Introduction to Research in Education

Second Edition

Donald Ary
Northern Illinois University

Lucy Cheser Jacobs
Indiana University

Asghar Razavieh
Pahlavi University,
Shiraz, Iran

Holt, Rinehart and Winston
New York Chicago San Francisco Dallas
Montreal Toronto London Sydney

Library of Congress Cataloging in Publication Data

Ary, Donald.
Introduction to research in education.

Includes index.
1. Educational research. I. Jacobs, Lucy Cheser, joint author
II. Razavieh, Asghar, joint author. III. Title.
LB1028.A7 1979 370'.78 78-10653
ISBN O-03-020606-5

Printed in the United States of America

90 1 039 9 8765 4 3 2 1

Preface

Our goals in preparing the first edition of *Introduction to Research in Education were to* provide a book that would enable readers to master the basic competencies necessary (I) to understand and evaluate the research of others and (2) to plan and conduct their own research with a minimum of assistance. The reception the first edition has received is an indication that we have been reasonably successful in achieving these goals. We hope both students and colleagues will find this second edition even more useful. The latter have provided suggestions we found very helpful in our attempts to provide a clearer and more complete text. For example, in response to many suggestions we have added study exercises for each chapter.

The sequence of topics discussed in this book begins with a general description of the scientific approach and the relevance of this approach to the search for knowledge in education. We assume that the reader is not familiar with the concepts, assumptions, and terminology of the scientific approach; therefore, these are explained as they are introduced. We have expanded the discussion of the roles deductive reasoning and inductive reasoning play in science. From this basis we proceed to suggestions for translating general problems into questions amenable to scientific inquiry. A section on the identification of the population and variables of interest has been added to the guidelines for developing problems for research.

Next we describe the role of previous research in the planning of a research project. We have updated the sources of related literature with particular emphasis on data bases that provide efficient access to relevant research and theory. We then proceed to investigate the ways in which theory, experience, observations, and related literature lead to hypothesis formation.

The more useful descriptive and inferential statistical procedures are included, with the emphasis on the role these procedures play in the research process and on their interpretation. The role of systematic observation and measurement is explored, and examples of useful measurement procedures are included. The chapter on validity and reliability has been extensively revised.

Following this, we discuss the various types of research that have proven useful in education, pointing out the advantages and disadvantages of various approaches without espousing any particular one as being superior to the others. The chapter on ex post facto research has been revised in order to show more clearly the strengths and weaknesses of this type of research.

We conclude by presenting the general rules for interpreting the results of research and the accepted procedures for reporting such results. We have added sections on procedures for meeting ethical and legal requirements in research and on the use of computers.

The focus of this edition remains the provision of a text designed for use in an introductory course in educational research. Its aim is to familiarize the beginning researcher with the procedures for conducting an original research project. We focus on the typical and practical problems encountered in research, beginning with the formulation of the question and continuing through the preparation of the final report.

Although *Introduction to Research in Education* is directed toward the beginning student in educational research, it is hoped that others who wish to learn more about the philosophy, tools, and procedures of scientific inquiry in education will find it useful. The principal criterion used in determining what to include has been the potential usefulness of various aspects of educational research to the educational practitioner.

To all of those teachers who used the first edition and have made very valuable suggestions for improving and updating the second edition, we are deeply grateful. We also thank Linda Burke for her many contributions. We are indebted to the Literary Executor of the late Sir Ronald A. Fisher, F.R.S., for permission to reprint Table A.5 in the Appendix from the book *Statistical Methods for Research Workers.* We are also indebted to the aforementioned and to Dr. Frank Yates, F.R.S., for permission to reprint Tables A.3 and A.7 (also in the Appendix) from their book *Statistical Tables for Biological, Agricultural, and Medical Research.*

D. A.
L. c. J.
A. R.

contents

Part 1 Foundations

1 The Scientific Approach in Education 3

Sources of Knowledge 3
The Nature of Science 12
The Nature of Research 20
Research Methodologies in Education 25
The Language of Research 26
History of Educational Research 30
Summary 35
Exercises 35

Part 2 Research Background

2 The Research Problem 41

The Nature of Problems 42
The Sources of Problems 42
Evaluating the Problem 47
Stating the Problem 50
Identifying Population and Variables 51
Summary 54
Exercises 55

3 Related Literature 57

The Role of Related Literature in a Research Project 57
Reference Sources in Education 58
Organizing the Related Literature 68

Summary 71
Exercises 7 1

4 **The Hypothesis 72**
Suggestions for Deriving Hypotheses 74
Characteristics of the Usable Hypothesis 76
Stating the Hypothesis 81
Testing the Hypothesis 81
The Research Plan 83
Summary 85
Exercises 85

Part 3
Statistical Analysis

5 Descriptive Statistics 91
Scales of Measurement 91
Organizing Research Data 95
Measures of Central Tendency 99
Measures of Variability 105
Standard Scores 109
The Normal Curve 111
Correlation 113
Correlation Coefficient 116
Summary 124
Exercises 125

6 **Sampling and Inferential Statistics 129**
Sampling 129
The Strategy of Inferential Statistics 140
Level of Significance 143
Significance of the Difference between Two Means 144
Analysis of Variance 152
The Chi-Square Test of Significance 162
Summary 166
Exercises 167

Part 4 Tools of Research

7 Fundamentals of Measurement 173
Measuring Instruments 173
Summary 194
Exercises 194

8 Validity and Reliability 196
Validity 196
Reliability 206
Summary 218
Exercises 2 I9

Part 5 Research Methods

9 **Experimental Research in Education 225**
Characteristics of Experimental Research 228
Experimental Comparison 237
Experimental Design 237
Summary 267
Exercises 268

10 **Ex Post Facto Research 271**
Ex Post Facto and Experimental Approaches Compared 271
Conditions Necessary for Inferring Causal Relationships 274
Partial Control in Ex Post Facto Research 279
Designing the Ex Post Facto Investigation 286
Procedure in an Ex Post Facto Investigation 287
The Role of Ex Post Facto Research 289

Summary 291
Exercises 292

11 Descriptive and Historical Research 295
Descriptive Research 295
Historical Research 312
Summary 314
Exercises 3 14

Part 6
Communicating Research

12 Guidelines for Writing Research Proposals 319
Writing a Research Proposal 319
Critiquing the Proposal 332
The Importance of Completing the Proposal before Collecting Data 334
Research Ethics 334
Summary 336
Exercises 337

13 Analyzing, Interpreting, and Reporting Results 339
Analysis of Data 339
Principles of Interpretation 342
The Research Report 347
Common Errors in Research Reports 354
Summary 356
References 357
Exercises 358

Appendix 36 1

Name Index 385
Subject Index 386

6 Sampling and Inferential Statistics

The statistics discussed in the previous chapter are used for organizing, summarizing, and describing data. In research, however, we often need to go further than describing data. After making observations of a sample, we employ induction or inference to generalize our findings to the entire population from which the sample was drawn. To do this we need techniques that enable us to make valid inferences from samples to whole populations.

Sampling

An important characteristic of inferential statistics is the process of going from the part to the whole. For example, we might study a randomly selected group of 500 students attending a university in order to make generalizations about the entire student body of that university.

The small group that is observed is called a *sample* and the larger group about which the generalization is made is called *apopulation.* A *population is* defined as “all members of any well-defined class of people, *events or* objects.’’[1] For example, in a study where American adolescents constitute the population of interest, one could define this population as all American boys and girls within the age range of 12-21. A sample is a portion of a population. For example, the students of Washington High School in Indianapolis constitute a sample of American adolescents. They are a portion of the large population in that they are both American citizens and within the age range of 12-21.

Statistical inference is a procedure by means of which one estimates *parameters,* characteristics of populations, from *statistics,* characteristics of samples. Such estimations are based on the laws of probability and are best estimates rather than absolute facts. In any such inferences a certain degree of error is involved.

RATIONALE OF SAMPLING

Inductive reasoning is an essential part of the scientific approach. The inductive method involves making observations and then drawing conclusions from these observations. If one can observe all instances of a population, one can with confidence base conclusions about the population on these observations (perfect induction). On the other hand, if one observes only some instances of a population then one can do no more than infer that these observations will be true of the population

[1] Fred N. Kerlinger, *Foundations of Behavioral Research* (New York: Holt, Rinehart and Winston, 1966), p. 52.

as a whole (imperfect induction). This is the concept of sampling, which involves taking a portion of the population, making observations on this smaller group, and then generalizing the findings to the large population.

Sampling is indispensable to the researcher. Usually the time, money, and effort involved do not permit a researcher to study all possible members of a population. Furthermore, it is generally not necessary to study all possible cases to understand the phenomenon under consideration. Sampling comes to our aid by enabling us to study a portion of the population rather than the entire population.

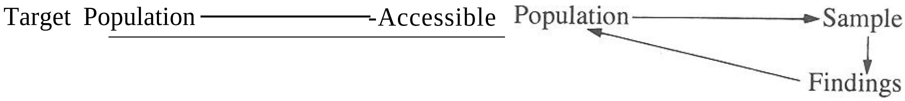
Since the purpose of drawing a sample from a population is to obtain information concerning that population, it is extremely important that the individuals included in a sample constitute a representative cross section of individuals in the population. That is, samples must be representative if one is to be able to generalize with confidence from the sample to the population. For example, the researcher might assume that the students at Washington High School are representative of American adolescents. However, this sample might not be representative if the individuals who are included have some characteristics that differ from the parent population. The location of their school, their socioeconomic background, their family situation, their prior experiences, and many other characteristics of this group might make them unrepresentative of American adolescents. This type of sample would be termed a ***biased sample.*** The findings of a biased sample cannot legitimately be generalized to the population from which it is taken.

Steps in Sampling

The first essential in sampling is the identification of the population to be represented in the study. If the researcher is interested in learning about the teachers in the St. Louis school system, all those who teach within that system constitute the target population. In a study of the attitudes and values of American adolescents, the target population would be all American boys and girls in the age range of 12-21, granted that adolescence is operationally defined as the period between ages 12 and 21.

However, since it is usually not possible to deal with the whole of the target population, one must identify that portion of the population to which one can have access--called the ***accessible population-and*** it is from this group that the researcher will take the sample for the study. The nature of the accessible population is influenced by the time and resources of the researcher. In a typical attitude study, for example, a researcher might designate all adolescent boys and girls in California or just those in San Francisco as the accessible population.

From the accessible population, one selects a sample in such a way that it is representative of that population. For example, the researcher would have to sample from adolescents all over the state of California if California adolescents are identified as the accessible population. Or if adolescents living in San Francisco are the accessible population, then the sample would be drawn from this particular group.



How safely can one generalize from a sample to the target population? If the sample selected is truly representative of the accessible population, then there is little difficulty in making this first step in the generalization process. The general principle is: If a sample has been selected so that it is representative of the accessible population, findings from the sample can be generalized to that population. For example, if one has selected a representative sample of California adolescents, then one could make generalizations concerning the attitudes and values of all adolescent boys and girls in California.

However, generalizing from the accessible population to the target population typically involves greater risk. The confidence that one can have in this step depends upon the similarity of the accessible population to the target population. In the example above, a researcher could have more confidence making generalizations about American adolescents if adolescents in several states throughout the country are designated as the accessible population rather than those in California alone. In this way all sections of the United States would be represented and a more adequate sampling of attitudes and values would be possible.

It is true that one must make an inferential "leap of faith" when estimating population characteristics from sample observations. The likelihood that such inferences will be correct is largely a function of the sampling procedure employed. Various sampling procedures are available to researchers for use in the selection of a subgroup of a population that will represent that population well and will avoid bias.

RANDOM SAMPLING

The best known of the sampling procedures is random sampling. The basic characteristic of random sampling is that all members of the population have an equal and independent chance of being included in the sample. That is, for every pair of elements X and Y, X's chance of being selected equals Y's chance, and the selection of X in no way affects Y's probability of selection. The steps in random sampling are:

1. Define the population.
2. List all members of the population.
3. Select the sample by employing a procedure where sheer chance determines which members on the list are drawn for the sample.

The most systematic procedure for drawing a random sample is to refer to a ***table of random numbers,*** which is a table containing columns of digits that have been mechanically generated, usually by a computer, to assure a random order. Table

A.6 in the Appendix is an example. The first step in drawing a random sample from a population is to assign each member of the population a distinct identification number. Then the table of random numbers is used to select the identification numbers of the subjects to be included in the sample.

Let us illustrate the use of this table to obtain a sample of adolescents from the population of students attending Washington High School. First it is necessary to enumerate all of the individuals included in the population. The principal's office could supply a list of all students enrolled in the school. One would then assign a number to each individual in the population for identification purposes. Assuming there were 800 students in the school, one might use the numbers 000, 001, 002, 003, , 799 for this purpose. Then one would enter a table of random numbers to obtain numbers of three digits each, using only those numbers that are less than or equal to 199. For each number chosen, the corresponding element in the population falls in the sample. One continues the process until the desired number for the sample has been chosen. It is customary, in using a table of random numbers, to determine by chance the point at which the table is entered. One way is to touch the page blindly and begin wherever the page is touched.

The generally understood meaning of the word *random* is "without purpose or by accident." However, random sampling is purposeful and methodical. It is apparent that a sample selected randomly is not subject to the biases of the researcher. When researchers employ this method, they are committing themselves to selecting a sample in such a way that their biases are not permitted to operate. They are pledging themselves to avoid a deliberate selection of subjects who will confirm the hypothesis. They are allowing chance alone to determine which elements in the population will he in the sample.

One would expect a random sample to be representative of the parent population sampled. However, random selection, especially with small samples, does not absolutely guarantee a sample that will represent the population well. Random selection does guarantee that any differences between the sample and the parent population are only a function of chance and not a result of the researcher's bias. The differences between random samples and their parent population are not systematic. For example, the mean reading achievement of a random sample of sixth graders may he higher than the mean reading achievement of the parent population, but it is equally likely that the mean for the sample will be lower than the mean for the parent population. In other words, with random sampling the sampling errors are just as likely to be negative as they are to be positive.

Furthermore, statistical theorists have, through deductive reasoning, shown how much one can expect the observations derived from random samples to differ from what would be observed in the population. All of the procedures described in this chapter have this aim in mind. Remember that characteristics observed in a small sample are more likely to differ from population characteristics than are characteristics observed in a large sample. When random sampling is used, the researcher can employ inferential statistics to estimate how much the population is likely to differ from the sample. The inferential statistics in this chapter are all based on random sampling and apply only to those cases in which randomization has been employed.

Unfortunately, random sampling requires an enumeration of all the individuals in a finite population before the sample can be drawn-a requirement that often presents a serious obstacle to the use of this method in practice.

STRATIFIED SAMPLING

When the population consists of a number of subgroups or strata that may differ in the characteristics being studied, it is often desirable to use a form of sampling called stratified sampling. For example, if one were conducting a poll designed to assess opinions on a certain political issue, it might be advisable to subdivide the population into groups on the basis of age or occupation because one would expect opinions to differ systematically among various age or occupational groups. In stratified sampling one first identifies the strata of interest and then draws a specified number of subjects from each stratum. The basis for stratification may be geographical or it may involve characteristics of the population, such as income, occupation, sex, age, year in college, or teaching level. In the study of adolescents, for example, one may be interested not merely in surveying the attitudes of adolescents toward certain phenomena, but also in comparing the attitudes of adolescents who reside in small towns with those who live in medium-size and large cities. In such a case, one would divide the adolescent population into three groups, based on the size of the towns or cities in which they reside, and then randomly select independent samples from each stratum.

An advantage of stratified sampling is that it enables the researcher to determine to what extent each stratum in the population is represented in the sample. One may either take equal numbers from each stratum or select in proportion to the sire of the stratum in the population. This latter procedure is known as ***proportional** stratified sampling;* that is, the stratum is represented in the sample in exact proportion to its frequency in the total population. If 10 percent of the voting population are college students, then 10 percent of one’s sample of voters to be polled would be taken from this stratum. The procedure used will be chosen according to the nature of the research question. If the emphasis is on the types of differences among the strata, one selects equal numbers of cases from each. If the characteristics of the entire population are the main concern, proportional sampling is more appropriate.

When applicable, stratified sampling may give us a more representative sample than simple random sampling. In simple random sampling certain strata may by chance be over- or underrepresented in the sample. For example, in the simple random sample of high school students it would be theoretically possible (though highly unlikely) to obtain female subjects only. This could not happen, however, if males and females are listed separately and a random sample is then chosen from each group.

The major advantage of stratified sampling is that it guarantees representation of defined groups in the population.

CLUSTER SAMPLING

As mentioned earlier, it is very difficult, if not impossible, to list all the members of a target population and select the sample from among them. The population of American high school students, for example, is so large that one cannot list all its members for the purpose of drawing a sample. In addition, it would be a very expensive undertaking to study a sample that is scattered all around the United States. In this case it would be more convenient to study subjects in naturally occurring groups or clusters. That is, the researcher would choose a number of schools randomly from a list of schools and then include all the students in those schools in the sample. This kind of sampling is referred to as *cluster* sampling since the unit chosen is not an individual but a group of individuals who are naturally together. These individuals constitute a cluster insofar as they are alike with respect to characteristics relevant to the variables of the study. Let us assume a public opinion poll is being conducted in Atlanta. The investigator would probably not have access to a list of the entire adult population; thus it would be impossible to draw a simple random sample. A more feasible approach would involve the selection of a random sample of, say, fifty blocks from a city map, and then the polling of all the adults living on those blocks. Each block represents a cluster of subjects, similar in certain characteristics associated with living in proximity.

It is essential that the clusters actually included in the study be chosen at random from a population of clusters. If only a single cluster were used-for example, one elementary school in a large city—one could not generalize to the population. Another procedural requirement is that once a cluster is selected, all the members of the cluster must be included in the sample. The sampling error in a cluster sample is much greater than in true random sampling.

SYSTEMATIC SAMPLING

Still another form of sampling is called systematic sampling. This procedure involves drawing a sample by taking every kth case from a list of the population.

One first decides how many subjects one wants in the sample (n). Since one knows the total number of members in the population (N), one simply divides N by n and determines the sampling interval (k) to apply to the list. The first member is randomly selected from the first k members of the list and then every kth member of the population is selected for the sample. For example, let us assume a total population of 500 subjects and a desired sample size of 50; thus, $k = N/n = 500/50 = 10$.

One would start near the top of the list so that the first case could be randomly selected from the first ten cases, and then every tenth case thereafter would be selected. Say the third name or number on the list was the first to be selected. One would then add the sampling interval *k, or* 10, to 3—and thus the thirteenth person falls in the sample, as does the twenty-third, and so on-and would continue adding the constant sampling interval until the end of the list is reached.

Systematic sampling differs from simple random sampling in that the various choices are not independent. Once the first case is chosen, all subsequent cases to be included in the sample are automatically determined.

If the original population list is in random order, systematic sampling would yield a sample that could be statistically considered a reasonable substitute for a random sample. However, if the list is alphabetical, for example, it is possible that every kth member of the population might have some unique characteristic that would affect the dependent variable of the study and thus yield a biased sample. Systematic sampling from an alphabetical list would probably not give a representative sample of various national groups because certain national groups tend to cluster under certain letters and the sampling interval could omit them entirely or at least not include them to an adequate extent.

It should be noted that the various types of sampling that have been discussed are not mutually exclusive. Various combinations may be used. For example, one could use cluster sampling if one is studying a very large and widely dispersed population. At the same time, one may be interested in stratifying the sample to answer questions regarding its different strata. In this case one would stratify the population according to the predetermined criteria and then randomly select the clusters of subjects from among each stratum.[2]

THE SIZE OF THE SAMPLE

One of the first questions to be asked concerns the number of subjects that need to be included in the sample. Technically, the sire of the sample depends upon the precision the researcher desires in estimating the population parameter at a particular confidence level. There is no single rule that can be used to determine sample size. An estimation of required sample size can be calculated algebraically if one defines precisely the variance of the population, the expected difference, and the desired probabilities of Type I and Type II errors (see 179-80). A number of statistics texts describe this procedure.

The best answer to the question of size is to use as large a sample as possible. A larger sample is much more likely to be representative of the population. Furthermore, with a large sample the data are likely to be more accurate and precise: which is to say, the larger the sample, the smaller the standard error. In general, the standard error of a sample mean is inversely proportional to the square root of n. Thus in order to double the precision of one's estimation, one must quadruple the sample size.

Some authors suggest that one include at least thirty subjects in a sample since this number permits the use of large sample statistics. In experimental research, one should select a sample that will permit at least thirty in each group. Descriptive research typically uses larger samples; it is sometimes suggested that one select 10 to 20 percent of the accessible population for the sample.

[2] For further discussion of specific sampling techniques, the reader is referred to W. G. Cochran, *Sampling Techniques,* 2nd ed. (New York: Wiley, 1963).

It must be emphasized, however, that size alone will not guarantee accuracy. Representativeness is the most important consideration in selecting a sample. A sample may be large and still contain a bias. The latter situation is well illustrated by the *Literary Digest* poll of 1936, which predicted the defeat of President Roosevelt. Although the sample included approximately two and a half million respondents, it was not representative of the voters; and thus an erroneous conclusion was reached. The bias resulted from the selection of respondents for the poll from automobile registrations, telephone directories, and the magazine's subscription lists. These subjects would certainly not be representative of the total voting population in 1936. Also, since the poll was conducted by mail, the results were biased by differences between those who responded and those who did not. Thus the researcher must recognize that sample size will not compensate for the bias that may be introduced through faulty sampling techniques. Representativeness must remain the prime goal in sample selection.

THE CONCEPT OF SAMPLING ERROR

When an inference is made from a sample to a population a certain amount of error is involved because even random samples can be expected to vary from one to another. The mean intelligence score of one random sample of fourth graders may be different from the mean intelligence score of another random sample of fourth graders from the same population. Such differences, called sampling errors, result from the fact that one has observed a sample and not the entire population.

Sampling error is defined as the difference between a population parameter and a sample statistic. For example, if one knows the mean of the entire population (symbolized μ) and also the mean of a random sample (symbolized $\bar{X}$) from that population, the difference between these two, $\bar{X} - \mu$, represents sampling error (symbolized e). Thus, $e = \bar{X} - \mu$. For example, if we know that the mean intelligence score for a population of 10,000 fourth graders is $\mu = 100$ and a particular random sample of 200 has a mean of $\bar{X} = 99$, then the sampling error is $\bar{X} - \mu = 99 \quad 100 = \text{I}$ Because we usually depend on sample statistics to estimate population parameters, the notion of how samples are expected to vary from populations is a basic element in inferential statistics. However, instead of trying to determine the discrepancy between a sample statistic and the population parameter (which is not often known), the approach in inferential statistics is to estimate the variability that could be expected in the statistics from a number of different random samples drawn from the same population. Since each of the sample statistics is considered to be an estimate of the same population parameter, then any variation among sample statistics must be attributed to sampling error.

THE LAWFUL NATURE OF SAMPLING ERRORS

Given that random samples drawn from the same population will vary from one another, is using a sample to make inferences about a population really any better than just guessing? Yes, it is, because sampling errors behave in a lawful and predictable manner. The laws concerning sampling error have been derived through deductive logic and have been confirmed through experience.

Although we cannot predict the nature and extent of the error in a single sample, we can predict the nature and extent of sampling errors in general. Let us illustrate this with reference to sampling errors connected with the mean.

Sampling Errors of the Mean

Some sampling error can always be expected when a sample mean X is used to estimate a population mean μ. Although, in practice, such an estimate is based on a single sample mean, assume that one drew several random samples from the same population and computed a mean for each sample. We would find that these sample means would differ from one another and would also differ from the population mean (if it were known). This variation among the means is due to the sampling error associated with each random sample mean as an estimate of the population mean. Sampling errors of the mean have been studied carefully and it has been found that they follow regular laws.

The Expected Mean of Sampling Errors Is Zero. Given an infinite number of random samples drawn from a single population, the positive errors can be expected to balance the negative errors so that the mean of the sampling errors will be zero. For example, if the mean height of a population of college freshmen is 5 feet 9 inches, and several random samples are drawn from that population, we would expect some samples to have mean heights greater than 5 feet 9 inches and some to have mean heights less than 5 feet 9 inches. In the long run, however, the positive and negative sampling errors will balance. If we had an infinite number of random samples of the same sire, calculated the mean of each of these samples, then computed the mean of all these means, the mean of the means would be equal to the population mean.

Since positive errors equal negative errors, a single sample mean is as likely to underestimate a population mean as to overestimate it. Therefore, we can justify saying that a sample mean is an unbiased estimate of the population mean, and is a reasonable estimate of the population mean.

Sampling Error Is an Inverse Function of Sample Size. As the size of a sample increases there is less fluctuation from one sample to another in the value of the mean. In other words, as the sire of a sample increases the expected sampling error decreases. Small samples are more prone to sampling error than large ones. One would expect the means based on samples of 10 to fluctuate a great deal more than the means based on samples of 100. In our height example it would be much more likely that a random sample of four had three above-average freshmen and only one below-average freshman than that of a random sample of 40 had 30 above average and ten below. As sample size increases the likelihood that the mean of the sample is near the population mean also increases. There is a mathematical relationship between sample size and sampling error. We will show later how this relationship has been incorporated into inferential formulas.

Sampling Error Is a Direct Function of the Standard Deviation of the Population. The more spread or variation we have among members of a population, the more spread or variation we expect in sample means. For example, the mean weights of random samples of 25 each selected from a population of professional jockeys would show relatively less sampling error than the mean weights of samples of 25 each selected from a population of school teachers. The weights of professional jockeys fall within a narrow range, the weights of school teachers do not. Therefore, for a given sample sire, the expected sampling error for teachers' weights would be greater than the expected sampling error for jockeys' weights.

Sampling Errors Are Distributed in a Normal or Near Normal Manner around the Expected Mean of Zero. Sample means near the population mean will occur more frequently than sample means far from the population mean. As we move farther and farther from the population mean we find fewer and fewer sample means **occurring.** Both theory and experience have shown that the means of random samples are distributed in a normal or near normal manner around the population mean.

Since a sampling error in this case is the difference between a sample mean and the population mean, the distribution of sampling errors is also normal or near normal **in** shape. The two distributions are by definition identical except that the distribution of sample means has a mean equal to the population mean while the mean of the sampling error is zero.

The distribution of sample means will resemble a normal curve even when the population from which the samples are drawn is not normally distributed. For example, in a typical elementary school we find about equal numbers of children of the various ages included, so a polygon of the children's ages would be basically rectangular. If we take random samples of 40 each from a school with equal numbers of children aged 6 through 11 we would find many samples with means near the population mean of 8.5, sample means of about 8 or 9 would be less common, and sample means as low as 7 or as high as IO would be rare.

Standard Error of the Mean

Since the extent and the distribution of sampling errors can be predicted, we can use sample means with predictable confidence to make inferences concerning population means. However, we need an estimate of the magnitude of the sampling error associated with the sample mean when it is used as an estimate of the population mean. An important tool for this purpose is the *standard error of the mean.*

It has been stated that sampling error manifests itself in the variability of sample means. Thus, if one calculates the standard deviation of a collection of means from random samples from a single population, one would have an estimate of the amount of sampling error. It is possible, however, to obtain this estimate on the basis of only one sample. We have seen that two things affect the size of sampling error, the size of the sample and the standard deviation in the population. When these two things are known, one can predict the standard deviation of sampling errors. This expected standard deviation of sampling errors of the mean is called the standard error of the mean and is represented by the symbol $\sigma_{\bar{X}}$. It has been shown through deductive logic that the standard error of the mean is equal to the standard deviation of the population (σ) divided by the square root of the number in each sample $(\sqrt{n})$. In formula form:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \tag{6.1}$$

where:
$\sigma_{\bar{X}}$ = standard error of the mean
σ = standard deviation of the population
n = number in each sample

In chapter 5 we saw that standard deviation (σ) is an index of the degree of spread among individuals in a population. In the same way standard error of the mean $(\sigma_{\bar{X}})$ is an index of the spread expected among the means of samples drawn randomly from a population. As we will see, the interpretation of σ and $\sigma_{\bar{X}}$ is very similar.

Since the means of random samples have approximately normal distributions we can also use the normal curve model to make inferences concerning population means. Given that the expected mean of sample means is equal to the population mean, and that the standard deviation of these means is equal to the standard error of the mean, and that the means of random samples are distributed normally, one can compute a z-score for a sample mean and refer that z to the normal curve table to approximate the probability of a sample mean occurring through chance that far or farther from the population mean. The z is derived by subtracting the population mean from the sample mean and then dividing this difference by the standard error of the mean. In formula form this becomes:

$$z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \tag{6.2}$$

To illustrate, let us consider a college admissions officer who wonders if his population of applicants is average or below average on the College Board examination. The national mean for College Board scores is 500 and the standard deviation is 100. He pulls a random sample of 64 from his population and finds the mean of the sample to be 470. He asks the question, “How probable is it that a random sample of 64 with a mean of 470 would be drawn from a population with a mean of 500?’’ Using formula (6.1), the admissions officer calculates the standard error of the mean as 12.5:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

$$= \frac{100}{\sqrt{64}}$$

$$= 12.5$$

Calculating the z-score for his sample mean with formula (6.2) he has:

$$z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$$

$$\frac{470 - 500}{12.5}$$

$$= -2.4$$

Thus, his sample mean deviates from the population mean by 2.4 standard error units. What is the probability of having a sample mean that deviates by this amount ($2.4\sigma_{\bar{X}}$'s) or more from the population mean? It is only necessary to refer to the normal curve in order to express this deviation (z) in terms of probability. Referring a z of -2.4 to the normal curve table, one finds that the probability of a z that low or lower is .0082. This means that a z-score that low or lower would occur by chance only about 8 times in 1,000. Since the probability of getting a sample mean that far from the population mean is remote, he concludes that his sample mean probably did not come from a population with a mean of 500 and therefore the mean of his population, applicants to his college, is probably less than 500.

The Strategy of Inferential Statistics

Inferential statistics is the science of making reasonable decisions with limited information. We use what we observe in samples and what is known about sampling error to reach fallible but reasonable decisions about populations. A basic tool of inferential statistics is the *null hypothesis.*

NULL HYPOTHESIS

Suppose we have 100 fourth graders available to participate in an experiment concerning the teaching of certain number concepts. We randomly assign 50 students to be taught these concepts by Method A and the other 50 to be taught by Method B. We arrange their environment in such a way that the two groups differ only in method of instruction. At the end of the experiment we administer an examination that is considered to be a suitable operational definition of mastery of the number concepts of interest. We find that the mean for those students taught by

Method B is higher than the mean for those taught by Method A. How do we interpret this difference?

Assuming we have been careful to make the learning conditions of the two groups equivalent except for the method of teaching, we could account for the difference by declaring that (1) the method of teaching caused the difference or (2) the difference occurred by chance. Even though the subjects were randomly assigned to the treatments, it is possible that through chance the Method B group had students who were more intelligent, more highly motivated, or for some other reason were more likely to learn the number concepts than the students in the Method A group, no matter how they were taught.

The difference between the groups therefore could be a result of a relationship between the variables-method of teaching and mastery of the concepts-r it could be the result of chance alone (i.e., sampling error). How are we to know which explanation is correct? In the ultimate sense we cannot know. What we do, then, is estimate the likelihood of chance alone being responsible for the observed difference and determine which explanation to accept as a result of this estimate.

The chance explanation is known as the null hypothesis, which, as you will recall from chapter 4, is a statement that there is no actual relationship between variables and that any observed relationship is only a function of chance. In our example the null hypothesis would state that there is no relationship between teaching method and mastery of the number concepts.

Another way of stating the null hypothesis in our example is to declare that the mean for all fourth graders taught by Method A is equal to the mean for all fourth graders taught by Method B. In formula form, using the symbol μ for population mean, this statement becomes

$$H_0: \mu_A = \mu_B$$

where
H_0 = the null hypothesis
μ_A = the mean of all fourth graders taught by Method A
μ_B = the mean of all fourth graders taught by Method B

Note that the assumption is made that the 50 pupils taught by Method A are a sample of all fourth graders who might be taught by Method A, and the 50 pupils taught by Method B are a sample of all those who might be taught by Method B. The investigator hopes to use the data from the experiment to infer what would be expected when other fourth graders are taught by Methods A or B.

In interpreting the observed difference between the groups, the investigator must choose between the chance explanation (null hypothesis) and the explanation that states that there is a relationship between variables (research hypothesis)-and must do so without knowing the ultimate truth concerning the populations of interest. This choice is based on incomplete information and is therefore subject to possible error.

TYPE I AND TYPE II ERRORS

The investigator will either retain or reject the null hypothesis. Either decision may be right or wrong. If the null hypothesis is true, the investigator is correct in retaining it and in error in rejecting it. The rejection of a true null hypothesis is labeled a Type I error.

If the null hypothesis is false, the investigator is in error in retaining it and correct in rejecting it. The retention of a false null hypothesis is labeled a Type II error. The four possible states of affairs are summarized in Table 6. 1.

Table 6.1 Schematic Representation of Type I and Type II Errors

		The real situation (unknown to the investigator) is that the null hypothesis is:	
		true	false
The investigator, after making a test of significance, concludes that the null hypothesis is:	true	investigator is correct	investigator makes Type II error
	false	investigator makes Type I error	investigator is correct

Let us consider some possible consequences of the two types of errors in our example.

Type I

The investigator declares that there is a relationship between teaching method and the mastery of the numerical concepts and therefore recommends Method B as the better method. Schools discard textbooks and other materials based on Method A and purchase materials based on Method B. In-service training is instituted to train teachers to teach by Method B. After all this expenditure of time and money, the schools do not observe an increase in mastery of the numerical concepts. Subsequent experiments do not produce the results observed in the original investigation. Although the ultimate truth or falsity of the null hypothesis is still unknown, the evidence supporting it is overwhelming. The original investigator is embarrassed and humiliated.

Type II

The investigator concludes that the difference between the two groups may be attributed to luck and that the null hypothesis is probably true. He declares that one method is as good as the other.

Subsequent investigators conclude that Method B is better than Method A, and schools that change from Method A to Method B report impressive gains in student mastery Although the ultimate truth still remains unknown, a mountain of evidence supports the research hypothesis. The original investigator is embarrassed (but probably not humiliated).

Type I errors typically lead to changes that are unwarranted. Type II errors typically lead to a maintenance of the status quo when a change is warranted. The consequences of a Type I error are generally considered more serious than the consequences of a Type II error, although there are certainly exceptions to this.

Level of Significance

Recall that all scientific conclusions are statements that have a high probability of being correct, rather than statements of absolute truth. How high must the probability be before an investigator is willing to declare that a relationship between variables exists? In other words, how unlikely must the null hypothesis be before one rejects it? The consequences of rejecting a true null hypothesis, a Type I error, vary with the situation. Therefore, investigators usually weigh the relative consequences of Type I and Type II errors and decide, before conducting their experiments, how strong the evidence must be before they would reject the null hypothesis. This predetermined level at which a null hypothesis would be rejected is called the level of *significance*.

Of course, one could avoid Type I errors by always retaining the null hypothesis or avoid Type II errors by always rejecting it. Neither of these alternatives is productive. If the consequences of a Type I error would be very serious but a Type II error would be of little consequence, the investigator might decide to risk the possibility of a Type I error only if the estimated probability of the observed relationship's being due to mere luck is one chance in a thousand or less. This is called testing the hypothesis at the .001 level of significance. In this case the investigator is being very careful not to declare that a relationship exists when there is no relationship. However, this decision means the acceptance of a high probability of a Type II error, declaring there is no relationship when in fact a relationship does exist.

If the consequences of a Type I error are judged to be not serious, the investigator might decide to declare that a relationship exists if the probability of an observed relationship's being due to mere luck is one chance in ten or less. This is called testing the hypothesis at the .10 level of significance. Here the investigator is taking only moderate precautions against a Type I error, yet is not taking a great risk of a Type II error.

The level of significance is the probability of a Type I error that an investigator is willing to risk in rejecting a null hypothesis. If an investigator sets the level of significance at .01, it means that the null hypothesis will be rejected if the estimated probability of the observed relationship's being a chance occurrence is one in a hundred. If the level of significance is set at .0001, the null hypothesis will be rejected only if the estimated probability of the observed relationship's being a function of mere chance is one in 10,000 or less. The most commonly used levels of significance in the field of education are the .05 and the .01 levels.

Traditionally, investigators determine the level of significance after weighing the relative seriousness of Type I and Type II errors, but before running the experiment. If the data derived from the completed experiment indicate that the probability of the null hypothesis being true is less than the predetermined acceptable probability, the results are declared to be statistically significant. If the probability is greater than the predetermined acceptable probability, the results are described as nonsignificant-that is, the null hypothesis is retained.

The familiar meaning of the word *significant* is "important" or "meaningful." In statistics this word means "less likely to be a function of chance than some predetermined probability." Results of investigations can be statistically significant without being inherently meaningful or important.

There are numerous ways of testing a null hypothesis. Among the most widely used are the t-test, analysis of variance, and the chi-square test.

Significance of the Difference between Two Means

THE t-TEST

We have shown that it is possible to make use of the normal probability curve to compare the mean of a sample with the population mean by using the z-score to see whether or not the sample mean is representative of the population mean. To demonstrate that point, we found the standard error of the mean for the sample distribution, then used the formula $(\bar{X} - \mu)/\sigma_{\bar{X}}$. Implied in using this procedure is the appropriateness of the normal probability curve.

However, it has been shown mathematically that the normal curve is appropriate for hypothesis testing only when the population standard deviation is known. In most research situations the population standard deviation is not known and must be estimated by the formula

$$s = \sqrt{\frac{\Sigma x^2}{n - 1}} \tag{6.3}$$

where

s = estimated population standard deviation
x^2 = sum of the squared deviations scores, $\Sigma(X - \bar{X})^2$
n = number in the sample

When this estimate (s) is substituted for the population standard deviation (a) in the calculation of the standard error of the mean, it is customary to express Formula (6.1) as

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} \text{ instead of } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

When $s_{\bar{x}}$ is used instead of $\sigma_{\bar{x}}$ each finite sample size has its own unique probability distribution. These distributions are known as the *t*-curves. These distributions become more and more similar to the normal curve as the size of the sample increases. A series of distributions called t-distributions has been developed for testing hypotheses concerning the population mean using small samples. When the sample size is infinite, the t-distribution is the same as the normal distribution. As the sample size becomes smaller, the t-distribution becomes increasingly different from the z-distribution. For our purposes it is not necessary to know how to calculate r-distributions since the most frequently needed results of these calculations are to be found in Table A.3 in the Appendix. The t-curve does not approach the base line as rapidly as does the normal curve. Some of the t-curves are shown in Figure 6.1 along with the normal curve, the solid line labeled ''*df* = ∞.''

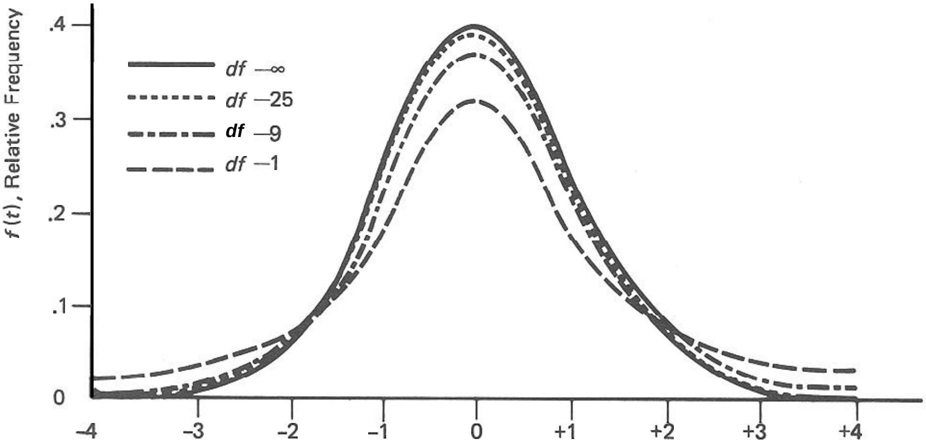


Figure 6.1 *t*-Curves for Various Degrees of Freedom

The r-curves arc labeled according to their degrees of freedom, abbreviated *"df."* Before further discussion of the characteristics of ~-curves, let us turn our attention to the concept of degrees of freedom.

Degrees of Freedom

The number of degrees of freedom refers to the number of observations free to vary around a constant parameter. To illustrate the general concept of degrees of freedom, suppose a teacher asks a student to name any five numbers that come into his head. The student would be free to name any five numbers he chooses. We would say that the student has five degrees of freedom. Now suppose the teacher tells the student to name five numbers but to make sure that the mean of these five numbers is equal to 20. The student now is free to name any numbers he chooses for the first four, but for the last number he must name the number that will make the total for the five numbers 100 in order to arrive at a mean of 20. If the student names, as his first four numbers, 16, 0.5, 1,000, and -65, then his fifth number must be -851.5. The student has five numbers to name and one restriction, so his degrees of freedom are five minus one equals four. We can show this in formula form as

$$\begin{aligned} df &= n - 1 \\ &= 5 - 1 \\ &\ \ 4 \end{aligned}$$

Now, suppose the teacher asks the student to name seven numbers in such a way that the first three have a mean of 10 and all seven have a mean of 12. Here we have seven numbers and two restrictions, so

$$\begin{aligned} df &= n - 2 \\ &= 7 - 2 \\ &\ \ 5 \end{aligned}$$

The concept of degrees of freedom is involved in most of the procedures in inferential statistics. There is an appropriate method of computing the degrees of freedom associated with each procedure.

The t-Test for Independent Samples

Research workers often draw two random samples from a population and assign a specific experimental treatment to each group. After being exposed to this treatment, the two groups are compared with respect to certain characteristics in order to find the effect of the treatments. A difference might be observed between the two groups after such treatments, but this difference might be statistically nonsignificant-that is, attributable to chance. The index used to find the significance of the difference between the means of the two samples in this case is called the t-test for *independent samples*. These samples are referred to as independent because they are drawn independently from a population without any pairing or other relationship between the two groups.

Let us use an example. Suppose a researcher is interested in finding out whether stress affects problem-solving performance. The first step is to randomly select two groups of 15 subjects from among the students in a course. The scores (X) on the dependent variable problem-solving performance are shown in Table 6.2, followed by the deviation scores (x) and the squared deviation scores (x^2). Since the members of the two groups are selected and assigned randomly, the mean performances of the two groups in a problem-solving task should not significantly differ prior to the treatment. After the treatment, however, the mean performance of the two groups should differ significantly if stress is actually related to problem-solving performance.

Table 6.2 The Computation of the r-Value for Two Sample Means

Group 1 (Nonstress Condition)			Group 2 (Stress Condition)		
X_1	x_1	x_1^2	X_2	x_2	x_2^2
18	+4	16	13	+3	9
17	+3	9	12	+2	4
16	+2	4	12	+2	4
16	+2	4	11	+1	I
16	+2	4	II	+1	I
15	+1	1	11	+1	1
15	+1	1	10	0	0
15	+1	I	10	0	0
14	0	0	10	0	0
14	0	0	IO	0	0
13	−1	I	9	−1	1
12	-2	4	9	−1	1
11	-3	9	8	−2	4
10	-4	16	7	−3	9
8	-6	36	7	−3	9
$\Sigma X_1 = 210$		$\Sigma x_1^2 = 106$	$\Sigma X_2 = 150$		$\Sigma x_2^2 = 44$
n_1 15			n_2 15		
$\bar{X}_1 = 14$			$\bar{X}_2 = 10$		

The data presented in Table 6.2 are the performance scores of the members of the two groups, one of which worked under stress conditions and the other, under relaxed (nonstress) conditions. The table shows that the mean performance score of the subjects in the stress group is 10 and the mean performance score of the nonstress group is 14. Clearly there is a difference. Now we need to determine whether or not this difference could easily occur by chance.

In order to do this we must estimate how much difference between the groups would be expected through chance alone under a true null hypothesis. An appropriate procedure for doing this is to calculate the *standard error of the difference between two means* ($s_{\bar{X}_1-\bar{X}_2}$). The formula for this in the case of independent samples is

$$s_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{\Sigma x_1^2 + \Sigma x_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{6.4}$$

where

$s_{\bar{X}_1-\bar{X}_2}$ = the standard error of the difference between two means
n_1 = the number of cases in Group 1
n_2 = the number of cases in Group 2

Σx_1^2 = the sum of the squared deviation scores in Group 1
Σx_2^2 = the sum of the squared deviation scores in Group 2

The standard error of the difference between two means is sometimes referred to as the error term for the t-test. In our example this would be calculated as follows:

$$\begin{aligned} s_{\bar{X}_1-\bar{X}_2} &= \sqrt{\frac{106 + 44}{15 + 15 - 2}\left(\frac{1}{15} + \frac{1}{15}\right)} \\ &= \sqrt{\frac{150}{28}\left(\frac{2}{15}\right)} \\ &= \sqrt{0.714} \\ &= 0.84 \end{aligned}$$

This calculation tells us the difference that would be expected through chance alone if the null hypothesis is true. In other words, the value 0.84 is the difference we would expect between the mean performance scores for our two groups if they are drawn at random from a common population and are *not* subjected to different treatments. Given an infinite number of samples in such circumstances, we would expect to observe a difference of less than 0.84 in 68 percent of the calculations of the differences between such random groups and a value of more than 0.84 in the other 32 percent. (It is beyond the scope of this text to discuss the reason why the application of the formula for the standard error of the difference between means yields the appropriate estimated difference that would be due to chance.)[3]

In our example for the data in Table 6.2 we should expect a difference of 0.84 through chance under a true null hypothesis. We observed a difference of 4.0. Is the observed difference sufficiently greater than the expected difference to enable us to reject the null hypothesis?

To answer this question we first make a ratio of the two numbers. This ratio is called the *t-ratio.* Its formula is

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1-\bar{X}_2}} \tag{6.5}$$

where
$\bar{X}_1 - \bar{X}_2$ = the observed difference between two means
$s_{\bar{X}_1-\bar{X}_2}$ = the standard error of the difference between two means (expected difference between the two means when the null hypothesis is true)

We can write the r-ratio formula in more complete form by including the formula for the standard error for the difference between two means:

[3] For a discussion of the rationale of this procedure, see Donald Ary and Lucy C. Jacobs, *Introduction to Statistics* (New York: Holt, Rinehart and Winston, 1976).

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{\Sigma x_1^2 + \Sigma x_2^2}{n_1 + n_2 - 2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

In our example the value of the t-ratio is

$$\frac{14 - 10}{0.84} = 4.76$$

Our observed difference is 4.16 times as large as the difference expected under a true null hypothesis. Is this large enough to reject the null hypothesis at the .05 level? To answer this we need only calculate the degrees of freedom and consult the t-table.

The degrees of freedom for an independent f-test are the number of cases in the first group plus the number of cases in the second group minus 2.

$$df = n_1 + n_2 - 2$$

In our example we have 15 + 15 2 = 28 degrees of freedom. We can now use Table A.3 in the Appendix to determine the significance of our results. The first column in this table is labeled "Degrees of Freedom." One finds the appropriate row in the table by locating the degrees of freedom in one's study. For our example we would consult the row for 28 degrees of freedom. The remaining columns show the t-values associated with certain probabilities. In the row for 28 degrees of freedom we find 1.701 in the column labeled .1, which tells us that with a true null hypothesis and 28 degrees of freedom a t-ratio of + 1.701 or more or − 1.701 or less will occur by chance one time in ten. The number 2.048 in the column labeled .05 indicates that under a true null hypothesis and 28 degrees of freedom a t-ratio of ±2.048 or more will occur by chance 5 percent of the time.

Our observed ratio of 4.76 is greater than 2.048, which means that the difference between our groups is greater than the value required to reject the null hypothesis at the .05 level of significance. The estimated probability of the null hypothesis being true is less than 5 percent ($p < .05$). Although we do not know for certain that the variables *stress* and *problem-solving performance* are related, the evidence is significant enough according to our previously set criteria to enable us to conclude that the observed relationship is not just a chance occurrence. If the observed t-ratio had been less than 2.048, we would have concluded that the evidence was not good enough to lead us to declare that a relationship exists between the variables. In other words, we would have retained the null hypothesis.

Notice that as we proceed from left to right in the t-table we find the t-values required for rejecting the null hypothesis at increasingly rigorous levels of significance. For 28 degrees of freedom a value of 2.763 or greater would lead to the rejection of a null hypothesis at the .01 level. A value of 3.674 or greater would lead to the rejection of the null hypothesis at the .001 level. So our value of 4.16 is significant not only at the .05 level ($p<.05$) but also at the .01 level ($p<.01$) and the .001 level ($p<.001$).

The t-Test for Nonindependent Samples

So far our discussion has centered around comparing the means obtained from two independent samples. In an independent sample each member is chosen randomly from the population, and the composition of one group has no bearing on the composition of the other group. Sometimes, however, investigators may wish to match the subjects of their two groups on some qualities that are important to the purpose of their research, or they may wish to compare the means obtained by the same group under two different experimental conditions. In such cases the groups are no longer independent, inasmuch as the composition of one group is related to the composition of the other group. Also we would expect the dependent variable scores to be correlated. Therefore the f-test for nonindependent or correlated means must be used. The measure to be analyzed by the nonindependent r-test is the difference between the paired scores.

Let us consider an example. Suppose we wish to know whether taking a research course affects the attitudes of the students toward research. To investigate this we select a research class and obtain attitude measures toward research from the students on the first and last days of class. Let us suppose we have collected such data and the results are as presented in Table 6.3. Columns (2) and (3) show the scores of each student in the first and second tests. Column (4) presents the difference between the first and second scores of each student. The sum of these differences amounts to $+30$. The mean of the differences, $+2$, is found by dividing $+30(\Sigma D)$ by n, the number of paired observations, or 15. Column (5) shows the squares of the differences.

The formula for the nonindependent f-test is

$$t = \frac{\bar{D}}{\sqrt{\dfrac{\Sigma D^2 - \dfrac{(\Sigma D)^2}{N}}{N(N-1)}}} \tag{6.7}$$

where

t = the t-value for nonindependent (correlated) means

D = the difference between the paired scores

$\bar{D}$ = the mean of the differences

ΣD^2 **= *the sum*** of the squared difference scores

N = the number of pairs

Substituting the values from Table 6.3, we obtain

$$t = \sqrt{\dfrac{\dfrac{30}{15}}{\dfrac{164 - \dfrac{(30)^2}{15}}{15(15 \quad 1)}}} = \sqrt{\dfrac{2}{\underline{164\ 210 - 60}}} \quad \dfrac{2}{\sqrt{\dfrac{104}{210}}} - \dfrac{2}{\sqrt{0.4952}} - \dfrac{2}{0.704} = 2.8$$

Table 6.3 Before-and-After Scores of 15 Students in an Introduction to Research Class

(1) Subject Number	(2) Pretest	(3) Posttest	(4) D	(5) D^2
1	10	12	+2	+ 4
2	9	13	+4	+16
3	8	12	+4	+16
4	11	9	−2	+ 4
5	10	8	−2	+ 4
6	7	9	+2	+ 4
7	10	12	+2	+ 4
8	9	11	+2	+ 4
9	8	10	+2	+ 4
10	6	10	+4	+16
11	10	12	+2	+ 4
12	7	13	+6	+36
13	10	6	−4	+16
14	9	13	+4	+16
15	10	14	+4	+16
			$\Sigma D = +30$	$\Sigma D^2 = +164$

The t-ratio tells us that the observed difference is 2.84 times as great as the difference that would be expected under a true null hypothesis. We must now consult the Table of t-Values to determine the statistical significance of our observed ratio.

The number of degrees of freedom for the nonindependent t-test equals N − 1, N being the number of pairs of observations. In our example we have 15 − 1 = 14 degrees of freedom. In the Table of r-Values we find that with 14 degrees of freedom a r-value of 2.145 is needed for the t to be significant at the .05 level and a z-value of 2.917, for significance at the .01 level. Our obtained value of 2.84 exceeds the given value for the .05 level but does not reach the given value for the .01 level. This means that the difference between the two means is significant at the .05 level but not at the .01 level. If we had set our level of significance at .05, we could conclude that taking a research course does change the attitude of the students toward research under the conditions present in our study.

The Logic of the t-Test

The numerator of the f-test is the actual difference that has been observed between two groups. The denominator $(s_{\bar{X}_1-\bar{X}_2})$ is an estimate of how much these two groups would be expected to differ by chance alone; that is, it indicates the difference to be expected between two groups selected by a random procedure from a single parent population. This denominator is based on: (1) the number in the samples, $n_1 + n_2$ (the larger the number, the less random differences to be expected between sample means), and (2) the variation within the groups, s_1 and s_2 (the greater the variation *within groups*, the greater the random differences to be expected between groups). Since the denominator is a measure of how much apparent difference can be expected through chance alone, it is called the *error* term of the f-test.

If the ratio of observed difference (numerator) divided by error term (denominator) equals or exceeds the value indicated in the Table of t-Values, the null hypothesis can be rejected at the indicated level of significance.

Analysis of Variance

In *analysis of variance* (ANOVA), as in the f-test, a ratio of observed differences/error term is used to test hypotheses. This ratio, called the F-ratio, employs the variance (σ^2) of group means as a measure of observed differences among groups. This means that ANOVA is a more versatile technique than thet-test. At-test can be used only to test a difference between *two* means. ANOVA can test the difference between two *or more* means. Some statisticians never use the f-test, since ANOVA can be used in any situation where at-test can be used and, moreover, can do many things the f-test cannot do.

The general rationale of ANOVA is that the *total* variance of all subjects in an experiment can be analyzed into two sources, variance *between* groups and variance *within* groups.

Variance between groups is incorporated into the numerator in the F-ratio. Variance within is incorporated into the error term or denominator, as it is in the t-test. As variance between groups increases, the F-ratio increases. As variance within increases, the F-ratio decreases. The number of subjects influences the F-ratio; the larger the number, the larger the numerator becomes. When the numerator and denominator are equal, the differences between group means are no greater than would be expected by chance alone. If the numerator is greater than the denominator, one consults the Table of F-Values (2.4 in the Appendix) to determine whether the ratio is great enough to enable one to reject the null hypothesis at the predetermined level.

COMPUTATION **OF F-RATIO (SIMPLE ANALYSIS OF VARIANCE)**

Suppose we have the three experimental conditions of high stress, moderate stress, and no stress, and we wish to compare the performance of three groups of individuals, randomly assigned to these three conditions, in a simple problem-solving task. Assume that the data presented in Table 6.4 summarize our observations of the performance of these three groups and we are now to test the null hypothesis that there is no significant difference among these observations.

Table 6.4 Measures Obtained in Three Random Samples after Performance of a Task under Conditions of Moderate Stress, High Stress, and No Stress

Group 1 High Stress		Group 2 Moderate Stress		Group 3 No Stress	
X_1	X_1^2	X_2	X_2^2	X_3	X_3^2
19	361	22	484	15	225
18	324	20	400	14	196
17	289	19	361	14	196
16	256	18	324	13	169
15	225	17	289	13	169
15	225	16	256	12	144
14	196	16	256	12	144
13	169	15	225	11	121
12	144	14	196	11	121
II	121	12	144	10	100
150	2310	169	2935	125	1585
ΣX_1	ΣX_1^2	ΣX_2	ΣX_2^2	ΣX_3	ΣX_3^2

$\bar{X}_1 = 15.0$ $\bar{X}_2 = 16.9$ $\bar{X}_3 = 12.5$ $\bar{\bar{X}} = 14.8$

The means can be seen to differ from each other and from the overall mean for all 30 subjects. Are the differences among these means great enough to be statistically significant or is it likely that they occurred by chance? To answer this, we compute the F-ratio.

The first step is to find the sum of the squared deviation of each of the individual scores from the grand mean. This index is called the total sum of squares and is found by applying the formula

$$\Sigma x_t^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{N} \tag{6.8}$$

In our example this value is

$$\Sigma x_t^2 = 6830 - \frac{(444)^2}{30} = 258.8$$

Then we find the part of the total sum of squares that is due to the deviations of the group means from the grand mean. This index is called the sum of ***the squares between groups. (To*** be grammatically correct, we should say the sum of squares among groups when more than two groups are involved. However, it is a long-standing tradition to use the term sum *of squares between groups*, and in order to be consistent with other texts, we are retaining this usage here.) This index is found by applying the formula

$$\Sigma x_b{}^2 = \frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \cdots - \frac{(\Sigma X)^2}{N} \tag{6.9}$$

In our problem this value is

$$\Sigma x_b{}^2 = \frac{(150)^2}{10} + \frac{(169)^2}{10} + \frac{(125)^2}{10} - \frac{(444)^2}{30} = 97.4$$

Then we find the part of the total sum of squares that is due to the deviations of each individual score from its own group mean. This index is called the sum *of the squares **within groups*** and is found by applying the formula

$$\Sigma x_w{}^2 = \Sigma X_1{}^2 - \frac{(\Sigma X_1)^2}{n_1} + \Sigma X_2{}^2 - \frac{(\Sigma X_2)^2}{n_2} + \tag{6.10}$$

In our problem this value is

$$\Sigma x_w{}^2 = 2310 - \frac{(150)^2}{10} + 2935 - \frac{(169)^2}{10} + 1585 \frac{(125)^2}{10} = 161.4$$

The sum of the squares within groups could also be found by subtracting the sum of squares between groups from the total sum of the squares, that is,

$$\Sigma x_w{}^2 = \Sigma x_t{}^2 \ \Sigma x_b{}^2 \tag{6.11}$$

In our case

$$\Sigma x_w{}^2 = 258.8 - 97.4 = 161.4$$

The F-Test of **Significance**

Table 6.5 summarizes the results of our calculations so far, together with the results of further calculations. Column (1) of the table lists the three sources of variance: between-groups variance, within-groups variance, and total variance. Column (2) contains the sums of squares, which we have already calculated. Column (3) lists the number of degrees of freedom associated with each source of variance. The number of degrees of freedom for between-groups variance is equal to $(G-1)$, G being the number of groups. In our example this value is $3 - 1 = 2$. The degrees of freedom for within-groups variance is $n_1 - 1 + n_2 - 1 + \cdot$. In our example this value is $10 - I + 10 - 1 + 10 - 1 = 27$. The number of degrees of freedom for total variance equals N − 1; in our example 30 − l = 29. This last value could also be obtained by adding the between-groups and within-groups degrees of freedom.

Table 6.5 Summary **of the Analysis of Variance of the Three Groups**

(1) **Source of** Variance	(2) *SS*	(3) *df*	(4) *MS*	(5) *F*	(6) Level **of** Significance
Between groups	91.4	2	48.70	8.14	0.01
Within groups	**161.4**	**27**	**5.98**		
Total	**258.8**	29			

The next step, then, is to find the *between-groups mean square* and the *within-groups mean square.* These values are obtained by dividing the between-groups and within-groups sums of squares by their respective degrees of freedom. The resulting values are the mean squares. In our example the mean square between groups is 97.412 = 48.7 and the mean square within groups is 161.4127 = 5.98. The mean square within groups is the error term for our F-ratio. By applying the following formula, we finally arrive at the end product of the analysis-of-variance procedure, the F-ratio:

$$F = \frac{MS_b}{MS_w} = \frac{SS_b/df_b}{SS_w/df_w} \tag{6.12}$$

In our example this value is

$$F = \frac{48.70}{5.98} = 8.14$$

We now consult Table A.4 in the Appendix to determine whether the F-ratio we have obtained is statistically significant. We find the column headed by the between-groups (numerator) degrees of freedom of our experiment and go down this column to the row entry corresponding to the number of our within-groups (denominator) degrees of freedom. At this point in the column we find two values, one in roman type and one in boldface type. If our F-ratio is equal to or greater than the value given in lightface, our F-ratio is significant at the .05 level. If our obtained F-ratio is equal to or greater than the value given in boldface, it is also significant at the .01 level. In our example, with 2 and 27 degrees of freedom, we need an F-ratio of 3.35 to reject the null hypothesis at the .05 level and an F-ratio of 5.49 to reject the null hypothesis at the .01 level. Since our obtained F-ratio is greater than both of these values, it is significant at the .01 level and the null hypothesis is rejected at that level.

The assumption underlying the analysis-of-variance procedure is that if the groups to he compared are truly random samples from the same population, then the between-groups mean square should not differ from the within-groups mean square by more than the amount we would expect from chance alone. Thus under a true null hypothesis we would expect the F-ratio to be approximately equal to one. On the other hand, if the null hypothesis is false, the difference among group means should be greater than what is expected by chance, so the mean square between would exceed the mean square within. In such cases the F-ratio, the mean square between divided by the mean square within, will have a value greater than one. We then consult the Table of *F*-Values to determine whether the ratio for our data is sufficiently greater than 1 .O to enable Us to reject the null hypothesis at our predetermined level. As the difference between these mean squares increases, the F-ratio increases and the probability of the null hypothesis being correct decreases.

When the null hypothesis is rejected as a result of this analysis-of-variance procedure, we cannot say more than that the measures obtained from the groups involved differ and the differences are greater than one would expect to exist by chance alone.

A significant F-ratio does not necessarily mean that all groups differ significantly from all other groups. The significant *F may* be a result of a difference existing between one group and the rest of the groups. For instance, in our problem it might be that Group 3 is significantly different from Group 1 and Group 2, but Groups 1 and 2 do not differ significantly from each other. There are several statistical tests that can be applied to find the location of significant differences. Those developed by Tukey and by Scheffé are particularly useful.[4]

In our example we selected our three groups randomly from the same population and thus we can assume that they did not differ beyond the chance expectation prior to our experimental treatments. The significance of the&ratio indicates that the differences found between these groups&r treatment are beyond chance expectation. We attribute this to our experimental treatment and conclude that the level of stress affects the performance of individuals in simple problem-solving tasks. This is as far as we can go in our interpretation of this F-ratio. If we need further statistical **analysis, we can use** Tukey's, Scheffé's, or other tests to determine the significance between pairs of individual measures. These techniques can tell us how specific stress conditions affect the performance and can answer such questions as, Is there any difference in performance scores under conditions of moderate and high stress? moderate and no stress? and high and no stress?

MULTIFACTOR ANALYSIS OF VARIANCE

We may wish to investigate the combined effect of stress level and task difficulty on performance in a problem-solving task. To investigate this problem we will vary both the level of stress and the difficulty of the task. The layout for an experiment

[4] See Gene V. Glass and Julian C. Stanley, *Statistical Methods in Education and Psychology* (Englewood Cliffs, N.J.: Prentice-Hall, 1970).

investigating the combined effects of two or more independent variables is called a factorial *design* and the results are analyzed by means of a *multifactor* analysis of *variance.*

Let us assume that we have carried out this experiment using five subjects in each group and that the data shown in Table 6.6 represent a summary of our observations of the performance of the subjects. Applying multifactor analysis of variance will enable us to find (1) whether there is a significant difference between the performance of the subjects under a moderate stress condition and under a high stress condition, (2) whether there is a significant difference between the performance of the subjects given an easy problem-solving task and those given a difficult task, and (3) whether or not the two variables, stress and task difficulty, have a combined effect on the performance of the subjects. The effects investigated by the first and second analyses are called *main effects,* whereas the third is referred to as the *interaction effect.* The end products of these analyses will be three F-ratios, two of which indicate the significance of the two main effects and the third, that of the interaction effect.

Table 6.6 Measures on Two Levels of Problem-Solving Tasks under Moderate and **High Conditions of Stress**

		Moderate	High	
	Simple	20 20 Group 1 19 19 $\bar{X} = 19$ 17 ΣX 95	23 22 Group 3 21 20 $\bar{X} = 21$ 19 ΣX 105	$\Sigma X_{r_1} = 200$ $\bar{X}_{r_1} = 20.0$
Task				
	Difficult	22 21 Group 2 20 19 $\bar{X} = 20$ 18 ΣX 100	18 16 Group 4 15 14 $\bar{X} = 15$ 12 ΣX 75	$\Sigma X_{r_2} = 175$ $\bar{X}_{r_2} = 17.5$
		$\Sigma X_{c_1} = 195$ $X_{c_1} = 19.5$	$\Sigma X_{c_2} = 180$ $\bar{X}_{c_2} = 18.0$	ΣX Total = 375 $\bar{\bar{X}}$ (Grand mean) = IX.75

The computation of these F-ratios involves the following steps:

1. Find the total sum of squares, the sum of squares between groups, and the sum of squares within groups using the same procedures and formulas applied in simple analysis of variance. These values, derived from the data in Table 6.6. are

$$\Sigma x_t^2 = 7181 - \frac{(375)^2}{20} = 149.75$$

$$\Sigma x_b^2 = \frac{(95)^2}{5} + \frac{(105)^2}{5} + \frac{(100)^2}{5} + \frac{(75)^2}{5} - \frac{(375)^2}{20} = 103.75$$

$$\Sigma x_w^2 = 149.75 \quad 103.75 = 46.00$$

2. Break down the sum of the squares between groups into three separate sums of squares: (a) the sum of squares between columns, (*b*) the sum of squares between rows, and (*c*) the sum of squares for interaction between columns and rows:

a. *The between-columns sum of squares* represents the sum of the squared deviations due to the difference between the column means and the grand mean. It is found by using the formula

$$\Sigma x_{bc}^2 = \frac{(\Sigma X_{c_1})^2}{n_{c_1}} + \frac{(\Sigma X_{c_2})^2}{n_{c_2}} + \cdots - \frac{(\Sigma X)^2}{N} \qquad (6.13)$$

Using this formula, the sum of squares between the columns for the data shown in Table 6.7 is

$$\Sigma x_{bc}^2 = \frac{(195)^2}{10} + \frac{(180)^2}{10} - \frac{(375)^2}{20} = 11.25$$

b. The between-rows sum of squares is the sum of the squared deviations due to the difference between the row means and the grand mean. It is found by applying the formula

$$\Sigma x_{br}^2 = \frac{(\Sigma X_{r_1})^2}{n_{r_1}} + \frac{(\Sigma X_{r_2})^2}{n_{r_2}} + \cdots - \frac{(\Sigma X)^2}{N} \qquad (6.14)$$

For the data presented in Table 6.6 this value is

$$\Sigma x_{br}^2 = \frac{(200)^2}{10} + \frac{(175)^2}{10} - \frac{(375)^2}{20} = 31.25$$

c. The sum-of-squares interaction is the part of the deviation between the group means and the overall mean that is due neither to row differences nor to column differences. In other words, this is the difference between the total of the sum of squares between groups and the sum of squares between rows, that is,

$$\Sigma x_{int}^2 = \Sigma x_b^2 - (\Sigma x_{bc}^2 + \Sigma x_{br}^2) \qquad (6.15)$$

Expressed in words, the interaction sum of squares is equal to the between-groups sum of squares minus the sum of the between-columns sum of squares and the between-rows sum of squares.

For the data presented in Table 6.6, this interaction value is

$$\Sigma x^2_{int} = 103.75 - (11.25 + 31.25) = 61.25$$

3. Determine the number of degrees of freedom associated with each source of variation. They are found as follows:

df for between-columns sum of squares = $C - I$
df for between-rows sum of squares = $R - 1$
df for interaction = $(C - 1)(R - 1)$
df for between-groups sum of squares = $(G - 1)$
df for within-groups sum of squares = $\Sigma(n - 1)$
df for total sum of squares = $N - 1$

where
C = the number of columns
R = the number of rows
G = the number of groups
n = the number of subjects in one group
N = the number of subjects in all groups

4. Find the mean-square values by dividing each sum of squares by its associated number of degrees of freedom.

5. Compute the F-ratios for the main and the interaction effects by dividing the between-groups mean squares by the within-groups mean square for each of the three components.

Table 6.1 Summary of a 2 × 2 Multifactor Analysis of Variance

Source of Variance	*SS*	***df***	*MS*	*F*	Level of **Significance**
Between columns (stress)	11.25	1	11.25	3.913	
Between rows (task)	31.25	1	31.25	10.869	.01
Columns by rows (interaction)	61.25	1	61.25	21.304	.01
Between groups	103.75	3	34.583		
Within groups	46.00	16	2.875		
Total	149.75	19			

6. The results of the calculations based on the data presented in Table 6.6 are summarized in Table 6.7. Three F-ratios are listed in this table. To find the significance of each of these values we consult the Table of F-Values as before. To

enter this table we use the number of degrees of freedom associated with each F-ratio ***(df*** for the numerator) and the number of degrees of freedom associated with the within-groups mean square ***(df*** for the denominator). For example, our between-columns F-ratio is 3.913. Consulting the table, we see that, with 1 and 16 degrees of freedom, an F-ratio of 4.49 or more is needed for significance at the .05 level. Since our F-ratio is smaller than the value shown in the table, it is not significant.

To be significant, the F-ratio for between rows, with 1 and 16 degrees of freedom, should reach 4.49 (.05 level) or 8.53 (.01 level). Since our obtained value of *F,* 10.869, exceeds both of these values, it is significant at the .01 level.

For the interaction between columns and rows, with 1 and 16 degrees of freedom, an F-ratio of 4.49 (.05 level) or 8.53 (.01 level) is needed. Our obtained value of ***F,*** 21.304, exceeds both of these values and thus is significant at the .01 level.

Interpretation of the F-ratios

The first F-ratio (between columns) in Table 6.7 is not significant and shows that the stress conditions do not differ significantly from one another in their effect on the performance of the subjects in the experiment. This analysis is a comparison of the combined performance of Groups 1 and 2 with the combined performance of Groups 3 and 4. We could have arrived at the same conclusion by using the f-test procedure.

The second F-ratio (between rows), which is significant at the .01 level, is based on the comparison of the performance of the subjects in Groups 1 and 3 with those in Groups 2 and 4. From the significance of this F-ratio we can infer that the difference between the performance of those subjects given an easy problem-solving task and those given a difficult problem-solving task is beyond chance expectation. Examining the data presented in Table 6.7 we see that those groups who performed simple problem-solving tasks have obtained a combined mean of 20 as compared with a mean of 17.5 for those groups who performed difficult tasks. Since we have a significant F-ratio for the difference, we conclude that under conditions similar to those of our experiment, a higher level of task performance can be expected when the task is simple than when it is difficult.

The third F-ratio shows the interaction effect between the two variables, stress level and the degree of task difficulty. The significance of the F-ratio in this case means that the effect of stress level on performance in a problem-solving task depends on the degree of difficulty of the task. We can see this phenomenon more clearly if we compare the observed results with the results that would be expected if there had been no interaction between the two independent variables.

Let us calculate what we would expect the means of the four groups to be if there had been no interaction. The mean for all subjects is 18.75. The mean for the ten subjects under moderate stress, 19.5, is .075 greater than this figure, whereas the mean of the ten subjects under high stress is 0.75 less. The mean for the ten subjects doing the simple task, 20.0, is 1.25 greater than the mean for all subjects, whereas the mean for the ten subjects doing the difficult task is 1.25 less.

For each group we can calculate the mean that would be expected for this group if there had been no interaction. We do this by adding to the grand mean the difference for the column that group is in and the difference for the row that group is in. If there had been no interaction, what would we expect the mean of Group 1 to be? Beginning with the total mean, 18.75, we would add 0.75 because the subjects were under moderate stress and another 1.25 because they did an easy task. This gives a total of 20.75.

Following this procedure for each of the four groups, we would have the following expected values:

	Overall Mean	+	Stress Difference	+	Task Difference	=	Expected Value
Group 1	18.75		+0.75		+1.25		20.75
Group 2	18.75		+0.75		− 1.25		18.25
Group 3	18.75		-0.75		+ 1.25		19.25
Group 4	18.75		−0.75		− 1.25		16.75

Now compare the actual group means with these expected group means:

	Actual				*Expected*		
	Moderate	High			Moderate	High	
Simple	Group 1 $\bar{X}$ = 19	Group 3 $\bar{X}$ = 21	$\bar{X}$ = 20	Simple	Group 1 $\bar{X}$ = 20.75	Group 3 $\bar{X}$ = 19.25	$\bar{X}$ = 20
Difficult	Group 2 $\bar{X}$ = 20	Group 4 $\bar{X}$ = 15	$\bar{X}$ = 17.5	Difficult	Group 2 $\bar{X}$ = 18.25	Group 4 $\bar{X}$ = 16.75	$\bar{X}$ = 17.5
	$\bar{X}$ = 19.5	$\bar{X}$ = 18.0	$\bar{X}$ = 18.75		$\bar{X}$ = 19.5	$\bar{X}$ = 18.0	$\bar{X}$ = 18.75

(Note that we could use the differences between expected and observed values to compute the sum of squares for interaction directly. Each group differs from its expected mean by 1.75. Square this value and multiply by the number of cases to get $1.75^2 \times 20 = 61.25$.)

We see that Group 1 actually did less well than we would expect, knowing they were under moderate stress and doing a simple task. Group 2, doing a difficult task under moderate stress, did better than we would expect. Considering the groups under high stress, we find that Group 3, with the simple task, did better than expected, whereas Group 4, with the difficult task, did less well than expected. Since our F-test indicated that the interaction was significant, we conclude that moderate stress produces higher scores when combined with a difficult task than with a simple task, whereas high stress produces higher scores when combined with a simple task than when combined with a difficult task.

The use of multifactor analysis has been of great value in educational research since many of the questions that educators need to investigate are inherently complex in nature. These techniques enable us to analyze the combined effects of two or more independent variables in relation to a dependent variable. For example, a simple comparison of the dependent variable means of two groups of pupils taught by different methods might yield insignificant results. But if intelligence is incorporated into the experiment as a measured independent variable, we might find that one method works better with the less intelligent pupils while the other works better with the more intelligent pupils.

Multifactor analysis of variance is not limited to two independent variables as in our example. Any number of independent variables may be incorporated in this technique. Several intermediate statistics books, including Edwards',[5] explain the computation and interpretation of these procedures.

The Chi-Square Test of Significance

Sometimes we need to find the significance of differences among the *proportions* of subjects, objects, events, and so forth, that fall into different categories. A statistical test used in such cases is called the chi-square (χ^2) test.

In the chi-square test two sets of frequencies are compared: *observed frequencies* and *expected frequencies*. Observed frequencies, as the name implies, are the actual frequencies obtained by observation. Expected frequencies are theoretical frequencies, which are used for comparison.

Consider the hypothesis that the proportion of male to female students in statistics courses is different from that of male to female students in a schoool of education as a whole. If we know that 40 percent of the total enrollment in the school is male and that 300 students are enrolled in statistics courses, our expected frequencies will be

$$\left.\begin{array}{ll}\text{Male students} & 120\\ \text{Female students} & 180\end{array}\right\}300$$

Now suppose that our observed frequencies are found to be

$$\left.\begin{array}{ll}\text{Male students} & 140\\ \text{Female students} & 160\end{array}\right\}300$$

[5] Allen L. Edwards, *Experimental Designs in Psychological Research,* 3rd ed. (New York: Holt, Rinehart and Winston, 1968), chs. 11 and 12.

We want to determine whether the difference between our expected and observed frequencies is statistically significant. To determine this we apply the chi-square formula, which is

$$\chi^2 = \sum\left[\frac{(f_o - f_e)^2}{f_e}\right] \tag{6.16}$$

where
χ^2 = the value of chi-square
f_o = the observed frequency in each cell
f_e = the expected frequency in each cell

Applying this formula to our data, we obtain

$$\chi^2 = \frac{(140 - 120)^2}{120} + \frac{(160 - 180)^2}{180} = 5.55$$

To determine whether this chi-square value is significant we consult the table of χ^2 values in the Appendix (AS). The first column in this table shows the number of degrees of freedom involved in any given chi-square problem. The remaining columns present the values needed for different levels of significance. The number of degrees of freedom, as we have discussed previously, is based on the number of observations that are free to vary once certain restrictions are placed upon the data. When we have a fixed number of observations divided into only two categories, as soon as the number falling into one category has been determined, the other is fixed. Thus, when we find that the number of male students is 140, the number of female students in the total of 300 must be 160. In this example there is only one degree of freedom. In problems such as this the number of degrees of freedom equals $\boldsymbol{K} - 1$, where $\boldsymbol{K}$ is the number of categories used for classification. By consulting the Table of χ^2 we find that our observed value of 5.55 is statistically significant at the .05 level.

Interpreting this result we can now state that the proportion of males who take statistics courses within our school is significantly greater than that of females at the .05 level of confidence. The significance level of .05 means that there are less than five chances in a hundred of observing such a difference between the proportions of male and female students through chance alone. Thus the data lend support to our research hypothesis that the proportion of male students who tend to take statistics courses is greater than that of female students.

The use of the chi-square test is not limited to situations in which there are only two categories of classification; this test can also be used to test a null hypothesis that there is no significant difference between the proportion of the subjects falling into any number of different categories. Suppose, for example, we have asked a sample of 120 undergraduate students to indicate whether they prefer to live in a dormitory or in town, or whether they have no preference, with the results shown in Table 6.8.

Table 6.8 The Observed Frequencies of Responses of 120 Undergraduate Students as to Their Preferences with Respect to Living Accommodations

Subject	Dormitory	Town	No Preference	Total
Undergraduate students	40	50	30	120

If there were no difference between the three categories of response, we would have 40 responses in each category. These would be our expected frequencies, as shown in Table 6.9.

Table 6.9 The Expected Frequencies **of Responses of 120 Undergraduate Students as to Their Preferences with Respect to Living Accommodations**

Subiect	Dormitorv	Town	No Preference	Total
Undergraduate students	40	40	40	120

A comparison of the two sets of frequencies presented in Tables 6.8 and 6.9 shows that there are differences between our expected and observed data. To find whether or not they are significant, we apply the chi-square test. The value of χ^2 for these data, using formula (6. 16), would be

$$\chi^2 = \frac{(40-40)^2}{40} + \frac{(50-40)^2}{40} + \frac{(30-40)^2}{40} = 5.00$$

The degrees of freedom, again, equal the number of categories minus one ($K -$ I) or, in this case, $3 - 1 = 2$. Referring to the Table of χ^2 we see that with two degrees of freedom a X^2 value of 5.991 or greater is required for significance at the .05 level. However, our obtained χ^2 value is smaller than this value and therefore is not statistically significant. This means that the observed differences between categories could easily have happened by chance. Consequently, the null hypothesis that there is no significant difference between the frequencies of the three categories, cannot be rejected. In other words, if the proportions of preferences for the three categories in the entire undergraduate population were equal, we would expect to observe sample differences as great as those in our sample more often than five times in a hundred through chance.

THE CHI-SQUARE TEST OF INDEPENDENCE

So far we have only considered examples in which observations were classified along a single dimension. Sometimes, however, we wish to use more than one dimension for classification. Suppose, for example, we add another dimension to the previous problem and ask both graduate and undergraduate students to state their preferences as to their living accommodations. Assume the frequencies as shown in Table 6.10 were the result.

Table 6.10 The Observed Frequencies of Responses of 200 Undergraduate and Graduate Students as to Their Preferences with Respect to Living Accommodations

Subjects	Dormitory	Town	No Preference	Total
Undergraduate students	40	50	30	120
Graduate students	20	40	20	80
Total	**60**	**90**	**50**	**200**

In this case our null hypothesis might be that the preference for living accommodations is the same for graduates as it is for undergraduates-that is, the variables student status and preference for living accommodations are unrelated. Our observations show that 30 percent of all students prefer dormitories, 45 percent prefer town, and 25 percent state no preference. If the null hypothesis is true, we would expect to find identical proportions among both graduates and undergraduates, as shown in Table 6.11. We can compute expected cell frequencies by multiplying the row frequency associated with a cell by the column frequency associated with that cell, then dividing this product by the grand total. For example, the expected

Table 6.11 The Expected Frequencies of Responses of 200 Undergraduate and Graduate Students as to Their Preferences with Respect to Living Accommodations

Subjects	Dormitory	Town	No Preference	Total
Undergraduate **students**	**36**	**54**	**30**	120
Graduate students	**24**	**36**	**20**	**80**
Total	**60**	**90**	**50**	**200**

frequency of response for undergraduate students who want to live in a dormitory is 120 × 60 ÷ 200 = 36, for those undergraduate students who prefer to live in town it is 120 × 90 ÷ 200 = 54, and for graduate students who want to live in a dormitory it is 80 × 60 ÷ 200 = 24. Using this approach, we find the expected frequencies for each cell.

Note that all the row and column totals in Table 6.11 are exactly the same as those shown in Table 6.10. We now ask if the observed frequencies differ enough from the expected frequencies to enable us to reject the likelihood that these differences could have occurred merely by chance. Applying the formula, we obtain

$$\chi^2 = \frac{(40-36)^2}{36} + \frac{(50-54)^2}{54} + \frac{(30-30)^2}{30} + \frac{(20-24)^2}{24} + \frac{(40-36)^2}{36} + \frac{(20-20)^2}{20}$$

$$\chi^2 = 1.8518$$

The number of degrees of freedom for a two-way table is found by applying the formula

$$df = (C-1)(R-1) \qquad (6.17)$$

where
df = the number of degrees of freedom
C = the number of columns
R = the number of rows

Applying this formula to the problem under consideration, we obtain

$$df = (3-1)(2-1) = 2$$

Referring to Table A.5 we see that with two degrees of freedom a χ^2 value of 5.991 is needed for significance at the .05 level. But our obtained χ^2 value of 1.85 18 is smaller than this table value and is therefore not significant. This means that the differences between expected and observed frequencies are not beyond what would be expected by chance. In other words, we do not have reliable evidence that there is a relationship between the variables student status and living accommodation preference in the population from which our sample was drawn.

Summary

Investigators hope to form generalizations about populations by studying groups of individuals selected from the populations. These generalizations will be sound only if the selected groups-the samples-used in these studies are representative of the larger groups-the populations-from which they arc chosen.

A sample is random if all the members of a population have an equal chance of being included within that sample. It is the preferred means of subject selection for behavioral research.

Sometimes it is important for the purpose of a specific study to choose independent samples from different subgroups or strata of a population and to obtain separate measures for each stratum. This is stratified sampling.

When the target population is unwieldy, an investigator may choose randomly a number of groups for study rather than individual subjects. This is called cluster sampling. A cluster sample is more subject to sampling errors than is simple random sample.

Inferential statistics provide tools by means of which researchers are able to estimate how confident they can be in inferring that phenomena observed in samples would also be observed in the populations from which the samples were drawn. In other words, inferential statistics enable us to estimate how reliable our observations may be.

A basic strategy in inferential statistics is to compute the extent of difference among observations that would be likely to arise by chance alone. The result of this computation is often called the error term. Then the observed differences among observations are compared with the error term. If the observed differences are similar to the differences that could arise by chance, the researcher cannot reject the likelihood that the observed differences were merely a function of chance. If the observed differences are greater than the error term, the researcher consults the tabled values of the statistic to determine whether the ratio of observation to error is great enough to reject the chance explanation at a predetermined level of confidence.

The indices most commonly used in inferential statistics are: the t-test, analysis of variance, and the chi-square test of significance. The f-test is used to find whether the difference between two sample means is statistically significant. There are two types of f-tests: (1) the t-test for independent groups, which is used to compare two sample means when the samples have been drawn independently from a population and (2) the f-test for nonindependent groups, which is employed with two samples in which the subjects are matched or with two repeated measures obtained from the same subjects.

Analysis of variance is used to compare the means of two or more samples and to test the null hypothesis that no significant differences exist between the means obtained from these samples. Multifactor analysis of variance enables us to test the effect of more than one independent variable and also the interaction effect of such variables.

The chi-square statistic is an index employed to find the significance of differences between proportions of subjects, objects, events, and so forth, that fall into different categories, by comparing observed frequencies and expected frequencies.

Exercises

1. Does the accuracy of a sample in representing the characteristics of the population from which it was drawn always increase with the size of the sample? Explain.

2. You have been asked to determine whether teachers in the Central School District favor the "year around school" concept. Because the district is rather large you are asked to contact only 500 teachers. Determine the number you would choose from each of the following levels to draw a proportioned stratified random sample:

Level	*Total Number*
Elementary	3,500
Middle School	2,100
High School	1,400
Total	7,000

3. You are asked to conduct an opinion survey on a college campus with a population of 15,000 students. How would you proceed to draw a representative sample of these students for your survey?
4. A national magazine has one million subscribers. The editorial staff wants to know which aspects of the magazine are liked and which are not. The staff decides that a personal interview is the best method to obtain the information. For practical and economic reasons only 500 people in five cities will be surveyed. In this situation, identify:
   a. the target population
   b. the accessible population
   c. the sample
5. Investigators wish to study the question, Do blondes have more fun?
   a. What is the null hypothesis in this question?
   b. What would be a Type I error in this case?
   c. What would be a Type II error in this case?
   d. If one investigator uses an .05 level of significance in investigating this question and another investigator uses an .001 level of significance, which would be more likely to make a Type I error?
   e. If one investigator uses an .05 level of significance in investigating this question and another investigator uses an .001 level of significance, which would be more likely to make a Type II error?
6. Inferential statistics enable researchers to:
   a. reach infallible conclusions
   b. reach reasonable conclusions with incomplete information
   c. add an aura of legitimacy to what is really sheer guesswork
7. What two conditions are necessary for a Type I error to occur?
8. Which of the following statements describes the role of the null hypothesis in research?
   a. It enables us to determine the probability of an event occurring through chance alone when there is no real relationship between variables.
   b. It enables us to prove there is a real relationship between variables.
   c. It enables us to prove there is no real relationship between variables.
9. A Type II error occurs when one:
   a. rejects a false null hypothesis
   b. rejects a true null hypothesis
   c. has already made a Type I error

d. retains a false null hypothesis
e. retains a true null hypothesis

10. The phrase level *of significance* refers to
a. the probability of an event being due to chance alone, which is calculated after the data from an experiment are analyzed
b. the probability of a Type I error that an investigator is willing to accept
c. the actual probability of a Type II error
d. the probability of a Type II error that an investigator is willing to accept

11. How does one determine the level of significance to use in an experiment?

12. A cigarette manufacturer has employed researchers to compare the rate of occurrence of lung cancer among smokers and nonsmokers. Considering the results of previous research on this question, the manufacturer would probably urge the researchers to be especially careful to avoid making a
a. Type I error
b. Type II error

13. What is indicated when the results of a study are not statistically significant?

14. You have a list of pupils in a high school who have been assigned the number 1 to 1,000. Use the table of random numbers in the Appendix to select a sample of 50 from the hypothetical list. List the numbers selected for the sample.

Answers

1. A larger ***randomly*** drawn sample is more likely to be representative of the population than is a smaller *random* sample. A large sample obtained with a method that permits systematic bias will not be any more representative than a small biased sample.

2. To obtain a proportional stratified sample, divide the 500 teachers in proportion to their representation in the population, as follows:

Elementary	$\frac{3500}{7000} \times 500 = 250$
Middle School	$\frac{2100}{7000} \times 500 = 150$
High School	$\frac{1400}{7000} \times 500 = 100$
Total Sample	500

3. Number a list of all students, then select a random sample of a given number by using a table of random numbers. Starting at a random point in the table, go up or down the column and include those students whose numbers are listed.

4. a. all subscribers to the magazine
b. the subscribers in the five cities
c. 500 individuals who are interviewed.

5. a. There is no relationship between hair color and fun.
b. The investigators make a Type I error if they declare that blondes have more fun than nonblondes or that blondes have less fun than nonblondes, when in fact the two groups have an equal amount of fun.
c. The investigators make a Type II error if they fail to conclude that blondes have more fun or less fun, when in fact they do.

d. the investigator with the .05 level of significance.
e. the investigator with the .001 level of significance.

6. b

I. The null hypothesis must be true and the investigator must reject it.

8. a
9. d
10. b
11. by weighing the consequences of Type I and Type II errors
12. a
13. The results could easily be a function of chance; the evidence is insufficient to justify a conclusion.