This chapter has been made available as a sample of our Maths notes. It is intended for personal use.

Statistics

Types of data

Statistics is a very large part of the Leaving Cert Higher Level course. Statistics deals with the collection, presentation, analysis and interpretation of data. It has a huge and varied use in real life: from insurance to election campaigns etc. Statistics is a large area of study within the course that involves lots of learning off of definitions and concepts, which makes it a good area to focus on, as it is not as ambiguous as other areas of the course and marks can be maximised with strategic learning and hard work.

There are many questions on Paper 2 particularly that are very doable once definitions and concepts are fully understood, so it is important to know these inside out.

Make sure you know how to represent each type of data on a graph, as often some graphs are more appropriate to represent certain types of data than others.

We will start with a few general definitions that are frequently found in statistics. This will help us get into the language of the area.

- **Data** can be facts and figures collected together for reference or analysis. It is often an unordered list.
- **Information** is the term we give 'ordered data'.
- **A variable** is the characteristic of interest being recorded in a sample or population (we will explore these two terms when we discuss sampling later on). A variable we may be interested in may be the age of Irish men when they get married.
- **An observation** is the actual value of the variable we are looking at in the population or sample. For example, the age that John Smith got married is 31 years old.
- **A data set** is all of the observations for the particular variable we were wishing to find out. For example, the ages that 10 men got married at were 31, 29, 29, 40, 18, 24, 32, 49, 30, 22.
- **Descriptive Statistics** describes the visual representation of data. We are all familiar with the use of bar charts, pie charts, histograms and even tables. Its main advantage is to organise and summarise information (data).
- **Inferential Statistics** refers to the use of statistics and data to predict outcomes. This is usually used when talking about a portion of the population and using data obtained from these members of the population to reach conclusions about the whole population. If you want to study people who have had a stroke, you can survey a small group of them (around 100) and then use the data collected from these 100 people to make assumptions/reach conclusions about everyone who has had a stroke (the population).
- **Explanatory variable** is something that has caused a change within the population. Think of the amount of hours spent studying for a test, often the more hours you study, the better your score in the exam. Therefore, in this case the amount of hours spent studying is the explanatory variable (it explains the better test results)
- **Response variable** is the effect of/the change caused by the explanatory variable**.** The better score in the exam is the response variable i.e it is the response of doing more hours of study.
- **Univariate data** refers to when only one pieces of information is collected from each member of the group being surveyed. For example, weight in kilograms.

- **Bivariate (paired) data** refers to when 2 pieces of information are collected from each member of the group being surveyed. For example, gender and species of pet belonging to those surveyed.

The next concept we need to explore is the difference between **primary** and **secondary** data.

- **Primary data:** This is data that has been collected first hand/collected by the person who intends to use the data. For example, someone collecting information themselves regarding favourite soccer teams by surveying on the street, then using the data they have collected to make a graph.
  - **Observational Studies** and **Designed Experiments** are used in collecting primary data.
  - First of all, we need to know what an **experiment** is; a controlled study in which the collector of the data understands the cause/effect relationship and wishes to investigate this further (and perhaps back it up with statistics)
  - **Observational Studies** refers to data that is obtained/collected by counting/noticing what things are happening. For example counting the amount of people attending a football match or traffic passing a certain point along a motorway. The researcher is not going out to effect events, they are simply observing what is going on and presenting this data on a graph.
  - **Designed experiments** refers to data obtained by experiment, in this case the researcher is setting up a 'set of conditions' or an event that they have designed themselves. They then record the outcome of the experiment and use the data obtained. For example, giving a group of people treatment to influence outcome on blood pressure and observing the effects (does the blood pressure increase or decrease etc.) It is important to remember that experiments must be reproducible and repeated a number of times in order to ensure accurate data collection.
- **Secondary data** refers to second hand data, i.e. this is data that has been collected by somebody else. The person using this data to create their graph has not collected it themselves. For example, getting figures regarding religions of the population of Ireland from the Central Statistics Office (CSO) to make a graph.

There is another way we can categorise data: into **Quantitative (or numerical) data** and **Qualitative (or categorical) data**

- **Quantitative (Numerical) Data** is data that can be measured, it is data that involves the use of numbers or questions that are answered in numerical terms. For example, the number of people in a family, the number of supporters attending the All Ireland Final in 2017 etc.
- Qualitative (Categorical) Data refers to data that cannot be measured, it is not based on numbers. For example, eye colour, country of birth, grade in a test.

Quantitative (Numerical) Data can be further divided into 2 more subtypes: **continuous and discrete data**

- **Continuous Data** refers to data which is measured or represented on a scale and the variable can take any value on that scale. The data can be within a range. For example, the amount of time (t) in hours spent on Snapchat in one day by students in a sixth year. The value of the variable can take any real number between 0 and 24 hours inclusive. It can be an infinite number of values within this range in this case.
  Histograms are usually used to display continuous data.

- **Discrete Data** refers to data that can only have certain specific values**.** For example, the number of students within a class of 26 who own a dog. The possible values run from 0,1,2 etc. all the way to 26. It cannot, however, be every number within this range, as 0.5 of a student cannot own a dog!
  Bar charts are used to represent discrete data

Qualitative (Categorical) data can also be divided into 2 subtypes: **Ordinal and Nominal**

- **Ordinal Data:** data within a category that can have an order put on it. For example, exam grades (H1, H2, etc…), pain level on a scale of 1 to 10, or position in race (first, second, etc…).
  Pie charts and bar charts (more frequently) are used to display ordinal data.
- **Nominal Data:** any categorical data that cannot be ordered. For example, nationality, favourite song or favourite food.
  Pie charts and bar charts are also used to display nominal data.

For exam questions make sure you know what types of graphs (bar chart etc.) are used for each type of data. Although it has never been asked in the Leaving Cert it is important it is understood as such a question would test for both knowledge and understanding of each type of data.

As you can see there are lots of different definitions and concepts to understand regarding types of data in statistics. See if you can categorize the following variables into discrete, continuous, categorical and ordinal.

1. Height of county football players.
2. Social class of teachers in a school.
3. Number of players on the panel for a basketball team.
4. Amount of rainfall in a week in Co. Laois.
5. Blood type of population
6. Colour of hair of population
7. Shoe size of a soccer team.
8. Number of days someone is on a drug

Exam questions on this topic would involve differentiating between types of data, as well as being able to classify variables like in the questions above. Although no questions have been directly asked in this area, understanding this part of statistics allows us to answer more complicated (and more frequently asked) questions better. Sometimes, at the end of statistics questions they ask questions asking for your interpretation of the data. If you do not understand/know the definitions above you will not be able to give a sufficiently in depth answer that shows you understand the words related to the topic.

Sampling

Similar to the last subchapter on types of data, we have a number of definitions that are important to the understanding of the topic. Understanding these will allow you to apply your knowledge to answer questions relating to sampling.

- **Population** is the complete group of survey who can be surveyed, a population is usually defined by the researcher. For example, all females who study an Irish module in UCD, or all males under the age of 18 in Ireland.
  Other definitions related to population are;
  Census: collection of data from all of the population being studied.
  Sampling frame: refers to every item in the population being studied.
- **Sample** is the group of people who are actually surveyed. They are meant to be representative of the population and allow inferences/conclusions to be drawn about the population. It is much more cost effective and time efficient to do a survey on a sample of a population as only a small group need to be surveyed. It is important to choose a sample representative of the population, otherwise we introduce bias into our survey.
  A random sample is a sample whereby everyone within the population being studied has an equal chance of being selected.
- A **population parameter** is a numerical measurement measuring a characteristic of the population. These can be measures of central tendency (like mode or median) or the range.

Bias is another important concept when it comes to statistics.

Bias occurs when the sample chosen is not representative of the population wishing to be studied. The data may be distorted and can result in a viewpoint or result being inappropriately represented in a survey. Examples of causes of bias include:

- Voluntary response samples can result in biases, as people who voluntarily choose to take part may have stronger opinions.
- If the population is not identified or defined correctly, it can lead to choosing a misrepresentative sample.
- Sample size may not be large enough to sufficiently represent the population being studied.
- Questions that are misleading or hard to understand
- Data not being recorded correctly, leading to misrepresentation.

Exam questions, although never asked regarding bias, may ask for examples so be sure to have some learned.

Types of Samples

There are also various **types of samples** that can be taken from a population. In an exam you may be asked what is the most suitable sample for a particular type of survey or may simply be asked to explain or differentiate one type of sample from another. Be sure to have these concepts straight in your head as they are frequently asked, and the similarity between them.

Simple Random Sample:

- A sample, n, is picked whereby everyone in the population has an equal chance of being picked.
- From a purely theoretical point of view this can be described as the most representative as everyone within the population has an equal likelihood of being picked.
- A simple random sample can be done by assigning everyone in the population a number and picking those numbers out of a hat, or using a random number generator to pick the sample.
- It must be remembered that although this sample is picked at random by a computer and therefore free from personal bias, the resulting sample is not necessarily free from bias.

Stratified Random Sample:

- The population is divided into at least 2 groups, which have common characteristics.
- These groups are called strata and usually are natural divisions of population (gender, age, etc.)
- A simple random sample (outlined above) is then drawn from each subgroup, and these subgroups together make up the entire sample.
- This method of sampling allows the researcher to examine certain subgroups in more detail and allows us to put them into comparison with one another.
- For example, if you want to survey secondary school students in Ireland, we could divide them into strata boys and girls. Every member of the population should fit into only one strata.

Stratified Sample:

- The population is again divided into groups which have common characteristics.
- The percentage/proportion of each subgroup in the population is then calculated. This means that if a certain subgroup or strata represents 20% of the population, the subgroup/strata must make up 20% of the total sample.
- It is essential to know the proportion of the population in each strata and to take this into account when choosing the amount of people from each strata. This will ensure the strata are representative within the overall sample.
- For example, the population is divided into age groups. If the 20 – 29 year old age group makes up 10% of the population, then the sample used for the survey will contain 10% of people within the 20-29 year old age group.

Systemic Sample:

- This involves using a patterned system to pick a sample as the name suggests.
- For example, picking a person at random and then picking every 40^{th} member of the population after that, until the sample size has been reached (it doesn't have to be every 40^{th}, any interval will do).

- Only the first person is really chosen randomly. Although this is an incredibly fast way of creating sample it may not always be representative as there is no real degree of randomness.

Cluster Sample:

- The population is divided into sections or clusters.
- A random cluster is selected and then every member in that cluster is surveyed.
- Every parameter in the survey is explored with everyone in the cluster looked at. It allows for reduction of cost and increases efficiency, and more than one cluster can also be surveyed.
- This is similar to Leaving Cert Biology, we did not survey every piece of land in the habitat, rather 10 different clusters or areas (1m squared), and looked at every item in that small area and used this data to draw conclusions about the whole habitat

Quota Sampling:

- This method allows the surveyor some discretion. The person conducting the sample is given a quota, or a specific number of people to survey.
- This may be specific, like a survey of 10 schoolchildren and 20 farmers or just one number (30) that have to be surveyed.
- The surveyor can therefore choose who they want to survey. There is no randomisation in this survey type.
- The surveyor can easily induce bias also as they are free to choose who they want to interview. Only the opinions of those chosen by the interview are conveyed on the survey.

Convenience Sampling:

- The surveyor chooses members of the population that are the most convenient for him to survey.
- This introduces obvious bias as there is no real randomness involved, it is just whatever suits the surveyor.

As you can see there are many different types of sampling and it is important to know how to differentiate one another. Thinking of your own example for each type of sample can be beneficial and can make it easier to learn the difference between each type of sample. Exam questions in the past have asked students to define one type of sampling (2013 P2 Q7(a)(i)), so it is important to know/be able to define types of sampling. Be sure to give a practical example (like the ones listed above) when writing a definition in an exam as it shows a complete understanding of the topic and that you have not just learned it off. For questions asking you to compare sampling types (or the difference between population and sample), just provide the definition for each type, you do not have to directly compare them.

Issues when Sampling

- **Ethics**
  - **Informed consent** – participants must have full knowledge of what they are participating in. They must be informed as to what the survey involves, and they must voluntarily consent to participate in the survey. If those surveyed feel they are not sufficiently informed about the survey they may decline to do the survey and also may discredit the

findings of the survey when the results do come out. It is important to insure those you take information are aware what they are giving their personal information to you for.

- **Confidentiality –** the information confided by participants in the survey must be confidential. It must not be divulged without the participant's consent. This preserves the integrity of both those surveyed and the survey itself. Ensuring confidentiality can also help remove an element of bias from the survey, as those surveyed will be more likely to supply an honest answer as they know their anonymity is being respected and preserved in the collection of data. Confidentiality benefits those being surveyed and those surveying also.

- **Reliability**
  - The sample must be
    - Large enough
    - Randomly selected from the population
    - Have a high response rate
    - All members of the population should have an equal chance of being picked
    - Must not be biased

Collecting data

- **Surveys** are the most common method of collecting data. This involves using a **questionnaire**.
- Surveys can involve...

  -face to face interview

  -telephone survey

  -email/postal survey

  -online survey

  -observation

- Each method has advantages and disadvantages which are explored below. General elements such as randomness, bias, how the participants are chose, how reflective participants are of the entire population, accuracy, cost, whether or not questions can be explained and anonymity are also good.

- A **questionnaire** is a set of questions created to get data from a population. They are very common in collecting data in surveys
- Questionnaires should be….
  - Easy to understand – written using **clear and simple words**
  - **Relevant** to the survey and to the information you want to find out, be clear who you want to complete
  - **Short** in length, make the questions as short as possible
  - Be clear as to **how and where** the answers are to be recorded
  - **Simple at first**, and the most **difficult or time-consuming questions left to the end** (to encourage participation)
  - Should **include all possible answers**, but only use open ended responses (where the participant writes their own response) when necessary. Open ended questions are hard to analyse and can be difficult to extrapolate data from so avoid if possible.
  - Start with tick the box questions, leave open ended questions till last
  - There should be **no leading questions**, for example, do you agree that soccer is way better than hurling?

An example regarding good questionnaire writing is as follows:

Asking directly “what age are you?” or “tick what age group you belong to” and then a list containing different age groups after to be ticked.

- The direct question is less likely to get an honest answer as people are not inclined to tell you their age straight out.
- The direct question is also more difficult for the surveyor to analyse, as they have to go through each survey and find the age and record it.
- The “tick the box” question sorts each category and makes it easier for the surveyor and the person being surveyed.

There is also some advantages and disadvantages to methods of collecting data. We will explore these now.

Face to face interview

Advantages:

- There is a high response rate, people are more likely to answer you as they can see the person asking them questions.
- The surveyor can use non-verbal and verbal reactions on their survey as they can physically see the person they are interviewing.
- The surveyor can ask more personal and in depth questions as they develop a rapport with the person they are interviewing.

Disadvantages:

- These surveys can be expensive and time consuming, as they individually interview each person.
- The surveyor can also influence the response in the manner they ask the questions, introducing bias.

Telephone survey:

Advantages:

- There is a high response rate, this is due mostly to almost everyone in a population having a phone.
- Questions can be more targeted and personal as the surveyor is having a conversation with the person they are interviewing.

Disadvantages:

- Similar to face to face interviews, these surveys are expensive and time consuming.
- Interviewer can also distort the question in order to influence the response of the person being surveyed.
- Timing must be taken into consideration as many people do not have access to their phones during the working day.

Postal/Email Survey:

Advantages:

- Not as time consuming as the previous two methods of surveying.
- Can have a large sample as you simply have to post out the survey.
- Can ask a large array of questions.

Disadvantages:

- Less likely to have an accurate survey, as survey may in inadequately completed or not completed at all.
- There is no opportunity to clarify the questions as the researcher is not physically there to answer questions from those surveyed.

Online questionnaires:

Advantages:

- Very cheap and data can be collected quickly
- Anonymous, those being surveyed do not have to identify themselves.
- If there is difficulty with a particular question, that question can be easily fixed or taken out of the survey.

Disadvantages:

- Only those with internet/or a computer can access these types of questionnaires, which can be problematic if a certain population is indicated.
- There is no interviewer to clarify questions that may not be understood.

To date no question has come up on the LC HL Maths paper on the above topic, however, if a question asks you to design your own survey, think about what method of collecting data is most suitable for the data you wish to collect. If you want to collect data from students, an online survey is probably more accessible for them. One may argue that an older population would probably be more comfortable in an interview setting. Again, knowing the strengths and weaknesses of each type will allow you to pick whichever surveying technique is most applicable to the circumstances the question describes. Issues when sampling is another area which has never been asked, but given GDPR having been brought into law recently it is topical at the moment, so important to understand.

Understanding the basic concepts and definitions of statistics is important in this section of the maths course as more complex exam questions assume familiarity with these terms. Although not many questions ask for the definition outright, every questions assumes you know the terms described above. Almost every question you complete in statistics will contain words like “population”, “sample”, “simple random sampling” etc. so make sure you learn all these definitions if you want to guarantee you will understand all statistic question on the Leaving Cert!

Normal distributions

- A normal distribution refers to the 'bell curve' graph.
- Normal distributions are found repeatedly in statistics eg. if you survey people's height, weight, grades etc.
- **In a normal distribution, the mode, the mean and the median are all equal**

Empirical rule

- The empirical rule is a rough estimate used in relation to normal distributions
- For Higher Level, it is unlikely you will be asked to use the empirical rule to solve a statistics question. However, a question could ask you to 'estimate' using the empirical rule and then compare it to your more accurate answer, so it is best just to learn off the following points:
- In any normal distribution,
    - 68% of the distribution lies within one standard deviation of the mean
    - 95% of the distribution lies within two standard deviations of the mean
    - 99.7% of the distribution lies within three standard deviations of the mean
- Example: 1000 people are surveyed about their weight. The mean weight is 60kg. The standard deviation is 3kg. According to the empirical rule, 68% of people in the survey (680 people) have a weight somewhere between 57kg and 63kg (within one standard deviation of the mean).

Standard normal distribution

- As there are numerous possible normal distributions, in order to compare them we use a 'standard normal distribution'
- A standard normal distribution involves the use of z-scores
- **Z-scores refer to how many standard deviations a variable is away from the mean**
- Z-scores can be positive or negative. For example, a z-score of 1 means that the value is one standard deviation above the mean. A z-score of -0.6 means that the value is 0.6 standard deviations below the mean.
- Knowing the z-score allows you to calculate certain areas under the curve in relation to that z-score, which allows you to figure out the probability that the value is above, below or within certain values.
- Remember that probability is always between 0 and 1, therefore the total area under the curve is 1.
- The mean lies in the middle of the curve, which is symmetrical ie. 50% of the data lies on either side of the mean.
- The area under the curve **behind** (to the left of) every possible z-score is given on pgs. 36-37 of the log tables book. It is important to be able to use these values in order to calculate the area behind and in front of any given z-score.
- In any question involving z-scores and normal distributions, it is a good idea to roughly sketch the normal distribution and shade in the area you are looking for.
- This is best understood using examples.
- (i) Find $P(z \leq 1.26)$
    - We are asked to calculate the probability that the z-score is less than or equal to 1.26
    - To do this, we go to pg. 37 of the log tables to find the area that is behind a z-score of 1.26

- Go down the column on the left-hand side until you reach 1.2. Then go across the row until you reach the column under 0.06.
- You should be at the value of 0.8962. This is the value of the area under the curve behind the z-score of 1.26, and hence it is our answer.

- (ii) Find P(z ≤ -1.52)
  - The values given in the log tables are for positive z-scores only. In this question we are asked to find the area behind a negative z-score.
  - Remember that the normal distribution is symmetrical.
  - Hence, the area behind -1.52 (what we are looking for) will be equal to the area in front of 1.52
  - If we go to pg. 37 of the log tables, we can see that the area behind 1.52 is 0.9357. We want the area in front of it.
  - Since the total area under the curve is 1, calculating (1 – 0.9357) will give us this area.
  - 1 – 0.9357 = 0.0643, which is our answer.
- (iii) Find P(z ≥ 0.42)
  - This question is similar to the last question as we are looking for the area in front of the z-score instead of behind it.
  - The area behind a z-score of 0.42 is 0.6628 according to pg. 36 of the log tables.
  - 1 – 0.6628 = 0.3372
- (iv) Find P(-0.3 ≤ z ≤ 1.68)
  - This question is slightly different to the others are we are asked to find the area between two z-scores.
  - The best way to approach this is to calculate the area behind the lower z-score (-0.3) and subtract it from the area behind the higher z-score (1.68).
  - Remember that the normal distribution curve is symmetrical. The area behind -0.3 is equal to the area in front of 0.3.
  - The area behind 0.3 is 0.6179 (pg. 36 of log tables), so the area in front of it is (1 – 0.6179) = 0.3821.
  - The area behind a z-score of 1.68 = 0.9535, according to the log tables.
  - 0.9535 – 0.3821 = 0.5532.

<u>Converting variables to z-scores</u>

- The question will not give you the z-scores, so you will have to calculate them yourself. The formula for this is on pg. 34 of the log tables (standardising formula)
- $z = \frac{x-\mu}{\sigma}$, where x = the variable, μ = mean and σ = standard deviation.
- It is essential to convert a variable to a z-score so that you can use the normal distribution. The values for area under the curve given on pgs. 36-37 are only relevant for z-scores.
- Example 1: In a class test, the mean score is 57%. The standard deviation is 6%. Find the probability that a student picked at random scored 62% or more in the test.
  - In this question, no z-scores are mentioned. We need to know what z-score corresponds to a score of 62% in order to use the normal distribution curve to solve the problem.
  - x = 62, μ = 57, σ = 3
  - $z = \frac{62-57}{3} = 1.67$
  - We now want to find out the area in front of a z-score of 1.67
  - From pg. 37 of the log tables, the area behind a z-score of 1.67 is 0.9525.

- 1 – 0.9525 = 0.0475.
- The probability that a student picked at random score 62% or above is 0.0475.

- Example 2: On a farm, the mean weight of pigs is 301kg. The standard deviation is 32kg. If 40 pigs are picked at random, how many of them can be expected to weigh between 290kg and 306kg?
  - Note that this question is asking for the number of pigs, not the probability.
  - We need to find out what z-scores correspond to 290kg and 306kg.
  - $z = \frac{306-301}{32} = 0.16$.
  - $z = \frac{290-301}{32} = -0.34$.
  - We need to find out the area between -0.34 and 0.16 on the normal distribution.
  - The area behind 0.16 is 0.5636 according to pg. 36 of the log tables.
  - The area behind -0.34 will be equal to the area in front of 0.34. The area behind 0.34 is 0.6331, so the area in front of it will be (1 – 0.6331) = 0.3669.
  - 0.5636 – 0.3669 = 0.1967.
  - The probability that a pig picked at random will weigh between 290kg and 306kg is 0.1967.
  - Therefore in a sample of 40 pigs, we would expect that (40 x 0.1967) ie. 7.868 pigs would weigh between 290kg and 306kg. However since 0.868 of a pig is not possible in this context, our final answer would be 7 pigs.

Central limit theorem and standard error

- In statistics, the term 'population' refers to all the possible relevant observations. For example, if you are studying the age of dogs, the population refers to the age of every single dog in the world.
- In conducting studies and surveys, it would be impossible to include the entire population. For this reason, samples are studied (data is collected from a subset of the population).
- However, because not every single item in the population is included, the sample data may not be representative of the entire population. For example, this is seen in elections when a survey of how a certain number of people voted does not reflect exactly the results of the election.
- To account for this, we use something called the standard error. There are two different formulae for the standard error, one for proportion and one for the mean (pg. 34 of the log tables).
- There is also a formula for an approximation of the margin of error, which should be learned off.
- The standard error of the mean is accounted for by the Central Limit Theorem.
- Standard error is important for constructing confidence intervals (next section).

Central Limit Theorem

- There are three main ideas outlined by the Central Limit Theorem that are important to know in order to understand Higher Level statistics.
- 1. The sampling distribution of the mean forms a normal distribution
    - This means that if you carry out lots of samples, calculate the means of all those samples and then plot the means, they will form a normal distribution curve.
- 2. The mean of all the samples = the mean of the original population
    - Once you calculate the means of all the individual samples and then calculate the 'mean mean', it will be the same mean as if you surveyed the entire population.
- 3. The standard error of the mean is equal to the standard deviation divided by the square root of the number in the sample.
    - Standard error of the mean = $\frac{\sigma}{\sqrt{n}}$
- This is important in relation to Leaving Cert. questions. For example, if the question mentions a distribution of sample means, you know it will be a normal distribution and therefore you can use z-scores.

Standard error of the proportion

- Population proportion
    - In a study of public transport use in Dublin, if you interview every single person in Dublin and 72% use the bus, this is referred to as the population proportion
    - It is denoted by p.
    - In statistics, percentages are typically written as decimals
    - In this case, p = 0.72
- Sample proportion
    - It would be difficult to survey every person in Dublin, so you might survey a sample.
    - In the sample, 70% of people use the bus.
    - This is referred to as the sample proportion
    - It is denoted by $\hat{p}$, referred to as 'p hat'.
    - In this case $\hat{p} = 0.7$.

- The formula for the standard error of the proportion is $\sigma_p = \sqrt{\frac{p(1-p)}{n}}$, where p = the population proportion and n = the number of people included in the sample.
- In many questions, p is not known and so $\hat{p}$ is used instead.

Approximation of the margin of error

- Margin of error = E = $\frac{1}{\sqrt{n}}$ where n = number of people in the sample.
- This is only an estimate and should not be used unless it is specifically asked in the question.
- The formulae for standard error discussed in the above sections are more accurate.
- However, this approximate formula for the margin of error highlights an important point.
- The margin of error is inversely proportional to the sample size.
- This means that the larger the sample size, the smaller the margin of error and vice versa.
- There is no 'rule of thumb' for what is a good sample size, however it is generally accepted that $n > 30$ is required for the principles of the Central Limit Theorem to apply. Generally, the bigger the sample size, the better.

Confidence intervals

- For Leaving Cert. Higher Level Maths, you must be able to construct 95% confidence intervals for the population proportion and for the mean

95% confidence interval for the population proportion

- This is used to estimate the population proportion from the sample proportion.
- With a 95% confidence interval, we are saying that we are 95% confident that the population proportion will lie between these two particular values.
- To construct a 95% confidence interval, we use the sample proportion and the standard error to find two values: (the sample proportion + the standard error) and (the sample proportion – the standard error)
- The standard error is calculated using the formula discussed in the previous section. However, because this is a 95% confidence interval, we multiply the formula for standard error of the proportion by 1.96.
- The reason we multiply it by 1.96 can be understood by thinking about a normal distribution curve:
    - If we want to be 95% confident that our answer lies between two values, then 95% of the data must lie between these two values.
    - If 95% of the data is between these two values, then 5% must be outside of these two values.
    - As the normal distribution curve is symmetrical, this means there is 2.5% of data outside the confidence interval at each end.
    - We can use the log tables (pgs. 36-37) to find out what z-score corresponds to having 2.5% or an area of 0.025 in front of it.
    - 1 – 0.025 = 0.9750. The value of 0.975 corresponds to the z-score of 1.96.
- The 95% confidence interval for the population proportion (p) is:
    - $\hat{p} - E \leq p \leq \hat{p} + E$ where $E = 1.96\sqrt{\frac{p(1-p)}{n}}$
- Example: In a health survey of 670 office workers, it was found that 230 were overweight. Construct a 95% confidence interval to determine the true proportion of office workers that are overweight.
- Answer:
    - We are trying to find two values that we can be 95% confident the population proportion will lie between: ($\hat{p}$ + E) and ($\hat{p}$ - E)
    - Therefore we need to find $\hat{p}$ and E
    - $\hat{p}$ is the sample proportion, which we can calculate from the given data.
    - $\hat{p} = \frac{230}{670} = 0.34$.
    - Next we need to find E. Remember $E = 1.96\sqrt{\frac{p(1-p)}{n}}$.
    - As we do not know p, the population proportion, in this question, we use $\hat{p}$ instead.
    - $E = 1.96\sqrt{\frac{(0.34)(1-0.34)}{670}} = 0.0359$
    - $\hat{p} + E = 0.34 + 0.0359 = 0.376$.
    - $\hat{p} - E = 0.34 - 0.0359 = 0.3041$.
    - We can be 95% confident that the proportion of office workers who are overweight lies between 0.3041 and 0.376.
    - $0.3041 \leq p \leq 0.376$.

95% confidence interval of the mean

- Constructing a 95% confidence interval for the mean involves finding two values, based on a sample mean, that we are 95% confidence the population mean lies between.
- It is similar to the population proportion confidence interval, except the formula for standard error is different.
- The population mean is denoted by μ.
- The sample mean is denoted by $\bar{x}$.
- In the 95% confidence interval for the mean:
  - $\bar{x} - E \leq \mu \leq \bar{x} + E$
  - $E = 1.96\frac{\sigma}{\sqrt{n}}$ (recall the standard error of the mean from the Central Limit Theorem, multiplied by 1.96 as explained above).
- Example: In a survey of the weight of sweets being bought in a sweet shop, it was found that the mean weight of sweets bought by 150 customers was 0.2kg. There was a standard deviation of 0.03kg. Form a 95% confidence interval for the weight of sweets being bought by all customers of the shop.
- Answer:
  - We are looking for $(\bar{x} + E)$ and $(\bar{x} - E)$
  - $\bar{x} = 0.2$kg (given in the question).
  - $E = 1.96\frac{\sigma}{\sqrt{n}} = 1.96(\frac{0.03}{\sqrt{150}}) = 0.005$.
  - $\bar{x} + E = 0.2 + 0.005 = 0.205$
  - $\bar{x} - E = 0.2 - 0.005 = 0.195$
  - We can be 95% confident that the mean weight of sweets bought in the sweet shop is between 0.195kg and 0.205kg.
  - $0.195\text{kg} \leq \mu \leq 0.205$.
- Note that in the above question we are not given the population standard deviation (σ), but the sample standard deviation (sometimes denoted by s). In questions where the population standard deviation is not given, it is acceptable to use the sample standard deviation, as we did in the above question.

Other confidence intervals

- The current Leaving Cert. syllabus states that students only need to be able to construct 95% confidence intervals.
- However, if you were asked to construct a different confidence interval, the only thing that would change would be the number (1.96) you multiply the standard error by.
- If it was a 90% confidence interval, there would be 10% of data outside the interval (5% on each side), which would correspond to a z-value of 1.64.
- The syllabus also says that students must understand that increased confidence levels means increased intervals.
- Consider a 98% confidence interval. If we wanted to be 98% confident that the proportion/mean was between two values, the standard error (E) would have to be larger in order to include more data to increase confidence levels.
- Similarly, in lower confidence intervals such as 90%, E is smaller as we are 'excluding' more data (1.64 is less than 1.96, hence E will be smaller because we are multiplying it by a smaller number).

Hypothesis testing

- Hypothesis testing involves using statistics to determine whether to reject a claim that is made about the proportion/mean
- The question will outline a claim that is made. There will then be a sample survey done to investigate this claim. The question will ask you to investigate whether the claim should be rejected or 'not rejected' based on the sample results.
- Leaving Cert. questions will ask you to test 'at the 5% level of significance'.
- When we test a hypothesis at this level of significance, we are investigating whether the claimed population proportion/mean lies in the most extreme 5% of results.
- You must be able to test hypotheses relating to the population proportion and to the mean. These vary slightly in their methods.

Null hypothesis and alternative hypothesis

- The claim that is made is the null hypothesis, denoted by H_0.
- For example, imagine the question states that 65% of children wear glasses. This is the population proportion, p. Our null hypothesis would be written as H_0: $p = 0.65$.
- The alternative hypothesis states that the null hypothesis is not true. It is denoted by H_A.
- The alternative hypothesis for the null hypothesis above would be written as H_A: $p \neq 0.65$.
- For every question on hypothesis testing, you must state both the null and alternative hypotheses.

Hypothesis testing of the population proportion

- There are 4 steps for hypothesis testing of the population proportion
- 1. State the null and alternative hypotheses
- 2. Construct a 95% confidence interval using the claimed population proportion.
- 3. Investigate whether the sample proportion lies within these values.
- 4. Conclude whether you reject or do not reject the null hypothesis.

Example: A pharmaceutical company claims that 22% of patients taking Drug A experience side effects. To investigate this claim, a survey of 500 patients is conducted. The survey finds that 135 of these patients experience side effects. Use a hypothesis test at the 5% level of significance to investigate if there is sufficient evidence to dispute the pharmaceutical company's claim.

Step 1: State the null and alternative hypotheses.

- The null hypothesis is the company's claim. H_0: $p = 0.22$.
- The alternative hypothesis disputes this. H_A: $p \neq 0.22$.

Step 2: Construct a 95% confidence interval using the claimed population proportion.

- Remember to construct a 95% confidence interval we find (p + E) and (p – E).
- $p = 0.22$.
- $E = 1.96\sqrt{\frac{p(1-p)}{n}} = 1.96\sqrt{\frac{0.22(1-0.22)}{500}} = 0.0363$.
- $p + E = 0.22 + 0.0363 = 0.2563$.
- $p – E = 0.22 – 0.0363 = 0.1837$.
- If the null hypothesis is true, the sample proportion should lie between 0.1837 and 0.2563.

Step 3: Investigate whether the sample proportion lies within these values.

- Sample proportion = $\hat{p} = \frac{135}{500} = 0.27$.
- Since $0.27 > 0.2563$, it does not lie within these values.
- This means it must lie outside the 95% confidence interval ie. within the most extreme 5% of results.

Step 4: Conclude whether you reject or do not reject the null hypothesis.

- As the sample proportion lies outside the critical values, we say it is significant.
- This mean that we reject the null hypothesis.

Hypothesis testing of the population mean

- There are 4 steps in hypothesis testing of the mean.
- 1. State the null and alternative hypotheses.
- 2. Calculate the test statistic.
- 3. Investigate if the test statistic lies within the critical values.
- 4. Conclude whether you reject or do not reject the null hypothesis.

Example: A tyre company claims that its tyres last a mean distance of 38,500km. To investigate this claim, a distributor conducts a study of 3000 tyres and finds they last a mean distance of 38,387km with a standard deviation of 3,450km. Is there sufficient evidence, at the 5% level of significance, to reject the company's claim?

Step 1: State the null and alternative hypotheses.

- Remember that the null hypothesis is the original claim ie. what the company says. H_0: μ = 38,500km.
- The alternative hypothesis simply states that the null hypothesis isn't true. A common mistake that students make is saying that the alternative hypothesis is what the sample finds eg. μ = 38,387km. This is incorrect. The correct alternative hypothesis is H_A: $\mu \neq$ 38,500km.

Step 2: Calculate the test statistic.

- The formula for the test statistic can be found on pg. 35 of the log tables (one-sample t-test).
- $T = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$ where $\bar{x}$ = the sample mean, μ = the original mean, s = the sample standard deviation and n = the number in the sample.
- $T = \frac{38{,}387 - 38{,}500}{\frac{3{,}450}{\sqrt{3000}}} = -1.79$.
- The test statistic is actually a z-score. Remember the formula for converting a variable to a z-score: $z = \frac{x - \mu}{\sigma}$. Recall the formula for standard error of the mean: $\sigma = \frac{\sigma}{\sqrt{n}}$ or $\frac{s}{\sqrt{n}}$. Combining the two gives us a formula for the test statistic.

Step 3: Investigate if the test statistic lies within the critical values.

- We want to investigate whether the sample mean lies in the most extreme 5% of possible values of the mean.

- Remember the 95% confidence interval for the mean. In order to be 95% confident that our value lies within a given range (and not the extreme 5%), the values have to lie between 1.96 and -1.96 z-scores on the normal distribution curve.
- The critical values for the hypothesis test at the 5% level of significance are -1.96 and 1.96.
- -1.79 is less than 1.96 and greater than -1.96, therefore it does lie within these critical values.
- $-1.96 \leq -1.79 \leq 1.96$.

Step 4: Conclude whether you reject or do not reject the null hypothesis.

- The sample test statistic lies within the critical values at the 5% level of significance, there we do not reject the null hypothesis.
- Note that you never say you 'accept' the null hypothesis. Just because there is not enough evidence to reject the null hypothesis does not mean we accept it. Therefore we say we do 'not reject' the null hypothesis.

Calculating *p*-values

- Some questions may ask you specifically to investigate a hypothesis using p-values.
- A p-value is the probability of getting a value 'more extreme' than the test statistic.
- This can be understood by thinking about the area under the normal distribution curve. If the test statistic is less extreme, so closer to the middle of the normal distribution curve, then the area outside of it (the p-value) is going to be larger.
- Conversely, if the test statistic is more extreme, ie. closer to the edge of the normal distribution curve, the p-value will be smaller.
- Remember that the normal distribution curve is symmetrical. If your test statistic is 1.96, for example, the p-value is comprised of the area in front of a 1.96 z-score and the area behind a -1.96 z-score added together.
- In order to use p-values to test a hypothesis, you must first calculate the test statistic as outlined in the previous section.

Example: Imagine the test statistic is -1.85. Calculate the associated p-value.

- The test statistic, which is a z-score, is equal to -1.85.
- We need to calculate the area behind a z-score of -1.85 and in front of a z-score of 1.85. These two areas will be equal (refer to the section on z-scores at the beginning of this chapter).
- Looking at pg. 37 of the log tables, the area behind a z-score of 1.85 is equal to 0.9678.
- The area in front of this = (1 – 0.9678) = 0.0322.
- This is also the area behind a z-score of -1.85.
- The p-value = 0.0322 + 0.0322 = 0.0644.

Using p-values to test a hypothesis

- Remember that in hypothesis testing at the 5% level of significance, we are investigating whether a certain value is within the 5% most extreme values.

- At the 5% level of significance, the p-value is equal to 0.05. This is because 5% of the area of the curve lies outside the 95% confidence interval. (5% = 0.05 in decimal form).
- If the p-value is less than 0.05, we must reject the null hypothesis. If the p-value is less than 0.05, this must mean that the test statistic is outside the 95% confidence interval since the area outside of the test statistic is less than 5% (less than 0.05).
- If the p-value is greater than 0.05, we do not reject the null hypothesis as this means that the test statistic is inside the 95% confidence interval.
- There are 4 steps when using a p-value to test a hypothesis:
- 1. State the null and alternative hypotheses.
- 2: Calculate the test statistic.
- 3: Calculate the p-value.
- 4: Conclude whether you reject the null hypothesis.

Example: A farmer claims that the eggs produced by his chickens have a mean weight of 56g. To investigate this claim, a distributor samples 100 eggs and finds that the mean weight is 54g with a standard deviation of 9g. Using p-values, investigate whether there is evidence at the 5% level of significance to dispute the farmers claims.

Step 1: State the null and alternative hypotheses

- H_0: $\mu = 56$g.
- H_A: $\mu \neq 56$g.

Step 2: Calculate the test statistic

- $T = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$
- $T = \frac{54-56}{\frac{9}{\sqrt{100}}} = -2.22$.

Step 3: Calculate the p-value.

- The test statistic (z-score) is -2.22.
- We need to calculate the area behind a z-score of -2.22 and in front of a z-score of 2.22.
- According to pg. 37 of the log tables, the area behind a z-score of 2.22 = 0.9868.
- The area behind a z-score of 2.22 = 1 – 0.9868 = 0.0132.
- This will be equal to the area behind a z-score of -2.22 since the normal distribution curve is symmetrical.
- Therefore the p-value = 2 x 0.0132 = 0.0264.

Step 4: Conclude whether you reject the null hypothesis.

- Remember the p-value at the 5% level of significance is 0.05.
- $0.0264 < 0.05$.
- This means that the test statistic is outside the 95% confidence interval ie. it is in the most extreme 5% of values.
- Hence we reject the null hypothesis.

p-values for other levels of significance

- A question could potentially ask if a result is significant at levels of significance other than 5% eg. 2%.

- If we were testing a hypothesis at the 2% level of significance, we would be investigating whether the value was in the most extreme 2% of values.
- This would mean that 98% would be inside the confidence interval, as 2% is outside of it.
- The critical p-value in this case would be 0.02 (2%).