CHAPTER TEN

Analyzing Quantitative Data

This chapter presents introductory information about statistics to enable the reader to begin to understand basic concepts. We focus on issues and methods of analysis that are common in second language research. The chapter deals with descriptive as well as analytic measures and also addresses concepts such as normal distribution, standard scores, probability, and effect sizes, all of which are necessary to an understanding of basic statistical procedures. This chapter serves only to briefly familiarize the reader with central topics; for more extensive information, it is recommended that statistics books be consulted.

10.1 INTRODUCTION

In Chapter 4, we considered issues of data coding and basic data description. These were important prerequisites to the topic of analyzing data. This chapter focuses on issues of analysis and, in particular, provides background information on statistical procedures commonly used in second language research.[1] We recommend that before conducting statistical analyses of data, researchers gain greater knowledge of statistics through academic coursework, statistical texts, and/or consultations with statistical experts.

10.2 DESCRIPTIVE STATISTICS

The first issue we deal with has to do with description and data display. Descriptive statistics can help to provide a simple summary or overview of the data, thus allowing researchers to gain a better overall understanding of the data set.

Because raw data are not in and of themselves revealing, they must be organized and described in order to be informative. In this section, we present an overview of three different types of descriptive statistics: (1) measures of frequency; (2) measures of central tendency; and (3) measures of variability or dispersion. We will also discuss ways of displaying these data visually to facilitate the exposition of summaries of findings. Data displays are generally prescribed by the particular style manual one is using. Many journals with a second language research focus require APA, but one needs to consult each journal for the required style sheet.

TIME TO DO ...

Look up two to three journals you are familiar with and find out the style formatting required.

10.2.1 Measures of Frequency

Measures of frequency are used to indicate how often a particular behavior or phenomenon occurs. For example, in second language studies, researchers might be interested in tallying how often learners make errors in forming the past tense, or how often they engage in a particular classroom behavior. One of the most common ways to present frequencies is in table format. For example, in Table 10.1 below, we present a sample frequency table from Storch and Tapper (1996), who provide the frequencies of different types of annotations that second language writers made on their own texts, indicating the areas where they felt they were having difficulty.

In addition to tables, frequencies may also be represented graphically in forms such as histograms, bar graphs, or frequency polygons. In these graphic representations, the categories are typically plotted along the horizontal axis (x-axis), while the frequencies are plotted along the vertical axis (y-axis). For example, if we were to convert Storch and Tapper's (1996) frequency table (Table 10.1) into a graphic representation, one possible way would be through the bar graph seen in Figure 10.1a, with the same data displayed using a line graph (Figure 10.1b).

Frequencies, as well as measures of central tendency (which are described below), are often presented in second language studies even when they do not relate directly to the research questions. This is because frequency measures provide a succinct summary of the basic characteristics of the data, allowing readers to understand the nature of the data with minimum space expenditure.

TABLE 10.1 Sample frequency table

Content of student annotations ***Content of annotation***	***Number***	***Total***
(i) Syntactic		
Preposition or verb + preposition	21	
Verb tense	17	
Word order/sentence structure	17	
Articles	10	
Singular/plural agreement	8	
Word form	6	
Other	37	116
(ii) Lexical	70	70
(iii) Blanket requests		
Tenses	13	
Grammar	7	
Sentence structure	6	
Punctuation	4	
Other	29	59
(iv) Discourse organization	5	5
(v) Ideas	5	5
TOTAL		255

Source: Storch, N., & Tapper, J. (1996). Patterns of NNS student annotations in identifying areas of concern in their writing. *System*, *24*(3), 323–336 (excerpt from p. 329). Copyright ©1996 by Elsevier Science Ltd. Reprinted with the permission of Elsevier Science Ltd.

TIME TO THINK ...

Consider Table 10.1 and Figures 10.1a and 10.1b, which represent three ways of displaying data. Which of these ways is most useful to you as you try to interpret the results of the study from which these data came? Why? Is a line graph a reasonable way of displaying these data? Why or why not? Consider whether there really is a relationship between the category types.

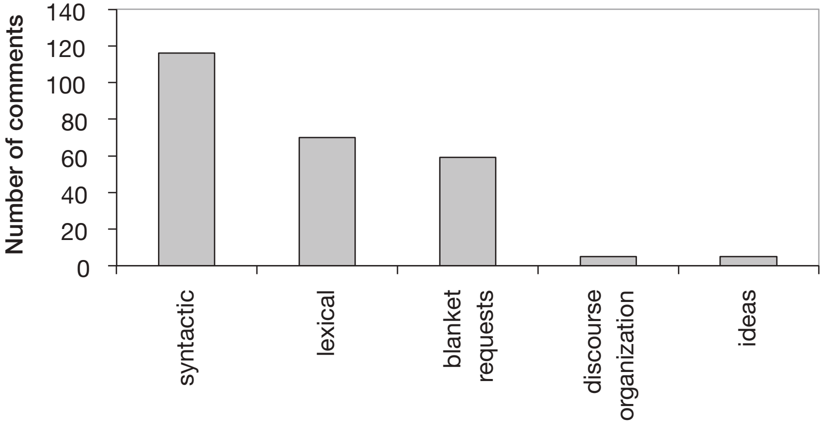

FIGURE 10.1a Sample frequency bar graph

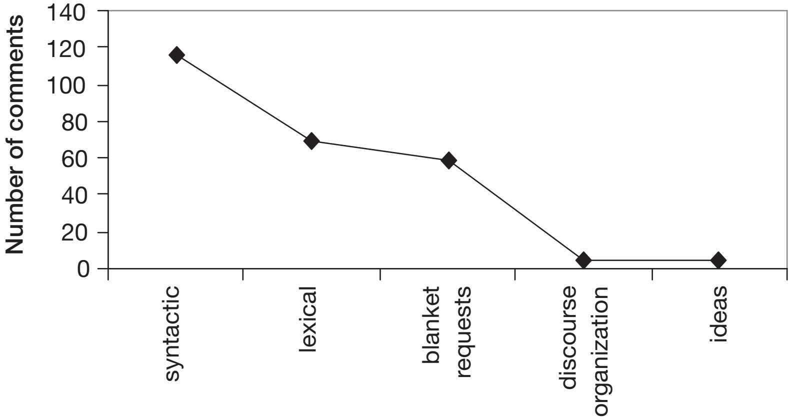

FIGURE 10.1b Sample line graph

Also, frequencies and measures of central tendency can help researchers determine which sorts of statistical analyses are appropriate for the data.

In order to visualize trends in the data, it is almost always useful to plot the data even before carrying out statistical analyses. In this section, we show various ways of visually representing data (e.g., see Figure 10.2 and bar graphs); these and other visual means of representation are useful in order to gain an overall impression of the patterns in the data. For example, creating a scatterplot (Figure 10.3) to assist with visualization of a data set (see correlation figures in section 10.12.1) can provide an early picture of any outliers in one's data (see, for example, the person who arrived at age 2, but had the lowest proficiency test

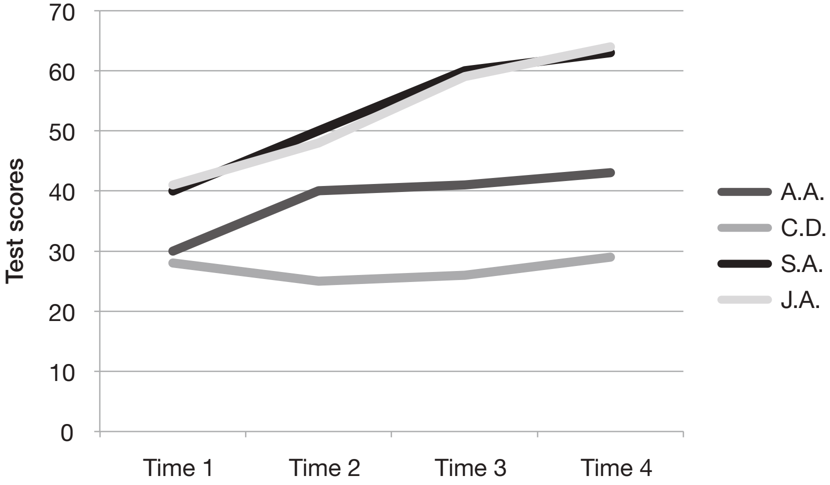

FIGURE 10.2 Sample line graph

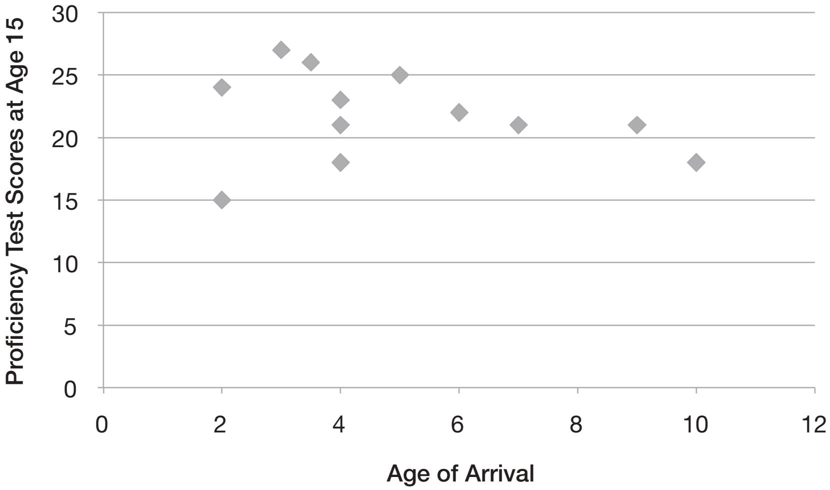

FIGURE 10.3 Scatterplot showing relationship between age of arrival and a proficiency test score (maximum score of 30)

TIME TO DO ...

Consider Figure 10.2. How would you verbally interpret this line graph? Which student(s) showed the greatest improvement over time? Which the least?

score [15]). Providing visual representations of results in graphical form can also contribute to a clearer understanding of any patterns confirmed through statistical testing.

10.2.2 Measures of Central Tendency

While simple frequencies are useful ways of providing an initial picture of the data, they are not as precise as other measures, particularly when the data are obtained from different groups. Second language researchers often use one or more measures of central tendency to provide precise quantitative information about the typical behavior of learners with respect to a particular phenomenon. There are three commonly used measures of central tendency, each of which will be discussed below.

10.2.2.1 Mode

Arguably the easiest measure of central tendency to identify is the mode. Simply put, the mode is the most frequent score obtained by a particular group of learners. For example, if the ESL proficiency test scores recorded for a group of students were 78, 92, 92, 74, 89, and 80, the mode would be 92 because two students in this sample obtained that score. Although this measure is convenient in that it requires no calculations, it is easily affected by chance scores, especially if the study has a small number of participants. For this reason, the mode does not always give an accurate picture of the typical behavior of the group and is not commonly employed in second language research.

10.2.2.2 Median

Another measure of central tendency that is easy to determine is the median. The median is the score at the center of the distribution—that is, the score that splits the group in half. For example, in our series of ESL proficiency test scores (78, 92, 92, 74, 89, and 80), we would find the median by first ordering the scores (74, 78, 80, 89, 92, 92) and then finding the score at the center. Since we have an even number of scores in this case (namely, six), we would take the midpoint between the two middle scores (80 and 89), or 84.5. This measure of central tendency is commonly used with a small number of scores or when the data contain extreme scores, known as outliers (see section 10.2.2.4 for an explanation of outliers).

10.2.2.3 Mean

The most common measure of central tendency is the mean, or the arithmetic average.[2] Furthermore, since the mean is the basis for many advanced measures (and statistics) based on group behavior, it is commonly reported in second language studies. For our scores (78, 92, 92, 74, 89, and 80), the mean would be the sum of all scores divided by the number of observations, or (Σx /n =) 84.2. It should be kept in mind that even though the mean is commonly used, it is sensitive to extreme scores especially if the number of participants is small.

The mean may be represented visually through the use of graphics, including a *bar graph*. For example, Toth (2000) created the graph in Figure 10.4 for his study of the role of instruction, L2 input, and Universal Grammar in the acquisition of the Spanish morpheme *se* by English-speaking adult learners. In this graph, he provides a visual representation of the means of three different groups on three acceptability judgment tests (pre-test, post-test, and delayed post-test). This visual presentation using a bar graph succinctly summarizes the information.

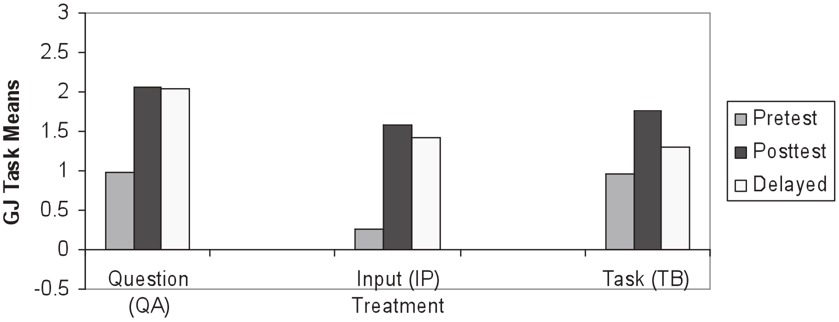

FIGURE 10.4 Visual presentation of group means: bar graph

Source: Toth, P. D. (2000). The interaction of instruction and learner-internal factors in the acquisition of L2 morphosyntax. *Studies in Second Language Acquisition*, *22*(2), 189. Copyright © 2000 by Cambridge University Press. Reproduced with the permission of Cambridge University Press.

TIME TO DO ...

Interpret this figure. How would you interpret the post-test results? Which group did better? For which group did learning persevere? Which group gained the most? At which point in time?

Alternatively, means may be shown through the use of a frequency polygon, also known as a line graph. For example, Zsiga (2003) compared patterns of consonant-to-consonant timing at word boundaries in Russian and English to investigate the roles of transfer and the emergence of linguistic universals in second language articulation. Zsiga provided the following graph to illustrate a significant interaction between L1 and language spoken, showing that the articulatory timing patterns of native Russian and English were different (Figure 10.5). The graph shows the mean duration ratios for English speakers and Russian speakers speaking their L1s and L2s, respectively.

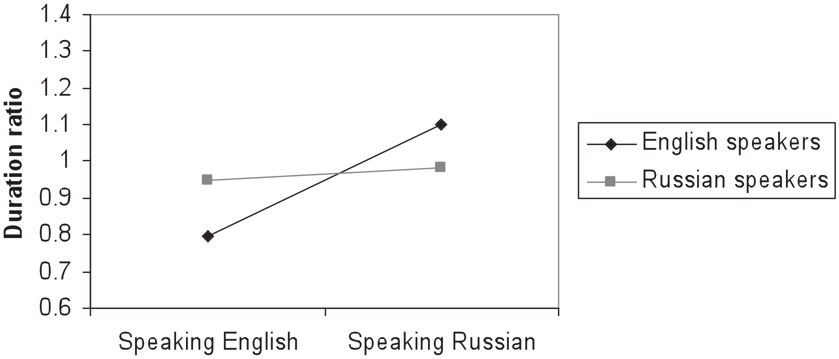

FIGURE 10.5 Visual presentation of means: line graph

Source: Zsiga, L. (2003). Articulatory timing in a second language. *Studies in Second Language Acquisition*, *25*(3), 413. Copyright © 2003 by Cambridge University Press. Reproduced with the permission of Cambridge University Press.

In terms of measures and displays of central tendency and summaries of the data, it is always important to be flexible to the needs of your particular research questions and data set. In the words of Woods, Fletcher, and Hughes (1986):

> Although it will usually be possible to display data using one of the basic procedures . . . you should always remain alive to the possibility that rather special situations may arise where you may need to modify or extend one of those methods. You may feel that unusual data require a rather special form of presentation. Remember always that the major purpose of the table or graph is to communicate the data more easily without distorting its general import.
>
> (pp. 20–21)

10.2.2.4 Outliers

Earlier in this section, we raised the concept of outliers. This refers to data that seem to be atypical of, or lying outside, the rest of the data set. The presence of outliers strongly suggests that the researcher needs to take a careful look at the data and determine whether the data collected from specific individuals are representative of the data elicited from the group as a whole. There are times when researchers may decide not to include outlier data in the final analysis, but if this is the case there needs to be a principled reason for not including them beyond the fact that they "don't fit right." Should researchers decide that there are principled reasons for eliminating outlying data, a detailed explanation in the research report needs to be provided. Below, we provide two hypothetical examples where a researcher might, after careful consideration, decide to eliminate some data.

Example 1:

Data elicitation: Sentence matching (see Chapter 3 for further discussion of this elicitation technique). Participants are instructed to press the Yes button (the J key on the keyboard) if the sentences match or the No button (the F key on the keyboard) if the sentences do not match.

Problem: One participant has: (a) pressed only the Yes button throughout the experiment for all the sentences; and (b) consistently pressed it very quickly (i.e., the reaction times are much faster than the average).

Possible reason: Participant was not attentive to the task and repeatedly pressed only one button, suggesting that there was little processing going on.

Decision: Delete this individual's data.

Justification: These data did not represent the processing that one has to assume for sentence matching.

Example 2:

Data elicitation: Child-child interactions in which the researcher is measuring feedback provided by children to their peers.

Problem: One child's behavior appears to be unlike the others in that no feedback is ever provided.

Further exploration: In talking to the teacher, it was found that the child had a severe learning disability. It was also typical for this child not to stay on task in other classroom activities.

Decision: Delete this individual's data.

Justification: This child was most likely not on task. This child did not represent the population from which data were being collected.

Both of the examples are based on data that appeared to be unlike the rest of the data set. The data were not immediately deleted, but because they were outliers, the researchers took a closer look at what might have been going on. It was only after careful consideration and a determination that these data did not reflect a valid characterization of the construct of interest that the researchers decided that it was not appropriate to include the data in the final data pool. As mentioned earlier, information and justification of decisions like this should be included in a final research report.

Examples 1 and 2 illustrate occasions where it may be necessary to eliminate all of an individual's data when the extent to which he or she was on task is questionable. There are also cases when it may be appropriate to remove a subset of the data. For example, Duffield and White (1999) excluded from analysis responses on a sentence-matching task from any participant who had an overall error-rate of greater than 15 percent (e.g., said sentences were different when they were actually the same, and vice versa). Or, in some reaction time experiments, such as the one described in Example 1, a researcher might eliminate responses greater than a certain length of time, known as a cutoff point (e.g., 5000 msec [Lotto & de Groot, 1998]). A researcher might also move responses longer than the cutoff time to that cutoff point. For example, Duffield and White (1999) calculated the mean response time on a sentence-matching task for each individual. All responses that "fell outside a cut-off of ±2 standard deviations (see below) of a particular subject's personal mean were corrected to the corresponding cut-off value" (p. 145). In other words, there was a maximum response time value that was used in their analysis.

TIME TO THINK ...

Think of a situation where you might want to eliminate data. What would your rationale be for eliminating those data? And think of a situation where you would not want to eliminate data. Why not?

10.2.3 Measures of Dispersion

Measures of central tendency are useful in that they give both the researcher and the reader an idea of the typical behavior of the group. However, the use of measures of central tendency alone may also obscure some important information. For instance, consider the hypothetical case of two groups of learners who take a final exam. One group of students obtains scores of 45, 99, 57, 17, 63, and 100, while the other group obtains scores of 66, 62, 65, 64, 63, and 60. Both groups have approximately the same mean (63.5 and 63.3, respectively). However, if you report only the mean, you will not be able to show that the groups have a fairly different dispersion of scores: one group's scores are all close to the mean; the other group's scores are more widely dispersed. How can we present this additional information on the dispersion, or variability, of scores?

One informal way to do so is by presenting the range of scores. The range is the number of points between the highest and lowest scores on the measure. For example, the range for the first group of test scores would be 83 (17–100), while the range for the second would be 6 (60–66). The range, though easy to calculate, is not commonly reported in second language studies because it is sensitive to extreme scores and thus is not always a reliable index of variability.

A more common way of measuring variability is through the calculation of the standard deviation. Simply put, the standard deviation is a number that shows how scores are spread around the mean; specifically, it is the square root of the average squared distance of the scores from the mean. In other words, one takes the differences between each score and the mean and squares that difference. The next step is to add up these squared values, and divide by the sample size. The resulting number is called the variance. The standard deviation is the square root of the variance. As an example, consider the scores given above, 45, 99, 57, 17, 63, and 100. To calculate the standard deviation, the following steps are taken:

1. Calculate the mean. $\Sigma x / n = 63.5$
2. Subtract the mean from each score and square the difference $(63.5–x)^2$.

Score (x)	*Mean*	*Difference*	*Difference squared*
49	63.5	–14.5	210.25
99	63.5	35.5	1260.25
57	63.5	–6.5	42.25
17	63.5	–46.5	2162.25
63	63.5	–0.5	0.25
100	63.5	36.5	1332.25

3. Sum the differences squared and divide by the number of scores (6) to arrive at variance.

Score (x)	*Mean*	*Difference*	*Difference squared*
49	63.5	–14.5	210.25
99	63.5	35.5	1260.25
57	63.5	–6.5	42.25
17	63.5	–46.5	2162.25
63	63.5	–0.5	0.25
100	63.5	36.5	1332.25

Σ = 5007.5
variance = 834.58

4. Take the square root of the variance.

   SD = 28.89

The second set of scores given above (66, 62, 65, 64, 63, and 60) are closer to one another. If we do the same calculation as above, we see that the variance is 3.89 and the standard deviation is 1.97. Thus, while the means are similar, the amount of dispersion from the mean is quite different.

The larger the standard deviation, the more variability there is in a particular group of scores. Conversely, a smaller standard deviation indicates that the group is more homogeneous in terms of a particular behavior. We return to standard deviations below in our discussion of normal distributions.

Because the mean does not provide information about how scores are dispersed around the mean, the standard deviation (SD) should always be reported in second language research, often in a table along with the mean (M) and the number of participants (n). An example of a table (Table 10.2) with this information comes from Rodríguez and Abreu (2003), who investigated the construct of *anxiety* in pre-service teachers (native speakers of Spanish) majoring in English and French. The teachers were at different proficiency levels and came from two universities. Below, we present only the results from one of the universities in their example. This shows a table that portrays descriptive information.

One can examine the SDs and means in relation to one another. All groups (Table 10.2) (with the exception of level 5 for French Anxiety) are more or less equally dispersed from the means. If SDs are consistently large compared to the mean, you have groups with little homogeneity. In general, researchers should closely examine data with SDs that are consistently larger than the mean. Measures of dispersion (particularly standard deviations) can serve as a quality control for measures of central tendency; the smaller the standard deviation, the better the mean captures the behavior of the sample.

TABLE 10.2 Sample mean and standard deviation table

	English Anxiety			*French Anxiety*		
Level	*M*	*SD*	*n*	*M*	*SD*	*n*
1	74.42	14.87	12	76.75	17.32	12
3	90.42	14.98	12	89.08	13.99	13
5	94.38	14.00	8	93.50	120.52	8
Overall	84.39	17.45	33	85.15	15.80	33

Note: Maximum Score = 165.

Source: Rodríguez, M., & Abreu, O. (2003). The stability of general foreign language classroom anxiety across English and French. *The Modern Language Journal*, *87*, 371. Copyright © 2003 by Blackwell. Reproduced with the permission of Blackwell.

As was the case with frequencies, this information can also be represented visually. For example, Morgan-Short et al. (2012) provided the graph shown in Figure 10.6 in their study of the effect of different instruction conditions on the ability of adult Japanese ESL learners to acquire a rule about verbs. In this graph, the mean scores are represented by the height of the bars, while the black line extending from the top of each bar represents the size of the standard deviation.

As we noted earlier, it is important to include measures of variability in descriptions of data. Each offers different information, but when taken together they provide a richer understanding of the data than when viewed alone. As will be seen later in this chapter, means and standard deviations figure prominently in many statistical analyses.

10.3 NORMAL DISTRIBUTION

A normal distribution (also known as a bell curve) describes the clusterings of scores/behaviors. In a normal distribution, the numbers (for example, scores on a particular test) cluster around the midpoint. There is an even and decreasing distribution of scores in both directions. Figure 10.7 shows a normal distribution. As can be seen, the three measures of central tendency (mean, mode, median) coincide at the midpoint. Thus, 50 percent of the scores fall above the mean and 50 percent fall below the mean. Another characteristic of a normal distribution relates to the standard deviation. In a normal distribution, approximately 34 percent of the data lie within one standard deviation of the mean (above and

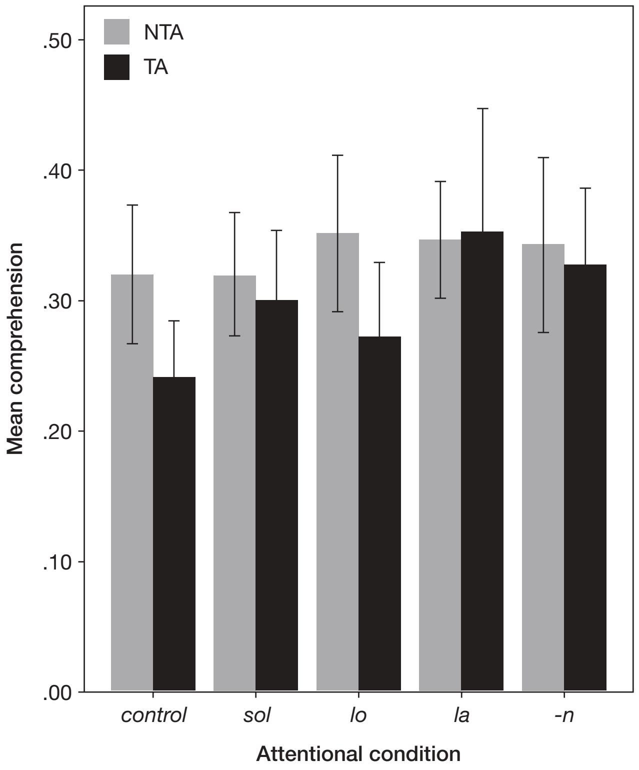

FIGURE 10.6 Sample visual representation of mean and standard deviation (comprehension scores per TA group per attentional condition)

Source: Morgan-Short, Heil, Botero-Moriarti, & Ebert (2012). Allocation of attention to second language form and meaning: Issues of think-alouds and depth of processing. *Studies in Second Language Acquisition*, *34*, 675. Copyright © 2012 by Cambridge University Press. Reproduced with the permission of Cambridge University Press.

below) and thus comprise 68 percent of the data. If we look at two standard deviations above and below the mean, we capture an additional 27 percent for a total of 95 percent. Thus, only 5 percent of the data in a normal distribution lies beyond two standard deviations from the mean. Finally, approximately 2.13 percent of the data fall between two and three standard deviations, leaving only approximately 0.3 percent of the data beyond three standard deviations above and below the mean. If we know that a group of scores is normally distributed and if we know the mean and the standard deviation, we can then determine where individuals fall within a group of scores. Many statistics assume normal distribution of scores. We return to this concept below. Figure 10.7 represents a normal distribution with the mean, mode, and median corresponding at the midpoint.

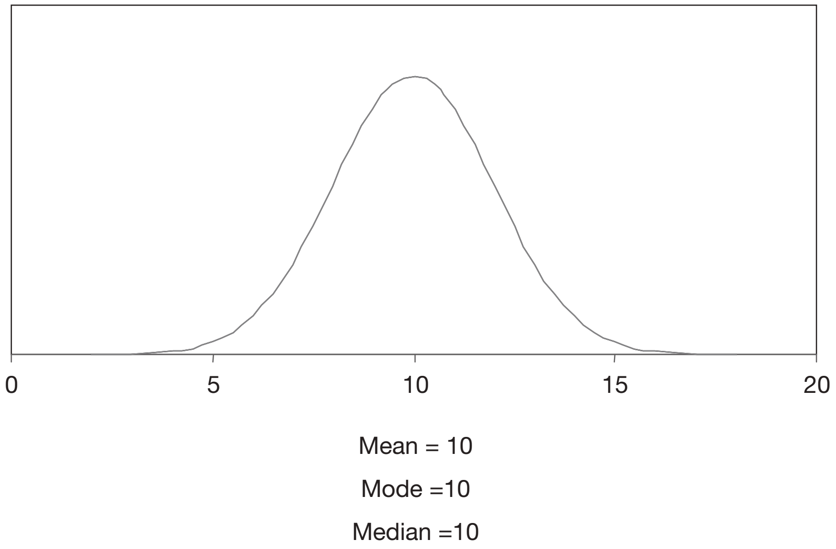

FIGURE 10.7 Normal distribution

10.4 STANDARD SCORES

There are times when we want to compare an individual's performance on different tests. For example, we might want to compare a score on a vocabulary test with a score on a test of grammar. Given the nature of the two tests, it is also likely that the maximum score on each is different. It would, of course, not be prudent to compare a score of 22 on one test with a score of 22 on another when one was based on a total possible score of 75 and the other based on a total possible score of 25. One way to make a more meaningful comparison is to convert these raw scores into standard scores.

The two most common standard scores are *z* scores and *T* scores. *Z* scores use standard deviations to reflect the distance of a score from a mean. If a score is one standard deviation above the mean, it has a *z* score of +1, a score that is two standard deviations above the mean has a *z* score of +2, and a score that is one standard deviation below the mean has a *z* score of –1. The calculation of a *z* score is straightforward: we subtract the mean from the raw score and divide the result by the standard deviation. The formula is given in Appendix G.

A second common standard score is the *T* score. In essence, it is a converted *z* score. *Z* scores are often expressed in negative terms (when they are below the mean) and in fractions. For certain manipulations of scores, negative scores are inappropriate. If non-negative standard scores are needed, *T* scores are commonly used. *T* scores are calculated by multiplying the *z* score

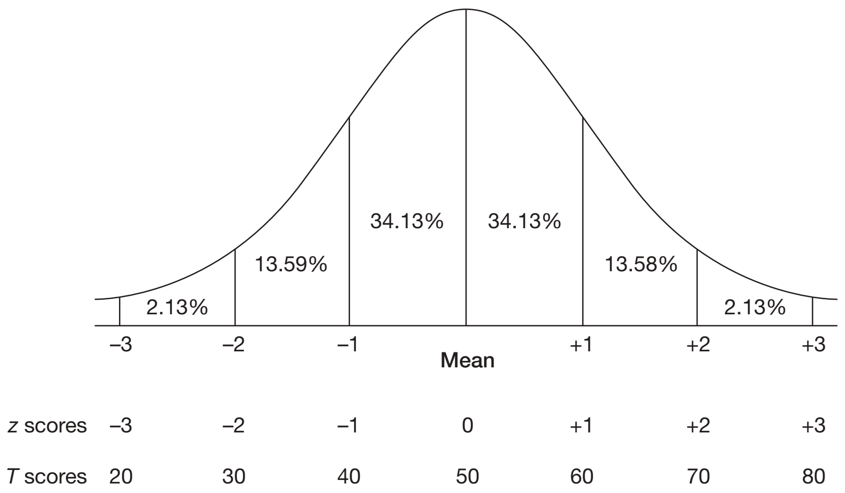

FIGURE 10.8 Means, standard deviations, *z* scores, *T* scores

by 10 and adding 50 ((z × 10) + 50). Consider a test with a mean of 60 and a standard deviation of 14. A learner who receives a score of 39 on this test has scored one and one-half standard deviations below the mean, and would have a *z* score of –1.5 and a *T* score of 35. The relationship between means and standard deviations and the two standard scores discussed here can be seen in Figure 10.8.

TIME TO THINK ...

Think of a circumstance when you might want to convert scores to a standard score.

10.5 PROBABILITY

The purpose of conducting statistical tests is to provide information about the likelihood of an event occurring by chance. The probability value (referred to in research reports as the *p*-value) that is reported is designed to provide confidence in the claims that are being made about the analysis of the data. We are all familiar with the concept of probability from everyday life. Insurance companies rely on the concept of probability to determine rates using various factors such as age, and health (for life insurance) or age and driving record (for automobile insurance). The general way of expressing probability is through a percentage (.20 = something occurring 20 percent of the time). Probability is an expression of the likelihood of something happening again and again. For example, if the probability is .05, there is a 5 percent possibility that the results were obtained by chance alone. If the probability is .50, there is a 50-50 possibility that the results were obtained by chance. The generally accepted *p*-value for research in second language studies (and in other social sciences) is .05. A *p*-value of .05 indicates that there is only a 5 percent probability that the research findings are due to chance, rather than to an actual relationship between or among variables. In second language research reports, probability levels are sometimes expressed as actual levels and sometimes as simply greater or less than .05 or some other probability level. Table 10.3 (modified from Ortega, 1999) shows actual *p*-values from a study on planning and focus on form with L1 English speakers learning Spanish. The column labeled *F*-value reflects the specific statistical procedure that Ortega used, analysis of variance (ANOVA), and will be discussed later in this chapter, a less preferred method of reporting given that best statistical programs present actual *p* values.

Table 10.4, from a study on word meaning by L1 English speakers learning Spanish (modified from Barcroft, 2003), shows *p*-values being expressed in relation to .05 and .01.

In Chapter 5, we introduced the concept of null hypotheses. Null hypotheses predict that there is no relationship between two variables; thus the statistical goal is to test the hypothesis and reject the null relationship by showing that there is a relationship. Let us take the following hypothesis: "Resumptive pronouns [*The man that I saw him is very smart*] will decrease with time." This hypothesis predicts change in a particular direction, that is, the occurrence will decrease over time. We could express this hypothesis as a null hypothesis as

TABLE 10.3 Example of expression of probability levels

Summary of findings from ANOVAs on IL measures

Measure	***F-Value***	***P-Value***
Words per utterance	8.444	.0002
Noun-modified TLU	5.8472	.0217
Pruned speech rate	16.0625	.0004
Type-token ratio	1.5524	.2221
Article TLU	4.3599	.0451

Source: Ortega, L. (1999). Planning and focus on form in L2 oral performance. *Studies in Second Language Acquisition*, *21*, 126. Copyright © 1999 by Cambridge University Press. Reprinted with the permission of Cambridge University Press.

TABLE 10.4 **Example of expression of probability levels**

Repeated measures ANOVA for effect of condition and time on cued recall

Source	F
Time	4.84*
Condition	9.06**

*p < .05. **p < .01

Source: Barcroft, J. (2003). Effects of questions about word meaning during L2 Spanish lexical learning. *The Modern Language Journal*, *87*, 557. Copyright © 2003 by Blackwell. Reproduced with the permission of Blackwell.

follows: "There is no relationship between the use of resumptive pronouns and the passage of time." We can then test whether the null hypothesis can be rejected. Consider the hypothetical scenarios below representing the number of instances of null subject use over time.

Time 1	**Time 2**	**Time 3**	**Time 4**
Scenario 1			
4	2	2	1
Scenario 2			
30	20	8	1

We can see that the difference in the number of instances in Scenario 1 is slight, suggesting that this may be a random finding and that were we to repeat this study many times, the results would be different. If we were to do a statistical test, we would probably come up with a high *p*-value and we would have little confidence that our results would be the same were the test to be repeated. On the other hand, the difference in the numbers in Scenario 2 is such that we would have more confidence in our results not being due to chance alone. A low level of probability would indicate this.

As noted above, probability is an estimation of the likelihood of something occurring due to chance. Two potential problems with such estimates, commonly referred to as Type I and Type II errors, are noteworthy. Type I errors occur when a null hypothesis is rejected when it should not have been rejected; Type II errors occur when a null hypothesis is accepted when it should not have been accepted. Examples are provided below.

Error Type	*Definition*	*Example*
Type I	Reject null hypothesis when it should not be rejected.	A statistical test shows a significant difference between an experimental and a control group ($p < .05$) and the researcher confirms that a treatment has been successful when in actuality it was unlikely that the two groups were different.
Type II	Accept null hypothesis when it should not be accepted.	A statistical test shows no significant difference between an experimental and a control group and the researcher confirms the treatment has not been successful when there really was a difference.

The approach discussed above is also known as null hypothesis significance testing (NHST) and relies on establishing a hypothesis that is then supported or rejected at some level of confidence (*p*-value). This discussion will be expanded when we return to the construct of effect sizes (section 10.11.2). Before moving to a discussion of statistics, we utter a word of caution about the difference between significance and meaningfulness. When we have a large sample size, it is often not difficult to get statistical significance. Assume, for example, that you are testing the effect of recasts versus models for the learning of irregular past tense verbs in English. Assume also that you have a sample size of 500 learners in the recast group and 500 in the model group (a highly unlikely event in second language research). Following the treatment, you give a post-test, and the model group has a mean score of 8.8 and the recast group has a score of 9.1. With such a large sample size, it is possible that this difference is significant, but given the small difference (.3), we might not want to make strong claims based on these results. However, in second language research we generally deal with much smaller sample sizes, making it difficult to get statistical significance.

As mentioned earlier, the commonly accepted level for significance in second language research is .05. This is known as the alpha (α) level; alpha levels should be established by the researcher at the onset of the research. For certain research, for example where high-stakes decisions will not be based on the analysis, the researcher may decide to set a less conservative alpha level. In fields such as medicine where the stakes are high, the alpha levels are much more conservative, thereby reducing the likelihood of chance occurrences. In sum, the *p*-value is the exact probability level matching the calculated statistic. The actual *p*-value must be lower than the predetermined alpha level for the results of the analysis to be considered significant. In second language research, even when the alpha level of .05 is used, researchers occasionally describe their findings in terms such as "approaching significance" or "demonstrating trends" when the *p*-value is between .05 and .075 or even .10.

In considering the difference between meaningfulness and significance (in the statistical sense), we need to recognize that second language learning is a slow and complex process often involving a period of production of correct forms only to be followed by a later period of production of incorrect forms. Therefore, we often need longer periods of observation, but the exigencies of research do not often allow long periods of time. Thus, it may be that meaningful trends are worthy of discussion, independently of statistical significance. As Gass et al. (1999, pp. 575–576) noted:

> The need to have all results adhere to a .05 standard may be questionable. Shavelson (1988) noted that the convention of using .05 or .01 "grew out of experimental settings in which the error of rejecting a true H_0 was very serious. For example, in medical research, the null hypothesis might be that a particular drug produces undesirable effects. Deciding that the medicine is safe (i.e., rejecting H_0) can have serious consequences. Hence, conservatism is desired" . . .
>
> (p. 248)

Given the essential arbitrariness in setting significance levels and given the constraints in conducting (second language research, particularly) classroom research, we feel that trends are important and at least point to the notion that experiments should be replicated, particularly when it is impractical or impossible for experiments to cover a long period. We also believe that trends may at times be as meaningful as statistical significance.

We are not suggesting that different levels or standards for significance should apply to second language research than those that apply to education, social, or cognitive sciences in general; what we are suggesting is that given the nature of second language research, it is not always necessary to completely discount trends in all data that do not fit within the narrow confines of the standard alpha level of .05. In fact, some researchers have argued that alpha levels are essentially unnecessary in the field of second language research, and instead we should focus on large sample sizes, and on other means of determining relationships.

10.6 INFERENTIAL STATISTICS

The goal of some types of second language research is to go beyond uncovering information about how a particular group of students, for example those enrolled in first-year Spanish, learn a particular part of the language. Rather, the goal is to generalize beyond the results. In other words, such researchers want to make inferences from the particular sample to the population at large. Given that it is impossible to gather data from all members of the population, inferential statistics can allow researchers to generalize findings to other similar language learners, that is, to make inferences. In the following sections, we deal with some of the most common inferential statistics that are used in applied linguistics and second language research.

10.6.1 Prerequisites

Before moving to present information about specific statistical analyses, we briefly discuss some basic concepts that relate to statistical procedures. While the first two, standard error of the mean and standard error of the difference between sample means, are not concepts that are presented in research reports, they are important for conceptualizing the statistics presented later in the chapter.

10.6.1.1 Standard Error of the Mean

Standard error of the mean (SEM) is the standard deviation of sample means. The SEM gives us an idea of how close our sample mean is to other samples from the same population. If we know that the mean for the total population is 50 and if we know that the SEM is 5, we also know that if our sample mean is 52, it is within one SEM of the population mean and is within 34 percent of all sample means taken from the population. Because we do not know the mean for the total population, this is not a precise measure, but is important in determining the standard error of the difference between sample means, discussed in the next section.

10.6.1.2 Standard Error of the Difference between Sample Means

Standard error of the difference between sample means (SED) is based on the assumption that the distribution of differences between sample means is normal. This distribution, because it is normal, will have its own mean and standard deviation. This standard deviation is known as the SED. In order to calculate the SED, one needs to know the SEM of the two samples in question (see Appendix G).

10.6.1.3 Degrees of Freedom

The concept of degrees of freedom is necessary as we consider the determination of significance of statistical tests. To put it simply, the degree of freedom is the number of scores that are not fixed. Suppose we know that our total value on a test adds up to 50 and we have five scores contributing to this value of 50. If we know what four of the scores are, the fifth one is fixed; it cannot vary. In other words, only one of the scores cannot vary. In this case, 4 represents the degree of freedom. This is important when we look up critical values on statistical tables. Statistics tables are organized by alpha-level (such as $p < .05$) and degrees of freedom and are expressed in terms of critical values. When a statistic is calculated, the numerical result of the calculation is compared against the statistical table, to determine whether it reaches the critical value. If the result of the calculation reaches or surpasses the appropriate value, the findings are considered statistically significant. In today's world of computer analyses, this information is provided in the output.

10.6.1.4 Critical Values

Researchers can look up critical values in a statistics table, although statistical packages that calculate statistics also provide the critical value. This is the value that we can use as a confidence measure to determine whether our hypothesis can be substantiated. The observed statistic (based on our statistical calculation) must exceed the critical value in order to reject the null hypothesis. This is further discussed below.

10.6.1.5 One-Tailed versus Two-Tailed Hypotheses

When we discussed hypotheses in Chapter 1, we presented some hypotheses that predicted differences in one direction or another and others that were neutral as to direction, that is, they predicted a difference but not in which direction the difference was expected. The former (those that predict a difference in one direction) are known as one-tailed hypotheses and require a different critical value than the "neutral" or two-tailed hypotheses. Examples of one-tailed and two-tailed hypotheses are provided below.

One-tailed hypothesis:

The group that received explicit grammar instruction before reading a passage with those grammatical elements will have higher comprehension scores than those who had vocabulary instruction before reading a passage with those vocabulary items.

This hypothesis clearly predicts which group will perform better.

Two-tailed hypothesis:

The group that received explicit grammar instruction before reading a passage with those grammatical elements will have a different level of comprehension than those who had vocabulary instruction before reading a passage with those vocabulary items.

This hypothesis predicts a difference in the performance of the two groups, but says nothing about which group will perform better. The critical value needed to reach significance depends on the concept of directionality.

TIME TO THINK ...

Think of an RQ and the corresponding hypothesis/prediction that would require a one-tailed test. And think of one that would require a two-tailed test.

10.6.2 Parametric versus Nonparametric Statistics

There are two broad categories of inferential statistics known as parametric and nonparametric tests. As the names suggest, they deal with the parameters of the population from which researchers have drawn samples.

With parametric statistics, there are sets of assumptions that must be met before the tests can be appropriately used. Some of the assumptions for parametric tests include the following:

Some assumptions of parametric tests

- The data are normally distributed and means and standard deviations are appropriate measures of central tendency.
- The data (dependent variable) are interval data (such as scores on a vocabulary test; see Chapter 4 for further information).
- Independence of observations – scores on one measure do not influence scores on another measure (e.g., a score on an oral test at Time 1 does not bias the score on an oral test at Time 2).

Again, we refer the reader to detailed descriptions in statistics books related to the specific sets of assumptions for each test.

The assumptions underlying nonparametric tests are minimal. Nonparametric tests are generally used with frequency data (such as the amount of other-correction in class discussion in different classrooms) or when the assumptions for parametric tests are not met.

Parametric tests have more power. This means that they are more likely to detect a genuine effect because they are more sensitive. Parametric tests are also more likely to detect an effect that does not really exist. One reason for the greater power of parametric tests is that there is more information that feeds into the statistic. If a statistical test lacks power, it may be difficult to detect the effect of the independent variable upon the dependent variable, resulting in a Type II error, or failure to reject the null hypothesis when it is incorrect. However, using a parametric statistic when it is not appropriate can lead to a Type I error, an incorrect rejection of the null hypothesis.

In the following sections, we briefly discuss some of the more frequently used parametric and nonparametric tests used in second language research.

10.6.3 Parametric Statistics

In this section, we deal with t-tests and analysis of variance.

10.6.3.1 T-Tests

The t-test can be used when one wants to determine if the means of two groups are significantly different from one another. There are two types of t-tests—one is used when the groups are independent and the other, known as a paired t-test, is used when the groups are not independent, as in a pre-test/post-test situation when the focus is within a group (a person's performance before treatment compared with his or her own performance after treatment). Below are examples of types of research in which a t-test and a paired t-test would be appropriate.

Example 1:

Description: You have completed a research study looking at the effectiveness of two kinds of feedback on learners' vocabulary test scores. Group 1 has 35 learners and Group 2 has 33. You have calculated the means and standard deviations of the end-of-semester exams of the two groups and you believe that all of the assumptions for a parametric test have been met.

Statistic: You compare the two groups using a t-test.

Example 2:

Description: You have conducted a study on the effectiveness of a particular way of teaching reading. You have given a pre-test and a post-test. Each individual has two scores (pre-test and post-test). You want to know if the improvement following the treatment was significant.

Statistic: A paired t-test is appropriate; each person is paired with him or herself on the two tests.

Example 3:

Description: You have conducted a study on the acquisition of relative clauses by Korean and Spanish learners of English. There are two groups, matched for native language and gender. Group 1 consists of three male native speakers of Korean, four female native speakers of Korean, five male native speakers of Spanish, and four female speakers of Spanish. Group 2 has the same profile. (Groups could be matched on a variety of factors, such as age, pre-experiment tests, reading tests, listening tests, and so forth.) Group 1 receives instruction on subject relative clauses; Group 2 receives instruction on indirect object relative clauses. You have pre-test scores and post-test scores for each individual on a range of relative clause types and calculate a gain score for each. You want to see if there are differences in learning between Groups 1 and 2.

Statistic: Paired (matched) t-test is appropriate because you have matched pairs. That is, Korean male #1 in Group 1 can be compared with Korean male #1 in Group 2, and so forth.

A word of caution about the use of t-tests is necessary. As noted, they are appropriate when comparing two groups, but there is a tendency in second language research to run t-tests on different parts of a data set in a way that overuses the test. Using an alpha level of .05 means that there is a 5 percent possibility of getting significance by chance. In other words one time out of 20 we might have a significant result that in actuality should not be significant. If one carries out 10 t-tests, for example, the odds are increased that a Type I error will be produced. For example, if you have conducted an experiment in which there were four groups, carrying out multiple two-way comparisons or carrying out multiple t-tests on sub-parts of the data (e.g., native speakers of one language versus native speakers of another language or males versus females) could be considered overusing the test. If there are multiple groups, rather than doing multiple two-way comparisons using t-tests, another statistic, such as an analysis of variance, may be more appropriate because analysis of variance calculations mathematically account for the increased chance of error that occurs as multiple comparisons are made.

TIME TO THINK ...

Come up with a small research project in which you might want to use a t-test (paired or not). What would you be comparing? And, why would a t-test be appropriate?

10.6.3.2 Analysis of Variance (ANOVA)

In the previous section, we discussed t-tests, which enable researchers to compare performance on two groups. Many research designs require comparisons with more than two groups, and ANOVA may be appropriate in this context. ANOVA results provide an F value, which is a ratio of the amount of variation between the groups to the amount of variation within the groups.

Example

You have conducted a study in which you are comparing the effectiveness of three different types of instruction. You are confident that the assumptions of a parametric test have been met. You want to compare the results and use an ANOVA to see if group differences are due to chance or are sufficient to reject the null hypothesis.

A sample result from an analysis of variance is presented in Table 10.5. This table includes the information that is relevant to understand an analysis of variance result (see Appendix G for formula).

TABLE 10.5 Example of an ANOVA results table

Source of variance	*SS*	*df*	*MS*	*F*
Between groups	521.43	2	260.71	53.98*
Within groups	202.66	42	4.83	
Total	724.09	44		

* $p < .01$

An ANOVA provides information on whether or not the three (or more) groups differ, but it provides no information as to the location or the source of the difference. That is, is Group 1 significantly different from 2 or 3, or is Group 2 significantly different from Group 3? To determine the location of the difference when the F value is significant, a post-hoc analysis is used. Common post-hoc analyses include the Tukey test, the Scheffé test, and Duncan's multiple range test. A typical display showing the source of a difference for the possible study described in the above example is presented in Table 10.6.

In this hypothetical example, differences were found between the groups who had Instruction 1 and Instruction 2, and between the groups who had Instruction 1 and Instruction 3. No other differences were found; the differences between Instruction 2 and Instruction 3, for example, were not significantly different.

10.6.3.3 Two-Way ANOVA

In second language research, there is often a need to consider two independent variables, for example instruction type and proficiency level. Where there is more than one independent variable, the results will show main effects (an effect of each independent variable without considering the effect of the other) and an interaction effect, which is the effect of one independent variable that is dependent on the other independent variable. In Figure 10.9, we can see that there is not a straightforward relationship between test scores and instruction type. Rather, there is an interaction between instruction type and proficiency level such that-high proficiency students do better with instruction type 1, whereas low-proficiency students perform better with instruction type 3.

10.6.3.4 Analysis of Covariance (ANCOVA)

There are times when there might be a preexisting difference among groups and the variable where that difference is manifested is related to the dependent

TABLE 10.6 Example of a post-hoc table

	Group		
Group	***Instruction 1***	***Instruction 2***	***Instruction 3***
Instruction 1		•	•
Instruction 2	•		
Instruction 3	•		

• Pairs where there was a significant difference at the .05 level

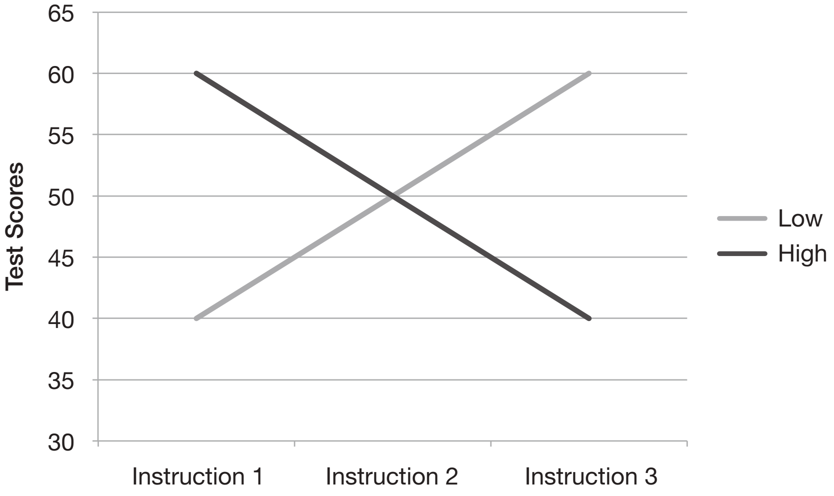

FIGURE 10.9 Instruction type as a function of proficiency level

variable. In other words, differences in means on variable X will show up on a pre-test. The preexisting difference will need to be controlled for and is referred to as the covariate. Because of differences among groups, the post-test results will need to be adjusted. The amount of adjustment will depend on two factors: (1) how large the difference is between the pre-test means; and (2) the change between the pre-test and the post-test (the dependent variable).

Example

You are testing three types of pedagogical treatments for learning the orthographic writing system of Arabic (explanation, visual repetition, practice). To do this, you use three separate first-semester university-level Arabic classes. You have a pre-test, treatment, post-test design. You find that your groups are not matched at the outset on language proficiency. Thus, your pre-test score, the covariate, will have to be adjusted to compensate for the fact that one group starts at a higher level than the other. If no adjustment is made, we would not know whether the group with the initial higher score learned more or appeared to learn more because of the higher initial score. An ANCOVA is appropriate.

10.6.3.5 Multivariate Analysis of Variance (MANOVA)

The MANOVA is part of the family of analyses of variance. It differs from an ANOVA in that it has more than one dependent variable. In order to appropriately use a multivariate analysis of variance, there has to be justification for believing that the dependent variables are related to one another.

Example

You have conducted a study of the effectiveness of different interlocutor types on learner performance (as measured by oral abilities and grammar). You devise a spoken proficiency test as well as an acceptability judgment task to measure learning. Because you are interested in the relationship between oral abilities and grammatical knowledge for the different interlocutor types, a multivariate analysis of variance is appropriate.

10.6.3.6 Repeated Measures ANOVA

There are times when we might want to compare participants' performance on more than one task.

Example

You have conducted a study of different writing prompts on learner performance as measured by writing accuracy. You have developed a measure of accuracy that includes length of essay, error-free T-units, and sophistication of vocabulary. You have carefully devised a counterbalanced design where each participant writes an essay under four conditions:

1. timed with a prompt;
2. untimed with a prompt;
3. timed with no prompt; and
4. untimed with no prompt.

Because each individual does all the tasks, we have a repeated measures design. And, because we have three different sets of results to be compared, a repeated measures ANOVA is appropriate.

10.6.4 Nonparametric Tests

As discussed above, nonparametric tests are generally used with frequency data or when the assumptions for parametric tests have not been met. In this section, we discuss some of the most frequently used nonparametric tests in second language and applied linguistics research.

10.6.4.1 Chi Square (χ^2)

Chi square tests are often used with categorical (i.e., nominal) data. Examples of categorical data include age groups (e.g., 8–10 years old, 11–13, 14–16), gender, native language, types of relative clauses, and so forth. The chi square statistic relies on observed frequencies and expected frequencies.

Example

You want to determine whether ESL learners from different L1 backgrounds and of different genders are more likely to use stranded prepositions (*That's the man I talked to you about*). You elicit this structure from 40 learners with the same gender distribution (20 L1 Japanese and 20 L1 Spanish—10 males and 10 females each). You construct a table that looks like the following:

Participants who use stranded prepositions

	Japanese	*Spanish*
Male	8	4
Female	6	10

If native language and gender did not matter in the use of stranded prepositions, we would expect the values in each square in the table to be equal. You can determine whether the actual values are different from the expected values by using a chi square analysis. If the actual values differ from the expected values, it can be assumed that at least one of the variables (native language or gender) influences the use of stranded prepositions. The expected frequency is determined by taking the sum total of the observations (in this case, 28) and dividing it by the number of cells (in this case, 4). So the expected frequency for each cell is 7. These are the values that feed into a chi square formula. Degrees of freedom are then determined by subtracting one from the number of columns. In this example, there is one degree of freedom. Degrees of freedom and corrections are generally built into computer programs that automatically calculate chi squares. When there is one degree of freedom, Yates' correction factor is often used.

Just as with many parametric statistics, chi square analyses rely on assumptions as to the type of data needed. Primary among these assumptions are the following:

- Each observation is independent, that is, it only falls in one cell. In the example above, an individual is either male or female and is either Japanese or Spanish.
- The data are raw frequencies (not relative frequencies or percentages).
- Each cell has an expected frequency of at least 5.

The Fisher's exact test, a variant of the chi square test, may be more appropriate than a chi square in some situations, including those in which there are several cells with expected frequencies that are less than 5, or where there are cells with a value of 0.

When chi square tests are calculated to determine the relationship among several variables, the results will indicate significant relationships. However, as with ANOVA tests, the location of the significance is not identified. There are procedures (e.g., Haberman's residuals) that can be used to locate the significant relationships.

10.6.4.2 Mann-Whitney U/Wilcoxon Rank Sums

Other nonparametric tests are used with ordinal or interval data rather than categorical data. Mann-Whitney U and Wilcoxon Rank Sums are two such tests; we discuss these together as they are essentially the same test. These are comparable to the t-test in that they compare two groups but are used when the results are rank scores (i.e., with ordinal scale dependent measures). Both sets of scores are pooled and the scores are compared in relation to the median.

Example

You want to determine the effects of interaction on the ability of the interlocutor to comprehend descriptions. You design a study with two groups, each made up of 10 dyads: in one group interaction is allowed, and in the other interaction is not allowed. You subsequently observe where objects were placed on a board (dependent variable, measure of comprehension). The object placement scores are quantified and converted into rank scores.

A Mann-Whitney U is appropriate because the interval data assumption of a parametric test was not met. Had that assumption been met, a t-test could have been used.

10.6.4.3 Kruskal-Wallis/Friedman

A Kruskal-Wallis is a nonparametric test comparable to an ANOVA, but used when parametric test assumptions are not met. It is used when a researcher wants to compare three or more independent groups. In other words, a between-groups comparison is being made. A Friedman test is the nonparametric counterpart to a repeated measures ANOVA. That is, when you have non-independent samples and need to compare within groups, a Friedman may be appropriate.

In the preceding sections, we dealt with some commonly used parametric and nonparametric statistics in second language research. Below, we summarize some of the different types of second language data together with possible statistical techniques in Table 10.7.

10.7 STATISTICAL TABLES

In closing this discussion on parametric and non parametric tests, we present two statistical tables to illustrate how they can be read and used. Most statistical textbooks include full versions of tables that can be consulted to determine if your test results are significant. If statistics are carried out using a computer-based statistical package (see below), the results will be provided for you and

TABLE 10.7 Summary of statistics

Type of comparison/type of test	*Parametric*	*Nonparametric*
Two independent samples	T-test	Mann-Whitney
Two related samples	Paired t-test	Wilcoxon
More than two independent samples	ANOVA	Kruskal-Wallis
More than two related samples	Repeated measures ANOVA	Friedman

there will be little need to consult a statistical table such as the ones given in this section. Table 10.8 provides a partial display of the distribution of t.

This table and other statistical tables display the minimum value (i.e., critical value—see section 10.6.1.4) based on the desired probability level and degrees of freedom that one must have to claim significance. There are two points to note about this table. First, to determine significance, one looks at the left-hand column at the relevant degrees of freedom. Second, one has to determine whether one has a one-tailed or a two-tailed hypothesis. The figures given across the top (*p*-levels) are for a two-tailed hypothesis. For a one-tailed hypothesis, one halves the probability level. For example, column 2 (headed by .1) is .05 for a one-tailed hypothesis. Thus, if one has 14 degrees of freedom on a two-tailed test, and a value of 2.98, one can claim that the significance is < .01.

The second table (Table 10.9) is from a nonparametric test χ^2.

The method for reading this table is the same as that for reading the t-test table. If one has 15 degrees of freedom and a χ^2 value of 33.021, one has a significance level of < .01.

TABLE 10.8 Distribution of t

p	*.1*	*.05*	*.02*	*.01*	*.001*
df					
11	1.796	2.201	2.718	3.106	4.437
12	1.772	2.179	2.681	3.055	4.318
13	1.771	2.160	2.650	3.012	4.221
14	1.761	2.145	2.624	2.977	4.140
15	1.753	2.131	2.602	2.947	4.073

Note: *p* refers to the probability level; *df* refers to the degrees of freedom

TABLE 10.9 Distribution of χ^2

p	*.1*	*.05*	*.02*	*.01*	*.001*
df					
11	17.275	19.675	22.618	24.725	31.264
12	18.549	21.026	24.054	26.217	32.909
13	19.812	22.362	25.472	27.688	34.528
14	21.064	23.685	26.873	29.141	36.123
15	22.307	24.996	28.259	30.578	37.697

Note: *p* refers to the probability level; *df* refers to the degrees of freedom

As mentioned above, many second language researchers will not often have to read a statistical table since most computer programs provide exact probability levels, for example in the form $p = .023$, rather than in the form of $p < .05$. Many journals require reporting of exact probability levels and the use of notations such as <.05 or <.01 is no longer accepted.

10.8 STRENGTH OF ASSOCIATION

There are times when we might want to determine how much of the variation is actually due to the independent variable in question (e.g., the treatment, the learner's language background, the learning context, etc.). That is, if we find a difference, for instance, in performance between native speakers of Japanese learning English and native speakers of Arabic learning English on some measure, we don't know how much of the difference is due to the fact that their native languages are different or to something else (which we probably cannot specify). The following sections discuss some statistical procedures that can help us address these questions.

10.9 ETA^2 AND OMEGA^2

The most common measurement that can be used after a t-test is eta^2 (expressed as $\eta2$), which goes beyond the fact that there is a significant difference and gives us an indication of how much of the variability is due to our independent variable. Consider Example 1 of a t-test in the study of two different types of vocabulary instruction. Suppose that the t-test indicates that the learners from Group 1 score significantly better on their end-of-semester exam than the learners from Group 2. You know that there is a difference between these groups, but you don't know how much of that difference can be explained by the independent variable (instruction type). You calculate η^2 and determine that $\eta^2 = .46$. That means that 46 percent of the variability in their scores can be accounted for by the instruction type.

The same reasoning applies for ANOVAs. Omega^2 (ω^2) is the statistic used when all groups have an equal n size. Otherwise, eta^2 is appropriate. The formulae for these tests are given in Appendix G.

10.10 CORRELATION

One differentiating factor between correlational research and what we have discussed in previous sections is that in correlational research, no variables are manipulated. Correlational research attempts to determine the relationship between or among variables; it does not determine causation. Consider the fictitious example below.

The relationship between infant-directed speech and growth spurts

Introduction

A research team believes that talking to young children (infants) is related to their growth; the more talk addressed to young children, the more they grow. To test this, they consider two mother/child pairs. They gather speech and growth data from children aged 6 months to 18 months (twice a month, 30 minutes each time). To measure the amount of talk, they count all words in that two-hour period. The table below shows the data for both mother/child pairs.

	Pair #1			*Pair #2*	
Month of data collection (week)	*# of words*	*Height in inches*	*Month (week)*	*# of words*	*Height in inches*
1(1)	72	24	1(1)	65	28
1(3)	75	24	1(3)	70	28
2(1)	75	25	2(1)	66	28
3(3)	70	25.5	3(3)	72	29
3(1)	90	25.5	3(1)	59	29.5
2(3)	92	25.5	2(3)	64	29.75
4(1)	89	26	4(1)	64	30
4(3)	90	27	4(3)	80	30
5(1)	91	27	5(1)	82	30
5(3)	102	27.5	5(3)	100	30
6(1)	93	28	6(1)	125	30.25
6(3)	94	28	6(3)	152	30.5
7(1)	91	28	7(1)	145	30.5
7(3)	121	28.5	7(3)	150	30.5
8(1)	132	29	8(1)	145	31
8(3)	120	29.5	8(3)	180	31
9(1)	145	30	9(1)	92	31.25
9(3)	145	30	9(3)	165	32
10(1)	120	30	10(1)	172	33
10(3)	105	31.5	10(3)	170	33.5
11(1)	75	31.5	11(1)	200	34
11(3)	105	32	11(3)	180	35
12(1)	190	32	12(1)	178	36
12(2)	190	32	12(2)	180	36

Problematic Interpretation

Using the data presented above, a correlation coefficient of .70 for Pair 1 ($p < .001$) and .82 (Pair 2) was obtained ($p < .001$). The team conclude that because there is a relatively high correlation, this proves that the number of words used was the source of the growth.

What is wrong with this picture? There are a number of issues that could be raised, but the important one for this section is the interpretation. The first is that nothing has been "proven." All that has been shown is that there is a relationship between the number of words used when addressing a child and a child's height. The relationship is not necessarily one of cause and effect. While it is true that there is a relationship, the source of each variable is different. Increased amount of talk to an infant is possibly due to a relatively high interactive capability of the infant; increased height is a natural part of increased age in most children.

The above example focused on the interpretation of correlational data. We now turn to how to determine the strength of a correlation. Correlations are calculated between two sets of scores (in the previous example, one score is the amount of talk and the second is height). We can plot this information on a graph. The amount of talk can be plotted along the x-axis and the height on the y-axis. The result would be a graph with many individual points. If there were a relationship between the two scores, the dots would cluster around an imaginary line. When we calculate a correlation, we come up with a correlation coefficient (*r*) that characterizes the direction of the line and how well the line represents the patterns in the data. Depending on the direction of the line, correlation coefficients can be expressed as positive and negative values. A positive value means that there is a positive relationship; for example, the more talk, the taller the child. Conversely, a negative value means a negative relationship—the more talk, the shorter the child. A value of zero means that there is no relationship between the variables. These three possibilities are illustrated in Figures 10.10–10.12. The first figure (Figure 10.10) comes from the data showing a positive relationship between amount of talk and height of child for Pair 2. Figure 10.11 represents a graph that would depict a negative relationship and Figure 10.12 is a graph showing no relationship between two variables.

10.10.1 Pearson Product-Moment Correlation

We now turn to the Pearson product-moment correlation, a common means for determining the strength of relations (see formula in Appendix G). There are four assumptions that underlie this particular statistic; these are provided below.

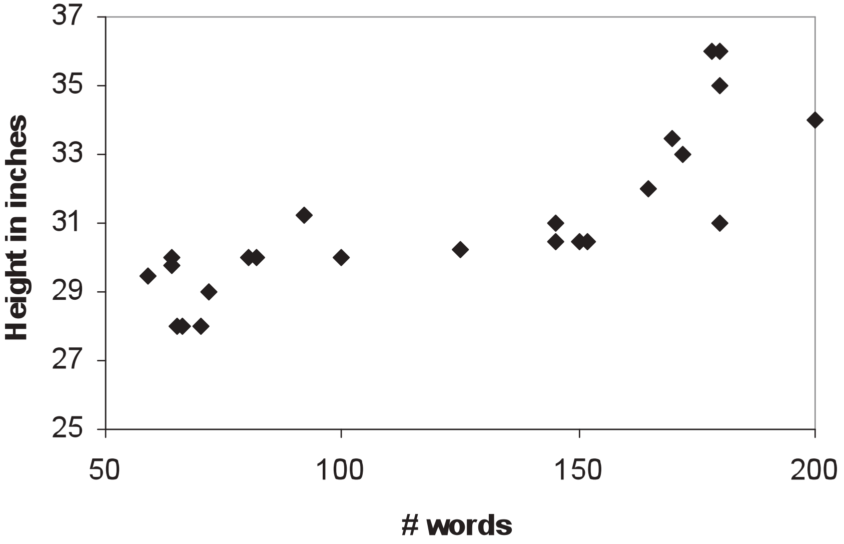

FIGURE 10.10 Positive relationship (*r* = .82, *p* < .001)

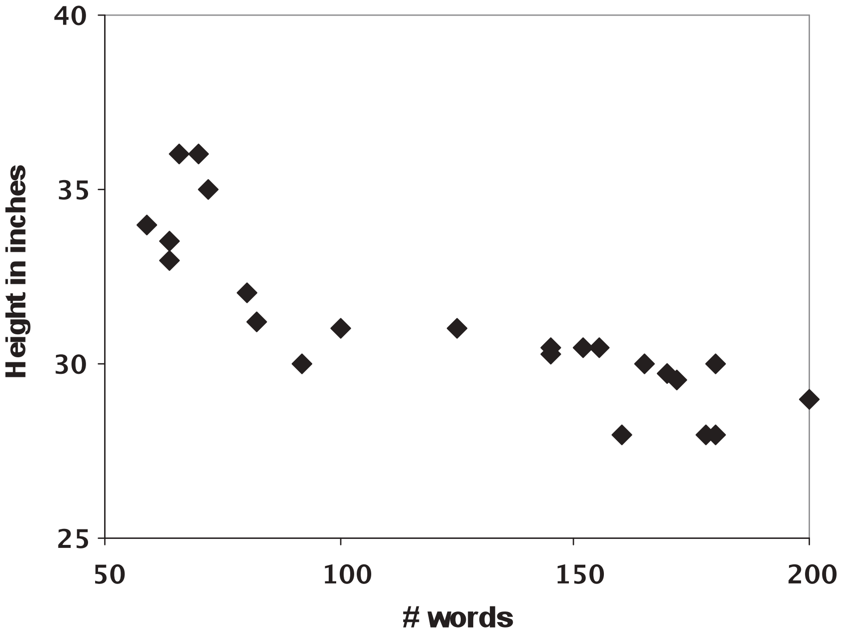

FIGURE 10.11 Negative relationship (*r* = −.85, *p* < .001)

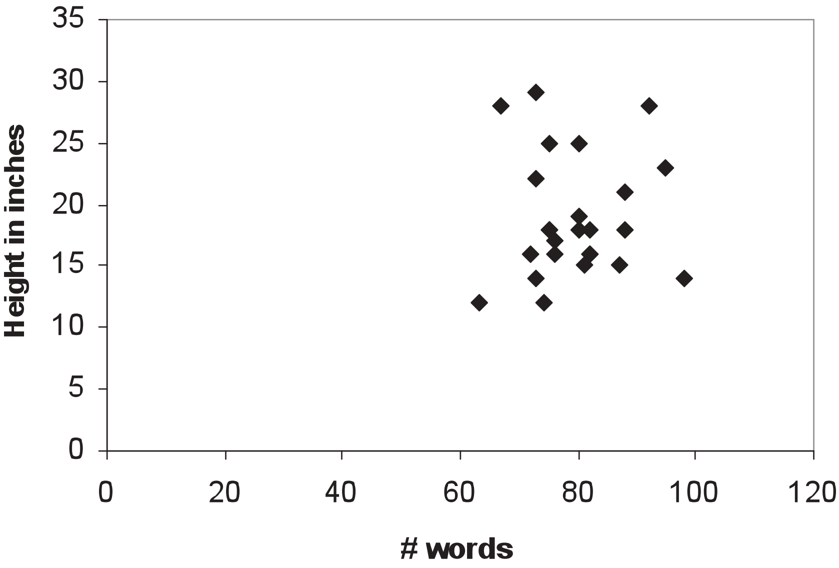

FIGURE 10.12 No relationship (*r* = –.051; *p* = .808)

Assumptions underlying the Pearson product-moment correlation

1. Normal distribution
2. Independence of samples
3. Continuous measurement scale (generally interval or sometimes ordinal if continuous)
4. Linear relationship between scores for each variable

The correlation coefficient gives information about the extent to which there is a linear relationship between the variables.

Frequently, correlations are calculated between multiple sets of scores in research studies. One concise way of presenting this data is in a correlation table, in which correlation coefficients for different sets of scores are listed. An example of a Pearson's correlation table (Table 10.10) comes from a study by de Graaff (1997) on the role of explicit instruction versus implicit instruction in an artificial language by native speakers of Dutch as it relates to language aptitude. This table is to be interpreted in such a way that if we look at T2 in task type 3 in the explicit condition, there is a .56 correlation between the mean aptitude score and the immediate post-test score on the gap-filling task. The probability level is based on the value of the correlation coefficient and the sample size.

TABLE 10.10 Pearson's product correlation table

Correlations (Pearson's r) between the language aptitude mean scores and the mean scores per task type and test session, under explicit and implicit conditions

Task Type	*Test Session*	*Explicit (n = 27)*	*Implicit (n = 27)*
1	T1	.38	.15
	T2	.47*	.42*
	T3	.50*	.55*
2	T1	.21	.02
	T2	.34	.34
	T3	.39	.34
3	T1	.52*	.32
	T2	.56*	.50*
	T3	.54*	.39
4	T1	.19	.36
	T2	.45*	.51*
	T3	.40	.50*

Note: Task type 1 = judgment task with time pressure; task type 2 = judgment task without time pressure; task type 3 = gap-filling task; task type 4 = correction task. T1 = mid-test; T2 = immediate post-test; T3 = delayed post-test

* $p < .01$

Source: de Graaff, R. (1997). The eXperanto experiment: Effects of explicit instruction on second language acquisition. *Studies in Second Language Acquisition*, *19*, 263. Copyright © 1997 by Cambridge University Press. Reprinted with the permission of Cambridge University Press.

10.10.1.1 Linear Regression

We now turn to another use of correlations, that of prediction. Again considering our fictitious study of the relationship between infant-directed speech and growth, we repeat Figure 10.13 below.

As can be seen from the figure and the correlation coefficient (.71), there is a positive relationship, but if we had reason to believe that the relationship was meaningful, we might want to make predictions. For example, if the amount of words addressed to one specific child was 145, what might we expect her height to be? A straight line, called a regression line, might help us to address this question. A prediction equation can be used once we know the slope of the line and the intercept. While the details of these calculations go beyond the scope of this chapter, it is useful to know that if we want to predict one variable from another, and we know details of the regression line, we can calculate, for any

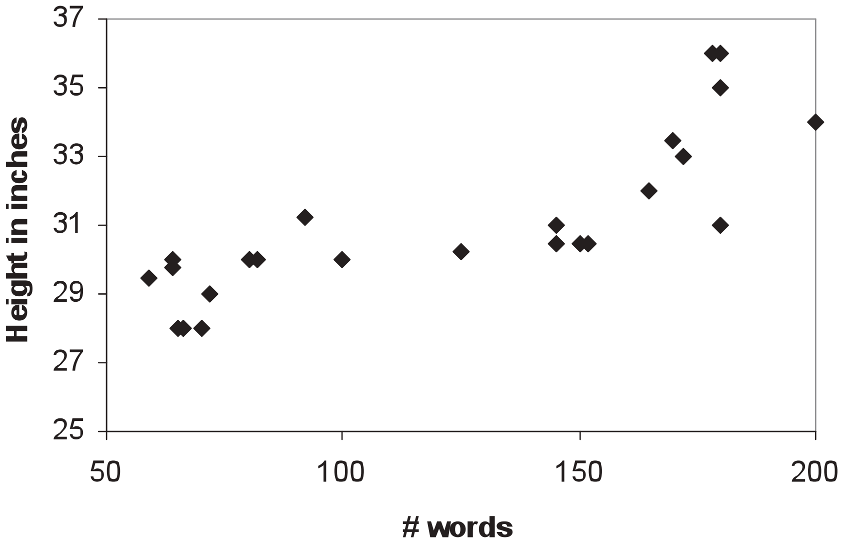

FIGURE 10.13 Relation between #words and height

given words addressed, the predicted height. Note that the validity of regression for prediction is dependent on the variables selected. A theoretically sound explanation for suspecting a relationship between the variables should be presented when regression is used to predict values.

10.10.1.2 Multiple Regression

This type of analysis is increasingly being used in the second language research field, probably because as studies get more methodologically sophisticated, our methods of analysis are also becoming more finely tuned. There may be instances when we want two or more variables to be used to predict a third variable. We can use multiple regression to do that. For example, we might want to predict how ESL learners will do in college based on two factors: (1) their results on a standardized test (e.g., TOEFL—Test of English as a Foreign Language); and (2) their performance in an Intensive English Program (IEP) on their own campus. A multiple regression prediction formula enables us to do this.

To test the validity of the predictor variables for predicting the third variable, data on the third variable should be collected from a subset of the population of study. For example, if we want to know how well TOEFL scores and IEP grades predict college grades, we could obtain actual college grades from a group of students and correlate our predicted grades based on a multiple regression formula with the actual grades. The resulting coefficient is called a coefficient of multiple correlation (*R*). The same idea applies as to *r*: *R* refers to the strength of the relationship among the variables in question (including the variable that is being predicted). Thus, as with other correlations, *R* can vary from +1 to –1, with +1 being a perfect positive correlation and –1 a perfect inverse relationship. The higher the absolute *R* value, the more confident we can be in our predictions.

TIME TO THINK ...

Which of the following research questions would lend themselves to a correlational analysis? Justify your answer.

a. Attitudes toward target culture and success in language learning.
b. Feedback type and language learning.
c. Attention paid to form and success in vocabulary acquisition.
d. Number of times male and female learners respond to feedback.

10.10.2 Spearman Rho/Kendall Tau

Both Spearman rho (ρ) and Kendall Tau are used for correlational analyses when there are ordinal data (or with interval data when converted to ranks). Spearman rho is more common, but the Kendall Tau is more suitable when there are many ties in the rankings.

10.10.3 Factor Analysis

Factor analysis is a complex procedure for determining common factors that underlie measures that test different (possibly related) variables. Researchers search for groups of variables that are correlated with one another; each group can be referred to as a factor. In doing a factor analysis, researchers take into account the variance that is common to all individual variables, the variance that is specific to each variable and sampling error. Factor analysis can be used to determine overall patterns found in correlation coefficients, and is particularly useful when analyzing results from surveys.

10.11 NEW APPROACHES[3]

We call this section “new approaches” (cf. Cumming, 2012, who refers to this suite of approaches as *new statistics*) recognizing that in fact we are not dealing with new statistics in the sense that they have just been invented, but rather in the sense that they represent a set of statistics that have recently entered into the common suite of approaches used in second language research. This is a reflection of the emphasis on study quality and methodological rigor that has become prevalent as part of accepted practice in SLA. Many of these issues were detailed in Plonsky and Gass (2011) and Plonsky (2013), and include such issues as reporting practices, design practices, and statistical power and significance. In what follows, we deal briefly with some of these constructs. Nick Ellis, then editor of the prominent journal *Language Learning*, raised the collective consciousness of the field in 2000 when he required that all articles submitted to the journal *Language Learning* include effect sizes which form the basis of power analyses and meta-analyses. According to Plonsky (2013), *TESOL Quarterly*, *The Modern Language Journal*, and *Language Learning and Technology*[4] also require the reporting of effect sizes along with "other reporting practices that include stating the hypotheses to be tested, describing whether or not assumptions for statistical tests were met, and using graphs and tables to complement in-text presentations and explanations of quantitative data" (p. 662). While other journals are not currently requiring reports of effect sizes, the sixth edition of the *Publication Manual of the American Psychological Association* (2010) strongly encourages the reporting of these statistics, emphasizing that "[f]or the reader to appreciate the magnitude or importance of a study's findings, it is almost always necessary to include some measure of effect size in the Results section" (p. 34).

The approaches discussed in this section take as a basis the inadequacy of null hypothesis significance testing (NHST). One of the main arguments in favor of alternative approaches to data analysis is that NHST is highly dependent on the sample size, as we saw in preceding sections. In general, NHST does not give an indication of the magnitude of a relationship. Rather, it relies on a yes/no (significant/not significant) decision. Further, the typical cutoff point in applied linguistics and second language research is .05, but when one considers this more carefully, one has to question whether there really is a difference between a result that is .054 (significant) and .055 (not significant).

TIME TO THINK ...

Think of a study you recently read. What were the research questions? Were they phrased dichotomously (Do . . .? Is there a difference?)

If so, what kind of an answer can come from such an RQ?

How might the findings and implications differ with an emphasis on magnitude rather than the presence/absence of a relationship or effect?

10.11.1 Power Analyses

Power analyses are associated with inferential statistics and are frequently used when we want to know how many participants are appropriate for a particular study. For this, we can conduct a power analysis, which is "a procedure designed to determine a priori the sample size necessary to reliably detect whether or not an effect will be found using inferential statistics" (Plonsky, 2012, p. 201). Plonsky (2013, p. 674) reported that only 1 percent of the studies he examined (6/606) actually conducted power analyses. Power analyses can be conducted relatively easily (e.g., http://danielsoper.com/statcalc3/calc.aspx?id=47, retrieved April 19, 2015).

10.11.2 Effect Size

An effect size is a measure that gives an indication of the strength of one's findings. It is not dependent on sample size and therefore can allow comparisons (meta-analyses) across a range of different studies with different sample sizes. A standard measure of effect size is Cohen's d (see Appendix G for formula), which can be used to test differences in means between two groups or differences in gain scores between two groups. A value of .2 is generally considered a small effect size, .5 a medium effect size, and .8 or more a large effect size. Effect size can be calculated based on a number of statistics (correlations, parametric, and nonparametric). Two useful references are: http://web.uccs.edu/lbecker/Psy590/es.htm (retrieved April 19, 2015) and Wilkinson et al. (1999). However, these interpretations have recently been challenged by Plonsky and Oswald (2014), who argue for a more nuanced and field-specific scale for interpreting effect sizes.

Consider the data in Table 10.11 below. Note how in the three hypothetical studies, the *p*-value changes based on variation in the sample size, whereas the effect size (*d*) is resistant to such variation. In other words, it is relatively easy to obtain significant results by increasing sample size.

TABLE 10.11 Effects of sample size and effect size

Study	N_1	N_2	M_1 (SD_1)	M_2 (SD_2)	*p*	*d*
1	5	5	15 (3)	18 (4)	.2265	.085
2	15	15	15 (3)	18 (4)	.0276	.085
3	45	45	15 (3)	18(4)	.0001	.085

TIME TO THINK ...

What are the implications of the results of Table 10.11 for consistency/lack of consistency in studies in second language? Does it make it more difficult to determine the real effects of a treatment? Why or why not?

Thus, effect sizes can be a useful tool for researchers who want to compare results with other research that addresses similar questions (see section on meta-analysis below) and who want to better understand the magnitude of a relationship.

TIME TO THINK ...

Compare the following:

1. Are there differences in the mean ratings that experienced ESL teacher raters and novice raters assign on measures of comprehensibility, accentedness, and fluency? (Isaacs & Thomson, 2013)
2. To what degree do (a) raters' background characteristics . . . (b) . . . affect students' ratings of ITAs' oral performances? (Kang, 2012)

Do you think that the analysis is based on NHST? Why? Which is more likely to have used effect sizes? Why? And which is more informative?

10.11.3 Confidence Intervals

A confidence interval (CI) is an estimate of the larger group from which a particular sample is taken. Assume that you know that the students in a fourth-semester Spanish class in the U.S. all have a particular proficiency level (on a scale from 1 to 10). What you really want to know is the proficiency level of all fourth-semester Spanish students within the state of California but for reasons of logistics and expense, you are unable to obtain that information. A CI is a way of determining the extent to which the proficiency level mean will actually contain the mean of the entire population of interest (students within the state of California) and thereby represents the boundaries within which we believe the actual population value falls. A 95 percent CI means that if one were to take multiple samples and compute confidence intervals, 95 percent would include the actual mean of the proficiency levels within the state of California. In other

TABLE 10.12 A way of expressing CIs in tabular form

Regression Output (Fixed Effects) for Finished Reading with Regression

Predictors	***Odds***	***OR***	***95% confidence interval***	***p***
Intercept	3.70	2.53	5.41	< .001
L1-L2 status	0.95	0.62	1.44	.801
Time	0.51	0.44	0.59	< .001
Grammaticality	0.95	0.83	1.10	.476
Sentence length (centered)	0.95	0.92	0.99	.013
Sentence-final length (centered)	1.00	0.93	1.07	.951
L1-L2 status × Time	0.35	0.25	0.51	< .001
L1-L2 status × Grammaticality	1.20	0.82	1.77	.351
Time × Grammaticality	1.98	1.32	2.98	.001
L1-L2 status × Time × Grammaticality	0.68	0.39	1.20	.187

Note: The intercept represents the odds for a NS reading a grammatical, untimed sentence of average total length (7.81 words) and average length of sentence-final region (0.61 words). OR = odds ratio

Source: A. Godfroid, S. Loewen, S. Jung, J-H. Park, S. Gass, & R. Ellis. (2015). Timed and untimed grammaticality judgments measure distinct types of knowledge: Evidence from eye-movement patterns. *SSLA*, *37*. Copyright © 2015 by Camgridge University Press. Reprinted with permission of Cambridge University Press.

words, it is a way of expressing our level of confidence that our interval includes the wider population mean. Many journals require or at least strongly recommend the inclusion of confidence intervals as part of the statistics presented. Table 10.12 shows a table display of confidence intervals.

10.12 META-ANALYSES

There are times when our research questions involve surveying a wide range of existing studies rather than collecting original data. In most instances, it will be difficult to directly compare studies given the unevenness of available data, size of experimental and control groups, and so forth. To make a meaningful comparison and to synthesize results, effect sizes become the main comparative tool. Norris and Ortega (2000) exemplify this in a study on the effectiveness of second language instruction. They outline the following five uses of effect sizes in their study.

- Average effect sizes across studies were calculated for specific predetermined categories.
- Average pre-test to post-test effect sizes were calculated.
- Average effect sizes across studies were calculated based on duration of treatment.
- Average effect sizes were calculated for delayed post-tests.
- Average effect sizes were calculated by type of dependent variable.

In essence, meta-analysis is an average of results from numerous studies. This is done by averaging effect sizes which are taken from original studies that focus on a particular research domain (e.g., interaction-based studies, strategy instruction, instructional effects). Effect sizes become the primary unit for analysis because they are not dependent on sample size. Meta-analyses provide an objective measure for determining the strength of an effect (for information on how to conduct a meta-analysis, see Plonsky & Oswald, 2012). One can see from this brief description why effect sizes are an integral part of what is reported in studies (see Plonsky, 2013; Plonsky & Gass, 2011, for a discussion of reporting practices).

In the preceding sections, we have dealt with statistical approaches. In the next sections, we deal with statistical packages that can assist in analyzing data and preparing those data for presentation.

10.13 STATISTICAL PACKAGES

There are commercially available statistical packages that can be used with a number of different operating systems, including Macs and PCs. In addition to the ones we deal with in this section, some basic statistics and graphing can be done using Excel. Others have been discussed elsewhere in this book. Two of the most common packages are SPSS and VARBRUL.[5] Learning to use either of these requires some initial effort and practice, but there are many courses or workshops in the use of these programs. Specific detail on statistical packages is beyond the scope of this book. Below, we give an indication of the use that can be made of each.

10.13.1 SPSS

SPSS is a basic analytic program. There are add-on packages for more sophisticated statistical use, but the standard statistical tests such as frequency statistics (chi square), t-tests, ANOVAs (with post-hoc tests), regression, correlations, and other more complex statistics are included. One can also convert raw data and output from SPSS to charts and graphs. Further details on SPSS are available through www.spss.com (see Larson-Hall, 2009, 2015) for detailed information on using SP55 in second language research).

10.13.2 VARBRUL

VARBRUL (Pintzuk, 1988; Rand & Sankoff, 1990) is a statistical package that is designed for analyzing variation data. For example, Young (1991) investigated –*s* plural marking in English by Chinese speakers. He wanted to know what the possible influence might be that would predict when the English plural form was used. He hypothesized 10 possible influences, including who the interlocutor was (Chinese or English), whether the noun was animate or inanimate, whether the noun was definite or not, and phonological surroundings. Each category could be divided into at least two levels. Because of the large number of factors and because many of the cells have a value of zero, an ANOVA is not appropriate for this comparison. VARBRUL, on the other hand, is designed to handle data of this sort. Young and Bayley (1996) provide detailed instructions on how to conduct a VARBRUL analysis and how to use data from this program.

10.14 CONCLUSION

This chapter has provided an overview of some of the statistical techniques commonly used in second language research. In the final chapter, we provide guidelines on what to do as you complete your research and prepare your findings for presentations and/or publication.

POINTS TO REMEMBER

- Numerous ways to report data exist (e.g., frequency, measures of central tendency, measures of dispersion).
- It is important to report standard deviations when reporting means.
- Determine carefully what you will do with outliers (if anything).
- At times, it is important to use standard scores when, for example, one wants to compare results on different tests.

- Probability is crucial in dealing with inferential significance and exact levels of probability should be reported.
- Inferential statistics allows generalizations beyond the specific group of participants being tested.
- Assumptions for parametric statistics should always be checked.
- Nonparametric tests are used with frequency data or when the assumptions for a parametric statistic have not been met.
- Correlations are used to determine relationships; no variables have been manipulated. Check assumptions.
- Power analyses are a way of determining the appropriate number of participants needed for a study.
- Effect size allows one to understand the strength of findings and is not dependent on sample size.
- Confidence intervals are a way of determining the range of values (with a certain level of certainty) within which the population value falls.
- Meta-analysis is a way of comparing studies and coming up with an average across many studies.

MORE TO DO AND MORE TO THINK ABOUT ...

1. You are in charge of student services in an intensive ESL program and notice that the students who are doing well in their reading class are not doing well in their writing class. This is surprising, but you want to make sure that there is a relationship before bringing it to the attention of the Director. Below are the two sets of scores from the 14 students enrolled in the class. You determine that the Pearson product-moment correlation is the most appropriate and can be used because you have interval data. Calculate the correlation coefficient (it has been started for you) and determine if the results are worth bringing to the attention of the Director. All scores have a maximum of 100.

	Reading Class	*Writing Class*
Maria	92	72
Juan	85	85
Toshi	78	93
Yoon-soon	61	51
Bob	25	32
Sachiko	87	62
Young-Ahn	67	78
Gunter	59	57
Angelika	84	72
Noriko	85	82
Jean-Marc	77	55
Antonio	62	77
Giovanna	88	87
Susana	87	88

$\Sigma X = 1037$ $\Sigma Y = 991$
$\Sigma X^2 = 81025$ $\Sigma Y^2 = 74035$
$\Sigma XY = 76336$

Using the formula given in Appendix G, calculate the correlation between the two class scores (you will either need a calculator or you can use a program such as Excel). Given your results, would you notify the Director? Why or why not?

2. The statistical table below was discussed in this chapter. You have just conducted a study and have compared two means. You had 13 degrees of freedom. What must your critical value be to claim that your results are significant at the .05 level? At the .01 level? If you had 15 degrees of freedom, what would the critical value have to be to claim significance at the .01 level? At the .05 level?

Distribution of t

p	*.1*	*.05*	*.02*	*.01*	*.001*
df					
11	1.796	2.201	2.718	3.106	4.437
12	1.772	2.179	2.681	3.055	4.318
13	1.771	2.160	2.650	3.012	4.221
14	1.761	2.145	2.624	2.977	4.140
15	1.753	2.131	2.602	2.947	4.073

3. Consider the data below from a study comparing two groups of second language learners: one who received feedback and one who received no feedback. Compare the means of these two groups using the formula given below for a t-test (maximum score was 50). (Note that there are somewhat different formulae that are to be used with different group sizes. Consult a statistics book for the precise formula.)

	Feedback group (n=9)	*No feedback (n=8)*
Mean	42	29
Standard deviation (SD)	1.23	1.59
Standard error of the mean (SEM)		
Standard error of the difference (SED)		

The t value is ________. With _______df, the results are (not) significant.

4. Assume that you have gathered data on use by English speakers of correct noun-adjective gender agreement in Spanish. You have four groups of learners ranging from first-semester to fourth-semester Spanish classes. You want to determine whether learners from different classes are more likely to have acquired gender agreement. You define acquisition as 90 percent suppliance in obligatory contexts. Based on this standard, you determine which learners have and have not acquired gender agreement. The data can be seen in the table below.

Number of correct instances	*Acquired Gender Agreement*	*Not Acquired Gender Agreement*	*TOTAL*
Semester of study			
First semester	5	25	30
Second semester	10	22	32
Third semester	25	4	29
Fourth semester	26	6	32
TOTAL	66	57	123

Calculate the value using the formula given in Appendix G where f_E is the expected frequency and f_O is the observed frequency. Using the table below, determine if the results are significant and then write a summary statement about how to interpret the results.

Distribution of *t*

p	***.1***	***.05***	***.02***	***.01***	***.001***
df					
1	2.71	3.48	5.41	6.64	10.83
2	4.60	5.99	7.82	9.21	13.82
3	6.25	7.82	9.84	11.34	16.27
4	7.78	9.49	11.67	13.28	18.46
5	9.24	11.07	13.39	15.09	20.52

5. Describe a study in which measures of central tendency would be the only necessary analysis.

6. A researcher is interested in whether there is a connection between native language and ESL reading. He administered a reading comprehension test to high intermediate-level learners in a community education program. Means and standard deviations were calculated for each L1 group and are displayed in the table below. How can these data be represented in a figure? Sketch two possible figures. Sketch one that also represents the standard deviation.

Hypothetical reading comprehension study

L1 Group	***Mean (x/100)***	***SD***
Spanish	75	8.5
French	78	2.6
Russian	67	5.4
Mandarin Chinese	60	7.3
Korean	61	11.7

7. Locate three articles that use statistical analyses (select articles that differ in the statistics used). For each:

a. What statistical analyses were used?
b. Why were these statistical tests used?
c. Were data presented in tables, graphs, or both?
d. If tables were used, describe the information presented in the tables.
e. If graphs were used, interpret the graph(s).

NOTES

1. This chapter deals with statistics rather than parameters. When researchers provide basic information about all members of a population (e.g., all first-year Spanish students at U.S. universities), they have information about the parameters of that population. It should be obvious that these data would be quite difficult to obtain; thus researchers draw information from a representative subset of that population, known as a sample. The information that we have about that population is referred to as statistics.
2. Butler (1985) suggests, "The 'mean' is what the layman means by an average although the statistician would regard all three measures of central tendency as types of average" (p. 27).
3. We acknowledge and are grateful to the contributions of Luke Plonsky to this section. He provided us with data and examples throughout.
4. *Studies in Second Language Acquisition* is now on this list.
5. SAS is another available statistical package (www.sas.com). It is more often used in business and in the hard sciences than in second language research, perhaps because it is perceived to be less user-friendly than SPSS, although it is used within the domain of language testing. SYSTAT (www.systat.com) is another program for statistical analysis and display. An example of its use in second language research is to investigate bilingual education. R, a free statistical package, has also gained traction in second language research (see Larson-Hall, 2015).