

Chapter Seven - Introduction to statistical inference

Simona Simona

4/3/2020

In this chapter we address the following topics:

- Statistical inference
- Random sampling methods
- The normal distribution
- The standard normal distribution
- The sampling distribution
- The central limit theorem
- The t-distribution

Inference

This topic can be rather daunting and it is quite tempting for new comers in quantitative social sciences to avoid it altogether. However learning inference unleashes the power of statistical theory that lies beneath the programming language you are using so that in instances when you are given the wrong calculation by the machine, it may ring a bell at the back of your mind that something is wrong. The computer is often times described as *GIGO*, for **Garbage In Garbage Out**. Simply meaning that if you feed wrong data into the computer, it will give you wrong results or whatever you give the computer it will produce results. Your understanding of inference and statistical theory in general will help you avoid this problem. So, it is important that you devote particular care to ensuring that you understand the materials presented in this chapter.

What is statistical inference?

We have already defined statistical inference as basically, the process of inferring properties of the population from that of a random sample. In other words we can say that statistical inference allows us to learn something about the population based on what we know about the sample. If you have been reading this book from the beginning you already know what the sample and population are. But in case you jumped the gun, we define a **sample** as a subset of the population that is collected using a standard procedure. A sample can be in form of objects, rocks, soil, blood, plants and almost anything depending on a field of study. In the social sciences however, almost always it will refer to a set of individuals because we often times study human beings. A numeric characteristic that describes a sample is referred to as a **sample statistic**. In terms of mathematical notation, Latin letters are often used to represent the sample statistics such as \bar{x} for the mean and s for the standard deviation. The **population**¹ on the other hand, denotes all possible elements of interest. This is the group about which social scientists want to generalise. A numeric characteristic that

¹Notice the stark difference from the everyday use of the word as all the inhabitants of a particular place!

describes a population is referred to as a **population parameter**. Greek letters μ and σ denote population parameters of population mean and standard deviation respectively.

In most cases we want to learn about population parameters. We hardly ever want to know much about the population. We are often interested in the sample because it tells us something about the population from which it was collected. If we are somehow able to collect all our population parameters, we would not be interested in the sample. In fact I would say that in that case, the subject of statistics and quantitative social science would be almost null and void. There are times when we are able to study all the elements of interest and this process is often called the **population census**. I am sure you have heard that term before. Every country in the world has some form of a population census and it typically takes place in ten year intervals. The population usually carried out by national governments. I am sure I don't really need to labour explaining why population censuses are most often unattainable especially in small to medium-scale research projects. They are resource constraining, time consuming and too complex to implement. In the end a more pragmatic choice we have is to rely on samples and look for credible ways of generalising from the sample to the population. This is why statistical inference is so important. It is one of the major concepts of the field of statistics and quantitative social science.

There are mainly two areas of inferential statistics, **parameter estimation** and **hypothesis testing**. Parameter estimation basically means taking the sample statistic such as the sample mean and using it to say something about the population parameter such as the population mean. Hypothesis testing is where we use sample data to assess the credibility of some claim made about the population. For example you may want to assess whether years of spent in education affects one's income. A hypothesis may claim that there is a relationship between education and income and you want to use your sample data to answer this research question. These two areas are actually the subjects of the subsequent chapter. For now we still have basic questions we need to address before moving to estimation and hypothesis testing. There are three important questions we need to deal with here. What kind of sample do we need for us to have the confidence of generalising sample data to the population? We know even from our common sense that the sample we get from a population is not the only possible sample we can collect from the same population. This goes with saying that the sample statistics from the one sample we have may be very far away from the population parameter of our interest. How do we go around this problem? Answering these two questions will draw us very close to understanding statistical inference.

The short answer to the first question is that we need a sample that is obtained by **random sampling**. Random sampling is also known as **probability sampling**. It can be defined as any method that utilises some form of random selection. Probability sampling is good for inferential statistics because it allows for the application of the **law of statistical regularity**. This law is derived from the mathematical theory of probability and it states that a moderately large random sample taken from a large population will have the same composition and characteristics as the population from which it was taken. The sample we obtain using random sampling is likely to be similar to the population because eliminates systematic error.

A random sample is the opposite of a biased sample as the latter favours the selection of some elements of the population of interest over others. Take for example you were doing a study on a university campus about student's perceptions on climate change and you recruit your participants only from social spaces where they socialise and drink beer. The problem with this sample is that it only picks students who drink beer. That is not the only problem, it also only picks students who are in attendance at the particular time when you conduct your visit. Which means that this sample will exclude all the other students who don't drink and it also exclude students who drink

but are not in attendance at the time of your visit. As such, it will not represent the entire student population. This is a biased sample and cannot be used to generalise to the entire population of interest. There are many other examples you can think about where the condition of randomness is violated. If you conduct a study using an internet experiment you are conducting perhaps on the consumption of fake news among the population of a particular city. This study is not generalisable because not everybody in that city has access to the internet, let alone subscribes to the social media platform you are researching on. This also is a biased study and cannot be generalisable. Now lets look at the types of probability sampling methods you can use with a justification of generalising sample data to the population.

Simple-random sampling

The purest type of probability sampling is the simple random sampling. In simple random sampling, each member of the population of interest has an equal chance of being selected into the sample. This is the ideal sampling method. It has the highest degree of eliminating bias and thus ensuring representativeness of the population. It is the default sampling design assumed by most software packages that are used in quantitative social science. However, one of the biggest caveats of simple random sampling is that it requires the existence of a sampling frame from which the sample should be derived. The **Sampling frame** is the list of all possible elements of the population. As you can imagine, organising this list can be time consuming and costly especially for small studies. Not to say that in some cases this comprehensive list may not even exist. In larger studies involving face-to-face interviews in extended geographical areas, simple random sampling may actually be impossible due to again time and resource limitations. In such cases, there are other probability sampling methods that can be utilised.

Systematic random sampling

Systematic random sampling is a type of probability sampling that involves the selection of every i th member of the population into the sample. Randomisation or randomness here is introduced in the selection of the first unit (i th unit) and after that a fixed sampling interval is followed until the desired sample size is reached. The sampling interval and selection of the first unit are determined by the **sampling fraction**, which is calculated by dividing the sample size by the of total population:

$$\text{Sampling fraction} = \frac{\text{Sample size}}{\text{Total population}}$$

The sampling fraction guides the sampling selection process. For example is the sampling fraction is $\frac{1}{4}$ then the starting unit should be randomly selected within the first 4 units and incremented by selecting every 4th unit of the population. The random aspect of this sampling method is in the selection of the first item of the population.

The importance of this methods mainly lies in the fact that it is easy and less costly. Also, the sampling process spreads the sample evenly across the population. It is more suitable in the collection of data from geographically disperse populations and if implemented properly, it can approximate the results of the simple random sampling method. Again just like in simple random sampling, it has a caveat of requiring a sampling frame, which is not always feasible.

Stratified random sampling

Stratified sampling method is a type of probability sampling method involving the division of the population into smaller sub-populations called strata and selecting elements from each strata in a proportionate fashion. This method is often used to obtain a representative sample from the population which does not constitute a homogeneous whole. In stratified sampling, we are able to represent and get more precise estimates of each stratum and if each stratum is correctly represented, we can provide a better estimate of the whole population of interest. The procedure for stratified sampling usually begins with determining the proportion of the total sample size to be drawn from each stratum. Followed by randomly selecting elements from each stratum. For example if we are interested in studying public perceptions on violence against women and we have anecdotal evidence suggesting that there are gender differences in public perception, we may want to stratify our sample in order to have an equal proportion of women and men. We will start with determining the proportion of women and men to be selected in our sample (in this case 50% women and 50% men) and then we randomly select the required number of women and that of men. Stratified sampling does not only reduce sampling error but also allows for more representation.

Cluster sampling

We have seen in the first and second sampling methods that they require a complete list of all possible cases in the population, which is sometimes unattainable due to the total area of interest being too large. In such situations, the researcher can opt for cluster sampling. This sampling method typically involves dividing the total population of interest into separate and relatively small groups known as clusters. Then a random sample of clusters is selected with the ultimate sample consisting of all the units in the sampled clusters. In cluster sampling, the study is only concentrated in the sampled clusters and this compromises precision as compared to simple random and stratified sampling. But cluster sampling has the advantage of time and cost-effectiveness especially in large studies. **Area sampling** is the most popular version of cluster sampling whereby a geographical area is sub-divided into primary sampling units (PSUs) which are considered as clusters. The sampling process in this case would involve randomly selecting PSUs and considering all the units falling within the selected PSUs as the ultimate sample.

Multistage cluster sampling

Multistage cluster sampling is an extension of cluster sampling and it contains two or more stages in the sampling process. In other words, multistage basically means the division of large clusters into smaller clusters at several stages to make data collection more feasible. Multistage cluster sampling is complex and may also compromise precision but it may be the only possible methods of sampling in larger studies which extends to big geographical areas like national surveys. In this case, clusters can be provinces, districts, constituencies, villages and households. Multistage cluster sampling is easier to administer because the sampling frame is developed in small units (per cluster). It is also cost and time effective in such large studies.

The normal distribution

Now that we are done exploring the kind of sampling method we need to make robust inference, we move to our next question which borders on the justification or rationale of relying on the one sample we have got to generalise to the entire population when this is but just one of the million possible samples that can be generated from the same population. To address this question, I will

develop the idea step by step and if you are able to follow it through it will make some sense in the end.

First of all, there is a simple fact that most of the continuous variables have an idea of normality embedded in them. What I mean is that when you use a histogram to plot a continuous variable, you will discover that the majority of data points will be clustered around the mid point with fewer points on the opposite ends of the spectrum. Take age from any dataset for example, and you are going to discover that most of the people in a particular are clustered around a given age interval. This age interval of course differs from one dataset to another. But the common denominator will always be that there are fewer people on the opposite tails – there are fewer extremely young babies on one side and fewer extremely old people on the other side and more people are clustered around the average. You can apply this principle to any continuous variable whether it be height, weight, income, life expectancy, gross domestic product (GDP) etc. In some universities is is actually a rule that results of a relatively large class of students should reflect a bell like shape bearing fewer students performing very poorly and fewer students performing exceptionally well.

The behaviour of most continuous variables of fanning out of the center and tailing out in opposite direction is called **normal distribution** or bell shaped distribution. This is a very important characteristic in continuous variables and crucial in inferential statistics. The normal distribution is bell shaped and perfectly symmetrical and its central tendency lies in the middle and at the highest point of the distribution. It is unimodal, meaning that the mean, median and mode lie at the same point. 50% of the sample lies to the left and the right side of the mean. The area under the curve is equal to 1 or 100% and because of this, the vertical segments of the area under the curve can be treated as probabilities.

The properties of the normal distribution makes us calculation on the distribution of the height variable above measured in centimeters between 100 and 200. The area under the curve to the right of the right of 155 tells us the proportion of people who are taller than 155 centimeters and the area to the left shows us the proportion of people shorter than 155 centimeters. For example if the area to the right of 155 is 0.15, we can say that 15% of the people in the sample are taller than 155 centimeters.

However, you should know that perfect normal distribution is hardly attainable in the social sciences. Actually I actually simulated the figure above, it doesn't come from real data that is probably why it is looking unrealistically perfect. Let's look at the following three plots using the variables (height, weight and BMI) from the UK housing market dataset. The three distributions are not exactly normally distributed. The peak points of each of the variables tend to orient themselves at different scales on the x-axis rendering differences in **skewness**. For example, height is negatively skewed meaning that the heights of most people are located near the upper end of the distribution. Weight and BMI on the other hand have most data points located near the lower end of the distribution and thus enlisting positive skewness.

It is not expected that all continuous variables in the social sciences will exhibit a normal distribution. Even if most of them were, they will not be normally distributed in the same way because they are measuring different things and doing so using different scales. In other words there is no one normal distribution but a family of them because they come in different varieties due to differences in means and standard deviations. But statisticians love the concept of the normal curve because of the properties it possesses enabling calculations of statistics and parameters as we would never be able to do without it. As such they invented a baseline normal distribution which would be used as the ideal type of all continuous variables to which any of the bell shaped distributions will be

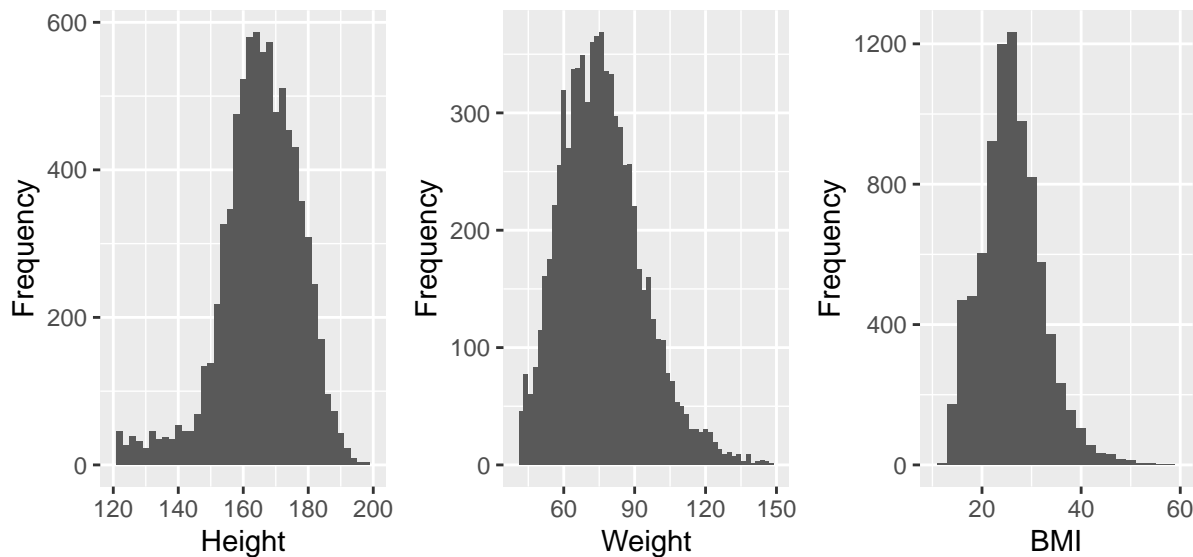
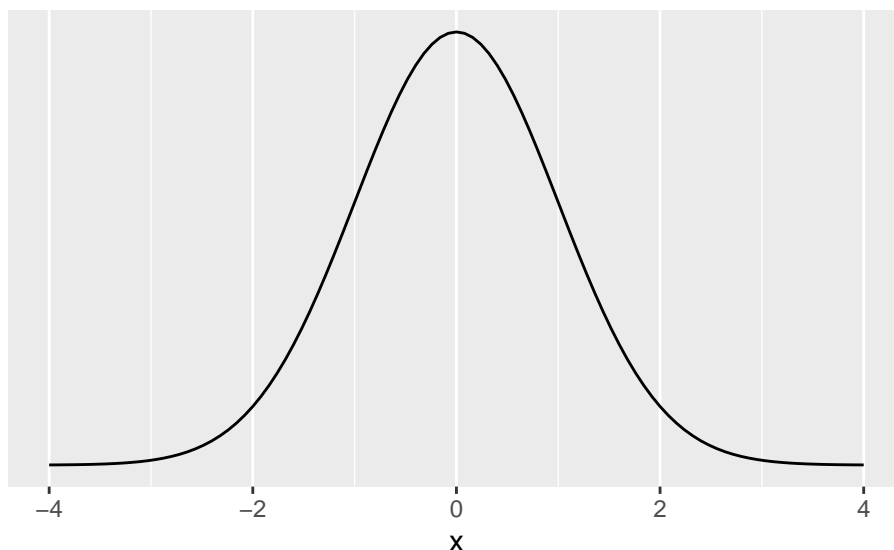


Figure 1: Distributions of continuous variables. Source: UK Housing makert data

compared. It is called the **standard normal distribution**

The standardised normal distribution

The standard normal distribution (also known as the z-distribution) allows us to convert any type normal distribution into a standardised one. We are also able to translate points on all other normal distribution into points on the the standard normal distribution. The standard normal distribution is like the common measurement standardising all continuous variables regardless of the scale or unit it was originally measured in. Whether the variable was about crime rates, housing prices, income levels or age, the standard normal distribution converts them all to a common unit. Each value on any normal distribution will lie exactly on the same position on the standardised normal distribution. The magic of the standard normal distribution is that it has a standard mean of 0 and a standard deviation of 1. This property permits us to calculate how far a value on any normal distribution falls relative to the mean measured in standard deviation. Values on the standard normal distribution are standard normal scores or z-scores. The value that is exactly the mean has the standard normal score or z-score of zero while. Values above the mean will have positive z-scores while those below the mean will have negative z-scores on a standard normal distribution. For example, if a data point has falls exactly one standard deviation above the mean, its z-score will be 1 while those located at exactly one standard deviation below the mean will have a score of -1.



Z-scores are formally calculated by subtracting the mean from each value and dividing the results by the standard deviation. Note that we are talking about the population mean and standard deviation here and not that of the sample. Yes indeed it assumes that we know the values of the population mean and standard deviation. We will shortly discuss what to do in case we don't know them. We use the equation below:

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

where z is the z-score we are seeking, x is the value of interest, μ is the population mean and σ is the population standard deviation. For example, suppose my yearly income from my salary is £36,000 and the mean income of university lecturers is £45,000 with a standard deviation of £15,000. We can calculate my z-score by imputing these values into the z-score equation as follows:

$$z = \frac{36000 - 45000}{15000} = -0.6$$

My z-score is -0.6 which means my income is -0.6 standard deviation away from the mean. My z-score is negative because it is less than the mean as we have already alluded to.

However, locating the position of standard scores relative to the mean is not the only use of the standard normal distribution. It can also allow us to calculate the proportion of cases falling above and below any given z-score. After we have calculated the z-score we are able to find the proportion of cases below and above that score. There is another property of the normal curve which we haven't looked before and it makes this calculation much easier. For every standard normal distribution:

- Approximately 68% of the area under the curve falls between -1 and +1 standard deviation from the mean.
- Approximately 95% of the area under the curve falls between -2 and +2 standard deviation from the mean.
- Approximately 97% of the area under the curve falls between -3 and +3 standard deviation from the mean.

Based on these properties of the standard normal curve, calculations of proportions of the cases has been made easy. For example, if we wanted to know the proportion of cases above the z-score of 2, we would just subtract 0.95 from 1 and divide the answer by 2 to get 0.025 or 2.5%. We have noted already that the area under the curve is a probability and it is equal to 1 and as indicated above, the area under the curve within 2 standard deviation from the mean is equal to 0.95 (or 95%). Therefore to get the proportion above 95%, we need subtract .95 from 1 and we get 0.5 (or 5%). But bare in mind that 5% is on both ends of the standard norm. To get a proportion from one end we need to divide the results by 2. The symmetrical nature of the standard normal curve means that the same principle applies if we wanted to calculate the proportion of cases below 2 standard deviation from the mean and we get the same answer. If instead our interest was on the proportion of cases above the z-score of 3, we would subtract 0.97 from 1 and divide the results by 2 to get 0.015 (or 1.5%).

We have this far just used the proportions that are part of the properties of the normal curve which have been easier to calculate because we have the default areas under the curve values for the z-scores. If we wanted to calculate the proportion of values falling above or below certain z-scores that we have calculated using the above equation, we would have to use the z-distribution probability tables for that. Almost all statistics books come with such tables in the appendix. We don't have them here because we are using R, which can produce the probability values we need or we just Google the tables and use them online. Let's look at the two examples below and find the solutions using the z-distribution tables and R.

Examples

1. If the average crime rates for all council areas in Scotland in 2019 was 545 with a standard deviation of 95 and Glasgow council recorded 401, what is the percentage of Scottish council areas who recorded fewer crime prevalence rates than Glasgow?
2. Highlands council area recorded 296 crimes in the same year. What percentage of council areas in Scotland reported crime rates between Highlands and Glasgow?

To answer the first question, we must start by calculating z-scores for Glasgow council area and then use the Z-distribution table to get areas under the curve to the left of Glasgow z-score, which is calculated by

$$z = \frac{401 - 545}{95} = -1.52$$

For the first question, we find that the z-score for Glasgow is -1.52 which we should know by now that it is less than mean. We look up the area under the curve to the left of -1.52 and we get 0.064 (or 6.4%). We say that 6.4% of Scottish council areas reported fewer crime rates than Glasgow. In question two, we first need to calculate the z-score for Highlands council, given by:

$$z = \frac{296 - 545}{95} = -2.62$$

When we look up the area under the curve to the left of the z-score of -2.62 and we find 0.0044. To get the percentage of Scottish council areas between Glasgow and the Highlands, we subtract the value associated with the Highlands score from the Glasgow score (0.064-0.0044) and we get 0.0596. Therefore there 6% of Scottish council areas whose crime rates lie between those of Highlands and Glasgow.

But since we are working with the powerful and versatile statistical programming language, we can actually calculate the areas under the curve using R using the `pnorm()` function in R. This function gives us the cumulative density function for a normal distribution. That is to say it will give us the probability of a distribution being equal to or less than a given z-score (in other words, the area under the curve to the left of the z-score). We just need to provide our z-score as an argument in the `pnorm()` function. Since we already know that R can work like a calculator, let's start by calculating the z-scores in our two examples, Glasgow and Highlands in R and then input the results into the our function:

```
score1 <- (401-545)/95
score2 <- (296-545)/95
```

Here I am using the knowledge of R that we have already learnt in Chapter Three by creating objects and naming them `score1` and `score2` to store our z-scores for Glasgow and Highlands respectively. We use brackets because we want the enumerators to be calculated first before the division is done (if you still remember that basic arithmetic principle you learnt in high school). You will get very strange results if you forget that small piece. In the next code, we supply those names in the `'pnorm()'` function to create our cumulative density function (CDF). We can see that apart from the rounding off, our area under the curve to the left of the z-scores are pretty the same as got from the probability table. We subtract the CDF of the Highlands from that of Glasgow to get the percentage of council areas falling between Glasgow and the Highlands.

```
pnorm(score1) # probability for Glasgow council
## [1] 0.0647863
pnorm(score2) # probability for the Highlands council
## [1] 0.004382936
```

Sampling error

In standard normal distribution, we were using one sample that we have gotten from the population to make statements about the population using the area under the curve. However, if we are making statements about the population using the sample, we can only be justified to do so if our sample statistic is exactly the population parameter that we are trying to estimate. But since we don't know the population parameter (we didn't study the whole population), we cannot be sure about how far away our sample static is from our population parameter. The difference between the value of our sample statistic and that of the population parameter is called **sampling error**. Sampling error causes the value of the sample statistic unlikely to be the same as the population we are estimating and that is a big problem.

We have already indicated that at any point, the one sample that we have collected from the population is just one of the infinite possible samples that we could have collected from the same population. Sampling error comes about if the sample we have collected is not exactly representative of the population from which it was drawn from. To solve this problem, we have to know what our population parameter looks like (of course in terms of approximation) and we need to have a way of figuring out how far our sample statistic is from the population and the probability that it could be wrong by various degrees. Once we know these things, we will know how much confidence we need to have in the results from our sample. Two more statistical inventions will help us in that endeavour, and they are **sampling distribution** and the **central limit theorem**

Sampling distribution

By now we know how to calculate statistics from one sample like the mean and standard deviation. Lets try to engage in a mental experiment to illustrate the concept of sampling distribution. Supposed we drew one random sample of 100 from a student population of 1000 and asked them how much money they earn in a month. We are going to end up with 100 answers giving us different income earnings of our sample group of students. If we calculated the mean of those 100 earnings we will have a single number representing the mean. If we then took our sample back into the population and then drew another sample of the same size (yes! with the possibility of some students from the first sample belonging to the second) and we calculated the mean of the second sample. The means from the two samples are likely to be different, but we would expect that they will be pretty similar. There is a chance, albeit a small one that the second random sample you have collected would have a sample mean that is very different from the first one. Now if drew every possible random sample from our population and calculated their means, we would end up with quite a large chunk of them². A frequency distribution summarising the means of our repeated samples of the same population is called the **sampling distribution of the mean**.

Like all other distributions, the sampling distribution has the central tendency (mean, median and mode), spread (standard deviation, variance) and shape (distribution). The nature of central tendency, spread and distribution of the sampling distribution give it its important properties that are crucial in statistical inference:

1. The mean of the sampling distribution is exactly the mean of the population. In other words, the mean of means from repeated random samples of relatively large sizes would be equal to the population mean. This gives us a way of approximating our population parameter (remember one of the issues we said we needed to solve in order to be confident of the results from the sample). We use probability for this and the only question we need to ask is “what is the probability of obtaining a sample static like ours in the sampling distribution?” If that probability is small then we wouldn’t be too confident that our sample statistic represents the population parameter. Don’t worry if this idea sounds rusty, we will come back to it repeatedly in this book.
2. The standard deviation (which is the measure of spread) of the sampling distribution is called the **standard error**. The standard error is dependent on the standard deviation of the population and the sample size. If the standard deviation in the population is larger, the standard error is equally expected to be larger and vise versa. If the sample size is small, the sample means are likely to have larger spread and this will be reflected in the standard error.
3. The sampling distribution is normally distributed. This simply means that if one draws samples of the same size repeatedly from a population and then computes the mean of each of the samples, the distribution of the means which is called sampling distribution would be approximately normal. The importance of this is highlighted below in a related concept that is equally important for inference

The central limit theorem

We have seen how magical the standard normal distribution is and how much we can accomplish with it. The only caveat with it of course is that the assumption of normality of the variable of interest has to hold for it to work. Of course not all continuous variables in our population of

²In fact, not a large chunk but you can do this infinite number of times

interest are normally distributed and this will cause problems in us using the properties of the standard normal curve for inference. Enter the **the central limit theorem**. It states that as the sample size approaches infinite, the sampling distribution of an infinite number of random samples of equal size selected from the population will approach normal distribution. The central limit theorem holds even if the values of the variables itself do not have a bell-shaped histogram.

As stated, the central limit theorem is important because it allows us to leverage the properties of the standard normal distribution even if the variable of interest is not normal. We don't usually know the distribution of the population, the only thing we will have is a random sample (for instance we have only interviewed 1200 people out of 5000 members of the population). But because we know from the central limit theorem that the means of the sample are normally distributed, we make use of the known values and probabilities associated with the z-distribution. We know that our sample mean is just one of the many possible sample means that make up the sampling distribution. To make inference, therefore, we only need to know how close our sample mean is to the population mean (which in this case is the mean of the sampling distribution). To do this, we take our sample mean and look at its relative position on the standard normal curve. Note that this time, the standard normal curve is acting as the sampling distribution with a mean that is equal to the mean of the population. If our sample mean is close to the mean of the sampling distribution, we will have greater confidence that our mean represents the population. The reverse is also true in that if it is far away from the mean of the sampling distribution we will have less confidence that our mean represents the population mean.

One of the basic limitations with the central limit theorem though is that of the sample size. For the central limit theorem to kick in, we need to have a sufficiently large sample. The central limit theorem will not work well with a small sample. The question is how large is the sample needed to be for it to be sufficient? There is no straightforward answer to this question but the rule of the thumb in statistics is 30 observations.

The question is what happens if the sample size is small? When the sample size is small we resort to using the t-distribution instead of the normal distribution. We will understand why this is the case in the next chapter. The t-distribution was invented by William Gosset, who published under the pseudonym "student" hence it is sometimes referred to as student's t-distribution. The t-distribution varies according to sample size. In fact there is a family of t-distributions whereby a different distribution for each sample size. The spread of the t-distribution is larger than the normal distribution for small samples but as the sample size becomes large, the t-distribution becomes closer to the normal distribution. It is because of this property of the t-distribution that it is often preferred in place of the normal distribution. We will also use it mostly in later topics in this book.

Conclusion

We have covered quite a lot in this chapter. The overarching idea is to provide the statistical theory of inference. Basically given the fact that at any one time, we will only have one sample which we must use to learn about our population of interest. In this regard, we raised two questions: what kind of sample do we need to make inference? The answer is a random one obtained by random sampling. We then explored different types of random sampling methods and the circumstances in which we would use them. The second question was the longer one. Given the fact that we can only have access to one sample and that our sample is just one of the samples that can be obtained from the population, how do we go around making inference. The answer is long but it is anchored on the sampling distribution and the central limit theorem which allow us to compare the one sample

we obtained with the population sample represented by the mean of the sampling distribution. If our sample is far from the mean of the sampling distribution, we can conclude that is unlikely to represent the population mean and if it is closer to the population mean we have more confidence that it represents the population mean.