Foundations of Medical Statistics

Statistical Inference 1: Introduction to Statistical Inference

1.1 Population and Sample

Much of statistical inference is concerned with making statements about properties of populations, based on properties of samples from the populations. Before considering the mechanisms by which this is done we need to understand what is meant in statistical terms by the terms *population* and *sample*.

In this session, and later sessions also, we will tend to assume and work with relatively simple concepts of population and sample, in order to facilitate our introduction: we will imagine that a population is a very large number of 'objects' contained in a large urn, from which we can randomly sample a relatively small number of the 'objects' at a time.

However, before we pursue this simplified picture, it is necessary briefly to point to some of the real-life subtleties and problems we will be ignoring, but which can affect a more complete account of the inferential process:

- Is the sample representative of the population?
  Clearly a sample can be chosen in may ways, and the way in which we are able to make inferences about the population depends critically on the way in which the sample is selected: it is hard to over-emphasize the importance and relevance of the sampling process to the meaning and validity of the subsequent inferences.

- Is the population well defined?
  A population can be a well-defined group of people, corresponding even to the colloquial meaning of the word: for example, all females aged between 15 and 30 years resident in the UK on a certain date; but in some situations defining a wider population can be problematic, both in time and space. Can we generalise geographically, for example, or only within a region or subgrouping? Can we generalise to other (including future) times, or only to the time period sampled?

- Is the population finite, or (effectively or potentially) infinite?
  For example, a study of a new treatment for a disease may wish to generalise to all potential patients.

- Does the population we wish to generalise to concern an intervention (e.g. when studying response to treatment in a randomised clinical trial), or is the sample used only observationally?

- Have we sampled **all** the population?
  For example, a study of leukemia in the years following a leak from a nuclear power station may sample all subjects developing leukemia within the relevant time period in the vicinity of the power station. In such an example it is not

clear how to define a wider population from which the sample can be considered to have been drawn. In these and other cases one approach is to consider a notional or counterfactual population, which can only have a conceptual existence.

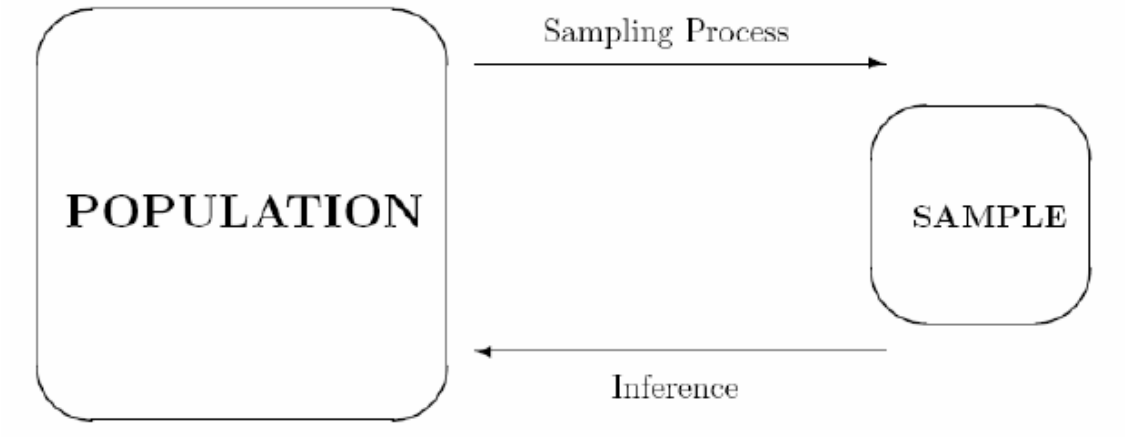
In general the issues can be complex and will not be considered further here.

1.2 Sample and Statistic

For now we can put aside the above complications, and assume that our population is something like a very large collection of stones in an very large urn, from which we are going to randomly take relatively small samples from whose measurements we are to infer something about the whole population. It is easier to visualise the process in terms of simple objects rather than complex human subjects, but the underlying picture is essentially the same. We may want to make inferences about very simple features of the population, such as a mean or proportion, or about the difference between the means of two groups into which the population can be subdivided, or about more complex features involving relationships between several variables. In all cases, inferences will be based on **statistics** calculated from the sample: a statistic is any quantity that can be calculated from the *known* measurements on the sample data. For an example, an intuitively 'obvious' statistic with which to estimate the population mean is the sample mean (in later sessions we will be able to give some justification for the most appropriate choice of statistic to use for a given inferential purpose). The unknown population quantity we wish to estimate is known as the population **parameter** (usually denoted with a Greek letter); and the statistic used to estimate the parameter is known as an **estimator**.

Before we consider the general way in which arguments are constructed in statistical inference, we should note at this point that there is more than one framework for statistical inference. The traditional and most widely used approach is termed the "classical" or **frequentist**, and this is the one pursued here. An important alternative, the **Bayesian** approach, is growing in influence and will be met later in the MSc course.

The basic structure of frequentist inference can be represented diagrammatically as follows:



To illustrate the process, we will imagine that our population is a very large collection of stones, and that we wish to estimate the mean weight μg of the population of these stones. Consider the extreme situation where although the population weight is unknown, we are certain that all the stones weighed the same: the variability in the population was zero. In this situation we could sample just one stone, observe its weight, say x — that is the top arrow in the diagram of the inferential process — and then we infer back — the bottom arrow — that the population weight $\mu = x$g, with absolute certainty. Notice that our logic here implicitly depends on what would happen if we were to take further samples from the population: these would add no additional information.

Suppose instead we have a more realistic situation where we knew that there was some small variability in the weight of the stones. In this case, from the weight of just one sampled stone, we could confidently give a small range within which we expect the population mean to fall. Moreover, if we sampled not one but 100 stones, we would expect the population mean to lie even closer to the sample mean; stones with relatively extreme weights would tend to cancel each other out: if we took a very large sample repeatedly, we would expect all the resulting sample means to lie much closer to each other and to the population mean than if we repeatedly sampled just a pair of stones.

Notice now that our intuitive inference is related to would happen if we repeated our sampling: so that if we knew precisely how the repeated sample means would be dispersed around the population mean, we could make precise inference to the population mean. In fact, the distribution of the sample means, and more generally the distribution of the statistic we have chosen to use, is known as the **sampling distribution** of the statistic. This distribution is usually hypothetical in the sense that we don't actually observe it; but if we know something about its form — even if, as above, we can only say that the statistic falls mostly in a certain range — we are able to make some form of inference about the population parameter.

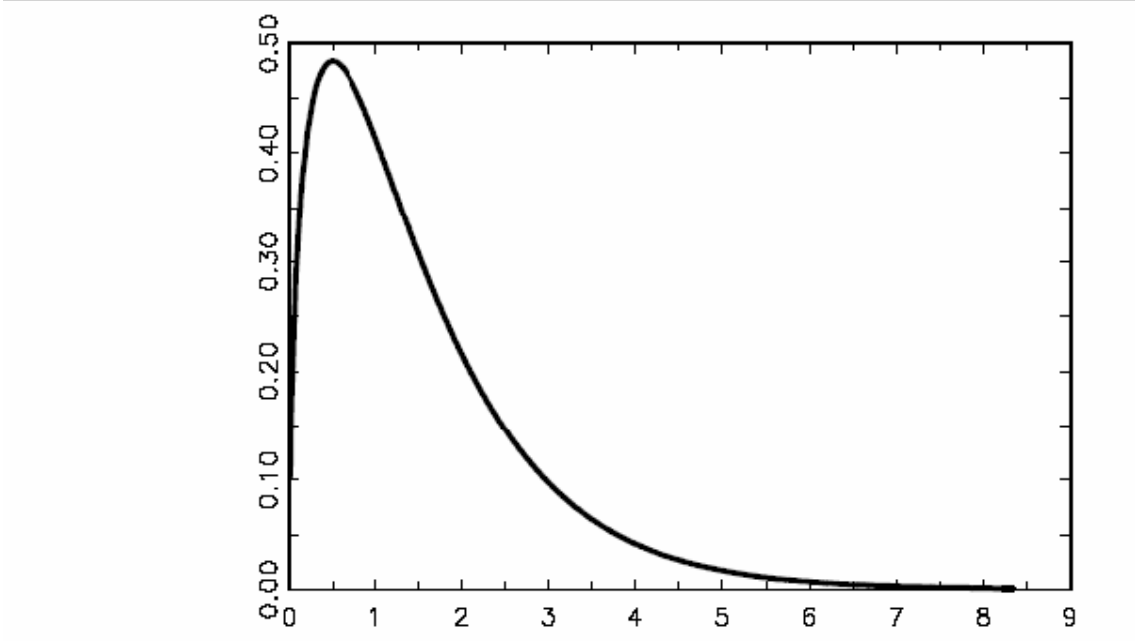
It turns out that in many cases we can describe the sampling distribution of our estimator so well that we can accurately assess our probability of being right or wrong in our estimate: by considering hypothetically what would happen in the long run of a large (possibly infinite) number of samples drawn from the population, we can often precisely calculate the proportion of sample statistics (e.g. sample means) in this long run which will fall within a given range of the population parameter (e.g. population mean).

The key to this process turns out to be the fact that from the variability within a single observed sample (e.g. of 100 stones), we can estimate the variability in the population as a whole; and hence, crucially, we can infer the actual shape of the sampling distribution of the statistic: in the present example, the shape of the distribution of sample means obtained from a large number of repeated samples.

We can review this process with a graphical example.

Suppose that the original population is infinite in size and can be presented by the following distribution (technically the **probability density function**: the area under a given x-axis range of the curve represents the probability that a sampled observation has

a value in that range). This is highly skewed to the right, the sort of distribution that might be produced by enzyme concentrations or economic variables.

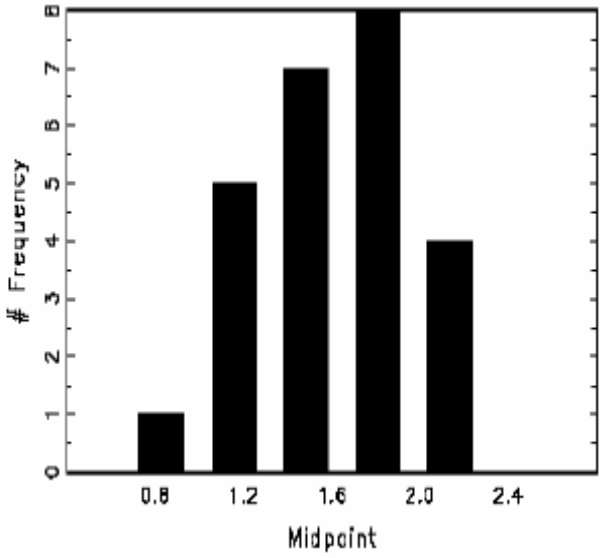


We happen to know in this case that the population mean is exactly 1.5; but suppose we wish to estimate it, and for this purpose we draw a random sample of 10 observations ('objects') from this population:

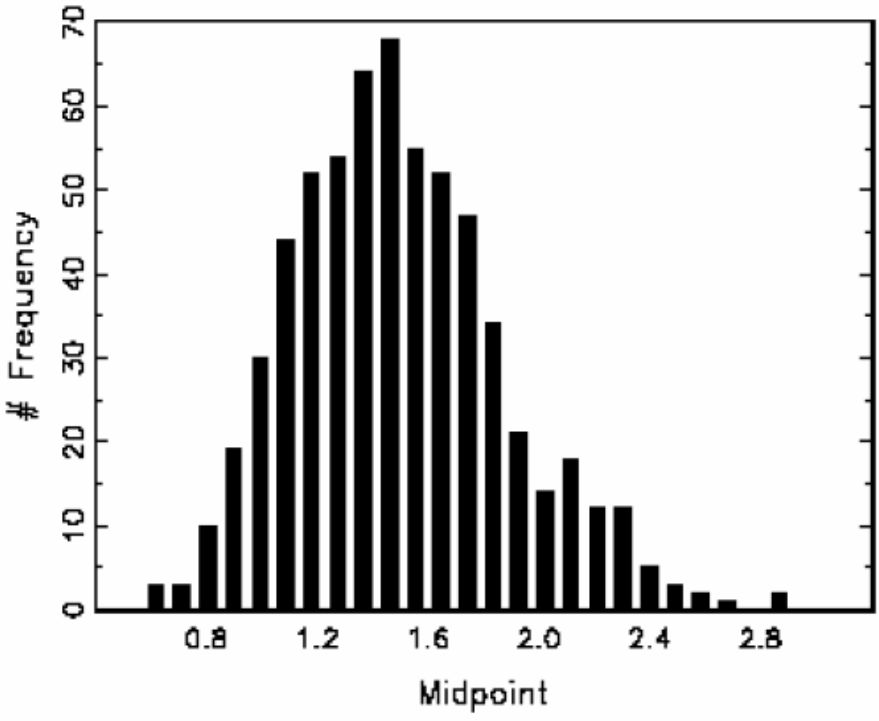
1.72, 1.32, 0.21, 2.31, 0.62, 2.72, 1.85, 0.97, 2.11, 1.92.

This has a mean of 1.575. If we didn't know the true (population) mean, to get an idea of how close this is to the population mean we would need to know the extent of variation of the sample means on repeated sampling: if on repeated sampling the sample means were very narrowly dispersed around the population mean, we could confidently infer the population mean was close to the single sample mean observed; if on the other hand the sampling distribution of sample means was widely dispersed around the population mean, then the population mean could easily be relatively far from the observed single sample mean of 1.575.

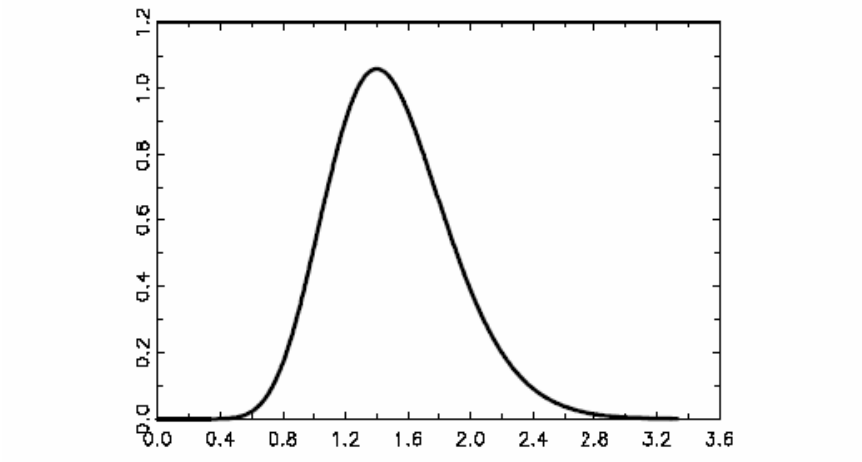
We can investigate the shape of the sampling distribution by taking 25 repeated samples of size 10, and plotting the sample means in a histogram. In most real cases where we need to make an inference we could not do this. Instead we would theoretically derive the sampling distribution (under certain plausible assumptions). In this case, however, we actually draw 25 repeated samples and plot the histogram:



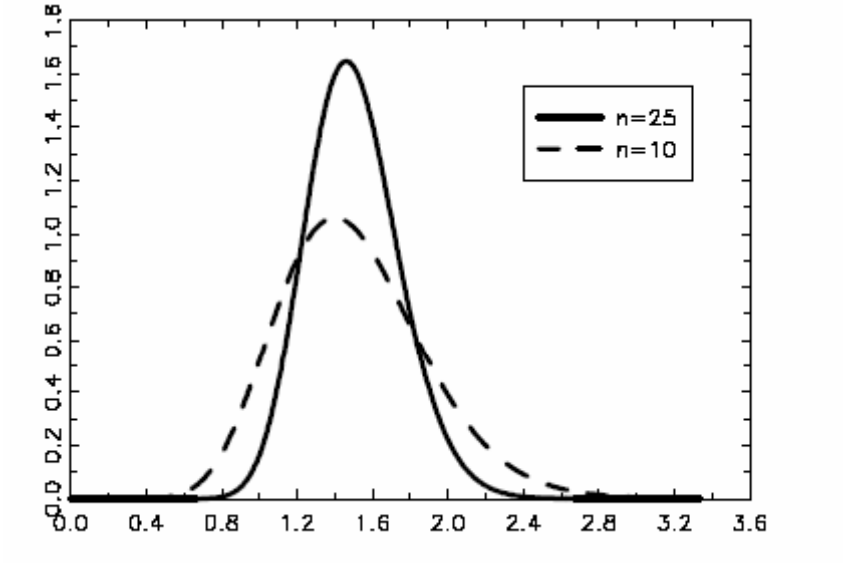
As we increase the number of samples we get a better picture of the sampling distribution of the mean; 625 samples produces the following histogram of sample means (but don't confuse the number of repeated samples, 625, with the number in each sample, 10):



In the limit, as the number of samples gets very large, the sampling distribution takes on its exact form:



This distribution tells us how the sample mean will behave over a large number of repetitions of the sampling process. It is an **operational** definition: the behaviour of the sample mean is described in long-run terms. For example, this distribution has the same mean as the original population. Hence, on average, the sample mean will be centred on the right value. The spread of the sampling distribution tells us how the sample mean will vary about the true value. We can say that on average it has a certain probability of being within a given distance of the mean. A "narrow" sampling distribution implies that the sample value is likely to be close to the mean of the sampling distribution; and in this respect, as we suspected, larger samples are better than small: their resultant sampling distributions show less variation. Suppose that the sample size in the present example is increased to 25, then the sampling distribution of the mean looks like this, compared with that for a sample size of 10:



You will also notice that the sampling distribution for the larger sample size is more Normal in shape. This is an example of the **central limit theorem**, which plays an important role in many parts of the course: informally and without rigour, the central limit theorem tells us that, no matter what the shape of the distribution of observations in the population, the sampling distribution of the statistic derived from the observations will tend to Normal as the size of the repeated samples increases.

Thus in the present example, since the population distribution is very skewed, we expect each individual sample to be skewed; but (increasingly as the size of the samples to be repeated is increased) the mean of each skewed sample tends to be as likely to fall above as below the population mean; and the probability of a sample mean having a given value decays exponentially the further that value is from the population mean; in other words, the sampling distribution tends to the symmetrical Normal shape. And this will apply, within reason, to any statistic derived from the observations in the sample.

To take another example, suppose the statistic we wish to derive is the proportion of stones in the sample with weight above a certain value: we wish to use this to estimate the population proportion. If we sampled repeatedly, the sampling distribution of the proportion will tend to Normal: the sample proportions will tend symmetrically and exponentially to fall on either side of the population proportion, provided the sampling is random.

In exceptional cases, however, the central limit theorem will not apply: where the statistic cannot be expressed as the sum of independent random variables (an example would be the sample maximum).

All statistics have sampling distributions, and it is through these distributions that we are able to relate a statistic to its corresponding population quantity and, most importantly, how "close" a statistic is on average to this quantity. This can guide us in the choice of appropriate statistic, given the features of the population that are of interest, and help us to choose between alternative statistics.

To summarise the process of inference from the sampling distribution:

1 We draw a random sample of size *n* from the population.
2 We calculate an appropriate statistic from the sample, in order to estimate the desired population parameter.
3 From this single sample we also need to determine the shape of the sampling distribution of the chosen statistic: we need to quantify how this statistic would be distributed if we were to resample infinitely.

> Note that although the statistic does not depend on any unknown quantities, this may not be true of the sampling distribution. Thus in the example of the stones above, it would appear that in order to know the precise sampling distribution of the mean, we first have to estimate the population variance using the variance in the single sample, and then from the *estimated* population variance derive the sampling distribution. However, as we shall see later in the course, by using the t-statistic (introduced in a later session) we can circumvent this problem, since provided the population is Normally distributed the sampling distribution of the t-statistic does not depend on the population variance.

4 Once we have the precise sampling distribution, we can quantify how close our statistic value is likely to be to the population parameter.

1.3 Sampling Methods

A sampling distribution can only describe the correct behaviour of the corresponding statistic if it is derived from a sampling process whose properties are known. The most common and best-known example of this is **random** sampling. In its most basic form this implies that all members of a population are equally likely to appear in a sample. In the example of the stones above, if there had been some malfunction in the way we sampled the stones so that only larger stones were able to be drawn, we would not expect the sample means to be centred around the overall population mean, nor the variability in the sampling distribution to be as predicted from unrestricted sampling. Such simple random sampling allows sampling distributions to be derived in a comparatively simple way, and inference procedures tend to be at their most straightforward in this setting. Such samples are easily generated using random number tables, or pseudorandom number generators on computers. The analogous situation in an intervention trial, or experiment, is the random allocation of treatments to subjects. All possible allocations of treatment are equally likely, and again sampling distributions can be obtained using simple rules.

Unfortunately (from a simple statistical point of view) many problems involve populations that have **structure**, for example groupings such as district or country. It may, for example, be necessary to sample randomly within such groups, but have the probability of selection differing among groups. It is not especially difficult to modify random sampling schemes for such structures, but this must then be reflected in the way the sampling distributions are derived. Situations like this are of particular relevance in the Unit on the Analysis of Hierarchical and Other Dependent Data.

There are other forms of sampling process that are not strictly random. Such systematic samples usually present problems in the analysis, in that it is difficult to justify the sampling distributions of the statistics (it may be difficult or impossible to calculate relative probabilities of being selected for sampling). Additional assumptions may be necessary. So-called "judgement samples", where some element of human judgement enters into the sample selection, are particularly notorious examples of this, from which it is almost impossible to justify statistically valid (*i.e.* generalisable) conclusions.

1.4 Statistical Models

To elucidate the information about population quantities contained in sample statistics we need a precise and formal description of the whole sampling process from population to sample. This description is called the **statistical model**. Relevant features of the population are represented by **parameters**, such as the mean, variance, or correlation. The structure of the population, together with the sampling process, allows a model to be formulated that describes the statistical behaviour of the sample, which in turn allows us to postulate sampling distributions for the relevant statistics.

The basic steps by which a population/sampling setup is formulated as a statistical model are described later in this course. Again the role of the sampling process needs emphasizing. If this is changed, then the statistical model needs to be changed accordingly. If the sampling distribution of a statistic is derived with the wrong assumptions about the sampling process then it is likely that the inferences derived from that statistic will be invalid.

The crucial importance of the statistical model is that, given a certain value of the population parameter (in the simple case where there is only one parameter of interest), it allows us to calculate the probability of drawing a sample with the properties we observe: this will allow us to quantify the compatibility between the observed data and possible values of the population parameter.

1.5 The Tools of Classical Inference

1.5.1 Estimation

We have already touched on **estimation** in the example above using the population mean. The aim is to calculate a value, from the sample, that is as close as possible to the unknown population parameter. We distinguish between the **estimator**, the general form of the statistic as calculated from any sample, and the **estimate**, the actual value the estimator takes when calculate for a particular sample. Two key properties of an estimator are (1) the **bias** and (2) the **precision**; both are obtained from its sampling distribution of the estimator.

1. The bias is the difference between the mean of the sampling distribution — the expected or average value of the estimator — and the population parameter being estimated. A small bias ensures that, on average, the estimator is centred in approximately the right place. Bias in itself is not a major problem, provided it disappears as the sample size increases. Many common estimators are biased in this sense. A more serious issue is bias that does not disappear with increasing sample size: this implies that the estimator is **inconsistent**. This is often associated with problems in the sampling process. For example, with simple random sampling the sample mean is an unbiased estimator of the population mean, but if smaller observations are less likely to be selected then the sample mean becomes an inconsistent estimator: it will be biased upwards.

2. The precision of an estimator is measured by the variance or standard deviation of the sampling distribution (in an inverse sense: low variance means high precision). The standard deviation of the sampling distribution is termed the **standard error** of the estimator. If the estimator is the sample mean then there is a simple relationship between the standard error of the sampling distribution and the standard deviation of the variable:

$$\text{true standard error of the mean} = \frac{\text{true standard deviation}}{\sqrt{\text{sample size}}}.$$

This result is a very important because if we use the sample standard deviation as an estimate of the true standard deviation, then the data from a single sample can be used to obtain not just an estimate of the population mean, but also an estimate of the precision of our estimate!

Note:

- The formula above is *not* a definition of standard error, but a formula for calculating the standard error of the **mean** of a sample of a given size. The general definition of standard error is, as given above, the standard deviation of

the sampling distribution – hence the square root of the variance of the sampling distribution.

- Although the standard deviation of an estimator, in contrast to the standard deviation of a variable, is given this special term, standard error, the term variance may be used to refer both to the estimator (as the square of its standard error) and to the square of the standard deviation of a variable. This should not cause confusion since it is usually clear from the context in which the term is used whether what is meant is the variance of an estimator or of a variable.

In simple settings the choice of estimator may be self-evident, for example a mean or proportion. In more complex settings there may be several alternatives with different advantages. Compromises may have to be made between bias and variance, or between simplicity and complexity, for example. The choice of estimator may also depend on the purpose to which the estimate is to be put. We may for example use different estimators for prediction, than for discrimination.

1.5.2 Confidence Intervals

An estimate, provided by the sample data, for an unknown population parameter is sometimes referred to as a **point estimate**, in contrast to a **confidence interval**, also estimated from the sample data, and which is a region around the point estimate which contains the population parameter with a certain conventional probability. Thus the confidence interval is a measure of how precisely we have estimated the parameter: the narrower the confidence interval, the more precisely we have narrowed down the range in which we have confidence that the parameter will lie.

The confidence interval consists of a pair of values (L, U), the confidence limits. This pair, being a statistic calculated from the sample, has a sampling distribution: each time we repeat the sampling process another pair of confidence limits is generated. The probability with which the interval contains the parameter, and which determines the **level** of the confidence interval (e.g. 95%), is based on this sampling distribution.

The confidence level, expressed as a percentage, is defined as the probability that the sample based interval contains the true population value: in other words, the proportion of such intervals, obtained from repeated sampling in the long run, which contain the population value. Conventionally, 90%, 95% and 99% intervals are often used in practice. Note that again this is an **operational** definition based on long run behaviour.

The appropriate confidence interval limits to achieve a given probability can be calculated from the sampling distribution of the estimator using the following argument. Suppose that the sampling distribution of the estimator implies that 95% of samples yield an estimator not more than a below the true population value μ, and not more than b above it. Then if our sample estimate (the value of the estimator for that sample) is m, we expect that in 95% of samples $m > \mu - a$ and $m < \mu + b$; or, conversely that $\mu < m + a$ and $\mu > m - b$, i.e. $m - b < \mu < m + a$. For symmetrical sampling distributions, $a = b = c$, so then $m - c < \mu < m + c$, or, as it is often expressed,

$$95\% \text{ confidence interval for } \mu = m \pm c.$$

In this case there is a 2.5% probability that upper limit falls below μ, and also a 2.5% probability that the lower limit falls above μ: in either case the interval fails to contain the true value.

When a calculated interval is wide, this implies that the data have provided little information about the parameter (given the statistical model and analysis); while a narrow interval results from the data giving more precise information about the parameter. Another way of expressing this is to consider that for a wide, say, 95% confidence interval, a wide range of parameter values are *supported by* or *consistent with the data at this level*; while for a narrow interval only a small range of parameter values are supported by the data at this level.

1.5.3 P-values

The confidence interval defines a region of parameter values consistent, at a given conventional level, with the sample data. However, parameter values within this range are not equally supported by the data: intuitively we feel that possible values of the parameter close to the sample estimate are better supported by that particular sample; and often there is interest in quantifying the support the data give to individual parameter values.

Consider, for simplicity, a symmetrical 95% confidence interval (CI) for $\mu = m \pm c$; and let's hypothesize that the true value of the parameter were actually $\mu_0 = m - c$. In that case, we know that there is a 2.5% probability, on repeated sampling, that we would obtain a sample estimate greater than μ_0 by c or more; and also a 2.5% probability that we would obtain a sample estimate lower than μ_0 by c or more: thus 5% of samples would yield an estimate which is as far or even further from this hypothesized value. In other words, 5% of possible samples would give as little or even less support to the hypothesized value μ_0 than the sample estimate we actually observe, m: of all the possible results we could have got, under the hypothesis that $\mu_0 = m - c$ is the true value, only 5% would be as or even less probable, while 95% would be more probable; and the same reasoning applies were $\mu_0 = m + c$.

We can use this idea to quantify the support given by sample estimate m for a given parameter value μ_0: conditional on this being the true parameter value, the proportion of results from repeated sampling (sample estimates) which would give the same or less support for this parameter value than the observed estimate m. This proportion is known as the **P-value**, from the sample data, for the hypothesis that $\mu = \mu_0$.

We see then that from the definition of the 95% CI we know straightaway the P-value for two hypothesized parameter values: the two confidence limits. For either $\mu_0 = m - c$ or $\mu_0 = m + c$ the P-value is exactly P=0.05. And though we cannot infer from the CI the P-values for other parameter values precisely, by extending the reasoning above we know that parameter values outside the 95% confidence interval will have $P<0.05$; while parameter values within the confidence interval will have $P>0.05$. In general, to obtain precise P-values for a given parameter value (and sample estimate), we need to apply the sampling distribution to the particular values involved.

In the next section we briefly discuss the role of P-values in formal hypothesis testing.

1.5.4 Hypothesis tests

When the P-value for a hypothesised parameter value is very small, this indicates that the sample data gives very little support to that hypothesised value; conversely, we can say, as the P-value becomes smaller, that the sample provides increasingly strong evidence *against* the hypothesised value: the smaller the P-value, the fewer the possible sample results which would count as even stronger evidence *against* the hypothesised parameter value.

For example, suppose we hypothesise a value $\mu = 0$, and that our sample estimate for the parameter, $m = -5$, yields a 'small' P-value, say P=0.001, for the hypothesised value. We can conclude, informally, that the sample provides evidence *against* the hypothesised value, and *in favour of* parameter values better supported by the sample: values in the region $\mu < 0$. (Clearly, if the P-value is small enough for us to conclude there is evidence against $\mu = 0$ from the sample $m = -5$, the sample would provide even stronger evidence against parameter values in the region $\mu > 0$, which would yield even smaller P-values.)

How small is 'small'? It is important to remember that the P-value provides a continuous measure of the compatibility between a particular hypothesis and the sample data. However, conventional values, such as P<0.05, are used in formal hypothesis testing, which we discuss in a moment, and these are often used as guidelines for assessing the strength of evidence which a sample provides against a hypothesised parameter value. Note that there is an essential asymmetry here: we can say that at the 5% (P<0.05) level there is evidence against the parameter taking values outside the 95% CI, but we cannot say that there is evidence for the parameter taking any particular value inside the 95% CI.

The need for conventional thresholds such as P<0.05 arises in the context of **formal hypothesis testing** which is essentially a *decision making* procedure. The target of the decision will be a population quantity that typically, for **parametric** tests, can be expressed as a population parameter. Two hypotheses are formulated, the **null** and the **alternative**, though the latter is sometimes only implicitly defined. The null hypothesis, H_0, makes some assertion about the population parameter, for example that it is zero, $H_0: \mu=0$, or less than some given quantity, $H_0: \mu<\delta$; while the alternative hypothesis, H_1, postulates alternative values, for example, $H_1: \mu\neq 0$ or $H_1: \mu\geq\delta$. A **test statistic** is calculated from the observed sample data, and based on the value of the statistic the decision is made either (1) to reject the null hypothesis in favour of the alternative, or (2) to not reject the null hypothesis. In practice the decision is generally based on the P-value for the null hypothesis, as calculated from the sample derived test statistic: if the P-value falls below a certain conventional threshold, and is thus said to be statistically **significant**, the null hypothesis is rejected and the alternative hypothesis accepted; while if the P-value falls above the threshold (indicating less evidence against the null), the null hypothesis is not rejected. The hypothesis testing procedure is thus not symmetric, in much the same way as noted just now: failure to reject the null hypothesis does not necessarily imply acceptance of the null hypothesis. Absence of

evidence against the null value is not evidence that the null value is true. Whether we make this additional step depends on the nature of the problem and the size and nature of the sample and testing procedure.

Two errors can be made when a hypothesis test is conducted, termed type I and type II errors.

1. A type I error occurs when we reject the null hypothesis and it is in fact true.

2. A type II error occurs when we fail to reject the null hypothesis when it is in fact not true (when the alternative hypothesis is true).

Decreasing the probability of one type of error increases the probability of the other, and in practice a compromise is reached. for a given type I error (the **size** of the test, not to be confused with sample size) a procedure is chosen that minimises, at least approximately, the probability of a type II error. The quantity 1 – probability of type II error is called the **power** of the test, and is the probability of rejecting the null hypothesis when the null hypothesis is false: essentially we are looking for procedures with good power for a given size of test. Note that the actual size and power of a test procedure are derived from the sampling distribution of the test statistic under the null and alternative hypotheses respectively: again, the sampling distribution of the relevant statistic is central to the inference procedure. This also implies that when doing a hypothesis test we need to be aware of the structure behind it: doing a hypothesis test is **not** like calculating a descriptive statistic.

Strictly, a test is only valid if the modelling assumptions behind it hold. For example, in a **parametric** test typically the null hypothesis can be formulated in terms of a population parameter whose estimator has a known sampling distribution under certain assumptions. Some testing procedures reduce the dependence of the test on specific assumptions: these include **non-parametric**, **distribution-free** and **permutation** tests. In spite of their names, these techniques are not usually free of parameters or distributions. As in many areas of statistics there is usually a tradeoff between the strength of the assumptions a test depends on and the sensitivity of the resulting power of the test.

Finally, note again the connection between confidence intervals and hypothesis tests: if a 95% CI calculated from the sample 'just' excludes the parameter value μ_0, then the hypothesis test H_0: $\mu=\mu_0$ versus H_1: $\mu\neq\mu_0$ yields a P-value 'just' below 0.05, and if the test has size 5% (i.e. adopts P=0.05 as the threshold) it will just reject the null hypothesis. More generally, if a (100-X)% confidence interval excludes a parameter value μ_0, then the test of the null hypothesis H_0: $\mu=\mu_0$ will reject this null hypothesis at the X% level; conversely, if that confidence interval includes a parameter value μ_0, then the null hypothesis H_0: $\mu=\mu_0$ will not be rejected at the X% level. For this reason confidence intervals are particularly useful following non-significant tests – tests which fail to reject the null hypothesis: how wide the interval is around the null value guides us as to whether we have failed to reject the null hypothesis because it is (approximately) confirmed by the data (when the interval is tight around the null value), or whether we have failed to reject because the data do not give precise enough information (power) to reject a false null hypothesis (when the interval is wide around

the null hypothesis, indicating that values quite different from the null are also compatible with the data).

The next few sessions are an introduction to probability. During these sessions you will be introduced to the Normal distribution, initially in the context of a Normally distributed *population*. However, it is very important to remember that the probability calculations that are most crucial to statistical inference are based on the Normality not of populations, but of the sampling distribution of statistics calculated on samples drawn from these populations.