Chapter Six
Statistical Inference and T-tests

Aims of the Chapter

In this chapter, you are introduced to inferential statistics. The chapter aims to:

- explain the nature of statistical reasoning
- explain how this abstract reasoning connects with real-world problems
- introduce the concept of the null hypothesis
- relate this concept to probability, chance, and statistical decision-making
- teach you how to use SPSS to conduct t-tests, as a procedure relevant to comparing scores from two groups

The datasets that the chapter draws on are from the Foster and Skehan datasets.

Associated Reading: Howell (1997), Chs. 3, 4, 5, and 7

Introduction

So far, we have looked at essentially a descriptive approach to statistics. Procedures such as the mean, standard deviation, scattergram and correlation coefficient are intended simply to condense data, and to thereby make it easier to talk about, and to communicate the data to other people. The essential approach is to say “Let’s look at this with a method of description which allows us to agree about what we are describing.” What the approach so far doesn’t allow us to do is to make decisions about the data we have examined.

In fact, there is one small exception to this. In the section on correlations, in Chapter Five, we did look at the concept of statistical significance. The SPSS printout showed the probability levels associated with the different correlations. These essentially enabled us to make the decision as to whether the correlation in question should be treated seriously and interpreted, or whether it should be regarded as having arisen by chance (in which case interpretation would not be justified).

But, except for this one case, everything in the course so far has been concerned with descriptive statistics. Chapter Six provides an introduction to inferential statistics. These are the sort of statistical techniques which enable decision making to take place. These techniques usually involve comparisons between the (mean) scores of two or more groups.

This chapter, compared to the others, may be heavier on exposition. There are references to other reading, and these are important. But in this chapter, the text itself takes you further, and avoids sending you off to other references for important explanations if the explanations can be given here.
Consider the following questions, drawn from different aspects of applied linguistics and teacher education:

- do students taught reading with extensive use of pre-reading activities (e.g. being required to ask questions or make predictions before reading a text) learn more than students who engage in the reading activities themselves without such preparation?
- will students taught in a communicative classroom do better than those taught in an audiolinguinal classroom?
- do students who learn vocabulary with some sort of mnemonic method (e.g. a mediating visual image) retain more words than students who use other methods (e.g. simply trying to memorise lists of target language words paired with L1 words)?
- do extroverts do better than introverts on communicative language tests?
- do introverts do better than extroverts on written multiple-choice grammar tests?

Imagine collecting data in relation to each of these questions. For example, with the fourth, extroverts and introverts with communicative (not multiple-choice) tests, you can imagine identifying a group of extroverts, and also a group of introverts, then giving them all a communicative test. This would give you two sets of scores, one set for each group. Suppose the extroverts got a mean score (out of 100) of 64.5, while the introverts got 61.7. What could this mean?

Clearly, 64.5 is more than 61.7. But does this allow you to conclude anything? What if the extrovert group score were 74.5 instead? Or 84.5? Would this ease your decision?

The key question is: what basis can you think of for deciding whether the extrovert mean score is “truly” higher than the introvert mean score? Or alternatively, deciding that it is just a little bit higher simply by chance?

The first thing to notice is that, in each of these cases, the question revolves around the comparison of two groups. Each of the two groups will have an average score on some measure or other, and the crux of the comparison is whether one of the groups obtains a score which differentiates it in some way from the other. One of the central problems in statistics is to find a method of establishing how such differentiation can occur and how a decision can be made as to whether a particular comparison represents a real difference or the sort of difference which could occur simply by chance. The key issue, and the one this chapter is centrally concerned with, is how to distinguish between real differences and chance differences. We will see, as the chapter goes on, that key elements here are:

- the amount of the difference between the two mean scores
- the amount of variation in the groups
- how big the groups are
We can instructively return to the main dataset (Display1.sav) that we are using to illustrate the discussion. In Foster and Skehan (1996), i.e. the dataset from the first of the Skehan and Foster studies, a major contrast is drawn between planned and unplanned performance. This comparison concerns:
• two different tasks (a decision making task and a narrative)
• three different measures (fluency, accuracy, and complexity)

### Task 6.2: Considering some comparison-based decisions

Table 6.1 shows the results of using the Means procedure (which you are now familiar with) with the data from this first study. The table shows the mean scores comparing the planned and unplanned groups for the three different measures for each of the two tasks. (If you want, as a refresher, to produce exactly these same results, first use the Data menu to select only the cases from the first study. Then use the Means procedure, selecting appropriate measures of pausing, accuracy and complexity as Dependent variables, and choose Plan cond as the Independent variable.)

<table>
<thead>
<tr>
<th>Decision-making</th>
<th>Unplanned</th>
<th>Planned</th>
<th>“Eyeballing” Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pauses</td>
<td>37</td>
<td>17.3</td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.63</td>
<td>0.72</td>
<td></td>
</tr>
<tr>
<td>Complexity</td>
<td>1.23</td>
<td>1.43</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Narrative</th>
<th>Unplanned</th>
<th>Planned</th>
<th>“Eyeballing” Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pauses</td>
<td>30.3</td>
<td>12.5</td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.61</td>
<td>0.62</td>
<td></td>
</tr>
<tr>
<td>Complexity</td>
<td>1.22</td>
<td>1.52</td>
<td></td>
</tr>
</tbody>
</table>

A first response is to “eyeball” the data and to make decisions quickly on the basis of a first impression. Clearly, the value in the Planned column is always “better” than the value in the unplanned column, i.e. performance is more fluent, more accurate, and more complex.

In the right hand column of the Table 6.1, write down whether you think that the figures reported could have arisen by chance, or whether you think that there is a real difference between the two groups.

For example, take the contrast, for the decision-making task, pauses, of 37 (unplanned) and 17.3 (planned). The mean score for the group of planners is clearly lower (i.e. they pause less, and are more fluent). But we also have to consider that when scores (on anything!) are compared, there will be what can be termed “chance” influences and factors as well as the controlled and intended influence (in this case planning). The question is: Does the difference between the two groups look of the scale that could occur by chance factors alone, or is there any indication of something more (which we hope will be linkable to pre-task planning)?
In the right hand column of Table 6.1, therefore, either write “Chance” or “Intended effect”.

DO NOT GO FURTHER UNTIL YOU HAVE COMPLETED THE EXERCISE

Feedback on Task 6.2

Actually, (and very irritatingly), there is no feedback to give here! The answer is: On the basis of “eyeballing”, we don’t know. The whole point in using statistics is to establish on some principled basis whether a difference could have arisen by chance. Although it is tempting to say that the Narrative Accuracy figure, for example, could easily have arisen by chance while none of the others could, that is not warranted. When you eyeball data, it is all too easy to see the results that you want to see in the data. Avoid doing this. Things are difficult enough as they are! It is appropriate statistical procedures that have the determining influence, and they, as we shall see, will involve taking into account the scale of the difference between the two mean scores in a particular comparison; the amount of variation in the scores in the two groups; and the size of the two groups.

A Crucial but difficult section on the nature of statistical reasoning

In this section you are going to be introduced to the way statisticians think. There may be time when things are difficult to follow. Please persevere, and perhaps re-read the section a time or two to help you to get there. As slight encouragement, it can be said that the statistical procedures you will later learn, with SPSS, are themselves very easy to use! But before we start the tough stuff, there is a small diversion that we have to go on:

A Necessary Diversion: The Normal Distribution Curve - A Strange Quirk of Nature

The problem we looked at just a few lines ago seemed very simple, but in fact, we have to follow some diversions before we have the tools and concepts to address it effectively. The first diversion is to understand the normal distribution curve, one of the foundation planks of statistics. Be forebearing while we cover this necessary material!

If you want to read more on the Normal Distribution Curve look at Howell, pp 72-87
We mentioned in an earlier section that human behaviour and attributes are variable. It is, pervasively. Consider an attribute such as height. (Decide here whether you want to look at male or female height, but for the sake of simplicity, chose just one!) Obviously, there will be an average height for the population (possibly 5’9” for men in the U.K. and 5’4” for women). What is also the case here is that the average height will be the commonest height, i.e. there will be more people with this average height than any other height. In addition, the further that you move away from this average height, the fewer people you will find. In regions immediately close to the population average height, e.g. 5’7” to 5’11” for men, and 5’2” to 5’6” for women, there will still be lots of cases. But as you move away even further from this average height, to regions such as 5’0” (or smaller) and 6’6” (or bigger) for men, the numbers of cases will diminish quite considerably. We can represent this discussion in the form of a graph, as Figure 6.1 shows.

**Figure 6.1: The Normal Distribution**

![The Normal Distribution](image)

The figure captures the way, in a normal distribution curve, that most cases occur towards the centre of the distribution, and that, as one moves away from the centre, progressively fewer cases occur. Notice, though, that running along the horizontal scale at the bottom of this distribution are two parallel scales. The first is the straightforward one of measurements of male height. This shows 5’9” as the height that occurs most often, (represented by the height of the graph at this point), while heights such as 6’2” and 5’4” occur less often, reflected in the lower height for the graph corresponding to those heights. (These actual heights are not given in the scale below, but you can infer where they are.) But the second “scale” is a little different. Recall, first of all, that in earlier chapters you learned how to calculate standard deviations, and that this is a measure reflecting the amount of variation/dispersion in a set of scores. Now, as it happens, the standard deviation has been computed for male height in the U.K. and it is (or at least is close to) 3”. (This value is not totally accurate, but it is close, and much easier to work with.) Clearly someone who is 6’0” tall is three inches above the average height, while a man who is 6’3” is six inches above the average height. We can express this in terms of the standard deviation measure, using the formula:

\[
\frac{\text{Actual Person’s Height} - \text{Average Height}}{\text{Standard Deviation}}
\]
The six-footer is equivalent to one standard deviation about average height \((6'0"-5'9")/3\) inches, while a 6’3” man is two standard deviations above.

### Task 6.3

Express the following heights in standard deviation terms.

<table>
<thead>
<tr>
<th>Height</th>
<th>Standard Deviation Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>6’6”</td>
<td></td>
</tr>
<tr>
<td>6’1”</td>
<td></td>
</tr>
<tr>
<td>5’6”</td>
<td></td>
</tr>
<tr>
<td>5’7”</td>
<td></td>
</tr>
<tr>
<td>5’9”</td>
<td></td>
</tr>
</tbody>
</table>

### Feedback on Task 6.3

In the table below, the heights are expressed in standard deviation units. Note this new term.

<table>
<thead>
<tr>
<th>Height</th>
<th>Standard Deviation Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>6’6”</td>
<td>+3</td>
</tr>
<tr>
<td>6’1”</td>
<td>+1.5</td>
</tr>
<tr>
<td>5’6”</td>
<td>-1</td>
</tr>
<tr>
<td>5’7”</td>
<td>-0.5</td>
</tr>
<tr>
<td>5’9”</td>
<td>0</td>
</tr>
</tbody>
</table>

The table shows that the tallest person in this group is taller, relatively, than the shortest person is short, so to speak, in that he is three standard deviations above the mean, whereas the shortest person is only one standard deviation below the mean. Notice also that it is easy to interpolate fractional values, so that the 6’1” person is one-and-a-half standard deviations above the mean.

The point, here, is that it is meaningful not simply to talk about an actual or “raw” measure, but also the same score expressed in standard deviation “units”. This latter score then portrays the original raw score in terms of all the other scores. It captures something additional beyond the raw score itself because it places an individual in relation to an entire group.

This shape and distribution of the normal curve are interesting enough in themselves, but three other aspects to the normal distribution curve make it one of the (unappreciated) wonders of the world.

First, it is crucial to restate that the normal distribution curve is pervasive. It crops up in an amazing number of situations. Essentially, it reflects the way variation patterns
in the world, with lots and lots of attributes and measures. So it is not confined to height or to unusual variables – on the contrary, where there is variation, there is likely to be variation shaped as in the normal distribution curve. It is a fact of nature which needs to be understood.

The second aspect of the normal distribution curve is that, because it is so pervasive, it has been studied intensively and described mathematically to a great level of exactness. The mathematical formula underlying the curve has been well explored, and it can be used to describe the curve to a considerable degree of precision. This becomes a powerful tool, as we shall see.

Which brings us to the third point. The normal distribution curve has been described very accurately in terms of the proportion of the total population which falls between different parts of the curve. Going beyond Fig 6.1 for a moment, it is useful to place over the curve some critical points. This is shown in Fig. 6.2.

![Figure 6.2](image)

This shows the position of the mean (at the peak or most frequently occurring point of the normal distribution) and the standard deviations above and below the mean. The crucial point here is that the standard deviations occur at very dependable points on the normal distribution curve in relation to the percentage of people (cases) involved. They, in effect, define areas of the normal curve with great exactness. To be more precise:

- 34% of cases fall between the mean and +1 standard deviations
- 34% of cases fall between the mean and -1 standard deviations
- 95% of cases fall between +2 and -2 standard deviations

A concrete example will make this a little clearer. Going back to male height in the U.K., with a mean of 5’9”, and a Standard Deviation of 3 inches, if we make the assumption that these scores are normally distributed, then we can say that:

- 34% of people will score between 5’9” and 6’0”
- 34% of people will score between 5’6” and 5’9”
- 95% of people will score between 5’3” and 6’3”

**Task 6.4**
Here are three questions:
1. what proportion of people will obtain scores between 5’6” and 6’0”?
2. what proportion of people will obtain scores less than 5’3” and above 6’3”?
3. what proportion of people will obtain scores above 6’3”? (hint: the normal distribution is symmetrical)

DO NOT READ ANY FURTHER UNTIL YOU HAVE ANSWERED THESE QUESTIONS

Feedback on Task 6.4

The answers are:
- 68%, which is obtained by adding the 34% of people who get between 5’9” and 6’0” and the 34% of people who get between 5’6” and 5’9”
- 5%, which is obtained by subtracting the 95% who get between 5’3” and 6’3” from the 100% which is the total number of people in the sample
- 2½, which is obtained by halving the answer to the previous question, on the assumption that the normal distribution is symmetrical, i.e. that of the 5% who obtain scores outside the -2 to +2 standard deviation range, half will be in the top tail of the distribution and half will be in the bottom tail.

In fact, the examples we have focussed on above are only the obvious ones, which are easy to compute using points like “+1 standard deviation” and so on. In fact, the proportion of cases falling between two points on the normal distribution curve can be computed for any range and with equal exactness. So, for example, one could find the proportion of cases falling between -1.24 and -0.65 standard deviations or +0.87 and +2.33 standard deviations. Statistics books contain tables giving the information which is necessary to answer such questions, and, although we don’t need to do this most days of our lives, it is an indication of the precision which can be achieved through the knowledge which is available on the normal distribution.

Note again that the central issue in what we are doing is that the normal distribution curve can be expressed in terms of standard deviation scores, and it is such an expression of the original scores which enables precision to be achieved in relation to the proportion of people who fall within a certain range.

A Thought Experiment

We are now going to set up a thought experiment, i.e. one that would never be carried out, but which will have the potential to explain some important aspects of statistical reasoning. The thought experiment is going to compare two groups of ESL learners. The first group consists of students who are taught by teachers who have done M.A. courses in English Language Teaching. The second group are taught by teachers who haven’t done M.As in English Language Teaching. As this is a thought experiment, we have to imagine two things:
Chapter Six Statistical Inference and T-tests

1. that we will be able to conduct this experiment without financial limitation (after all it is a thought experiment!). This means that we can imagine having lots and lots of comparisons between pairs of teachers (matched for age, teaching experience etc), one with an M.A. and one without.

2. that we know the answer. (O.K. it’s perverse to say this, but a useful device.) And the answer is that there is no difference between students taught by teachers with an M.A. and students taught by teachers without an M.A.

Thought experiment or no, let’s imagine some of the details. We’ll start with one particular comparison, between two groups, one taught by the +M.A. teacher and one taught by the -M.A. teacher. We imagine, perhaps unrealistically, that these learners of English are tested so that scores are out of 100. Let’s suppose we get the following results:

<table>
<thead>
<tr>
<th>Group taught by teacher without M.A.</th>
<th>Group taught by teacher with M.A.</th>
<th>Difference in average achievement score</th>
</tr>
</thead>
<tbody>
<tr>
<td>65.4</td>
<td>66.7</td>
<td>+1.3</td>
</tr>
</tbody>
</table>

Clearly the group taught by the teacher with the M.A. obtained a slightly higher average score than the group which was taught by the teacher without the M.A., with the difference being +1.3. Obviously there is no way of knowing whether this difference is due to the M.A.-teacher-taught-group being taught better or simply chance factors operating. Improbably, though, let’s now imagine that we have been funded to extend this study, and to study 100 pairs of classes, with each pair of classes coming from comparable schools, and with one of each pair being taught by an M.A.-qualified teacher and one by a non-M.A teacher. Let’s imagine that we get results similar to those shown below: (note that the results are not shown in full - enough are given to convey the idea, which is appropriate for a thought experiment).

<table>
<thead>
<tr>
<th>Group taught by teacher without M.A.</th>
<th>Group taught by teacher with M.A.</th>
<th>Difference in average achievement score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 65.4</td>
<td>66.7</td>
<td>+1.3</td>
</tr>
<tr>
<td>2 62.8</td>
<td>61.5</td>
<td>-1.3</td>
</tr>
<tr>
<td>3 66.0</td>
<td>66.0</td>
<td>0</td>
</tr>
<tr>
<td>4 60.5</td>
<td>65.2</td>
<td>+4.7</td>
</tr>
<tr>
<td>5 67.4</td>
<td>67.6</td>
<td>+.02</td>
</tr>
<tr>
<td>6 64.3</td>
<td>62.4</td>
<td>-1.9</td>
</tr>
<tr>
<td>7 62.5</td>
<td>62.6</td>
<td>+.01</td>
</tr>
<tr>
<td>8 63.7</td>
<td>63.2</td>
<td>-.05</td>
</tr>
<tr>
<td>etc</td>
<td>etc</td>
<td>etc</td>
</tr>
<tr>
<td>99 62.5</td>
<td>63.9</td>
<td>+1.4</td>
</tr>
<tr>
<td>100 66.0</td>
<td>64.5</td>
<td>-1.5</td>
</tr>
</tbody>
</table>

The first column of scores shows the results for the group in each paired comparison which was taught by the teacher without the M.A. and the second column of scores shows the results for those classes of students taught by teachers with M.A.s. Then,
the rightmost column, the most interesting column of all, shows the difference between the two class scores in each comparison.

Remembering that we “know” that there is no difference between the students taught by teachers without M.A.s and teachers with M.A.s, it is not surprising that the values in the right hand column (only partially shown, since all this data is made up, and so it is hardly worth including even more fictitious data) are a mixture of plus values and minus values.

**Task 6.5**

*Assuming no real difference between these two conditions, what will the average be for the values in the right-hand column?*

**Feedback on Task 6.5**

The answer is zero.

Following the assumption that there is no difference between the two groups, we expect, assuming further that the pairs of comparisons are randomly obtained, and that there are enough of them to avoid being overly influenced by any particular chance values, that the pluses should, over the long run, be cancelled out by the minuses, giving a total for the right hand column of zero. Then, dividing zero by the number of cases in the study, i.e. 100, we obtain the answer zero.

Surprisingly, then, we can “calculate” the mean for the right hand column by simply using logic. We are not so lucky with the standard deviation of the right-hand column. Clearly, there are 100 scores here, and since they do vary around the central value of zero, there will be an actual value for the standard deviation. We cannot know this value, however, without actually calculating it, and to be able to do this, we would have to do the actual research.

We are not going to pretend that we will calculate this value. Let’s assume, for the sake of this argument, that the value of the standard deviation of this set of scores is 3.00. (Why not, after all? It’s got to be something!). Now, we make a major assumption. We can ask, of the set of scores in the right hand, “difference” column, what sort of distribution of scores is involved. We would obtain this distribution by, for example, producing a graph showing the frequency of the possible difference scores, e.g. how many cases have a difference between –1 and 0; how many between zero and +1, and so on. And the answer to this question that we will assume is that
such a distribution would be normally distributed. In other words, for this particular set of scores, this would mean that most values will be close to zero, (the average score) and that as we move away from this average score, there will be fewer and fewer cases, i.e. following the same pattern shown in Figure 6.1 above. More than that, since we are equipped with the knowledge that the standard deviation is 3 (not 5, not 1) we can say some more things about the distribution of differences between means. You can certainly answer a number of questions.

<table>
<thead>
<tr>
<th>Task 6.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) How many scores will be more than 6 away from the average difference of 0?</td>
</tr>
<tr>
<td>b) Given that we know that there is no difference between the two groups really, in our sample of 100 paired comparisons, how many would we expect to be in this range (i.e. more than 6 away from the mean) simply by chance?</td>
</tr>
<tr>
<td>c) How many will be 6 above the mean?</td>
</tr>
<tr>
<td>d) Given that we know that there is no difference between the two groups really, in our sample of 100 paired comparisons, how many would we expect to be in this range (i.e. more than 6 away from the mean) simply by chance?</td>
</tr>
</tbody>
</table>

DO NOT READ FURTHER UNTIL ………!!!

<table>
<thead>
<tr>
<th>Feedback on Task 6.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) 5% of scores (one in twenty) will be 6 or more above or below the mean. The standard deviation is 3, and so a difference of 6 (plus or minus) will be two standard deviations above or below the mean. Since we know that 95% of cases fall between plus 2 and minus 2 standard deviations, we can deduce that 5% will have differences from the mean more extreme than this.</td>
</tr>
<tr>
<td>b) If 5% represents one in twenty, and we have 100 comparisons, then we would expect 5 of these comparisons to be more than 6 points or more away from the mean.</td>
</tr>
<tr>
<td>c) 2½% (one in forty). We assume that distributions are symmetrical (although this has to be checked on), and so we assume that half the scores more than 6 points or more away are above the mean and half below. As a result we can divide 5% by 2, and obtain the answer 2½%.</td>
</tr>
<tr>
<td>d) Well, two and a half. I leave you to figure out what half a person looks like!</td>
</tr>
</tbody>
</table>
Reflecting on what we have done, we have made a number of assumptions:

- that we have “gathered” data in our thought experiment
- that we know that the difference between the two groups concerned (plus or minus M.A.s for the teachers) is zero, i.e. there is no influence on the results whether the teacher has an M.A. or not
- that the standard deviation of the difference between the scores in these paired comparisons is 3

Armed with these assumptions, we are able, through our knowledge of the normal distribution curve, to say precise (if not particularly interesting) things about the distribution of scores. In particular, we are able to tease out statements about the proportion of times scores (i.e. differences between two means) of a particular size will occur.

*The Critical Ratio*

Now, inevitably, we will start removing some of these assumptions!

First of all, we don’t do 100 comparisons in real life! We don’t have the money or the time to do this.

**How, then, does statistics proceed?** Essentially, it is able to proceed by estimating the standard deviation of the differences between two means from large sets of observations “the population” on the basis of just one set of observations “the sample”. More accurately, the “population” here would be all the classes taught by M.A.-qualified teachers and all the classes taught by non M.A.-qualified teachers. The “sample” would be *one* class taught by a teacher with an M.A. and *one* class taught by a teacher without an M.A.. From this sample, of two actual classes, matched for composition of students, we would then have two means (the Class A mean and the Class B mean) and two standard deviations, (the Class A standard deviation and the Class B standard deviation). In other words on the basis of these statistical values from our sample of two classes, we make an estimate of the population variation, i.e. the population standard deviation.

This is the point where you simply have to take things on trust. That it is possible to make such an estimate of population values based on values obtained from samples is down to the work of mathematical statisticians (rather than the statisticians within the social sciences who simply regard statistics as a tool for problem solving). Mathematical statisticians, by carefully studying distributions such as the normal distribution, are able to develop formulae which estimate values such as, in the example given above, the standard deviation of a large number of differences between two mean scores, very effectively.

As a result, we only conduct one study, based on two groups, and then *use the variation in the scores in our sample data (or more exactly, we use the two standard deviation scores) to estimate what is called the population standard deviation.* When our observed set of scores contains a lot of variation, the population standard deviation will be estimated as larger. When our observed scores contain less...
variation, the population standard deviation is estimated as smaller. Then, equipped with our estimate, we try to decide whether the actual difference between two sets of scores is within the range that we could expect by chance alone, or whether it is greater than it is reasonable to expect from chance.

At this point, you are waiting for the formula in question. In response to this, here is a formula. The reason that it is not the formula is that, in fact, a family of formulae are used, and the one which is given below is (a) easier to understand (although you may find this difficult to believe at the moment), and (b) dependent on there being a very large number of cases/subjects in the sample, as though in our thought experiment, with only one pair of groups (with one taught by a teacher with an M.A. and one without), the number of students involved was well above 100. For the flow of the argument, these are reasonable assumptions to make for now. So here’s the formula:

\[
\text{Critical Ratio} = \frac{M_1 - M_2}{\sqrt{\frac{SD_1^2}{N_1 - 1} + \frac{SD_2^2}{N_2 - 1}}}
\]

In the numerator of this formula (the top part) we see the difference between the two average scores, represented as \(M_1\) and \(M_2\). The formula given is a general formula. In our hypothetical case, we will take \(M_1\) to be the average of the scores in the class taught by the teacher with the M.A. and \(M_2\) would be the average from the class taught by the non-M.A.-qualified teacher. In the lower part of the formula, within the square root expression, there are two components. The two parts are based on the two sets of scores respectively. The first takes the standard deviation of the first set of scores, and divides it by the number of cases, minus one. (Simply accept that the difference between “N” and “N-1” is not usually very large. Statisticians have reasons here for dividing by N-1 rather than N, so simply accept that this is so.) The second term in the denominator (the bottom bit) does exactly the same for the second set of scores. So you can see that the formula relates the difference between the two mean scores to a denominator which in turn relates the amount of variation in the scores (i.e. the standard deviation) to the number of cases the data is based on (i.e. N-1).

Formulae are forbidding at first. But it is possible to break them down, and understand more precisely how they work by looking at their different components. Then, even if you cannot see exactly what is going on, you may understand enough to see how the formula is operating at a more general level.
Task 6.7

Complete the following table to show what sorts of things make the answer to this formula get bigger, and what make it smaller. (The purpose of this task is to prepare the ground for what comes next.)

<table>
<thead>
<tr>
<th>Type of Change</th>
<th>Effect on answer or formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) The difference between means gets bigger.</td>
<td></td>
</tr>
<tr>
<td>b) The standard deviations of the two sets of scores get bigger.</td>
<td></td>
</tr>
<tr>
<td>c) The number of cases gets larger.</td>
<td></td>
</tr>
<tr>
<td>d) The standard deviations of the two sets of measures get smaller</td>
<td></td>
</tr>
<tr>
<td>e) The number of cases gets smaller</td>
<td></td>
</tr>
</tbody>
</table>

DO NOT READ ANY FURTHER UNTIL YOU HAVE COMPLETED THE EXERCISE!!

Feedback on Task 6.7

As you read this feedback, and relate it to your answers, you should be thinking about the sense that the answer makes, and the extent to which what is happening through the formula relates to what, intuitively, you would expect to be happening in a formula whose purpose is to establish whether there is a difference between two sets of scores.

**Type of change**

a) The difference between the means gets bigger

**Effect on answer to formula**

The answer to the formula will get larger. The larger the numerator, the top bit, in other words, the larger the eventual answer. This is desirable, since we know that the larger the difference between the means, the more obvious the effect of whatever caused the two groups to be considered worth studying. And, as we shall see, the larger the difference, the more likely it is that we will find an “effect”, where “effect” will be explained later.
b) The standard deviations of the two sets of measures get bigger

This is slightly trickier. The standard deviations are in the denominator of the formula. The first thing to realise is that, in general, the larger a denominator, the smaller the answer to an equation will be. So in this case, the larger the standard deviations, other things being equal, the larger the denominator will be and so the larger the amount that will be used for division. *As a result, the smaller the eventual answer.*

The reasonableness of this comes from what variation itself means. Standard deviations get larger when there is more variation. But more variation means that the difference between means, if it is to be regarded as “real”, has to be large enough to *stand out from the variation.* The more variation, the more difficult it is to stand out. Think of this as connected with the amount of overlap between the two sets of scores. The more variation, the more overlap. The result is that it is more difficult to establish that the two groups are “truly” different.

c) The number of cases gets larger

The first thing to note is that the number of cases (N in the formula) is in the denominator of the denominator! So, following the chain of reasoning closely, the larger the N, the lower the answer to *just the denominator.* Which makes the denominator smaller, and the eventual answer larger. So a larger N makes for a larger answer. (Work out some simple arithmetic answers to help you follow this sequence if you find it difficult. Concrete examples help a great deal.)

Again, thinking this through reveals the desirability of this outcome. It is saying that when you base your data analysis on more cases, you are
more likely to get a larger answer to the Critical Ratio formula, and therefore you are more likely to detect a real difference between two sets of scores. At a common sense level, it is reasonable to say that the larger the database one uses, the more robust and defensible the answer that one obtains. In the present case, what the formula is doing is reflecting this common sense intuition, leading to the precept “Always collect as much data as possible”.

d) The standard deviations of the two sets of measures get smaller

This is the reverse of the last but one question. Looking at it arithmetically, smaller standard deviations will make the denominator smaller, and so the eventual answer (and the possibility of detecting differences) larger.

In common sense terms, the smaller the variation in a set of scores, the more easily can the first set of scores be distinguished from the second set, because the amount of overlap, other things being equal, will be smaller.

e) The number of cases get smaller

Again, a reverse situation to an earlier question. Arithmetically, a smaller N (or number of cases) will make the denominator larger, which in turn will make the eventual answer (and the scope for detecting a difference) smaller.

In common sense terms, it is as if the investigator is being “punished” for using small sample sizes for research. The smaller the sample size, the more difficult it is to detect differences.

Let’s now have some examples of putting this formula to work. Here is an example of some data, and the calculation which follows:

*A study was conducted of the effects of small quantities of alcohol on ratings of pronunciation of those who had the alcohol. One group (the controls) were simply required to pronounce a series of sentences in a target language, and their pronunciation was rated by a group of expert judges. The judges gave higher*
numerical ratings for people whose pronunciation they thought was better. The experimental group were given a small quantity of alcohol thirty minutes before the sentence pronunciation and were also rated by the same judges. The results of the ratings were:

<table>
<thead>
<tr>
<th></th>
<th>Controls</th>
<th>Experimental</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>13.4</td>
<td>15.6</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>2.6</td>
<td>3.1</td>
</tr>
<tr>
<td>Number of cases</td>
<td>24</td>
<td>20</td>
</tr>
</tbody>
</table>

We can “plug” these values into our formula. Obviously we have two groups here, and the formula requires that one of these groups accounts for $M_1/SD_1$ and the other $M_2/SD_2$. In this case, it seems sensible to make the experimental group $M_1$ and the Control Group $M_2$. This arranges the numerator so that if the experimental group mean is higher than the control group, the result will be positive, rather than negative. But it should be realised that this is rather arbitrary. The eventual statistical result we will arrive at will be exactly the same, whether the numerator result is positive or negative. It is the size of the difference between the two means that really counts. We will return to this issue later in the chapter. For now, plugging in the values gives the following result:

\[
\text{CriticalRatio} = \frac{M_1 - M_2}{\sqrt{\frac{SD_1^2}{N_1 - 1} + \frac{SD_2^2}{N_2 - 1}}}
\]

\[
\text{CriticalRatio} = \frac{15.6 - 13.4}{\sqrt{\frac{3.1^2}{19} + \frac{2.6^2}{23}}}
\]

\[
\text{CriticalRatio} = \frac{2.2}{\sqrt{9.61 + 6.76}}
\]

\[
\text{CriticalRatio} = \frac{2.2}{\sqrt{16.37}}
\]

\[
\text{CriticalRatio} = 2.46
\]

**Task 6.8**

Let’s imagine that another researcher decided to do essentially the same study, but with one major difference: the amount of alcohol was doubled. The following results were obtained:

<table>
<thead>
<tr>
<th></th>
<th>Controls</th>
<th>Experimental</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>13.8</td>
<td>12.7</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>2.6</td>
<td>3.4</td>
</tr>
<tr>
<td>Number of cases</td>
<td>24</td>
<td>30</td>
</tr>
</tbody>
</table>

Calculate the answer to the formula. (Note that either having a Calculator, or using the Calculator supplied with Windows would be a help here.)

What provisional conclusions do you draw from these two calculations?
Obviously, now, you are in a position to calculate the value to a formula, and in a position to obtain an arithmetic answer. The small problem remains of what on earth this answer means!

In fact, we are now close to having all the elements to deal with this question of interpretation. But there is one more concept that has to be faced, and a little more patience is required before meaning can be brought to these infernal numbers. Stick with things just a little longer.

The Null Hypothesis

It may have occurred to you that the essence of statistical reasoning is back-to-front! It would be much nicer to look at two averages based on two sets of scores and simply say “This difference is a real difference.” This would be a direct form of reasoning. Unfortunately, it is not possible to follow it. The basic problem, mentioned earlier but worth restating, is the difficulty in separating out the effects of a real difference (say connected with alcohol before pronunciation) and the effects of random variation, of the sort that arises because human beings are human beings, and they don’t all behave the same way as one another, or even the same way as themselves at different times!

In other words, the problem is one of separating what might be called “systematic” variation from “common or garden”, or random variation. The thing we are trying to avoid is mistaking ordinary variation for real differences - we have to find ways of making decisions about:

- when a difference is likely to be down to something systematic versus when a difference is:

- based on nothing more than the non-constancy of human behaviour even under the same conditions.

So, what statisticians do is to reverse what might be considered to be normal reasoning, and say: Let’s calculate how improbable a result is purely by chance, and then make decisions about a proposed real difference based on this view of improbability, i.e. let’s see if an interpretation based on chance is improbable, forcing us to accept the alternative – a difference is based on something real, not on chance. All of this is expressed with one of statistics’ most wonderful terms, the null hypothesis.

The null hypothesis is (rather boringly) always the same. It runs:
Chapter Six Statistical Inference and T-tests

The null hypothesis states that the difference between the groups is due to random variation only.

In effect, it constitutes a challenge which says: **So prove me wrong!**

Mostly, if you gather research data, you do so because you are motivated to expect some sort of difference between groups. What you are interested in is establishing that your data support the existence of such a difference. So it is something of a deflator to discover that when it comes to the crux of statistical reasoning you have to use the formulaic (incantatory?) form of the null hypothesis to say exactly the opposite of what you are hoping to find. But that’s the way it goes. In a perverse way, you have to state the null hypothesis about your data in the fervent hope that what your data will do is deliver results in such a way that you can reject the null hypothesis.

If you do succeed, and your results allow you to reject the null hypothesis, then you are ready to tango. Rejecting the null hypothesis means that you are concluding that your results could not have arisen by chance alone. This in turn means that your results have to be accounted for in some other way, which brings you back to your original motivation for doing the study, and the explanation/theory which led you to design your data collection in the way that you did.

By eliminating chance as an explanation of your results, you justify the way you interpret your results with reference to other (infinitely more interesting) factors. By rejecting the null hypothesis, in other words, the real process of interpretation begins. Hence the need to engage in this rather back-to-front reasoning. It allows you to get to the place you wanted to, but you achieve this by elimination, rather than direct decision.

Now let’s return to the pronunciation under alcoholic influence example we were working on earlier. The figures, from the lower alcohol condition, are given again below:

<table>
<thead>
<tr>
<th></th>
<th>Controls</th>
<th>Experimental</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>13.4</td>
<td>15.6</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>2.6</td>
<td>3.1</td>
</tr>
<tr>
<td>Number of cases</td>
<td>24</td>
<td>20</td>
</tr>
</tbody>
</table>

These were plugged into the formula:

\[
CriticalRatio = \frac{M_1 - M_2}{\sqrt{\frac{SD_1^2}{N_1 - 1} + \frac{SD_2^2}{N_2 - 1}}}
\]

and yielded an answer to the Critical Ratio formula of 2.46.

Look again at the denominator to the formula. Essentially by combining the two sets of standard deviation scores in this way, it gives the standard deviation of the difference between the means. This was exactly the figure we were concerned with in
our thought experiment (where we imagined that it would be “3”). The denominator in the Critical Ratio is the method that statisticians give us of estimating this value when, necessarily, we are equipped only with the information from the two groups in our paired comparison. The answer to the Critical Ratio formula is then expressing the difference between means that we have found in terms of the standard deviation of the differences between means. If the answer to our formula had been 1.00, it would have meant that the difference between the two mean scores was exactly equal to one standard deviation of the differences between means. (This would be slightly analogous to someone getting a score in one of the earlier examples of 72, when the mean was 60 and the standard deviation 12.) In the present case, the difference between means was 2.46 times one standard deviation.

Recall that earlier we said that:
- the distribution of the differences between means in our thought experiment is normally distributed

We are now in a position to recreate the reasoning we used then, but this time we have our estimated standard deviation of the differences between means, rather than the made-up value we worked with earlier. In fact, to put this properly we should speak not of the standard deviation of the differences between means (which are, after all, the hypothetical result of our thought experiment). Instead we should describe the denominator of the formula by its proper statistical term - the standard error of the mean.

Now let’s restate the null hypothesis and apply it to our alcohol study.

*The null hypothesis states that there is no difference between the pronunciation scores of the group which were given a small amount of alcohol and the group which were not given any alcohol.*

The major item of information that we now have is that the formula has given the value of 2.46. Making the assumption that the values of differences between means are normally distributed, we are, in effect saying:

*Assume that there is no difference between the two groups. Now answer the question: How likely (or rather unlikely) is it that a value as large of 2.46 will be found when the null hypothesis is true?*

Recall that only 5% of cases lie in the region of the normal curve beyond 2 standard deviations. If we had obtained a value of 2, in other words, we would be dealing with a result that would occur by chance no more than 5% of the time. But our value is even greater than this, at 2.46, and so we can deduce that even fewer than 5% of cases would be beyond this value. In fact, this value would only be obtained by chance just a little more than 1 per cent of the time. For now we will treat this as 1 in a 100, and use rounding to justify this temporary decision. In other words, we would only get a result as large as this by chance very rarely. As a result, and very crucially, we have to decide:

(a) is this a freak, chance, occurrence?
or

(b) is it more plausible to reject the null hypothesis, and so conclude there is something happening which distinguishes between the two groups?

It is conventional in statistics to regard a chance of 1 in 100 as remote, and, in these circumstances, to reject the null hypothesis, i.e. to conclude that there is a systematic difference between the two groups.

**Task 6.8 (Restated)**

Earlier, in the first presentation of Task 6.8, you had to compute the formula for another set of data, fictitiously based on the results obtained when a higher dose of alcohol was given. The figures are shown below:

<table>
<thead>
<tr>
<th></th>
<th>Controls</th>
<th>Experimental</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>13.8</td>
<td>12.7</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>2.4</td>
<td>3.4</td>
</tr>
<tr>
<td>Number of cases</td>
<td>30</td>
<td>30</td>
</tr>
</tbody>
</table>

And the answer obtained was 1.42. What can you conclude about this number?

*Do not agonise over this task.*

**Feedback on Task 6.8**

Actually, not a lot! Clearly, the answer to the formula, at 1.42 is much smaller here than the previous value of 2.46. The answer is based on the difference between the two means in question, divided by the standard error of the mean (i.e. the standard deviation of the differences between means). The obtained means generate a Critical Ratio equivalent to 1.42 standard deviation units. This value must be associated, using the normal distribution curve, with some particular level of chance. But at the moment, you do not know what that level of chance is. To make progress, you need to refer this number to a table which does tell you the probability of occurrence, by chance, of a number of this size. For a normal distribution, i.e. large sample sizes, the critical values that you are looking for are as follows:

<table>
<thead>
<tr>
<th>Significance Level</th>
<th>Expression as chance value</th>
<th>Critical Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>.05 level</td>
<td>one in twenty</td>
<td>1.96</td>
</tr>
<tr>
<td>.01 level</td>
<td>one in a hundred</td>
<td>2.58</td>
</tr>
</tbody>
</table>
Even though in our line of research we don’t often deal with proper normal distributions (because our sample sizes are too small), these numbers from a genuine normal distribution are very important ones, and even worth trying to commit to memory, because they help you to see immediately with many statistical results whether a serious result is in prospect. They represent the numbers which have to be matched or exceeded if a significant result is going to be found and if the null hypothesis is going to be rejected.

We can now answer the question as it was posed. An answer of 1.42 is clearly below the value of 1.96 required. Because it doesn’t reach this minimum value, we have to conclude that the likelihood of the result in question occurring by chance are more probable than one in twenty (in fact, it’s about three in twenty). Therefore, we have to accept the null hypothesis in this case, and so draw the conclusion that these results could have arisen by chance. In other words, these results require us to conclude that there is no evidence that the amount of alcohol in question makes any difference to rated pronunciation. The difference in favour of the sober group could have arisen by chance.

Note also here that the value required to reach the .01 level of significance is 2.58, which is slightly more than the 2.46 we obtained in the other comparison. So, strictly speaking, we didn’t quite make the .01 level there. Even so, based on the two sets of results, it appears that a little alcohol significantly improves pronunciation, whereas the larger amount of alcohol does not!

Significance Levels

We need to return now to a discussion which started towards the end of the unit on correlation. The correlation output from SPSS, you’ll recall, showed a panel of numbers which clarified whether particular correlations could have arisen by chance, or whether they were of a level which required us to treat them as significant. The discussion in the text then went on to consider some of the reasoning associated with significance, and the levels of significance that have achieved consensus amongst social scientists. We need to reprise that discussion now, because the concepts introduced then are even more important for the present case.

Computing the chances that something will occur is the centre of statistics. In the case of correlations, a formula was used, based on mathematical statistical work, which enables us to assess, through knowledge of the correlation size and the number of cases in the sample, the likelihood of the correlation occurring by chance. In the present case, we have a different formula, but the process otherwise is exactly the same. Now we are looking at the difference between two mean scores. But once again, we are assessing the likelihood of a particular result by chance.
But there is also the issue that the “scale of chance”, as it were, is continuous. We can express chance with all sorts of fine gradations, from one in two, to one in ten million. Along the way, we can talk about one in sixty-seven and one in sixty-eight. In fact, we can make endlessly fine distinctions, if we so choose.

But we saw with the correlations that certain values on this scale of chance are privileged. In particular, we saw that three particular values are privileged in this way:

- one in twenty
- one in a hundred
- one in a thousand

These values have been accepted as the benchmarks to which we attach particular importance in this scale of chance. Essentially, inferential statistics, with which we are now dealing, aims at making decisions. And what these particular points on our scale contribute is the elements of our decision making. They are the standard we have to conform to if there is to be a principled method of decision making which avoids a situation where different researchers, possibly biased because they are pursuing ideas important to them personally, want to accept different points on the scale for decision making. *The three consensus decision points force researchers to make decisions in the same way, and so enable consistency in the way research proceeds.*

Essentially, this analysis provides you with the heart of statistical reasoning and the philosophy of science underlying quantitative investigations in our field. The approach is very Popperian in terms of philosophy of science (see, for example, Bryan Magee’s *Popper* (1973), in the Fontana Modern Masters Series, if you want to read more about this). It is consistent with an approach in which researchers make predictions, and then examine their results against these pre-set decision points. The decision points are unforgiving and provide a framework and a set of principles by which rational enquiry can proceed. So, while there is an element of arbitrariness in the choice of the particular privileged points, there have to be some privileged points if effective decision making is to be possible.

In a strange way, it can be said that what we have covered in the last few pages is the essence of using statistics in systematic quantitative inquiry. The rest is detail. You may need to learn another formula here and there, perhaps, and also some different procedures. But the central reasoning is always the same. The null hypothesis is expressed, and then procedures are followed which are appropriate to a particular data set to enable decision making to proceed about the original predictions, but always following the same set of decision criteria.
In the remainder of this chapter we are going to start with some of this detail. In particular, we are going to review the limitations of the formula you have seen. The explanation that you have worked through already of how to make decisions about the differences between two groups is essentially valid, but only if certain assumptions are met. And, in our field, these assumptions are often a little shaky.

The central (and most troublesome) of these is that the data we are working with is normally distributed. Unfortunately, this is only true for very large samples, e.g. more than 200 cases in each group. In applied linguistics and teacher education, it is often difficult to conduct research with the very large sample sizes that one would need to justify using techniques based directly on the normal distribution. So, some other approach needs to be used.

At this point, it should be said, strongly, that the underlying reasoning is always the same:
- a statistical distribution is postulated, e.g. the normal distribution
- formulae are developed based on this distribution
- effective decision-making procedures are followed

So, faced with the need to develop procedures appropriate to smaller group sizes, the “solution” to our problem is to use a slightly different distribution to the normal distribution, and one which is more appropriate to the sample sizes that characterise our field. The answer is the Student’s t distribution, so called because its author, who went by the pseudonym of “Student”, explored the use of a range of distributions, giving each of the distributions a name in the shape of a letter of the alphabet. When he got to “t” he was satisfied, and that is the distribution we still work with!

Earlier, we looked at a formula appropriate to the normal distribution, in the shape of the Critical Ratio test. This formula is not appropriate for the t-distribution. In fact, and totally unsurprisingly, the comparable formula which is appropriate is known as Student’s T, or the t-test. The formula is as follows:

\[
t = \frac{M_1 - M_2}{\sqrt{\frac{\sum d_1^2 + \sum d_2^2}{N_1 + N_2 - 2} \left( \frac{1}{N_1} + \frac{1}{N_2} \right)}}
\]

As you can see, the formula is not so directly interpretable as that for the Critical Ratio (which is one of the reasons the Critical Ratio was presented first). But you can see that the essential components are pretty similar (even if the order and arrangement are not!). The formula still contains, as numerator, the difference between the means. It also contains, within the denominator, standard deviation-like elements, and the sample sizes. The denominator is basically the standard error of the mean but for an underlying t-distribution rather than a normal distribution. The formula and the test based upon it are, basically, slightly more conservative than the Critical Ratio we
looked at earlier. Conservative, in this respect, means that it is slightly harder to achieve the significance levels of .05, .01, and .001, and this conservatism is more evident the smaller the sample size we use. This is an altogether sensible change given that we are essentially dealing with the problem of fewer people in the groups we typically work with, and what this means is that we need to compensate for not having such large groups by being more stringent in the decision-making that we do.

But it is important to restate that all the rest of the statistical reasoning we employed earlier is exactly the same. The purpose of the formula is still to calculate improbability of occurrence, judged by the remorseless scale of what sort of results could simply happen by chance. And we still think in terms of the null hypothesis, and generally are interested in disproving it, rather than accepting it. And we still use the same benchmarks of one in twenty, one in a hundred, and one in a thousand as the way we express results which are very unlikely to have arisen by chance.

With this introduction, it is appropriate now to use SPSS, and to see how easy it is to use the t-test procedure. In other words, you are now in a position to do some real calculation of statistics using the central datasets provided for this course. First the text will “walk you through” the calculation of some statistics, and then you will need to calculate some of your own.

### Task 6.9

You are going to calculate t-tests based on the Study One data.

1. First of all, start SPSS for Windows, and load the datafile Display1.sav.
2. Next, use the Data menu, and the Select Cases line to set up an If condition so that only cases from Study One are selected. (Look back at Chapter Five if you are a little unsure of how to do this. Recall that Study One is essentially concerned with exploring the effects of planning on second language performance.)
3. Next, use the Statistics menu to choose Compare Means and then choose Independent Samples t-test. (This choice will be explained more fully later.)
4. Select the Decision making Pauses (dpause) as the test variable by clicking on the variable name, and then clicking on the arrow. (Notice that the arrow becomes active (i.e. undimmed) only when you have clicked on a variable name.
5. Select Planning Condition (Plancond) as the grouping variable, using the same general procedure, but using the lower arrow. With Steps 4 and 5 you have told SPSS that the variable that you want to compare is the measure of pausing on the decision making task. You have also told SPSS that the basis for the comparison is a separation of the subjects into the two groups of the planners and the non-planners.
6. But you still have to tell SPSS how to form the groups that you want to investigate. Accordingly, click on the Define Groups box, and then, in the new dialogue box which opens up, type in “1” for Group 1, tab, and type in “2” for Group 2.
7. Click Continue, note that the values you have just typed in have been registered in the previous dialogue box, and then click OK.
8. SPSS will click and whirr for a moment or two, and then generate an output screen.
Feedback on Task 6.9

There are two parts of the output to this task. First you have the relevant descriptive statistics, showing mean, standard deviation, std. error and group size information. This is as follows:

<table>
<thead>
<tr>
<th>PLANCOND</th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Std. Error Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>DPAUSE</td>
<td>16</td>
<td>37.00</td>
<td>12.80</td>
<td>3.20</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>17.25</td>
<td>9.00</td>
<td>2.25</td>
</tr>
</tbody>
</table>

The group statistics table gives you the fundamental information, and in this case, allows you to see that the non-planners (Group 1) obtained a mean score of 37.00, clearly well above the corresponding score obtained by the planners (Group 2), with a mean of 17.25. The table also shows that the number of people in each group are equal in this case, and there is also information about the amount of variation. This shows a slightly greater amount of variation in the non-planner group, an unsurprising result given that the numbers involved are larger, as reflected by the higher mean. The Standard Error information can be safely ignored.

The key output is in the second table. This part of the output has been slightly re-edited, not for content, but only for presentation. This has been done because the default SPSS output is unnecessarily wide on the page, and the edited version is simply easier to read.

The first part of the Independent Samples section of the table, under the heading “Levene’s test for equality of variances” checks that a vital assumption for the calculation is met, i.e. that the amount of variation in the two groups (non-planners and planners) is sufficiently comparable. In this case, there is no problem. The significance of .21 that is given here is not alarming, and this “variation in variation” could easily have arisen by chance. (F values above 2, and significance levels lower than .05, on the other hand, would require a bit of attention.)

<table>
<thead>
<tr>
<th></th>
<th>Levene’s test for equality of variances</th>
<th>t-test for equality of means</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F</td>
<td>Sig.</td>
</tr>
<tr>
<td>DPAUSE</td>
<td>Equal variances assumed</td>
<td>1.665</td>
</tr>
<tr>
<td></td>
<td>Equal variances not assumed</td>
<td>5.05</td>
</tr>
</tbody>
</table>
Next we have the major portion of the output, the portion that is going to enable us to make decisions. The output is provided in two lines, one for the situation where equal variances are assumed and one for the situation where they are not. Since the two lines are virtually identical, we shall simply discuss the top line.

The first column is headed by the label “t”. It gives the value 5.05. Pause for just a moment to relate this to figures we have looked at earlier. Recall that when we were looking at the normal distribution, we saw that the figures of 1.96 (.05) and 2.58 (.01) were particularly important. The value we have here is considerably higher. Now, while in general, results for t need to be higher than those for the normal distribution (remember the conservatism of the t-distribution), a value of 5.05 is very clearly higher. So, on the face of it, simply by the experience of looking at values of t, this result looks promising, even though we haven’t considered probability yet. We return to this issue in just a few pages, when we link the idea of sample size, (represented as degrees of freedom) to the different values of t that are required for significance as sample sizes change.

The next column simply reports the number of degrees of freedom. For the moment, simply regard this figure as a reflection of the sample size we are dealing with (which is 31). The degrees of freedom figure is obtained by subtracting 1 from the total sample size (i.e. 31-1=30). Don’t worry about the details here - simply accept that it is necessary if we are to use the t-test procedure. T-tests have to adjust to the number of people in a study, as we shall see.

We move on to the key part of the output - the significance level. As you see, the significance level shown is .000. Essentially this means that the significance is beyond the .001 level, (so large is the value of t which has been obtained). In other words, the difference between means which was found (37.00 vs. 17.25, i.e. 19.75, see next column), is very unlikely to have arisen by chance. In fact, it would occur by chance less than 1 time in 1000. (You may also have noticed the bracket “2-tailed” in the heading of this part of the table. We will return to this below.)

Surprising as it may seem, we can ignore the remainder of the output. It consists of unnecessary detail which is not central to the interpretation we will make. The difference between means (19.75) has already been covered. Little would be added by discussing here the Standard Error figure, or the Confidence Interval information. These are important for certain purposes, but our purpose has been to establish significance, and it is the other cells in the table which have allowed us to do that.
After this detailed examination of output, we now come to the important, hard-
earned stage. Given a value of $t$ of this size, and of the associated probability
of $p < .001$, we can now make the decision to reject the null hypothesis. In
other words, having carefully followed statistical reasoning to this point, we
can now claim to have a result which merits interpretation.

We can now re-express all this in plainer English, and, finally, conclude that
providing learners with the opportunity to plan prior to their doing a decision
making task leads to their producing language with significantly fewer pauses.
In other words, with much greater fluency. In the study in question, the
interpretation offered is that when learners are given planning time, they use
this to assemble language so that during the task itself, it is more accessible,
and as a result, they are not so frequently required to stop to regroup and to
plan "on-line", so to speak. In other words, they are able to organise the
content of what they are going to say, and also bring to mind relevant
language to express such content without needing to regroup.

Task 6.10

Now, still with Study One only, do exactly the same, with the decision making task,
for the accuracy and complexity figures ($d_{accurac}$ and $d_{complex}$ respectively). You
can do the two in the same SPSS “run” simply by clicking on each, successively, as
test variables, so that the two variables are moved across to the upper box in the
independent t-tests screen. Then, generate the output, and write down the answers to
the following questions:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Value of $t$</th>
<th>Significance</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complexity</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Feedback on Task 6.10

Regarding the preliminaries, if you made a check for equality of variation, there was a simple result for accuracy, i.e. a low value for F (0.98) and a high probability (this could have happened 33 times out of a hundred), i.e. indicating that the difference in the amount of variation is consistent with chance factors only. With complexity the F value is 3.95 and the probability value, 0.056, rounded to 0.06, is just outside the important decision point of 0.05. So the difference in the amount of variation between the groups is very close to requiring a bit of attention, but luckily, can just about be ignored. Incidentally, if the values had been more extreme, it would only really cause you to be more cautious. It wouldn’t really stop you proceeding.

Regarding the main output, your table should resemble that shown below. It doesn’t have to be exact in the interpretation column, but if you carried out the procedure correctly, it should be exactly the same in the first two numerical columns.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Value of t</th>
<th>Significance</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>-3.42</td>
<td>.002</td>
<td>The null hypothesis can be rejected as the value for t could only have arisen by chance 2 times in 1000, which is beyond the criterion level of 1 in a hundred. Planners achieve a higher level of accuracy in their performance on a decision making task than do the non-planners.</td>
</tr>
<tr>
<td>Complexity</td>
<td>-4.20</td>
<td>.000</td>
<td>The null hypothesis can be rejected, as the value for t is well beyond the 1 on a thousand level. Planners achieve a higher level of complexity in the language they use on a decision making task than do the non-planners.</td>
</tr>
</tbody>
</table>

If these are the basic statistical decisions, then in effect they licence us to go on to offer an interpretation of the results. We could propose that planning time, which leads to the significantly greater complexity, does so because it is, at least partly, directed at generating ideas which feed into what the learners will say in this decision making task. As a result, they may mobilise more complex ideas which push learners to use more complex language to do justice to such ideas.
Regarding accuracy, we might say that learners (a) also use planning time to rehearse what they are going to say and so have available a repertoire they can draw on with less error, and (b) that engaging in pre-task thinking of ideas eases how much attention needs to be directed during the task to the cognitive component with the result that more attention is left over for to focus on avoiding error.

Additional Features of Interpreting T-test Output

There are a number of lessons to be learned from what you have just done, so we will now work through a more extended discussion of a number of relevant (and general) points.

The Sign of $t$: Although we discussed the sign of $t$ earlier, it is worth a brief mention again. In the first t-test that was done as an example (for fluency), the sign for $t$ was positive, because the non-planners value was put into the equation as $M_1$ and the planners as $M_2$. In contrast, the non-planners value for accuracy (0.63) was lower than that for the planners (0.67), generating a negative sign. The same thing applied for the complexity scores (1.30 - 1.53).

The nature of the sign is not especially important here, as the value that enters the numerator of the equation first (as $M_1$) is essentially arbitrary. What really counts is whether the difference is significant - which it was.

How many tails?: You may recall that in the t-test output that SPSS provides you with, the heading for significance also uses the bracketed expression “(2-tailed)”. Statistics always distinguishes between one- and two-tailed approaches. (Never more than this!) It is tempting to say: ignore this and no harm will come to you! What follows here is no more than a brief bit of coverage. Essentially, when we have looked at the t-tests for Study 1 for fluency, accuracy, and complexity, we have avoided the notion of direction. In other words, we have said that a difference between the two groups was predicted, but the researchers avoided committing themselves to saying whether it would be planners who would be more fluent (or accurate or complex) and the non-planners less so, or vice versa. So either group having a different score from the other would have generated a significant result. But in fact it is perfectly possible to predict not only that there will be a difference between two groups, but also the direction of difference that will occur. We have been proceeding as if no specifically directional prediction was made, and because of this we have worked with the two-tailed approach, i.e. either end of the normal distribution would be relevant. The value that SPSS automatically outputs is such a two-tailed result, and we have simply ridden with this. But if a researcher does also make a directional prediction, it would be more appropriate for them to use a one-tailed approach. We will not go into this here, and if you want to follow it further, Howell, (1997), pp101-103 is a good source. Note in passing, though, that Foster and Skehan, who gathered the data on which this analysis is based, did make directional predictions, e.g. that planners would not just get different scores, but that they would be more fluent, more accurate, and more complex. In the actual study, therefore, one-tailed tests were appropriate. None of this,
surprising though it may seem, influences the preceding discussion. But it is nice to have a handle on why the terms one-tailed and two-tailed do keep cropping up.

*After establishing significance:* In Task 6.10 we saw that the t-test results justified rejection of the null hypothesis, for all three measures, and that therefore, the researchers were justified in moving on to the stage of interpretation. Recall that there are three important benchmarks to orient towards (the .05, the .01, and the .001). This makes it clear that each of our results surpasses the minimum level of .05. In fact, the results for complexity and fluency go well beyond that, reaching the most demanding level of all at .001. Rejecting the null hypothesis is therefore very confidently done for the complexity and fluency scores. It is certainly done for the accuracy scores too, but recall that the exact significance value output by SPSS was p < .002. We may have views here about the strength of the effect, and may wonder whether there is a greater impact of planning on the complexity of the language which is produced than on the accuracy. But each can now be interpreted, and, very importantly, related to the hypotheses and theorising which motivated the study in the first place. In the original article from which this data is drawn, Foster and Skehan (1996) went on to talk about how planning is used to establish priorities and to reinterpret a task so that learners push themselves to express more complex propositions, (higher complexity) as well as incorporate rehearsed language (higher accuracy). To put this another way, establishing significance, which for a beginner may appear to be an enormous hurdle to overcome, is in fact only the beginning. If significance is established, this becomes a licence to discuss interesting things which connect with the real ideas which have motivated the study.

*Degrees of Freedom:* SPSS computes the significance level for accuracy as 0.002, which is remarkably exact, and also hides a step in statistical decision-making. In the past, decisions were made not by reporting significance directly, but instead by computing a value for *t*, (using the formula), and then taking the value obtained to a table of probabilities, *which only gave critical values for the designated significance levels*. Here is an extract from the table that was used:

<table>
<thead>
<tr>
<th>Degrees of freedom</th>
<th>.05 level</th>
<th>.01 level</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>2.57</td>
<td>4.03</td>
</tr>
<tr>
<td>10</td>
<td>2.23</td>
<td>3.17</td>
</tr>
<tr>
<td>15</td>
<td>2.13</td>
<td>2.95</td>
</tr>
<tr>
<td>20</td>
<td>2.09</td>
<td>2.85</td>
</tr>
<tr>
<td>25</td>
<td>2.06</td>
<td>2.79</td>
</tr>
<tr>
<td>26</td>
<td>2.06</td>
<td>2.78</td>
</tr>
<tr>
<td>27</td>
<td>2.05</td>
<td>2.77</td>
</tr>
<tr>
<td>28</td>
<td>2.05</td>
<td>2.76</td>
</tr>
<tr>
<td>29</td>
<td>2.05</td>
<td>2.76</td>
</tr>
<tr>
<td>30</td>
<td>2.04</td>
<td>2.75</td>
</tr>
</tbody>
</table>

Recall that the degrees of freedom is really a reflection of the sample size. (The degrees of freedom in our present case are 30.) You can see that as the degrees of freedom increases, (i.e. the sample size gets bigger), the critical value for *t* reduces. In other words, unlike the case for the normal distribution, there are no absolute
unchanging values (e.g. 1.96 and 2.58) that have to be reached for \( t \) to become significant. The level of \( t \) required for significance changes as the sample size changes. And inspecting the table shows some dramatic reductions in the value of \( t \) required as the sample size increases from 5 to 15. Then some stabilisation occurs, and the value for 30 is getting close to that for the normal distribution (2.04 vs. 1.96 to reach the .05 level). Bear all this in mind when you interpret the helpful SPSS output.)

**Avoiding overclaiming:** Inferential statistics, of which the t-test is your first example, are concerned with decision making, and with examining the (im)plausibility of the null hypothesis. What you are required to do when you use inferential statistics is to be open and fair in the decision making that is involved. Typically, this involves not focussing on minute changes in significance levels (e.g. reporting that you result was significant at the .036285 level), but instead accepting the benchmark values we spoke of earlier, and reporting your results in terms of them. To put this more directly, the accuracy result you have obtained above is significant at the .002 level. (You know this thanks to SPSS’s considerateness and precision.) But you should report this result as significant at the .01 level, **because that is the conventional benchmark that you set yourself before you started.** Reporting results in terms of exact significant levels is not just tacky - it misrepresents the nature of the decision-making process and the benchmark standards that investigators have to accept. Claim the strongest significance level you can, but remember, there are effectively only three possibilities to choose from.

**Alternative notations:** The final point is only one of labelling. We have spoken in terms of one-in-twenty levels and so on, which is reasonable enough. The convention, though, when reporting statistical results, is to represent them in terms of probability figures. Conventionally, this is done as follows:

<table>
<thead>
<tr>
<th>Verbal description</th>
<th>Arithmetic Description</th>
<th>Statistical description</th>
</tr>
</thead>
<tbody>
<tr>
<td>one in twenty</td>
<td>.05</td>
<td>( p &lt; .05 )</td>
</tr>
<tr>
<td>one in a hundred</td>
<td>.01</td>
<td>( p &lt; .01 )</td>
</tr>
<tr>
<td>one in a thousand</td>
<td>.001</td>
<td>( p &lt; .001 )</td>
</tr>
</tbody>
</table>

When you read journal articles, you will typically read sentences or tables which use the formulaic presentation of “\( p < .01 \)” as the conventionalised form for saying that significance was achieved, and that it is justifiable to reject the null hypothesis. The above three versions (verbal, decimal, and statistical reporting) all mean the same thing, but would be differently appropriate for different contexts. In general, in writing, use the last of them.
Task 6.11

Now return to the dataset for Study One, and calculate t-tests for the narrative task for:

- fluency
- accuracy
- complexity

Lay out the results in the following table:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean Non-planners</th>
<th>Mean Planners</th>
<th>t</th>
<th>Sig.</th>
<th>Null Hypothesis</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>fluency</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>accuracy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>complexity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

It would also be good if you could give some thought to interpretation.

Feedback on Task 6.11

What we get here is a similar, yet slightly contrasting pattern of results to those which were obtained for the decision-making task. The values that you should have obtained are as follows:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean Non-planners</th>
<th>Mean Planners</th>
<th>t</th>
<th>Sig.</th>
<th>Null Hypothesis</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>fluency</td>
<td>30.25</td>
<td>12.46</td>
<td>4.138</td>
<td>.001</td>
<td>Reject</td>
<td></td>
</tr>
<tr>
<td>accuracy</td>
<td>0.61</td>
<td>0.62</td>
<td>-0.371</td>
<td>.71</td>
<td>Accept</td>
<td></td>
</tr>
<tr>
<td>complexity</td>
<td>1.22</td>
<td>1.52</td>
<td>-3.312</td>
<td>.004</td>
<td>Reject</td>
<td></td>
</tr>
</tbody>
</table>

The results for complexity (p < .01) and fluency (p < .001) are highly significant and allow the null hypothesis to be rejected in each case. The results for accuracy contrast with this (p > .05) and so in this case the null hypothesis has to be accepted.

Before we move on to a brief interpretation, there is one complication that needs to be addressed. You may have noticed in the SPSS output that our luck has run out with the issue of equal variances for the non-planners and the planners, at least for the complexity scores. (The fluency and accuracy scores were fine in this regard and we can continue to assume equal variances for them.) The result for the Levene test was significance at the .001 level. For this reason, the figures given in the above table are from the lower line in the output, and so .004 is given for complexity, rather than the .001 which is given in the top line, but which assumes equal variances. In
other words, we have had to claim the lower significance level for this particular measure. Note also, in the previous paragraph, this was reported as "p < .01", since we have to avoid overclaiming.

Regarding interpretation, it appears that the effects of planning on complexity and fluency are similar for the narrative task to those found for the decision-making task. As a result, the previous interpretations can apply. But there is a contrast with the narrative task regarding accuracy, in that there is no significant difference. As one of the researchers involved, I have to say that this was disappointing. The provisional interpretation I would offer (but this has to connect with a range of other research findings which are not discussed here) is that the narrative task is more monologic in nature and is generally more difficult. In particular, the need to handle longer turns, spanning several utterances, seems connected to learners having less attention left over for a focus on error avoidance. Hence the lack of scope for planning to exert an influence on accuracy.

Further Interpretation: The results from Task 6.11 are for the Narrative task, and so it is now worth trying to get a wider perspective on what is going on here, and to incorporate discussion based on results for both tasks, decision-making and narrative. It appears that giving learners planning time prior to their completing tasks is a good thing. It produces very clear improvement in the quality of performance for both a decision making and a narrative task when the measure of performance is fluency and complexity. Planned performance avoids pauses, and uses a greater degree of subordinated language. Both these measures are taken to reflect wider concepts of fluency and complexity. The results, though, are less clear cut in the case of accuracy. Significance is achieved with the decision making task (albeit at “only” the p < .05 level), but the result for the narrative does not show any significant effect for accuracy. We can conclude here that while planning is generally beneficial, the strength of the effect is found more clearly in certain areas of performance (fluency and complexity) than others (accuracy). In a case such as this, this intriguing pattern of results might lead to further studies designed to probe precisely which variables control the nature of the accuracy effect. One could speculate, for example, that it is the more interactive nature of the decision making task which supports an accuracy effect, while the more monologic narrative draws people into expressing greater degrees of complexity, but somewhat at the expense of accuracy. Speculations at this point are normal. Results nearly always come out as partially confirming original hypotheses and partly not. The consequence is almost invariably that the original motivating ideas have to be modified and enriched as a result. Good questions, in other words, always lead to further questions - that is the nature of research.
Chapter Six Statistical Inference and T-tests

Conclusion

You have now learned to operate two inferential procedure – the Critical Ratio and the t-test. In fact, the first of these has really been covered only for pedagogic reasons. Practically, it is not used, since in our field we do not work with large group sizes. In any case, as group sizes do get as large as they get, the Critical Ratio in effect approximates the t-test, and it is this latter which SPSS uses, and which you have learned to work with.

The t-test procedure is appropriate when you have two groups to compare. Although this is only one procedure, it is fortunate that the “two-groups” problem is pervasive in applied linguistics, and teacher education. This one statistical procedure will, in other words, take you a long way, since many interesting questions in our field can be cast as a comparison between two groups.

Unfortunately, there are many questions which are different, or more complex. However, if that’s the bad news, there is also good news! You have learned to operate within the following parameters:

- Two groups of different people
  - t-distribution
  - Between subjects t-test and t
  - Decision-making procedures with the null hypothesis and significance levels

Look at the following table, showing three slightly different statistical scenarios juxtaposed:

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Two groups, different people in each group</th>
<th>Two groups, the same people in each group</th>
<th>Multiple groups, e.g. 3 or more</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distribution</td>
<td>t-distribution</td>
<td>t-distribution</td>
<td>F-distribution</td>
</tr>
<tr>
<td>Statistical Procedure or Formula</td>
<td>t-test, between subjects</td>
<td>t-test, within subjects</td>
<td>ANOVA (Analysis of Variance)</td>
</tr>
<tr>
<td>Decisions</td>
<td>Decision-making procedures with the null hypothesis and significance levels</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The two groups, different people column addresses the same situation we have been looking at for the last few pages, and the one you have learned to handle with SPSS. The next column, two groups but the same people, handles a related situation. Although we are still dealing with two groups, we now imagine that we have two measures on each member of a single group. For example, we might have a score at
the beginning of a course of some sort, and also a score (on the same test or sort of test) from the end of a course. We may want to compare the mean score from the beginning with the mean score from the end. Alternatively, and drawing on the data from Display1.sav, we could think of the example of a complexity score on the decision task and a complexity score on the narrative. Once again, there would be two scores, but only one group of people who had in this case completed two different language learning tasks. The question here would be whether the average decision making complexity score is different from the average narrative complexity score. In this case, the appropriate statistical test would be the within-subjects version of the t-test. This uses a slightly different formula, but is otherwise very similar. It, too, is found in the Compare Means set of choices in SPSS.

Finally, in Table 6.2, we have the multiple groups column. This might involve three groups of the same subjects, e.g. with complexity scores for decision-making, narrative, and personal information exchange tasks. Or it might be more complex with what is called a factorial design, such as:

<table>
<thead>
<tr>
<th></th>
<th>No Post Task</th>
<th>Post Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planners</td>
<td>Group A</td>
<td>Group B</td>
</tr>
<tr>
<td>Non-Planners</td>
<td>Group C</td>
<td>Group D</td>
</tr>
</tbody>
</table>

In this case we would have two “dimensions” working simultaneously, allowing us to compare planners vs non-planners (i.e. Groups A and B vs. Groups C and D), and to compare Post Taskers with Non Post-taskers (i.e. Groups B and D vs. Groups A and C). We could even compare whether there is something special about the combination effect of, for example, planners who also get the chance to do a post-task. The point, though, is that this set of more complex comparisons take us beyond t-tests, and require different statistical approaches. In fact, as you can see, a technique known as Analysis of Variance is needed, based on a different distribution, the F distribution.

Although the first three rows of the table are different, though, the crucial point is that the final row is exactly the same, whatever the statistical technique that is used. Although there may be different statistical problems to solve, and although there may be different distributions and formulae, the manner in which decisions are made, and the manner in which these decisions are expressed (e.g. p<.01) are exactly the same. What you have learned about is statistical reasoning in general, and the particular use of the t-test. The general statistical reasoning applies everywhere – what differs is simply the way in which particular statistical problems are handled. But the ultimate expression of these activities is always the same – whether or not the null hypothesis is accepted or rejected, and the significance levels associated with this. You are now in a better position to understand how this is done.

References
