INFERENTIAL STATISTICS

The material outlined in earlier sections has been concerned primarily with **describing** the distributions of variables and the way in which distributions covary, either via graphical representations or with the help of descriptive statistics. Usually, however, researchers are interested in going beyond a mere description of their data; they wish to make generalizations about their "meaning'. For instance, if I collected anxiety scores of a group of first year psychology students as well as their statistics marks, I might not be satisfied simply to draw a graph of the scores showing how many people obtained scores between 50 and 60, etc, or even to draw a scatterplot of the joint distribution of anxiety and statistics scores. In research, the goal is usually to make some **generalizable inference:** in the present example, for instance, I might be interested in offering an answer to the more general question, "Do highly anxious students obtain worse statistics scores than less anxious people?" If I had reason to expect that people with high levels of anxiety would obtain different scores from non-anxious people, I might form in my mind a **hypothesis,** for instance, the hypothesis that the mean statistics score of highly anxious students differs from the mean score of less anxious people. The task of **inferential** statistics is to test hypotheses of this kind.

Inferring population parameters from sample statistics

In order to determine whether or not first year psychology students with high anxiety scores perform differently on statistics exams from people with low anxiety scores, it would be necessary ,(at least in theory) to obtain the anxiety scores and the statistics grades of all first year psychology students everywhere (i.e., the scores of the population). This is obviously impossible, since it would involve hundreds of thousands of people in many countries. For this reason, researchers interested in testing the working hypothesis stated above are forced by harsh reality to obtain data only with a few members of the population. For instance, I could

not obtain the scores of all first-year students in the world (the population), but I could persuade some students to fill out the anxiety test and tell me their statistics marks. This subgroup of the population constitutes a sample**.** The task of inferential.

A separate but important point should be noted here; If I divide my students into a group of highly anxious people and a contrasting group of low anxious individuals, the anxiety scores of the groups will be fixed **to** some degree, since I put all the high scorers into one group, all the low scorers into the other. It is impossible for a low anxious person to get into the high anxious group, and vice versa. However, the statistics grades can vary freely, depending on the characteristics of the people in the **two** groups. For this reason, statistics scores are referred to as the **dependent variable**. The variable anxiety, which formed the basis of the deliberate division of the subjects into two (or more) groups, is called the **independent variable.**

Statistics may be seen as inferring (estimating) the scores of the population (which we can never know) on the basis of the known scores of a sample.

If I did test the **population, any numerical difference** between the group means would be sufficient to indicate that the mean for anxious people was different from that for non-anxious individuals. Usually, however, **I** have only a sample or, strictly, as in the present example, two samples - a sample of highly anxious people and another sample of low anxious individuals. Suppose that these samples have numerically different means. The question now arises of whether these numerically different **sample** means indicate that the **population** means are different. Suppose that the anxious sample obtained a mean statistics score of 55.60 and the non-anxious sample one of 65.82. The question which inferential statistics can answer is the following: "Is this difference between the sample means so large that I can infer that the samples come from two different populations with different mean statistics scores?"'

Significant Differences

The problem faced by researchers is that of knowing how large the difference between sample means needs to be before it is possible to say that the samples come from different populations. In the present case the question is: How large must the difference between the mean statistics scores of the anxious and non-anxious students in my samples be, before I can conclude that in the population, anxious and non-anxious people obtain different statistics scores? If we know that the difference is, sufficiently large, we refer to it as significant; we speak of a significant difference, or a statistically significant difference. (If sample means are different, but the difference is not large enough to be statistically significant, we can speak of numerically different means.) Generally, however, a mere numerical difference is uninteresting to researchers, unless it is significant. Inferential statistics provides techniques for testing whether numerical differences in samples are significant.

PROBABILITY

The first element in explaining how to determine whether or not differences between samples are significant is that of **probability.** In general the rule is that improbable events are significant.

Subjective probability

The basic unit in probability theory is the **event.** An event might be the achievement of a high distinction in introductory psychology, the winning of a horse race by a particular animal, a temperature of 35°C in Bendigo, my car breaking down during my next trip to Adelaide, a particular student becoming rich in the next 20 years, or whatever. In inferential statistics we are interested in the question: "How likely is it that a particular event will occur?"

Take the example of my car breaking down on my next trip to Adelaide. I do not know what will happen on my next trip, since it has not yet occurred (analogous to the situation with population parameters). Therefore, I must **estimate** the probability of a breakdown. Available to me is

information on the age and state of repair of my car, on the roads between Bendigo and Adelaide, on the distance, on the perfounance of my car on long runs in the past, and so on. On the basis of such experience, I might come to the conclusion that there is about 1 chance in 50 of a breakdown. Probabilities are usually expressed in decimal fractions, in the present example .02. This probability might seem to me to be so low as to be as good as zero, and I might set off on the journey.

Several things should be noted, however:

a. The estimate of the probability is very subjective (i.e., more on the basis of my mood, general level of optimism, willingness to take risks, etc than on the basis of "hard" evidence). Here we are dealing with **subjective probability,** which is often wide of the mark, as in the **gambler's fallacy** (a roulette player, to take one example, believes that because the number 24 has not come up for 500 spins, it must come up on the next spin, whereas the probability on any individual spin is always one in 32).

b. The probability of .02 does not mean that I will experience 0.02 of a breakdown on every trip; I will either experience a breakdown or I will not. The probability simply says how likely it is that this all or nothing event will actually occur on a specific trip. Where the probability is very low, I may conclude that it is as good as zero, although, in actual fact, it will eventually occur. On the occasion when it does occur, I will make an **error** in estimating that a breakdown will not occur. This event is called-error or Type I error.

c. This probability of .02 assumes that only **chance** is at work. For instance, if I already knew that the coil in my car's engine was certain to break down in 300km, the probability of a breakdown would no longer depend upon **chance or random** events, but would be affected by a systematic factor (the certain breakdown of the coil), and the likelihood of a breakdown would not be .02.

Mathematical Probability

Mathematical probabilities are somewhat different from those people experience subjectively. They often run completely counter to the intuitive estimates of, for instance, gamblers.

Consider the following examples: If I throw a coin, it can fall either heads or tails; there is no other possibility. Thus, since it is 50-50 that a particular throw will yield, let us say, a head, the probability of a head is **1** in 2, or .50 (unless the coin has been tampered with, but in this case, we are no longer dealing with **random or chance** events). If a dice is fair (i.e. not loaded), the likelihood of, let us say, a three being thrown **is** the result of chance (it is a random event). The likelihood is 1 in 6 or .17. Even if I had just thrown 10 threes in a row, the probability of a three on a new throw would be .17 (despite the gambler's subjective feeling that it could not possibly be another three).

Several things should be remembered:

(a) An event that is certain to occur has a probability of 1.00 (the probability of each of us dying, for instance is 1.00), an event that cannot possibly occur has a probability of 0.00. All other probabilities lie between 0.00 and 1.00.

(b) The probability of two or more events occurring simultaneously is the sum of their individual probabilities. The probability of **either a head or a tail** occurring on the next spin of a coin is .50 + .50 = 1.00, i.e., it is certain that, on the next throw either a head or a tail will be thrown. The probability of either a three or a four occurring on the next throw of a dice is .17 + .17 = .34.

(c) The probability of a string or sequence of events occurring, **prior to the commencement of the sequence,** is the product of the individual probabilities. For instance, the likelihood of 3 threes in a row, **prior to commencing throwing the dice,** is .17 x .17 x .17 = 0.004913. (However, once 2 threes **have already been thrown** the likelihood of a third is the usual .17.) The likelihood of 10 heads in succession is .00097656, **prior to**

commencement of spinning the coin. After 9 heads have already been thrown, the probability of a 10th head is .50!

In inferential statistics, we need to know the mathematical probability with which certain events will occur, in order to ascertain if various results are significant or not. (Remember: improbable events are statistically significant.) For instance, we need to know the probability of a difference between the anxiety means of our two samples of anxious and non-anxious students being, let us say, 5 points or more. If the numerical difference between the means turns out to \ have-a-low probability, we call it significant. In general, very large differences between means have a low probability

Normal curve and probability

The general nature of the normal curve has already been pointed out. The vital property of this curve, for present purposes, is that the frequency with which any score in a normal distribution occurs is known. This means that we can say that there is a certain **probability** that a particular score will occur, just as with a dice. By adding up the probabilities of all scores below a particular score, we can work out the probability of that score. In any normally distributed variable, a standard score (z score) of +1.00 or lower will occur 84.13% of the time. This is a mathematical property of the normal curve. If you picked any score at random out of a normally distributed variable, the probability of that score being + 1.00 or lower would be .8413. Even more interesting is the fact that scores equal to or greater than +1.00 would occur 15.87% of the time, i.e., the probability of a score greater than +1.00 is less than .1587. (If the probability of a score up to +1.00 is .8413, the probability of a score greater than +1.00 is 1.00 - 0.8413 = 0.1587.) In statistics, **events with a probability less than .05 or .01 are regarded as significant.** Thus, it is important to look at scores on the normal curve which occur with a probability of less than .05 or .01. Figure 11.1 shows such scores.

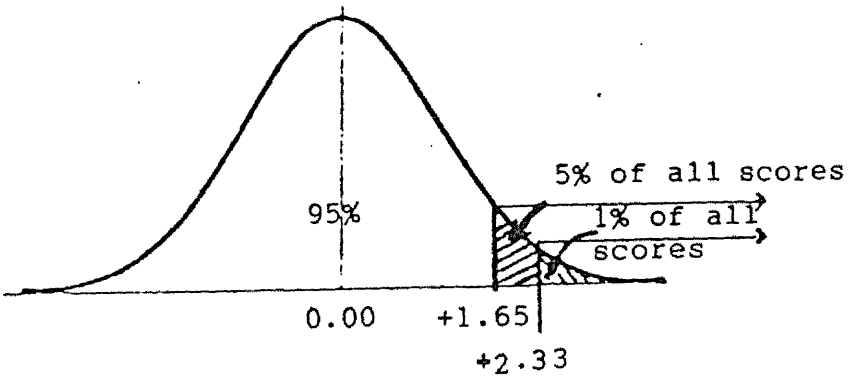


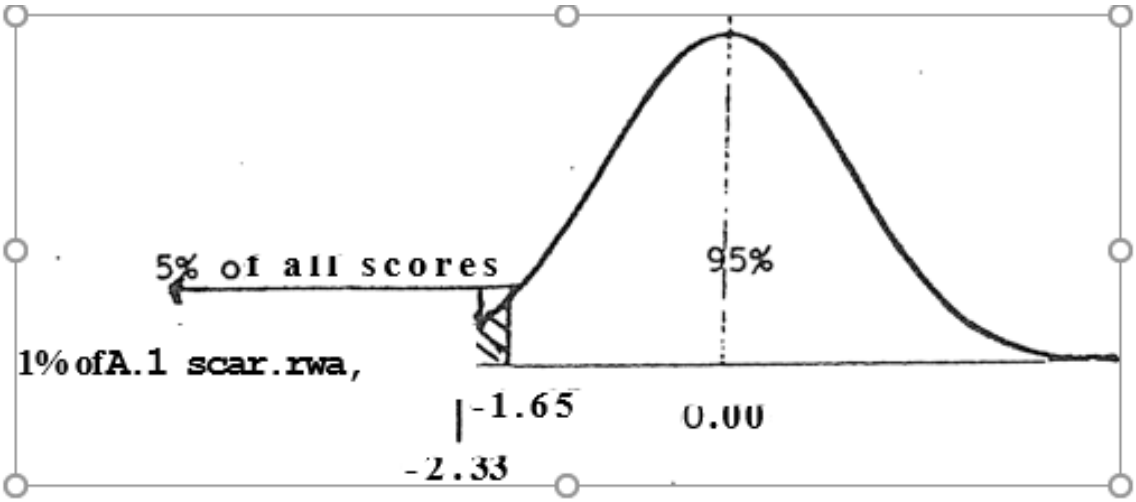
Figure 11.1: 5% and 1% zones on the normal curve

As can be seen, a standard score (often called a z-score) of + 1.65 has a cumulative probability of .95; i.e., 95% of all scores on a normally distributed variable are equal to or less than +1.65. This means that a score greater than +1.65 has a probability less than .05. In a similar way, a score greater than +2.33 has a probability of less than .01.

One and two tailed tests

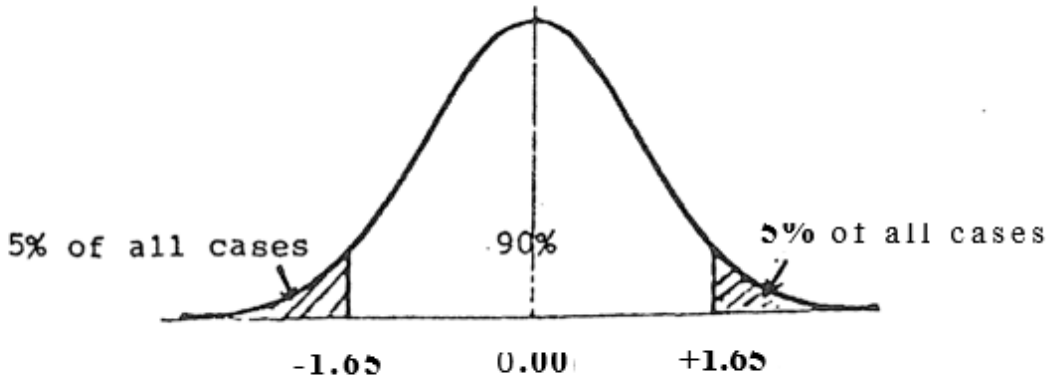
It is very important to note that we have only been looking at scores at the high end of the normal distribution, to date - only one **tail** of the distribution has been considered. The normal curve is symmetrical, i.e., the left-hand tail of the curve, where negative scores are found, has the same properties as the right-hand tail. This means that it is also possible to find scores in the negative range which have low probabilities. In fact, just as scores **greater than +1.65** have a probability less than .05, scores **lower than -** 1.65 have a probability less than .05, scores lower than -2.33 a probability of less than .01 (see Figure 11.2).

Figure 11.2: 5% and 1% zones at the lower end of normal curve



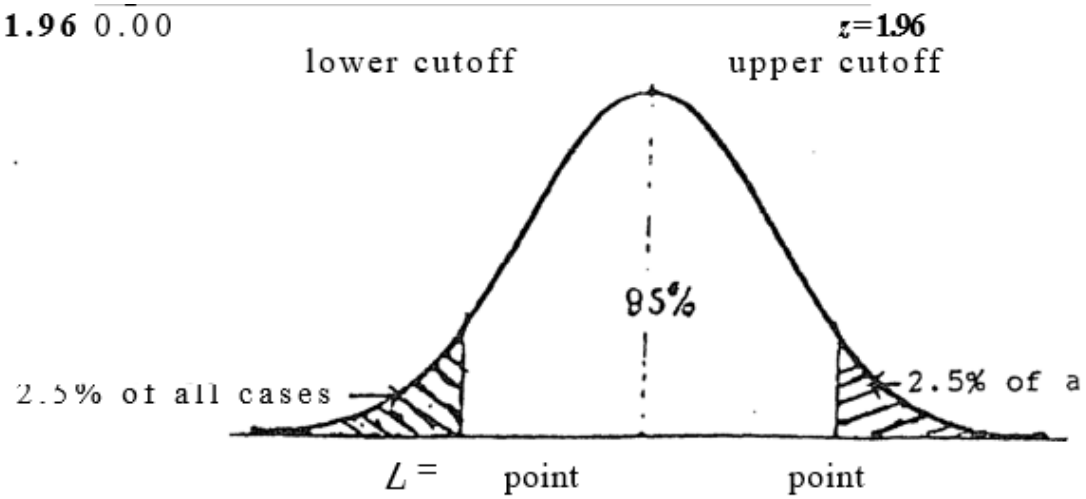
It is possible to say, when referring to the normal curve, that scores greater than +1.65 (i.e., at the right-hand tail) have a probability of less than .05, or that the same scores at the left-hand tail (lower than -1.65) also have a probability of less than .05. These are, in fact, two regions on the normal curve containing scores with a probability of less than .05 - those at the right-hand tail, and those at the left-hand tail (see Figure 11.3).

Figure 11.3: 5% zones at both ends of normal curve



Often, however, we want to know which scores have a probability of less than .05, regardless of which tail they lie at. If you look at Figure 11.3 again, you will see that 5% of all scores are greater than +1.65 (right hand tail) and a further 5% are lower than -1.65 (left hand tail). This means that 90% of all scores lie between -1.65 and +1.65. The interesting question here is which cutoff scores would contain 95% of all scores, in the same way as -1.65 to +1.65 contains 90%. What is required is the scores defining the negative and positive ends of the shaded area marked on Figure 11.4. It should be apparent that, if 95% of all scores lie within this range, 5% must lie outside it; i.e., the total percentage of all scores lying outside the limits (adding both tails together) will be 5%.

Figure 11.4: Two tailed 5% zones on the normal curve



The scores in question are -1.96 and +1.96. A score greater than +1.96, or less than -1.96, has a probability less than .05. Because this cutoff score takes account of both tails, it is referred to as involving a **two tailed** probability. The score which focuses on high positive scores only or on high negative scores only (at one of the two tails) refers to **one tailed** probability. To summarize: using one tailed probability, a z-score greater than +1.65 has a probability of less than .05. The same is true of a score less than -1.65. Using two tailed probability, a score greater than +1.96 or less than -1.96 has a probability of less than .05. These cut off scores, defining a probability of less than .05 or .01 are referred to as **critical values.**

It is important to notice that the two tailed critical values of 1.96 (probability less than .05) and 2.57 (probability less than .01) involve z-scores. If you think back to earlier sections you should recall that z-scores are calculated by calculating the **deviation** of a raw score from the mean of the distribution in question and dividing this **deviation score** by the **standard deviation** of the variable

In other words, raw scores are transformed into a certain number of standard deviations above or below the mean: a z-score of +1.96, for instance, means that the original raw score was 1.96 standard deviations above the mean. When we say that a score in a normal distribution of, let us say -0.75, is exceeded by 72.66% of all scores, we mean that a score of 0.75 **standard**

deviations below the mean has the properties just mentioned. The standard deviation has special importance in calculating probability, since it is the number of standard deviations a raw score lies above or below the mean which determines the probability of that raw score.

Confidence intervals and probability

The first application of the fact that there is a known relationship between scores on a normally distributed variable and the probability of the scores involves estimating population parameters from sample statistics. Since in the real world we never know the population scores, we must estimate them on the basis of sample scores. The first step involves, not surprisingly, working out rules for obtaining the estimated population scores from the sample scores. You should recall that in regression we also have to estimate scores, and that there are rules for obtaining the "best" estimates (least squares method). In the case of inferential statistics, the approach to estimating population scores is different; we calculate the **probability** that, given a particular sample statistic, the population parameter will lie within a particular range of values. For instance, if the, mean statistics score of our sample of anxious students were 55.6, it is possible to calculate the probability that the mean in the population of anxious students lies between, let us say, 50 and 60. This range of scores within which the population parameter is estimated to lie is called the **confidence interval.**

Confidence interval of sample means

In order to calculate the confidence interval, we need to know the standard deviation of the sample statistic (not of the raw scores in the sample, but of the sample statistic). For instance, if we wanted to know the confidence interval of the sample mean of 55.60, we would need to know the standard deviation of the sample mean. The **deviation** of the sample mean from the population mean (population parameters are always unknown) is calculated in the usual way: subtract the population mean from the sample mean.

Sampling distributions

Sampling distribution of the arithmetic mean

The crucial element in the procedures outlined in the sections preceding this one is the standard deviation of sample statistics; for instance, the standard deviation of a sample mean or the standard deviation of a sample correlation coefficient. You should recall that the standard deviation is a measure of **distribution.** When we concern ourselves with, for instance, the standard deviation of sample means, we are looking at the **distribution** of sample means.

Suppose that we had a population consisting of all psychology students presently at a university. I give the anxiety test already referred to the students; these students however constitute only simply a sample, since **I** cannot test the entire population. Suppose that another psychologist tests a group of psychology students in a different university, a third a group in Hamburg, a fourth a group in Regina, Canada, and so on. In each case, the psychologist has tested a different sample. We could test thousands of such samples if we had enough psychologists who were interested in the project. Suppose that each of these psychologists calculated the mean anxiety score of the sample which he or she tested. These means would probably not be numerically equal, but would vary. Thus, it would be possible to calculate the mean of the sample means (add up the means of the thousands of samples and divide by the number of samples), as well as the standard deviation of the sample means. It is known mathematically that when this is done, the distribution of means from the various samples is **normal,** regardless of the distribution of the raw data from which the means were calculated. Thus, qualities of the normal curve can be applied to the distribution of sample means, as was done in the section on confidence intervals.

In real life, however, we normally have only one sample, not thousands. Thus, we cannot actually calculate the standard deviation of the sample mean directly, because we only have one sample. As a result, since the standard deviation of the sample mean is so important, we have to **estimate** it, on the basis of a single sample. What we want is the **best estimate** of the

standard deviation of the sample mean, given that we are not in a position actually to test thousands of samples.

Sampling distribution of correlation coefficients

As was pointed out in earlier sections of the lecture, the correlation coefficient causes certain difficulties, because increases in the magnitude of the correlation coefficient are not related in a linear way to increases in the covariance of the variables in question. A linear relationship is reached by transforming correlation coefficients using Fisher's z (not to be confused with z-scores). This same problem arises when the sampling distribution of correlation coefficients is to be calculated. The solution, however, is equally simple--transform correlation coefficients into Fisher's z equivalents.