mean, \( \bar{X} \) of a random sample X₁, ... Xₙ is given by:

\[
\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \quad \text{for } i = 1, 2, ..., n
\]

The sample mean has a normal distribution with mean \( \mu \) and variance \( \sigma^2/n \), in other words:

\[
\bar{X} \sim N(\mu, \frac{\sigma^2}{n})
\]

Expectation and Variance

\( \bar{X} \) is itself a random variable and so has an expectation and variance.

These are easy to calculate, for example:

\[
E(\bar{X}) = E\left(\frac{\sum X_i}{n}\right) = \frac{1}{n}E(X₁ + X₂ + ... + Xₙ) = \frac{1}{n}[E(X₁) + E(X₂) + ... + E(Xₙ)]
\]

If each of the random variables X₁, ... Xₙ have expectation \( \mu \), then this is equal to

(1/n)(n \mu) = \mu.

• \( E(\bar{X}) = \mu \)

Similarly, if each of the random variables X₁, ... Xₙ has variance \( \sigma^2 \), it can be shown that:
- $\text{Var}(\overline{X}) = (\sigma^2)/n$

**The Sample Variance**

The sample variance is given by:

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

---

**Estimators**

Probability distributions depend upon parameters. For example, the normal distribution depends upon the parameters $\mu$ and $\sigma^2$ (the mean and variance). In some situations, these parameters may be unknown and we may wish to estimate them. An **estimator** is a statistic which is used to estimate a parameter.

**Desirable Characteristics**

Good estimators are those which have a small variance and small bias. The **bias** of an estimator $\theta$ which is estimating a parameter $p$ is $E(\theta) - p$.

An estimator is **unbiased** if the bias is zero. The sample mean and sample variance are unbiased estimators of the mean and variance.

So the best estimator for a particular parameter is that for which $B(\theta) + V(\theta)$ is smallest, where $B(\theta)$ is the bias of $\theta$.

**Example**

$X_1, X_2, ..., X_n$ is a random sample taken from a normal distribution with mean $\mu$ and variance $\sigma^2$, where $\mu$ is unknown. Show that the sample mean is an unbiased estimator for $\mu$.

We calculated that the expectation of the sample mean is $\mu$. Hence $E(\overline{X}) - \mu = 0$. So the sample mean is unbiased.
NB: \( \text{Var}(\bar{X}) = \sigma^2/n \), so as \( n \) gets large, \( \text{Var}(\bar{X}) \) gets closer to zero and so the sample mean, \( \bar{X} \), has a small variance for large values of \( n \). It turns out that \( \bar{X} \) is the best estimator for \( \mu \).

### The Central Limit Theorem

#### Linear Combination of Normals

Suppose that \( X \) and \( Y \) are independent normal random variables.

- Let \( X \sim N(\mu_1, \sigma_1^2) \) and \( Y \sim N(\mu_2, \sigma_2^2) \), then \( X + Y \) is a normal random variable with mean \( \mu_1 + \mu_2 \) and variance \( \sigma_1^2 + \sigma_2^2 \).

We can go a bit further: if \( a \) and \( b \) are constants then:

- \( aX + bY \sim N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2) \)

#### The Central Limit Theorem

The following is an important result known as the central limit theorem:

If \( X_1, \ldots, X_n \) is are independent random variables random sample from any distribution which has mean \( \mu \) and variance \( \sigma^2 \), then the distribution of \( X_1 + X_2 + \ldots + X_n \) is approximately normal with mean \( n\mu \) and variance \( n\sigma^2 \).

In particular, the distribution of the sample mean, which is \( (X_1 + X_2 + \ldots + X_n)/n \), is approximately normal with mean \( \mu \) and variance \( n\sigma^2/\sigma^2 \) (since we have multiplied \( X_1 + X_2 + \ldots + X_n \) by \( 1/n \) and multiplying by a constant multiplies the mean by that constant but does not change the variance). This important result will be used in constructing confidence intervals.
Confidence Intervals

An interval is a set of (real) numbers between two values. So, for example, (0, 1) is all of the real numbers \( x \) such that \( 0 < x < 1 \).

If we have an unknown parameter, we may find an estimator for this parameter and use for the parameter. However, how reliable this estimate is we do not know. This is where confidence intervals come in. Instead of estimating the parameter, we say that there is a 95% (or some other percentage) chance that a given interval contains the parameter.

For example, suppose we have an unknown parameter \( \theta \). If the probability that the interval \( (a,b) \) contains \( \theta \) is 0.95, i.e. if \( P(a < \theta < b) = 0.95 \), then \( (a,b) \) is known as the 95% confidence interval for \( \theta \).

N.B. \( \theta \) is fixed and it is the interval which varies. It is therefore incorrect to say that there is a 95% chance that \( \theta \) lies in the interval. Rather, there is a 95% chance the interval contains \( \theta \) (a subtle, but important difference!).

Constructing Confidence Intervals

If the random variable \( X \) has a normal distribution with mean \( \mu \) and variance \( \sigma^2 \), then the sample mean \( \bar{X} \) also has a normal distribution with mean \( \mu \), but with variance \( \sigma^2/n \) (see random samples). In other words, \( \bar{X} \sim N(\mu, \sigma^2/n) \).

In fact, if we have a random variable \( X \) which has any distribution (not necessarily normal), by the central limit theorem the distribution of \( \bar{X} \) will be approximately normal with mean \( \mu \) and with variance \( \sigma^2/n \), for large \( n \).

So standardising this, we get:

\[
Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}
\]

(since the mean is \( \mu \) and the standard deviation is \( \sigma / \sqrt{n} \)).

From the Normal Distribution section, we know that \( P(-1.96 < Z < 1.96) = 0.95 \).
Hence \( P(-1.96 < \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} < 1.96) = 0.95 \)

Rearranging this we get \( P[\bar{X} - 1.96(\sigma / \sqrt{n}) < \mu < \bar{X} + 1.96(\sigma / \sqrt{n}) ] = 0.95 \)

Hence the 95% confidence interval for \( \mu \) is:
\( (\bar{X} - 1.96(\sigma / \sqrt{n}) , \bar{X} + 1.96(\sigma / \sqrt{n}) ) \)

N.B. in the confidence interval we use \( x \) not \( X \).

### The Geometric Distribution

A discrete random variable \( X \) is said to have a geometric distribution if it has a probability density function (p.d.f.) of the form:

\[
P(X = x) = q(x-1)p, \quad \text{where } q = 1 - p
\]

If \( X \) has a geometric distribution with parameter \( p \), we write \( X \sim \text{Geo}(p) \)

### Expectation and Variance

If \( X \sim \text{Geo}(p) \), then:

- \( E(X) = \frac{1}{p} \)
- \( \text{Var}(X) = \frac{q}{p^2}, \quad \text{where } q = 1 - p \)

### The Exponential Distribution

A continuous random variable having p.d.f. \( f(x) \), where:

- \( f(x) = le^{-lx} \)

is said to have an exponential distribution, with parameter \( l \).
The cumulative distribution is given by:

- \( F(x) = 1 - e^{lx} \)

**Expectation and Variance**

- \( E(X) = 1/l \)
- \( V(X) = 1/l^2 \)

---

**Bayes' Theorem**

This may not be on your syllabus

Bayes' Theorem says that:

\[
P(A_i | B) = \frac{P(A_i | B)P(A_i)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + ... + P(B|A_n)P(A_n)}
\]

Note that the union of all of the As (A1, A2, ... An) = the total sample space, so they cover every possibility.

**Example**

There are two bags containing balls of various colours. A bag is selected at random and a ball taken from it at random. The probability of picking a blue ball out of bag 1 is \( \frac{1}{2} \). The probability of picking a blue ball from bag 2 is \( \frac{1}{4} \). If the experiment is carried out and a blue ball is selected, what is the probability that bag 2 was selected?

Let A2 be the event that bag 2 was selected and let A1 be the event that bag one was selected. Let B be the event that a blue ball is chosen. Then, using Bayes' Theorem:

\[
P(A2|B) = \frac{P(B|A2)P(A2)}{P(B|A1)P(A1) + P(B|A2)P(A2)}
\]

Now, \( P(B|A2) = \) probability of picking a blue ball given that bag A2 is selected = \( \frac{1}{4} \) from the question.
Similarly, \( P(B|A_1) = \frac{1}{2} \).

\[ P(A_1) = \text{probability of selecting bag 1} = \frac{1}{2} = P(A_2) \]

Hence \( P(A_2|B) = \frac{1}{4} \times \frac{1}{2} = \frac{1}{2} \times \frac{1}{2} + \frac{1}{4} \times \frac{1}{2} = 1/3 \).

**Permutations and Combinations**

**Arranging Objects**

- The number of ways of arranging \( n \) unlike objects in a line is \( n! \) (pronounced ‘\( n \) factorial’). \( n! = n \times (n - 1) \times (n - 2) \times \ldots \times 3 \times 2 \times 1 \)

**Example**

How many different ways can the letters P, Q, R, S be arranged?

The answer is \( 4! = 24 \).

This is because there are four spaces to be filled: _, _, _, _

The first space can be filled by any one of the four letters. The second space can be filled by any of the remaining 3 letters. The third space can be filled by any of the 2 remaining letters and the final space must be filled by the one remaining letter. The total number of possible arrangements is therefore \( 4 \times 3 \times 2 \times 1 = 4! \).

- The number of ways of arranging \( n \) objects, of which \( p \) of one type are alike, \( q \) of a second type are alike, \( r \) of a third type are alike, etc is:

\[
\frac{n!}{p! \cdot q! \cdot r! \ldots}
\]

**Example**

In how many ways can the letters in the word: STATISTICS be arranged?
There are 3 S’s, 2 I’s and 3 T’s in this word, therefore, the number of ways of arranging the letters are:

\[ 10! = 50400 \]
\[ 3! 2! 3! \]

**Rings and Roundabouts**

- The number of ways of arranging \( n \) unlike objects in a ring when clockwise and anticlockwise arrangements are different is \((n - 1)!\)

When clockwise and anti-clockwise arrangements are the same, the number of ways is \(\frac{1}{2} (n – 1)!\)

**Example**

Ten people go to a party. How many different ways can they be seated?

Anti-clockwise and clockwise arrangements are the same. Therefore, the total number of ways is \(\frac{1}{2} (10-1)! = 181440\)

**Combinations**

The number of ways of selecting \( r \) objects from \( n \) unlike objects is:

\[ \binom{n}{r} = \frac{n!}{r! (n-r)!} \]

**Example**

There are 10 balls in a bag numbered from 1 to 10. Three balls are selected at random. How many different ways are there of selecting the three balls?

\[ _{10}C_3 = \frac{10!}{3! (10-3)!} = \frac{120}{3 \times 2 \times 1} = 20 \]

**Permutations**

A permutation is an ordered arrangement.
The number of ordered arrangements of \( r \) objects taken from \( n \) unlike objects is:

\[
^n\text{P}_r = \frac{n!}{(n - r)!}.
\]

**Example**

In the Match of the Day’s goal of the month competition, you had to pick the top 3 goals out of 10. Since the order is important, it is the permutation formula which we use.

\[
^{10}\text{P}_3 = \frac{10!}{7!} = 720
\]

There are therefore 720 different ways of picking the top three goals.

**Probability**

The above facts can be used to help solve problems in probability.

**Example**

In the National Lottery, 6 numbers are chosen from 49. You win if the 6 balls you pick match the six balls selected by the machine. What is the probability of winning the National Lottery?

The number of ways of choosing 6 numbers from 49 is \( ^{49}\text{C}_6 = 13\,983\,816 \).

Therefore the probability of winning the lottery is \( \frac{1}{13983816} = 0.000\,000\,071\,5 \) (3sf), which is about a 1 in 14 million chance.