Statistics in Perspective

12

OBJECTIVES **Studying this chapter should enable you to:**

- Apply several recommendations when comparing data obtained from two or more groups.
- Apply several recommendations when relating variables within a single group.
- Explain what is meant by the term "effect size."
- Describe briefly how to use frequency polygons, scatterplots, and crossbreak tables to interpret data.
- Differentiate between statistically significant and practically significant research results.

INTERACTIVE AND APPLIED LEARNING After, or while, reading this chapter:

Go to the Online Learning Center at www.mhhe.com/fraenkel7e to:

- Learn More About Statistical Versus Practical Significance

Go to your Student Mastery Activities book to do the following activities:

- Activity 12.1: Statistical vs. Practical Significance
- Activity 12.2: Appropriate Techniques
- Activity 12.3: Interpret the Data
- Activity 12.4: Collect Some Data

Well, the results are in," said Tamara Phillips. "I got the consultant's report about that study we did last semester."

"What study was that?" asked Felicia Lee, as the two carpooled to Eisenhower Middle School, where they both taught eighth-grade social studies.

"Don't you remember? That guy from the university came down and asked some of us who taught social studies to try an inquiry approach?"

"Oh, yeah, I remember. I was in the experimental group—we used a series of inquiry-oriented lessons that they designed. They compared the results of our students with those of students who were similar in ability whose teachers did not use those lessons. What did they find out?"

"Well, the report states that those students whose teachers used the inquiry lessons had significantly higher test scores. But I'm not quite sure what that suggests."

"It means that the inquiry method is superior to whatever method the teachers of the other group used, doesn't it?"

"I'm not sure. It depends on whether the significance they're talking about refers to practical or only statistical significance."

"What's the difference?"

This difference—between statistical and practical significance—is an important one when it comes to talking about the results of a study. It is one of the things you will learn about in this chapter.

Now that you are somewhat familiar with both descriptive and inferential statistics, we want to relate them more specifically to practice. What are appropriate uses of these statistics? What are appropriate interpretations of them? What are the common errors or mistakes you should watch out for as either a participant in or consumer of research?

There are appropriate uses for both descriptive and inferential statistics. Sometimes, however, either or both types can be used inappropriately. In this chapter, therefore, we want to discuss the appropriate use of the descriptive and inferential statistics described in the previous two chapters. We will present a number of recommendations that we believe all researchers should consider when they use either type of statistics.

Approaches to Research

Much research in education is done in one of two ways: either two or more groups are compared or variables within one group are related. Furthermore, as you have seen, the data in a study may be either quantitative or categorical. Thus, four different combinations of research are possible, as shown in Figure 12.1.

Remember that all groups are made up of individual units. In most cases, the unit is one person and the group is a group of people. Sometimes, however, the unit is itself a group (for example, a class of students). In such cases, the "group" would be a collection of classes. This is illustrated by the following hypothesis: "Teacher

	Data	
	Quantitative	Categorical
Two or more groups are compared		
Variables within one group are related		

Figure 12.1
Combinations of Data and Approaches to Research

friendliness is related to student learning." This hypothesis could be studied with a group of classes and a measure of both teacher "friendliness" and average student learning for each *class*.

Another complication arises in studies in which the same individuals receive two or more different treatments or methods. In comparing treatments, we are not then comparing different groups of people but different groups of scores obtained by the same group at different times. Nevertheless, the statistical analysis fits the comparison group model. We discuss this point further in Chapter 13.

Comparing Groups: Quantitative Data

TECHNIQUES

Whenever two or more groups are compared using quantitative data, the comparisons can be made in a variety of ways: through frequency polygons, calculation of one or more measures of central tendency (averages), and/or calculation of one or more measures of variability (spreads). Frequency polygons provide the most information; averages are useful summaries of each group's performance; and spreads provide information about the degree of variability in each group.

When analyzing data obtained from two groups, therefore, the first thing researchers should do is construct a frequency polygon of each group's scores. This will show all the information available about each group and also help researchers decide which of the shorter and more convenient indices to calculate. For example, examination of the frequency polygon of a group's scores can indicate whether the median or the mean is the most appropriate measure of central tendency to use.

When comparing quantitative data from two groups, therefore, we recommend the following:

Recommendation 1: As a first step, prepare a frequency polygon of each group's scores.

Recommendation 2: Use these polygons to decide which measure of central tendency is appropriate to calculate. If any polygon shows extreme scores at one end, use medians for all groups rather than, or in addition to, means.

INTERPRETATION

Once the descriptive statistics have been calculated, they must be interpreted. At this point, the task is to describe, in words, what the polygons and averages tell researchers about the question or hypothesis being investigated. A key question arises: How large does a difference in means between two groups have to be in order to be important? When will this difference *make a difference?* How does one decide? You will recall that this is the issue of practical versus statistical significance that we discussed in Chapter 11.

Use Information About Known Groups. Unfortunately, in most educational research, this information is very difficult to obtain. Sometimes, prior experience can be helpful. One of the advantages of IQ scores is that, over the years, many educators have had enough experience with them to make differences between them meaningful. Most experienced counselors, administrators, and teachers realize, for example, that a difference in means of less than 5 points between two groups has little useful meaning, no matter how statistically significant the difference may be. They also know that a difference between means of 10 points is enough to have important implications. At other times, a researcher may have available a frame of reference, or

"That's the gist of what I want to say. Now get me some statistics to base it on."

standard, to use in interpreting the magnitude of a difference between means. One such standard consists of the mean scores of *known groups*. In a study of critical thinking in which one of the present authors participated, for example, the end-of-year mean score for a group of eleventh-graders who received a special curriculum was higher than is typical of the mean scores of eleventh-graders in general *and* close to the mean score of a group of college students, whereas a comparison group scored lower than both. Because the special-curriculum group also demonstrated a fall-to-spring mean gain that was twice that of the comparison group, the total evidence obtained through comparing their performance with other groups indicated that the gains made by the special-curriculum group were important.

Calculate the Effect Size. Another technique for assessing the magnitude of a difference between the means of two groups is to calculate what is known as **effect size** (ES).*

Effect size takes into account the size of the difference between means that is obtained, regardless of whether it is statistically significant. One of the most commonly used indexes of effect size is obtained by dividing the difference between the means of the two groups being compared by the standard deviation of the comparison group. Thus:

$$ES = \frac{\text{mean of experimental group} - \text{mean of comparison group}}{\text{standard deviation of comparison group}}$$

When pre-to-post gains in the mean scores of two groups are compared, the formula is modified as follows:

$$ES = \frac{\text{mean experimental gain} - \text{mean comparison gain}}{\text{standard deviation of gain of comparison group}}$$

The standard deviation of gain score is obtained by first getting the gain (post – pre) score for each individual and then calculating the standard deviation as usual.†

While effect size is a useful tool for assessing the magnitude of a difference between the means of two groups, it does not, in and of itself, answer the question of how large it must be for researchers to consider an obtained difference important. As is the case with significance levels, this is essentially an arbitrary decision. Most researchers consider that any effect size of .50 (that is, half a standard deviation of the comparison group's scores) or larger is an important finding. If the scores fit the normal distribution, such a value indicates that the difference in means between the two groups is about one-twelfth the distance between the highest and lowest scores of the comparison group. When assessing the magnitude of a difference between the means of two groups, therefore, we recommend the following:

Recommendation 3: Compare obtained results with data on the means of known groups, if possible.

Recommendation 4: Calculate an effect size. Interpret an ES of .50 or larger as important.

Use Inferential Statistics. A third method for judging the importance of a difference between the means of two groups is by the use of **inferential statistics**. It is common to find, even before examining polygons or differences in means, that a researcher has applied an inference technique (a *t*-test, an analysis of variance,

*The term *effect size* is used to identify a group of statistical indices, all of which have the common purpose of clarifying the magnitude of relationship.

†There are more effective ways to obtain gain scores, but we will delay a discussion until subsequent chapters.

CONTROVERSIES IN RESEARCH

Statistical Inference Tests—Good or Bad?

Our recommendations regarding statistical inference are not free of controversy. At one extreme are the views of Carver* and Schmidt,† who argue that the use of statistical inference tests in educational research should be banned. And in 2000 a survey of AERA members (American Educational Research Association) indicated that 19 percent agreed.‡

At the other extreme are those who agree with Robinson and Levin that "authors should *first* indicate whether the observed effect is a statistically improbable one, and *only if* it is should they indicate how *large or important* it is (is it a difference that *makes* a difference)."§

Cahan argued, to the contrary, that the way to avoid misleading conclusions regarding effects is not by using significance tests, but rather using confidence intervals accompanied by increased sample size.‖

In 1999 the American Psychological Association Task Force on Statistical Inference recommended that inference tests not be banned, but that researchers should "always provide some effect size estimate when reporting a p value," and further that "reporting and interpreting effect sizes in the context of previously reported research is *essential* to good research."#

What do you think? Should significance tests be banned in educational research?

*R. P. Carver (1993). The case against statistical significance testing revisited. *Journal of Experimental Education, 61:* 287–292.
†F. L. Schmidt (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1:* 115–129.
‡K. C. Mittag and B. Thompson (2000). A national survey of AERA members' perceptions of statistical significance tests and other statistical issues. *Educational Researcher, 29*(3): 14–19.
§D. H. Robinson and J. R. Levin (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher, 26* (January/February): 22.
‖S. Cahan (2000). Statistical significance is not a "Kosher Certificate" for observed effects: A critical analysis of the two-step approach to the evaluation of empirical results. *Educational Researcher, 29*(5): 34.
#L. Wilkinson and the APA Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54:* 599.

and so on) and then used the results as the *only* criterion for evaluating the importance of the results. This practice has come under increasing attack for the following reasons:

1. Unless the groups compared are random samples from specified populations (which is unusual), the results (probabilities, significance levels, and confidence intervals) are to an unknown degree in error and hence misleading.
2. The outcome is greatly affected by sample size. With 100 cases in each of two groups, a mean difference in IQ score of 4.2 points is statistically significant at the .05 level (assuming the standard deviation is 15, as is typical with most IQ tests). Although statistically significant, this difference is so small as to be meaningless in any practical sense.
3. The actual magnitude of difference is minimized or sometimes overlooked.
4. The purpose of inferential statistics is to provide information pertinent to generalizing sample results to populations, not to evaluate sample results.

With regard to the use of inferential statistics, therefore, we recommend the following:

Recommendation 5: Consider using inferential statistics only if you can make a convincing argument that a difference between means of the magnitude obtained is important (Figure 12.2).

Recommendation 6: Do not use tests of statistical significance to evaluate the magnitude of a difference between sample means. Use them only as they were intended: to judge the generalizability of results.

Recommendation 7: Unless random samples were used, interpret probabilities and/or significance levels as crude indices, not as precise values.

Recommendation 8: Report the results of inference techniques as confidence intervals rather than (or in addition to) significance levels.

Example. Let us give an example to illustrate this type of analysis. We shall present the appropriate calculations in detail and then interpret the results. Imagine that we have two groups of eighth-grade students, 60 in each group, who receive different methods of social studies instruction for one semester. The teacher of one group uses an inquiry method of instruction, while the teacher of the other group uses the lecture method. The

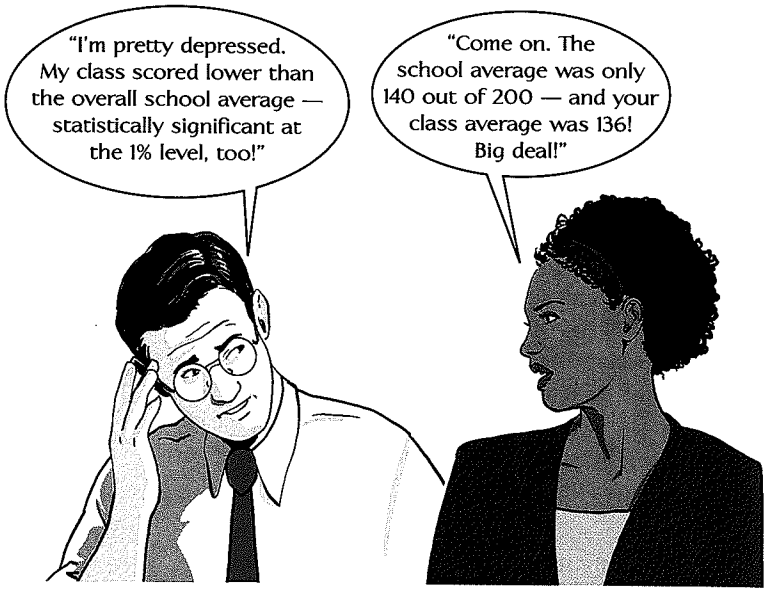


Figure 12.2 *A Difference That Doesn't Make a Difference!*

researcher's hypothesis is that the inquiry method will result in greater improvement than the lecture method in explaining skills as measured by the "test of ability to explain" (see page 151) in Chapter 8. Each student is tested at the beginning and at the end of the semester. The test consists of 40 items; the range of scores on the pretest is from 3 to 32, or 29 points. A gain score (posttest–pretest) is obtained. These gain scores are shown in the frequency distributions in Table 12.1 and the frequency polygons in Figure 12.3.

These polygons indicate that a comparison of means is appropriate. Why?* The mean of the inquiry group is 5.6 compared to the mean of 4.4 for the lecture group. The difference between means is 1.2. In this instance, a comparison with the means of known groups is not possible, since such data are not available. A calculation of effect size results in an ES of .44, somewhat below the .50 that most researchers recommend for significance. Inspection of Figure 12.3, however, suggests that the difference between the means of the two groups should not be discounted. Figure 12.4 and Table 12.2 show that the number of students gaining 7 or more points is 25 in the inquiry group and 13 (about half as many) in the lecture group. A gain of 7 points on a 40-item test can be considered substantial, even more so when it is recalled that the range was 29 points (3–32) on the pretest. If a gain of 8 points is used, the numbers are 16 in the inquiry group and 9 in the lecture group. If a gain of 6 points is used, the numbers become 34 and 20. We would argue that these discrepancies are large enough, in context, to recommend the inquiry method over the lecture method.

*The polygons are nearly symmetrical without extreme scores at either end.

TABLE 12.1 *Gain Scores on Test of Ability to Explain: Inquiry and Lecture Groups*

	Inquiry		Lecture	
Gain Scores[a]	Frequency	Cumulative Frequency	Frequency	Cumulative Frequency
11	1	60	0	60
10	3	59	2	60
9	5	56	3	58
8	7	51	4	55
7	9	44	4	51
6	9	35	7	47
5	6	26	9	40
4	6	20	8	31
3	5	14	7	23
2	4	9	6	16
1	2	5	4	10
0	3	3	5	6
−1	0	0	1	1

[a]A negative score indicates the pretest was higher than the posttest.

The use of an inference technique (a *t*-test for independent means) indicates that $p < .05$ in one tail

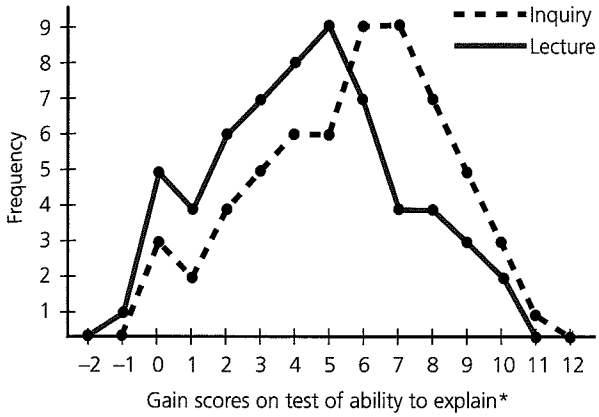


Figure 12.3 *Frequency Polygons of Gain Scores on Test of Ability to Explain: Inquiry and Lecture Groups*

*A negative score indicates the pretest was higher than the posttest.

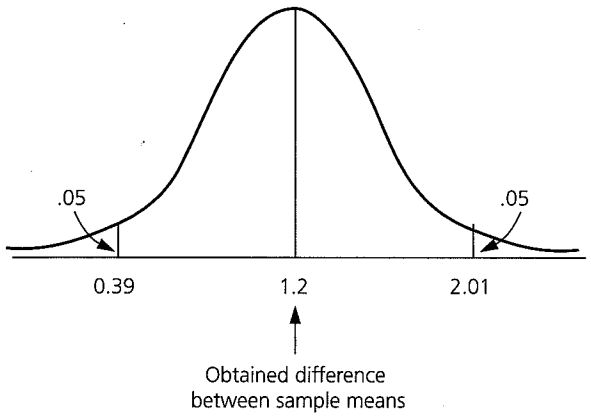


Figure 12.4 *90 Percent Confidence Interval for a Difference of 1.2 Between Sample Means*

(Table 12.2).* This leads the researcher to conclude that the observed difference between means of 1.2 points probably is not due to the particular samples used. Whether this probability can be taken as exact depends primarily on whether the samples were randomly selected. The 90 percent confidence interval is shown in Figure 12.4.† Notice that a difference of zero between the population means is not within the confidence interval.

*A directional hypothesis indicates use of a one-tailed test (see p. 227).

†1.65 SED gives .05 in one tail of the normal curve. 1.65 (SED) = 1.65(.49) = .81. 1.2 ± .81 equals .39 to 2.01. This is the 90 percent confidence interval. Use of 1.65 rather than 1.96 is justified because the researcher's hypothesis is concerned *only* with a *positive* gain (a one-tailed test). The 95 percent or any other confidence interval could, of course, have been used.

Relating Variables Within a Group: Quantitative Data

TECHNIQUES

Whenever a relationship between quantitative variables within a single group is examined, the appropriate techniques are the **scatterplot** and the **correlation coefficient.** The scatterplot illustrates all the data visually, while the correlation coefficient provides a numerical summary of the data. When analyzing data obtained from a single group, therefore, researchers should begin by constructing a scatterplot. Not only will it provide all the information available, but it will help them judge which correlation coefficient to calculate (the choice usually will be between the Pearson *r*, which assumes a **linear,** or **straight-line relationship,** and eta, which describes a **curvilinear,** or curved, **relationship**).‡

Consider Figure 12.5. All of the five scatterplots shown represent a Pearson correlation of about .50. Only in (*a*), however, does this coefficient (.50) completely convey the nature of the relationship. In (*b*) the relationship is understated, since it is a curvilinear one, and eta would give a higher coefficient. In (*c*) the coefficient does not reflect the fan-shaped nature of the relationship. In (*d*) the coefficient does not reveal that there are two distinct subgroups. In (*e*) the coefficient is greatly inflated by a few unusual cases. While these illustrations are a bit exaggerated, similar results are often found in real data.

When examining relationships within a single group, therefore, we recommend the following:

Recommendation 9: Begin by constructing a scatterplot.

Recommendation 10: Use the scatterplot to determine which correlation coefficient is appropriate to calculate.

Recommendation 11: Use *both* the scatterplot and the correlation coefficient to interpret results.

INTERPRETATION

Interpreting scatterplots and correlations presents problems similar to those we discussed in relation to

‡Because both of these correlations describe the magnitude of relationship, they are also examples of effect size (see footnote, page 244).

TABLE 12.2 *Calculations from Table 12.1*

Inquiry Group						Lecture Group					
Gain Score	f[a]	fX[b]	$X - \bar{X}$[c]	$(X - \bar{X})^2$[d]	$f(X - \bar{X})^2$[e]	Gain Score	f	fX	$X - \bar{X}$	$(X - \bar{X})^2$	$f(X - \bar{X})^2$
11	1	11	5.4	29.2	29.2	11	0	0	6.6	43.6	0.0
10	3	30	4.4	19.4	58.2	10	2	20	5.6	31.4	62.8
9	5	45	3.4	11.6	58.0	9	3	27	4.6	21.2	63.6
8	7	56	2.4	5.8	40.6	8	4	32	3.6	13.0	52.0
7	9	63	1.4	2.0	18.0	7	4	28	2.6	6.8	27.2
6	9	54	0.4	0.2	1.8	6	7	42	1.6	2.6	18.2
5	6	30	−0.6	0.4	2.4	5	9	45	0.6	0.4	3.6
4	6	24	−1.6	2.6	15.6	4	8	32	−0.4	0.2	1.6
3	5	15	−2.6	6.8	34.0	3	7	21	−1.4	2.0	14.0
2	4	8	−3.6	13.0	52.0	2	6	12	−2.4	5.8	34.8
1	2	2	−4.6	21.2	42.4	1	4	4	−3.4	11.6	46.4
0	3	0	−5.6	31.4	94.2	0	5	0	−4.4	19.4	97.0
−1	0	0	−6.6	43.6	0.0	−1	1	−1	−5.4	29.2	29.2
−2	0	0	−7.6	57.8	0.0	−2	0	0	−6.4	41.0	0.0
Total	$\Sigma = 338$				$\Sigma = 446.4$		$\Sigma = 262$				$\Sigma = 450.4$

$$\bar{X}_1 = \frac{\Sigma fX}{n} = \frac{338}{60} = 5.6 \qquad \bar{X}_2 = \frac{\Sigma fX}{n} = \frac{262}{60} = 4.4$$

$$SD_1 = \sqrt{\frac{f(X - \bar{X})^2}{n}} = \sqrt{\frac{446.4}{60}} = \sqrt{7.4} = 2.7 \qquad SD_2 = \sqrt{\frac{f(X - \bar{X})^2}{n}} = \sqrt{\frac{450.4}{60}} = \sqrt{7.5} = 2.7$$

$$SEM_1 = \frac{SD}{\sqrt{n-1}} = \frac{2.7}{\sqrt{59}} = \frac{2.7}{7.7} = .35 \qquad SEM_2 = \frac{SD}{\sqrt{n-1}} = \frac{2.7}{\sqrt{59}} = \frac{2.7}{7.7} = .35$$

$$SED = \sqrt{(SEM_1)^2 + SEM_2)^2} = \sqrt{.35^2 + .35^2} = \sqrt{.12 + .12} = \sqrt{.24} = .49$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{SED} = \frac{1.2}{.49} = 2.45 \qquad p < .05$$

$$ES(\Delta) = \frac{\bar{X}_1 - \bar{X}_2}{SD_2} = \frac{1.2}{2.4} = .44$$

[a] f = frequency
[b] fX = frequency × score
[c] $X - \bar{X}$ = score − mean
[d] $(X - \bar{X})^2$ = (score − mean)^2
[e] $f(X - \bar{X})^2$ = frequency × (score − mean)^2

differences in means. How large must a correlation coefficient be to suggest an *important* relationship? What does an important relationship look like on a scatterplot?

As you can see, doing or evaluating research is not cut and dried; it is not a matter of following a set of rules, but rather requires informed judgment. In judging correlation coefficients, one must first assess their appropriateness, as was done with those in Figure 12.5. If the Pearson correlation coefficient is an adequate summary (and we have shown in Figure 12.5 that this is not always the case), most researchers would agree to the interpretations shown in Table 12.3 when testing a research hypothesis.

As with a comparison of means, the use of inferential statistics to judge the importance of the magnitude of a relationship is both common and often misleading. With a sample of 100, a correlation of only .20 is statistically

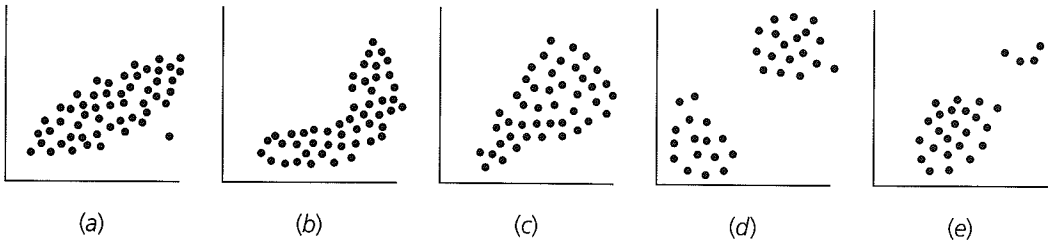


Figure 12.5 *Scatterplots with a Pearson* r *of .50*

significant at the .05 level with a two-tailed test. Accordingly, we recommend the following when interpreting scatterplots and correlation coefficients:

Recommendation 12: Draw a line that best fits all points in a scatterplot, and note the extent of deviations from it. The smaller the deviations all along the line, the more useful the relationship.*

Recommendation 13: Consider using inferential statistics only if you can give a convincing argument for the importance of the size of the relationship found in the sample.

Recommendation 14: Do not use tests of statistical significance to evaluate the magnitude of a relationship. Use them, as they were intended, to judge generalizability.

Recommendation 15: Unless a random sample was used, interpret probabilities and/or significance levels as crude indices, not as precise values.

Recommendation 16: Report the results of inference techniques as confidence intervals rather than as significance levels.

*Try this with Figure 12.5.

TABLE 12.3 *Interpretation of Correlation Coefficients when Testing Research Hypotheses*

Magnitude of r	Interpretation
.00 to .40	Of little practical importance except in unusual circumstances; perhaps of theoretical value.*
.41 to .60	Large enough to be of practical as well as theoretical use.
.61 to .80	Very important, but rarely obtained in educational research.
.81 or above	Possibly an error in calculation; if not, a very sizable relationship.

*When selecting a very few people from a large group, even correlations this small may have predictive value.

Example. Let us now consider an example to illustrate the analysis of a suspected relationship between variables. Suppose a researcher wishes to test the hypothesis that, among counseling clients, improvement in marital satisfaction after six months of counseling is related to self-esteem at the beginning of counseling. In other words, people with higher self-esteem would be expected to show more improvement in marital satisfaction after undergoing therapy for a period of six months than people with lower self-esteem. The researcher obtains a group of 30 clients, each of whom takes a self-esteem inventory and a marital satisfaction inventory prior to counseling. The marital satisfaction inventory is taken again at the end of six months of counseling. The data are shown in Table 12.4.

The calculations shown in Table 12.4 are not as hard as they look. Here are the steps that we followed to obtain $r = .42$.

1. Multiply n by ΣXY: $30(7{,}023) = 210{,}690$
2. Multiply ΣX by ΣY: $(1{,}007)(192) = 193{,}344$
3. Subtract step 2 from step 1: $210{,}690 - 193{,}344 = 17{,}346$
4. Multiply n by ΣX^2: $30(35{,}507) = 1{,}065{,}210$
5. Square ΣX: $(1{,}007)^2 = 1{,}014{,}049$
6. Subtract step 5 from step 4: $1{,}065{,}210 - 1{,}014{,}049 = 51{,}161$
7. Multiply n by ΣY^2: $30(2{,}354) = 70{,}620$
8. Square ΣY: $(192)^2 = 36{,}864$
9. Subtract step 8 from step 7: $70{,}620 - 36{,}864 = 33{,}756$
10. Multiply step 6 by step 9: $(51{,}161)(33{,}756) = 1{,}726{,}990{,}716$
11. Take the square root of step 10: $\sqrt{1{,}726{,}990{,}716} = 41{,}557$
12. Divide step 3 by step 11: $17{,}346/41{,}557 = .42$

Using the data presented in Table 12.4, the researcher plots a scatterplot and finds that it reveals two things. First, there is a tendency for individuals with higher initial self-esteem scores to show greater improvement in

TABLE 12.4 *Self-Esteem Scores and Gains in Marital Satisfaction*

Client	Self-Esteem Score before Counseling (X)	X^2	Gain in Marital Satisfaction after Counseling (Y)	Y^2	XY
1	20	400	−4	16	−80
2	21	441	−2	4	−42
3	22	484	−7	49	−154
4	24	576	1	1	24
5	24	576	4	16	96
6	25	625	5	25	125
7	26	676	−1	1	−26
8	27	729	8	64	216
9	29	841	2	4	58
10	28	784	5	25	140
11	30	900	5	25	150
12	30	900	14	196	420
13	32	1024	7	49	219
14	33	1089	15	225	495
15	35	1225	6	36	210
16	35	1225	16	256	560
17	36	1269	11	121	396
18	37	1396	14	196	518
19	36	1296	18	324	648
20	38	1444	9	81	342
21	39	1527	14	196	546
22	39	1527	15	225	585
23	40	1600	4	16	160
24	41	1681	8	64	328
25	42	1764	0	0	0
26	43	1849	3	9	129
27	43	1849	5	25	215
28	43	1849	8	64	344
29	44	1936	4	16	176
30	45	2025	5	25	225
Total (Σ)	$\Sigma = 1{,}007$	$\Sigma = 35{,}507$	$\Sigma = 192$	$\Sigma = 2{,}354$	$\Sigma = 7{,}023$

$$r = \frac{n\Sigma XY - \Sigma X \Sigma Y}{\sqrt{[n\Sigma X^2 - (\Sigma X)^2][n\Sigma Y^2 - (\Sigma Y^2)]}} = \frac{30(7023) - (1007)(192)}{\sqrt{[30(35507) - (1007)^2][30(2354) - (192)^2]}}$$

$$= \frac{210690 - 193344}{\sqrt{(1065210 - 1014049)(70620 - 36864)}} = \frac{17346}{\sqrt{(51161)(33756)}}$$

$$= \frac{17346}{\sqrt{1726990716}} = \frac{17346}{41557} = .42$$

marital satisfaction than those with lower initial self-esteem scores. Second, it also shows that the relationship is more correctly described as curvilinear—that is, clients with low *or* high self-esteem show less improvement than those with a moderate level of self-esteem (remember, these data are fictional). Pearson r equals .42.

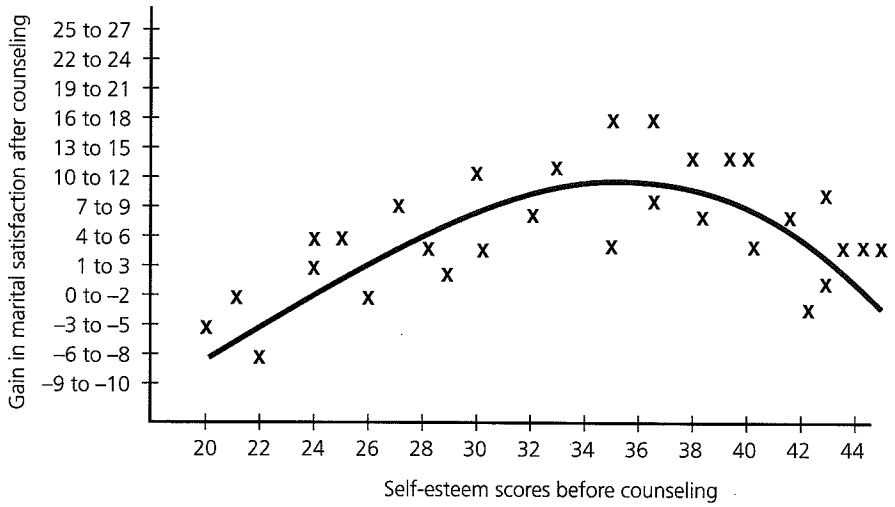


Figure 12.6 *Scatterplot Illustrating the Relationship Between Initial Self-Esteem and Gain in Marital Satisfaction Among Counseling Clients*

The value of eta obtained for these same data is .82, indicating a substantial degree of relationship between the two variables. We have not shown the calculations for eta since they are somewhat more complicated than those for r. The relationship is illustrated by the smoothed curve shown in Figure 12.6.

The researcher calculates the appropriate inference statistic (a t-test for r), as shown, to determine whether $r = .42$ is significant.

$$\text{Standard error of } r = \text{SE}_r = \frac{1}{\sqrt{n-1}}$$

$$= \frac{1}{\sqrt{29}} = .185$$

$$t_r = \frac{r - .00}{\text{SE}_r} = \frac{.42 - .00}{.185}$$

$$= 2.3; p < .01$$

As you can see, it results in an obtained value of 2.3 and a probability of $p < .01$, using a one-tailed test. A one-tailed test is appropriate for r if the direction of the relationship was predicted before examining the data. The probability associated with eta would (presumably) be obtained using a two-tailed test (unless the researcher predicted the shape of the curve from Figure 12.6 before examining the data). An eta of .82 is also statistically significant at $p = .01$, indicating that the relationship is unlikely to be due to the particular sample studied. Whether or not these probabilities are correct depends on whether or not the sample was randomly selected. The 95 percent confidence interval around the obtained value for r is shown in Figure 12.7.

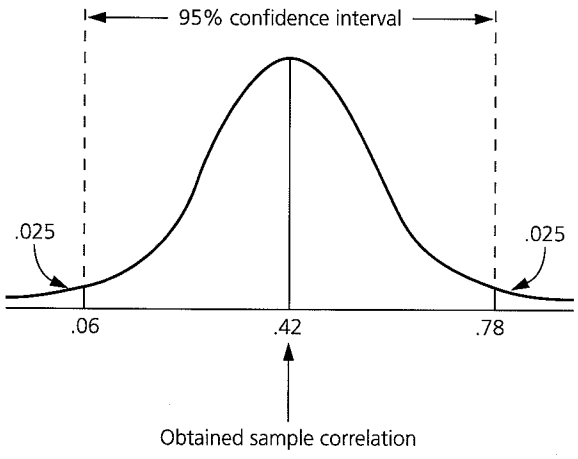


Figure 12.7 *95 Percent Confidence Interval for r = .42*

Comparing Groups: Categorical Data

TECHNIQUES

When the data involved are categorical data, groups may be compared by reporting either percentages (or proportions) or frequencies in crossbreak tables. Table 12.5 gives a fictitious example.

TABLE 12.5 *Gender and Political Preference (Percentages)*

	Percentage of Males	Percentage of Females
Democrat	20	50
Republican	70	45
Other	10	5
Total	100	100

TABLE 12.6 *Gender and Political Preference (Numbers)*

	Males	Females
Democrat	2	30
Republican	7	27
Other	1	3

TABLE 12.7 *Teacher Gender and Grade Level Taught: Case 1*

	Grade 4	Grade 5	Grade 6	Grade 7	Total
Male	10	20	20	30	80
Female	40	30	30	20	120
Total	50	50	50	50	200

TABLE 12.8 *Teacher Gender and Grade Level Taught: Case 2*

	Grade 4	Grade 5	Grade 6	Grade 7	Total
Male	22	22	25	28	97
Female	28	28	25	22	103
Total	50	50	50	50	200

INTERPRETATION

Once again, we must look at summary statistics—even percentages—carefully. Percentages can be misleading unless the number of cases is also given. At first glance, Table 12.5 may look impressive—until one discovers that the data in it represent 60 females and only 10 males. In crossbreak form, Table 12.6 represents the actual *numbers,* as opposed to percentages, of individuals.

Table 12.7 illustrates a fictitious relationship between teacher gender and grade level taught. As you can see, the largest number of male teachers is to be found in grade 7, and the largest number of female teachers is to be found in grade 4. Here, too, however, we must ask: How much difference must there be between these frequencies for us to consider them important? One of the limitations of categorical data is that such evaluations are even harder than with quantitative data. One possible approach is to examine prior experience or knowledge. Table 12.7 does suggest a trend toward an increasingly larger proportion of male teachers in the higher grades—but, again, is the trend substantial enough to be considered important?

The data in Table 12.8 show the same trend, but the pattern is much less striking. Perhaps prior experience or research shows (somehow) that gender differences become important whenever the within-grade difference is more than 10 percent (or a frequency of 5 in these data). Such knowledge is seldom available, however, which leads us to consider the summary statistic (similar to the correlation coefficient) known as the *contingency coefficient* (see Chapter 11). In order to use it, however, remember that the data *must* be presented in crossbreak tables. Calculating the contingency coefficient is easily done by hand or by computer. You will recall that this statistic is not as straightforward in interpretation as the correlation coefficient, since its interpretation depends on the number of cells in the crossbreak table. Nevertheless, we recommend its use.

Perhaps because of the difficulties mentioned above, most research reports using percentages or crossbreaks rely on inference techniques to evaluate the magnitude of relationships. In the absence of random sampling, their use suffers from the same liabilities as with quantitative data. When analyzing categorical data, therefore, we recommend the following:

Recommendation 17: Whenever possible, place all data into crossbreak tables.

Recommendation 18: To clarify the importance of relationships, patterns, or trends, calculate a contingency coefficient.

Recommendation 19: Do not use tests of statistical significance to evaluate the magnitude of relationships. Use them, as intended, to judge generalizability.

Recommendation 20: Unless a random sample was used, interpret probabilities and/or significance levels as crude indices, not as precise values.

Example. Once again, let us consider an example to illustrate an analysis, this time involving categorical data when comparing groups. Let us return to Tables 12.7 and 12.8 to illustrate the major recommendations for analyzing categorical data. We shall consider Table 12.7 first. Because there are 50 teachers, or 25 percent,

TABLE 12.9 *Crossbreak Table Showing Teacher Gender and Grade Level with Expected Frequencies Added (Data from Table 12.7)*

	Grade 4	Grade 5	Grade 6	Grade 7	Total
Male	10 (20)	20 (20)	20 (20)	30 (20)	80
Female	40 (30)	30 (30)	30 (30)	20 (30)	120
Total	50	50	50	50	200

of the total of 200 teachers at each grade level (4–7), we would expect that there would be 25 percent of the total number of male teachers and 25 percent of the total number of female teachers at *each* grade level as well. Out of the total of 200 teachers, 80 are male and 120 are female. Hence, the expected frequency for male teachers at each of the grade levels would be 20 (25 percent of 80), and for female teachers 30 (25 percent of 120). These expected frequencies are shown in parentheses in Table 12.9. We then calculate the contingency coefficient, which equals .28.

By referring to Table 11.1 in Chapter 11, we estimate that the upper limit for a 2 by 4 table (which we have here) is approximately .80. Accordingly, a contingency coefficient of .28 indicates only a slight degree of relationship. As a result, we would not recommend testing for significance. Were we to do so, however, we would find by looking in a chi-square probability table that three degrees of freedom requires a chi-square value of 7.81 to be considered significant at the .05 level. Our obtained value for chi square was 16.66, indicating that the small relationship we have discovered probably does exist in the population from which the sample was drawn.* This is a good example of the difference between statistical and practical significance. Our obtained correlation of .28 is statistically significant but practically insignificant. A correlation of .28 would be considered by most researchers as having little practical importance.

If we carry out the same analysis for Table 12.8, the resulting contingency coefficient is .10. Such a correlation is, for all practical purposes, meaningless, but should we (for some reason) wish to see if it was statistically significant, we would find that it is not significant at the .05 level (the chi-square value = 1.98, far below the 7.82 needed for significance).

Again, the calculations from Table 12.9 are not difficult. Here are the steps we followed:

*Assuming the sample is random.

1. For the first cell above (Grade 4-male), subtract E from O: $= 10 - 20 = -10$
2. Square the result: $(O - E)^2 = (-10)^2 = 100$
3. Divide the result by E:

$$\Sigma\frac{(O-E)^2}{E} = \frac{100}{20} = 5.00$$

O	E	$O - E$	$(O - E)^2$	$\frac{(O-E)^2}{E}$	
10	20	−10	100	100/20	= 5.00
40	30	10	100	100/30	= 3.33
20	20	0	0	0	= 0
30	30	0	0	0	= 0
20	20	0	0	0	= 0
30	30	0	0	0	= 0
30	20	10	100	100/20	= 5.00
20	30	−10	100	100/30	= 3.33

4. Repeat this process for each cell. (Be sure to include *all* cells.)
5. Add the results of all cells:

$$5.00 + 3.33 + 5.00 + 3.33 = 16.66 = \chi^2$$

6. To calculate the contingency coefficient, we used the formula:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{16.66}{16.66 + 200}} = .28$$

Relating Variables Within a Group: Categorical Data

Although the preceding section involves comparing groups, the reasoning also applies to hypotheses that examine relationships among categorical variables within just one group. A moment's thought shows why. The procedures available to us are the same—percentages or crossbreak tables. Suppose our hypothesis is that among college students, gender is related to political preference. To test this we must divide the data we obtain from this group by gender and political preference. This gives us the crossbreak in Table 12.6. Because all such hypotheses must be tested by dividing people into groups, the statistical analysis is the same whether seen as one group, subdivided, or as two or more different groups.

A summary of the most commonly used statistical techniques, both descriptive and inferential, as used with quantitative and categorical data, is shown in Table 12.10.

MORE ABOUT RESEARCH

Interpreting Statistics

- Suppose a researcher found a correlation of .08 between drinking grapefruit juice and subsequent incidence of arthritis to be statistically significant. Is that possible? (Yes, it is quite possible. If the sample had been randomly selected, and the sample size was around 500, a correlation of .08 would be statistically significant at the .05 level. But because of the small relationship—and many uncontrolled variables—we would not stop drinking grapefruit juice based on an r of only .08!)
- Suppose an early intervention program was found to increase IQ scores on average by 12 points, but that this was not statistically significant at the .05 level. How much attention would you give to this report? (We would pay considerable attention; 12 IQ points is a lot and could be very important if confirmed in replications. Evidently the sample size was rather small.)
- Suppose the difference in polling preference for a particular candidate was found to be 52 percent for the Democrat as opposed to 48 percent for the Republican, with a margin of error of 2 percent at the .05 level. Would you consider this difference important? (One way of reporting such results is that the probability of the difference being due to chance is less than .01.* In addition, a difference of only 4 points is of great practical importance since the winner in a two-person election needs only 51 percent of the vote to win. A very similar prediction proved wrong in the 1948 presidential election, when Truman defeated Dewey. The usual explanations are that the sample was not random and thus not representative, and/or that a lot of people changed their minds before they entered the voting booth.)

*The SE of each percentage must be 2.00 (the margin of error) divided by 1.96 (the number of standard deviations required at the 5 percent level), or approximately 1.00. The standard error of the difference (SED) equals the square root of $(1^2 + 1^2)$ or 1.4. The difference between 48 percent and 52 percent—4 percent—divided by 1.4 (the SED) equals 2.86, which yields a probability of less than .01.

TABLE 12.10 *Summary of Commonly Used Statistical Techniques*

	DATA	
	Quantitative	**Categorical**
Two or more groups are compared:		
Descriptive Statistics	• Frequency polygons • Averages • Spreads • Effect size	• Percentages • Bar graphs • Pie charts • Crossbreak (contingency) tables
Inferential Statistics	• *t*-test for means • ANOVA • ANCOVA • MANOVA • MANCOVA • Confidence intervals • Mann-Whitney *U* test • Kruskal-Wallis ANOVA • Sign test • Friedman two-way ANOVA	• Chi square • *t*-test for proportions
Relationships among variables are studied within one group:		
Descriptive Statistics	• Scatterplot • Correlation coefficient (*r*) • eta	• Crossbreak (contingency) tables • Contingency coefficient
Inferential Statistics	• *t*-test for *r* • Confidence intervals	• Chi square • *t*-test for proportions

A Recap of Recommendations

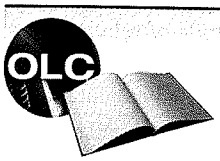
You may have noticed that many of our recommendations are essentially the same, regardless of the method of statistical analysis involved. To stress their importance, we want to state them again here, all together, phrased more generally.

We recommend that researchers:

- Use graphic techniques before calculating numerical summary indices. Pay particular attention to outliers.
- Use both graphs and summary indices to interpret results of a study.
- Make use of external criteria (such as prior experience or scores of known groups) to assess the magnitude of a relationship whenever such criteria are available.
- Use professional consensus when evaluating the magnitude of an effect size (including correlation coefficients).
- Consider using inferential statistics only if you can make a convincing case for the importance of the size of the relationship found in the sample.
- Use tests of statistical significance only to evaluate generalizability, not to evaluate the magnitude of relationships.
- When random sampling has not occurred, treat probabilities as approximations or crude indices rather than as precise values.
- Report confidence intervals rather than, or in addition to, significance levels whenever possible.

We also want to make a final recommendation involving the distinction between parametric and nonparametric statistics. Since the calculation of statistics has now become rather easy and quick owing to the availability of many computer programs, we conclude with the following suggestion to researchers:

- Use *both* parametric and nonparametric techniques to analyze data. When the results are consistent, interpretation will thereby be strengthened. When the results are not consistent, discuss possible reasons.



Go back to the **INTERACTIVE AND APPLIED LEARNING** feature at the beginning of the chapter for a listing of interactive and applied activities. Go to the **Online Learning Center** at **www.mhhe.com/fraenkel7e** to take quizzes, practice with key terms, and review chapter content.

Main Points

APPROACHES TO RESEARCH

- A good deal of educational research is done in one of two ways: either two or more groups are compared, or variables within one group are related.
- The data in a study may be either quantitative or categorical.

COMPARING GROUPS USING QUANTITATIVE DATA

- When comparing two or more groups using quantitative data, researchers can compare them through frequency polygons, calculation of averages, and calculation of spreads.
- We recommend, therefore, constructing frequency polygons, using data on the means of known groups, calculating effect sizes, and reporting confidence intervals when comparing quantitative data from two or more groups.

RELATING VARIABLES WITHIN A GROUP USING QUANTITATIVE DATA

- When researchers examine a relationship between quantitative variables within a single group, the appropriate techniques are the scatterplot and the correlation coefficient.
- Because a scatterplot illustrates all the data visually, researchers should begin their analysis of data obtained from a single group by constructing a scatterplot.
- Therefore, we recommend constructing scatterplots and using both scatterplots and correlation coefficients when relating variables involving quantitative data within a single group.

COMPARING GROUPS USING CATEGORICAL DATA

- When the data are categorical, groups can be compared by reporting either percentages or frequencies in crossbreak tables.
- It is a good idea to report *both* the percentage and the number of cases in a crossbreak table, as percentages alone can be misleading.
- Therefore, we recommend constructing crossbreak tables and calculating contingency coefficients when comparing categorical data involving two or more groups.

RELATING VARIABLES WITHIN A GROUP USING CATEGORICAL DATA

- When you are examining relationships among categorical data within one group, we again recommend constructing crossbreak tables and calculating contingency coefficients.

TWO FINAL RECOMMENDATIONS

- When tests of statistical significance can be applied, it is recommended that they be used to evaluate generalizability only, not to evaluate the magnitude of relationships. Confidence intervals should be reported in addition to significance levels.
- Both parametric and nonparametric techniques should be used to analyze data rather than either one alone.

Key Terms

correlation coefficient 247
curvilinear relationship 247
effect size (ES) 244
inferential statistics 244
linear relationship 247
scatterplot 247

For Discussion

1. Give some examples of how the results of a study might be significant statistically yet unimportant educationally. Could the reverse be true?
2. Are there times when a slight difference in means (e.g., an effect size of less than .50) might be important? Explain your answer.
3. When comparing groups, the use of frequency polygons helps us decide which measure of central tendency is the most appropriate to calculate. How so?
4. Why is it important to consider outliers in scatterplots?
5. "When analyzing data obtained from two groups, the first thing researchers should do is construct a frequency polygon of each group's scores." Why is this important—or is it?
6. Why is it important to use *both* graphs and summary indices (e.g., the means) to interpret the results of a study—or is it?
7. A picture, supposedly, is worth a thousand words. Would this statement also apply to analyzing the results of a study? Can numbers alone ever give a complete picture of a study's results? Why or why not?

Research Exercise 12: Statistics in Perspective

Using Problem Sheet 12, once again state the question or hypothesis of your study. Summarize the descriptive and inferential statistics you would use to describe the relationship you are hypothesizing. Then tell how you would evaluate the magnitude of any relationship you might find. Finally, describe the changes in techniques to be used from those you described in Problem Sheets 10 and 11, if any. If your study is qualitative, you will probably omit question 3.

Problem Sheet 12

Statistics in Perspective

1. The question or hypothesis of my study is: _____

2. My expected relationship(s) would be described using the following descriptive statistics: _____

3. The inferential statistics I would use are: _____

4. I would evaluate the magnitude of the relationship(s) I find by: _____

5. The changes (if any) in my use of descriptive or inferential statistics from those I described in Problem Sheets 10 and 11 are as follows: _____

An electronic version of this Problem Sheet that you can fill in and print, save, or e-mail is available on the Online Learning Center at www.mhhe.com/fraenkel7e.

A full-sized version of this Problem Sheet that you can fill in or photocopy is in your Student Mastery Activities book.