CHAPTER 6

DATA ANALYSIS: COMPARING GROUPS AND MAKING INFERENCES

Chapter 3 was on the descriptive analysis of data and Chapter 4 was on the different types of graphics illustrating them. Chapter 5 was on the computing relationship between variables focused in the study. This chapter will be dealing with the comparison of different groups for the purpose of making inferences. Therefore, statistical procedures applied for this purpose are inferential. Inferential statistics is used to reach conclusions that extend beyond the immediate data alone. A sound conclusion cannot be arrived at unless the findings obtained from different groups involved in the research are compared and contrasted, and their significance is tested to infer the generalizations for a larger population than the sample used in the study. In other words, inferences are drawn from the findings of the sample data to apply to the population.

Issues to be Considered in Inferential Statistics

There are certain concepts that need to be clarified to make the inferential statistics more comprehensible. The most common ones are (a) probability and the level of significance, (b) standard error of the mean, (c) Type I and Type II errors, (d) one-tailed and two-tailed tests, and (e) parametric and nonparametric tests.

Probability and the Level of Significance

Inferential statistics is also used to make judgments about the *probability* of whether an observed difference between groups is a dependable one or one that might have happened by chance in that particular study. When formulating a null hypothesis, the researcher assumes a difference between groups and expects the results obtained from the study to reject that null hypothesis so that the researcher's hypothesis would be accepted. In Moore's terms, "the decision to reject or accept a null hypothesis is based on a value of probability called the *level of significance*."

The reason for calculating the probability level of the chance is due to the fact that in a piece of research, it is not always possible to reach the whole population; therefore, a sample is randomly chosen to conduct the study. The technique used in random sampling may cause sampling errors. As a result of sampling errors, subjects chosen for the study may not represent the population. Thus, an error made in the sampling may reflect on the results; consequently, the obtained results would not reflect reality. Therefore, it is important to indicate the significance of the results. In other words, when the finding is stated to be statistically significant, the obtained results do not seem to have occurred by chance. Therefore, the chance of error is operationally defined as the probability level. In order to be able to report the statistical significance of the obtained results, the probability level should be calculated. "Chance is operationally defined by some alpha (ax) level" (Newman & Newman, 1994, p. 61). In psychology and education, the alpha levels generally used are .05, .01 and .001. If the relationship is calculated at the .05 level, this indicates that "the relationship is only likely to occur by chance 5 times out of 100" (Newman & Newman, 1994, p. 61).

The probabilities of the events are computed with the assumption that if events occurred within a certain frequency in the past, it will occur within the same frequency in the future. Thus, the probability of an event is the relative frequency, which is computed by dividing the number of occurrence (f) by the size of the sample (N). For example, if out of 100 products manufactured within the boundaries of a limited time 30 had defects, the relative frequency of the defects of these products is calculated as 30 percent.

Thus, inferential statistics is used to determine to what degree the results obtained from a sample can be attributed to the entire population. Since the generalizations are based on inferences, they are only probability statements. This is because there might be errors in sampling. For example, when a difference is found between the means of the sample groups, it needs to be questioned as to whether the difference is a result of sampling error, or whether it really reflects a true difference between the two samples.

Newman & Newman (1994), in reference to research design mention two methods. The first method focuses on "the conclusions that can be made from the analysis of the design, the limitations, and the population to whom the results can or cannot be generalized" (pp. 94-95). The second method focuses on the statistical procedures for the analysis of the design. Thus, the rest of the chapter will focus on the procedures applied in the second method.

Standard Error of the Mean

If we apply a standard error of the mean, we can estimate how much the sample means would differ from other samples if such sampling from the same population were arranged. According to the standard distribution, we say that 68 percent of the sample means fall between $\pm$ 1 of the standard error (which is simply the standard deviation (SD)).

If we do not know the standard deviation of the whole population, we can estimate the standard error of the mean "by dividing the standard deviation of our sample group by the square root of our sample size " (Hatch & Lazaraton, 1991, p. 253). Since the standard deviation for the whole population is generally unknown, the estimation of standard error is done by dividing the standard deviation of the sample by the square root of the sample minus one:

$$SE_X = \frac{SD}{\sqrt{N - 1}}$$

When the obtained standard error of the mean is small, it indicates less sampling error. Usually, as the size of the sample increases, the standard error of the mean decreases. Therefore, the larger the number of subjects in a sample, the more reliable the results are. It is still not possible to attribute the difference calculated between the experimental and control groups to the independent variable (e.g. treatment) as stated by the null hypothesis. Thus, a test of significance can be applied to test the validity of the null hypothesis for that particular study.

Type I Error and Type II Error

At the beginning of the research, the researcher may formulate a null hypothesis, which begins with an assumption that there is no relationship between groups (see Chapter 3). If a significant difference is calculated as a result of the findings, the researcher rejects the null

hypothesis. In rejecting the hypothesis, the researcher would not, however, know for sure whether the differences found are actual differences, or whether they have resulted from some kind of error made during the research, or the cause of these differences was a sheer chance. There is a probability of the occurrence of error even when the difference is calculated at a significant level of .05. Therefore, statisticians talk about two types of error in testing hypotheses: Type I error and Type II error.

Statisticians talk about Type I error if there is a likelihood of rejecting the null hypothesis although the statement in the null hypothesis is true. Type II error is just the opposite of the case in Type I error. In other words, Type II error is said to have occurred if the null hypothesis is accepted when it should be rejected.

The probability of making Type I error is inversely proportionate to the probability of making a Type II error. In other words, "as the probability of making a Type I error increases, the probability of making a Type II error decreases; and as the probability of making a Type I error decreases, the probability of making a Type II error increases" (Newman & Newman, 1994, p. 62). For instance, holding the sample size constant, the probability of making a Type I error can be decreased by accepting the significance level at .01; in this case, the probability of making a Type II error increases. When a Type II error increases, the researcher tends not to reject the null hypothesis when he/she should.

If the researcher has a powerful and reliable instrument to implement in collecting data, and if the study is conducted on a fairly large sample, it is less likely to make Type II error or Type I error. This is expressed by the term *power*, which is defined as "the probability of making a correct decision to reject a null hypothesis when differences do exist" (Moore, 1983, p. 273).

One-Tailed and Two-Tailed Tests

Before computing a test of significance on the findings derived from the sample, the researcher has to decide whether to apply a one-tailed or two-tailed test. This depends on the way the hypothesis is formulated. Let us discuss this issue on the following hypotheses:

Hypothesis I: It is hypothesized that girls' rate of acquisition of their native language will be significantly higher than the boys'.

Hypothesis II: It is hypothesized that there is a significant difference between the way boys (Group I) and girls (Group II) acquire their native language.

For instance, if Hypothesis I is adopted for the study, a one-tailed test is used because the hypothesis comprises a statement in favor of one group (girls). In other words, a prediction has been made as to which group will prove to be superior regarding language acquisition. In such a case, we see that in this hypothesis the direction of the prediction has been indicated. Therefore, it is a directional hypothesis and requires a one-tailed test. For this very reason, a one-tailed test is also called a directional test.

In Hypothesis II, however, the direction of the prediction has not been made in favor of any of the groups. Therefore, this is a non-directional hypothesis and thus requires a two-tailed test (a non-directional test).

A one-tailed test as a requirement of a directional hypothesis is considered to be more powerful because a significant relationship or difference is expected to be observed in the direction stated in the hypothesis. In some cases, however, the results of the test indicate a significant difference in the opposite direction. Under these circumstances, "the study needs to be replicated on a new sample … " (Newman & Newman, 1994, p. 64). In other words, in the

replication of the study, the researcher cannot use the same data, and the direction of the hypothesis needs to be considered well in advance. If a decision cannot be made as in which direction the hypothesis should be formulated, then, the researcher is suggested to use a two-tailed test.

Parametric and Nonparametric Tests

Parametric tests enable the researcher to compare the means of the groups concerning the study. These means are obtained from continued data formulated in ordinal scale illustrated in interval measurement (see Chapter 3). ANOVA is the most powerful parametric test.

When the linearity of the scale is known for sure, or when the distribution of data does not fall in even intervals on the scale, a nonparametric test is applied. Hatch & Lazaraton (1991, p. 332) suggest the use of Kruskal-Wallis Test for this purpose. If significant differences are obtained from this test, application of a test like the Ryan procedure is required to determine the precise location of this difference.

The difference between parametric (P) and non parametric (NP) tests can be indicated by the assumptions underlying each type. The existence of the assumption is indicated by (X).

A S S U M P T I O N S	P	NP
Indication of dependent variables on an interval scale	X	-
Use of ordinal scale		
Use of frequency counts	X	-
Use of rank-order scale	-	X
Normal distribution of data	-	X
Estimating the distribution of the population from the distribution	X	-
obtained from the sample	X	-

X = assumption exists
- = assumption does not exist

Tests of Significance

The purpose of research design is to control "certain variables while testing others" (Newman & Newman, 1994, p.94). In other words, with this type of research design, the researcher tries to

- control the variance between the variables;
- set the variables so that the relation between them can be tested in an adequate and reliable way;
- determine the limitations of the study.

The following are the most commonly used tests of significance:

1. The *t*-test
2. One-way analysis of variance (ANOVA)

3. Factorial analysis of variance (ANOVA)
3. Chi-square ($\mathbf{X^2}$)

Gay (1987, p. 417) suggests that in selecting an appropriate test of significance, the researcher should first decide on which scale the data are going to be represented. In other words, whether to use parametric or nonparametric tests should be the concern of the researcher. If the data are represented in ordinal or nominal scales as in Chi-square tests, nonparametric tests are used. If the data are represented in interval or ratio scales as in the rest of the statistical tests stated above, parametric tests are applied.

In order to be able to apply the appropriate test for the statistical analysis of the study, the researcher also takes into consideration

- the number of independent and dependent variables and the number of levels of each variable,
- the type of comparison adopted for the study,
  For example comparing the performance of
  a) individuals within groups on different tasks,
  b) the same task at different times,
  c) the same task with or without a treatment,
- the way the variables are measured (e.g. ordinal, interval etc.),
- the continuity of the collected data by examining the normality and the central tendency of the distribution (see Chapter 3),
- the shape of the distribution in the population from which the samples have been drawn. In other words, does the sample adequately represent the population?

The *t*-test

The *t*-test is used to calculate the degree of significance of two means at a selected probability level. In Trochim's terms "the *t*-test assesses whether the means of two groups are statistically different from each other" (1997). The *t*-test "makes adjustments for the fact that the distribution of scores for small samples becomes increasingly different from a normal distribution as sample sizes become increasingly smaller" (Gay, 1987, p. 418).

This test is especially suitable in analyzing data obtained from post-tests only (see Vol. 1 Chapter 5, Part II).

The *t*-test, as in other parametric tests, requires data with
- interval and ratio scale variables (see Chapter 3),
- normal distribution,
- homogeneity of the variances (see Figures 6.1- 6.3).

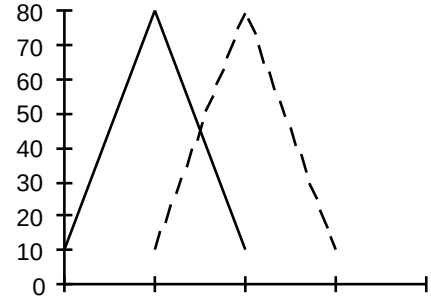

Figure 6.1 Low variability between means

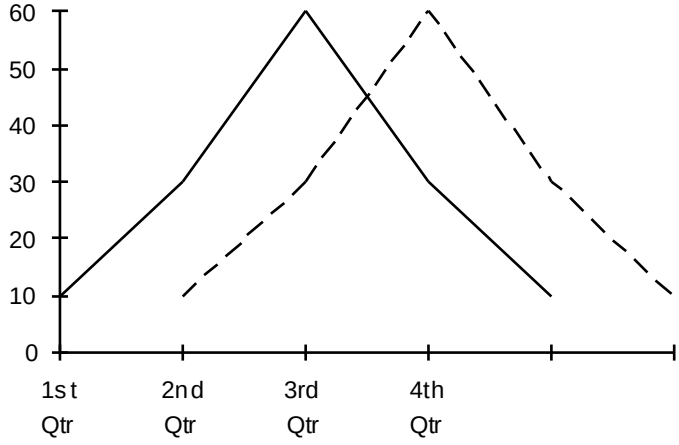

Figure 6.2 Medium variability between means

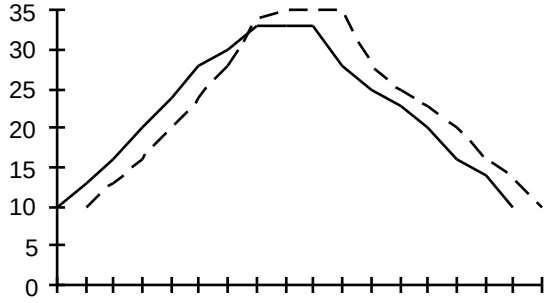

Figure 6.3 High variability between means

For instance, the effectiveness of a computer software in teaching grammar in ELT classes is to be investigated. After the application of the software program with the experimental group, another class of the same level that receives regular instruction is chosen as a control group, and the means of the final examination scores for these groups are compared. The results obtained from this comparison, gains importance only when they are tested by a test of significance. If the applied test proves the significance of the difference, the researcher can become confident of the claims made in the hypothesis. However, the most important issue is the choice of the statistical tests. Hatch & Lazaraton (1991, p. 237) suggest the consideration of certain factors in determining which statistical procedure to apply. These factors are parallel to the five assumptions (Hatch & Lazaraton, 1991). They warn the researchers to check that
the following assumptions are met before applying any statistical test:

> *Assumption 1: There are only two levels (groups) of one independent variable to compare.* … This means you cannot compare group 1 and 2, 1 and 3, and then 2 and 3, etc. If you try to use the *t*-test for such comparison, you make it very easy to reject the null hypothesis.
> *Assumption 2: Each S (or observation) is assigned to one and only one group.* That is, the procedure is not appropriate for repeated-measures designs.
> *Assumption 3: The data are truly continuous (interval or strongly continuous ordinal scores).* This means you cannot do a *t*-test on raw frequencies.
> *Assumption 4: The mean and standard deviation are the most appropriate measures to describe the data.* If the distribution is skewed, the median is a more appropriate measure of central tendency.
> *Assumption 5: The distribution in the respective population from which the samples were drawn is normal, and variances are equivalent.* It is , indeed, difficult to know if the distribution in the population is or is not normal. This is a special problem when Ss are not randomly selected but come from intact groups (pp. 263-64).

These assumptions are very important for the researcher to see whether the treatment has caused a favorable difference regarding the performance of the subjects under treatment. If such a difference is observed, a *t*-test is applied to test its significance.

A statistical software program would do the calculation of *t*-test when the scores of individuals are separately given for both groups. Depending on the claim made, the level of significance of the obtained *t*-test score is checked from the *t*-table. For instance, if, in research hypothesis, it is claimed that there will be a significant difference between the two groups without indicating any preference in favor of either of the groups, then, on the critical value of the t table, the column indicating the significance of the two tailed test is referred to in order to find out the *t*-value.

The most important task is to make a statistical decision based on the obtained results. Moore (1983, pp. .284-85) suggests a researcher proceed the following steps :

1. Find the degrees of freedom (df) :
   $df = N_1 + N_2 - 2$
   [According to this formula, you add the number of subjects in each group and subtract 2 from the total.]
2. State the obtained *t*-value: [This is the score you find as a result of the *t*-test.]
3. Find *t*-value from the provided table. [See Appendix C: Table C3].
4. If your obtained *t*-value is equal to or larger than the tabled *t*-value, the results indicate statistical significance; therefore, the null hypothesis is rejected. If the obtained *t*-value is smaller than the tabled *t*-value, the null hypothesis is accepted; in other words, it is not possible to reject the null hypothesis.

5. If the obtained *t*-value has statistical significance, then the highest level of significance is determined for that particular value (pp. 284-285).

Let us assume that

- a new treatment on teaching reading is applied to one group and another group of the same level is chosen as a control group.
- there are 10 students in Group I and 10 students in Group II.
- the research hypothesis states that there will be a significant difference between these two groups.
- the alpha level is .05.

In order to test the significance, we apply the *t*-test using the five steps mentioned above:

1. We add 10 and 10 and subtract 2 from the total.
   (10+10 = 20 - 2 = **18)**
   This way we calculate the degrees of freedom (df).
2. We enter the scores for the posttest measure for each group to the computer and calculate the *t*-value (e.g. $t = 4.42$). In calculating the t-value, we subtract the mean of one sample mean from another sample mean ($X_1 - X_2$). Then we divide the outcome by the standard error of the difference between means (Newman & Newman, 1994, p. 65).
3. We refer to the table (see Appendix C: Table C3) to find the critical value for *t*. The assumed research hypothesis stated above does not make any prediction in favor of any of the two groups; therefore, it is nondirectional. This indicates that we have a two-tailed test. On the table given in Appendix C: Table C3, we find the df column and go down the list till we find the df (18) that we calculated as a result of Step 1. Since our test is nondirectional (two-tailed test) and our alpha level is .05, we look under the appropriate level column. We find that when df is 18, the critical value of *t* is 2.101 at .05 level.
4. Our obtained *t*-value ($t = 4.42$) is larger than the tabled *t*-value ($t = 2.101$), so we claim that the difference between the two groups is statistically significant at .05 level. This leads us to reject the null hypothesis which claims that there is no difference between the groups.
5. With df = 18 for a two-tailed test, we look across the columns for the tabled t values at higher levels of significance. We find the following:
   $t = 2.552$, $p < .02$
   $t = 2.878$, $p < .01$
   $t = 3.922$; $p < .001$
   Since our obtained *t*-value ($t = 4.42$) is larger than any of the given tabled values, we can claim that our finding is statistically significant at the highest significant level presented, $p < .001$.

If we had made a directional hypothesis stating that the treatment given to Group I will yield better results than the treatment given to Group II, we would look for the values given under a one-tailed test. The df would still be 18 at .05 level. In this case, the critical value of *t* is 1.734. Since the calculated *t*-value is larger than the critical value indicated on the table, we could claim that the difference is statistically significant at .05 level. In fact, we can claim that it is even significant at .0005 level because 4.42 is larger than 3.922 (the tabled value of df 18 at .0005 level). Again, this would lead us to reject the null hypothesis since a significant difference has been found.

One Way Analysis of Variance (ANOVA)

"Most published research tends to use analysis of variance (the F-Test) to analyze factorial designs. This is used so frequently that people have confused the two. Analysis of variance is a statistical procedure, and factorial design is a design" (Newman & Newman, 1994, p. 94).

One-way of analysis is the simplest design applied to find out if there is a significant difference between **two means/groups** at a selected probability level. This is similar to a *t*-test between two groups. For instance, if we want to see if there is any difference
- between two groups on one treatment or
- between two treatment on one group.

Factorial Analysis of Variance (ANOVA)

Other tests of significance compare only two variables. For example, if we had 4 variables, we would be obliged to perform 8 tests of significance in order to compare each variable with one another. Statisticians have solved this problem by developing a method called analysis of variance.

ANOVA is used to investigate the relationship between one dependent variable and two or more independent variables each of which may have several levels. These designs are called factorial because they involve two or more factors. For example, we have two assumptions about language learning. One is that different methods yield different results in proficiency. The second is that men versus women in responding to these different methods of learning. In such a case, one of our independent variables would be methodology with two levels (e.g., audiolingual and cognitive code) and the other independent variable would be sex.

M E T H O D	S E X Female	Male
Audiolingual		
Cognitive		

In order to investigate the effect of methodology factor (Factor A = audio-lingual vs. cognitive), the effect of sex factor (Factor B = male vs. female), and the effect of an interaction of methodology and sex (Factor A x B), we form four groups:

Group 1: Audio-lingual, females
Group 2: Audio-lingual, males
Group 3: Cognitive Code, females
Group 4: Cognitive Code, males

Our aim is to find answers to questions of the following nature:

- Was there a difference according to the method applied? (Factor A)
- Did women learn more than men, or vice versa? (Factor B)

- Did men and women show greater gains when taught using one method vs. the other ? (Factor A x B)

The assumption is that the variance within groups, as well as variance between groups, may cause the difference. Thus, a ratio is formed with group differences as the numerator and an error term as the denominator to indicate the variance within groups. A formula of this kind can be utilized to compute the difference for the variances between and within groups at the end of the study or the treatment.

Therefore, in order to answer the above questions, the total variance (see Figure 3.10) (within variance, between-groups variance) needs to be calculated. The within-group variance represents error variance. The between-group variance can be the result of any of the three factors (Factor A, Factor B, or A x B). In order to find out the source of between-group variance, we must have a variance component for each of these factors. Then, we have to compute the values of these components and test the significance of each factor by using F-ratio formulas:

$$F_{\text{Factor A}} = \frac{S^2_{\text{Factor A}}}{S^2_{\text{within}}} \quad \text{(effect of methodology)}$$

$$F_{\text{Factor B}} = \frac{S^2_{\text{Factor B}}}{S^2_{\text{Factor B}}} \quad \text{(effect of sex)}$$

$$F_{\text{Interaction}} = \frac{S^2_{\text{interaction}}}{S^2_{\text{within}}} \quad \text{(effect of interaction of sex and method)}$$

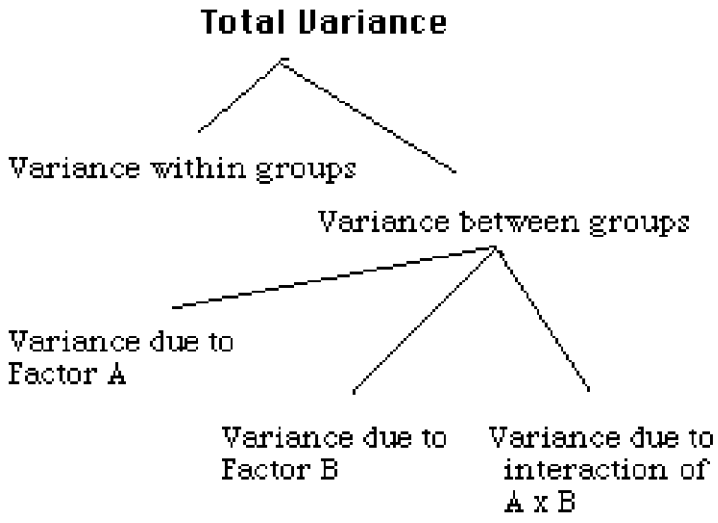

Fig. 3.10 Total variance

Steps to calculate ANOVA after the null hypothesis is formulated:

1. Compute sum of squares total (SST)
2. Compute sum of squares between (SSB)
3. Compute sum of squares within (SSW)
4. Compute sum of squares for Factor A (SS_a)
5. Compute sum of squares for Factor B (SS_b)
6. Compute sum of squares for interaction (SS_{ab})

Let us apply steps to calculate ANOVA on the following hypothetical data:

	Sex (Factor B)			
	Male		Female	
Method (Factor A)	X	X^2	X	X^2
Audio-lingual	6	36	12	144
	7	49	10	100
	5	25	7	49
	4	16	8	64
	8	64	13	169
Mean for Factor A X = 8	$\sum X=30$ X = 6	$\sum X^2= 190$	$\sum X =50$ X = 10	$\sum X^2 = 526$

Cognitive Code	15 225 14 196 20 400 13 169 13 169	10 100 9 81 8 64 7 49 6 36
Mean for Factor A X = 11.5	$\sum X=75$ $\sum X^2= 1159$ X = 15	$\sum X=40$ $\sum X^2= 330$ X = 8

X for Factor B (sex) X = 10.5 X = 9.0

Xg (mean of the levels) = 9.75

1. To compute the SST, we apply the following formula:

$$SST = \frac{\sum X^2 - (\sum X)^2}{N}$$

First, we square each individual score (in this example 20 scores = 5 in each of the four groups) , then take the sum of squares.

$$\sum X^2 = (6)^2 + (7)^2 + (5)^2 + ... + (6)^2$$
$$= 36 + 49 + 25 + ... + 36$$
$$(\sum X)^2 = 2205$$

To find $(\sum X)^2$ we first add up all the individual scores and then square that figure.

$$(\sum X)^2 = (6+7+5+4 ... 6)^2$$
$$= 195^2$$
$$= 38025$$

$$SST = \frac{2205 - 38025}{20}$$

$$= 2205 - 1901.25$$
$$= 303.75$$

2. We compute the sum of squares between (SSB).
The SSB will have to be divided into method, sex, and interaction later.

$$SSB = \frac{(\sum X_1)^2}{n_1 \; n_2} + \frac{(\sum X_2)^2}{n_3} + \frac{(\sum X_3)^2}{n_4} + \frac{(\sum X_4)^2}{} - \frac{(\sum X)^2}{N}$$

$30)^2$ $(50)^2$ $(75)^2$ $(40^2$ $(\sum X)^2$

$$= \frac{}{5} + \frac{}{5} + \frac{}{5} \quad \frac{}{5} - \frac{}{20}$$

$$= [180 + 500 + 1125 + 320\ ] - 1901.25$$

$$= 223.75$$

3. We compute the SSW using the following formula.

$$SSW = SST - SSB$$
$$= 303.75 - 223.75$$
$$= 80.0$$

4. We compute the sum of squares for Factor A (SS_a). To eliminate the effect of method, we must find the sum of squares for that factor. Therefore, we add the total score for each level of Factor A (audio-lingual and cognitive code) and divide by the number of observations and then subtract $(\sum X)^2$.

$$SS_a = \frac{(\sum \text{scores level 1})^2}{n_{\text{level 1}}} + \frac{(\sum \text{scores level 1})^2}{n_{\text{level 2}}} - \frac{(\sum X)^2}{N}$$

$$= \frac{80^2}{10} + \frac{115^2}{10} - 1901.25$$

$$= 61.25$$

5. We compute the sum squares for Factor B (SS_b). To eliminate the effect of the of the moderator variable, sex, we need to sum the squares for that factor and then add up the total score for each level of Factor B.

$$SS_b = \frac{(\sum \text{scores level 1})^2}{n_{\text{level 1}}} + \frac{(\sum \text{scores level 1})^2}{n_{\text{level 2}}} - \frac{(\sum X)^2}{N}$$

$$= \frac{105^2}{10} + \frac{90^2}{10} - 1901.25$$

$$= 11.25$$

6. We compute SS for interaction (SS_{ab}). We have already calculated $SS_a + SS_b$, so we subtract them from SSB to find the remainder.

SS_{ab} = SSB - (SS_a + SS_b)
= 223,75 - (61.25 + 11.25)

SS_{ab} = 151.25

7. We find Variance Estimates (Mean Squares). Variance estimate is indicted by degrees of freedom (df). Degrees of freedom represent the variance in the number of values. In ANOVA, there are two types of degrees of freedom: within and between. As seen in Figure 3. 11, degrees of freedom within groups is calculated by subtracting the number of groups (K) from the total number of sample size (N). For the calculation of df between groups, we subtract 1 from K. To calculate Factors A and B , we subtract 1 from the number of levels (g). In the calculation of Factor A x B, we multiply the degrees of freedom of Factor A with Factor B.

df total = N - 1 (20 Ss total - 1) = (20 - 1) = 19
df within = N - K (20Ss total - 4 groups) = (20 - 4) = 16
df for A = g - 1 (2 levels for method - 1) = (2 - 1) = 1
df for B = g- 1 (2 levels for sex - 1) = (2 - 1) = 1
df for AB = (df A) (df B) = (1) (1) = 1

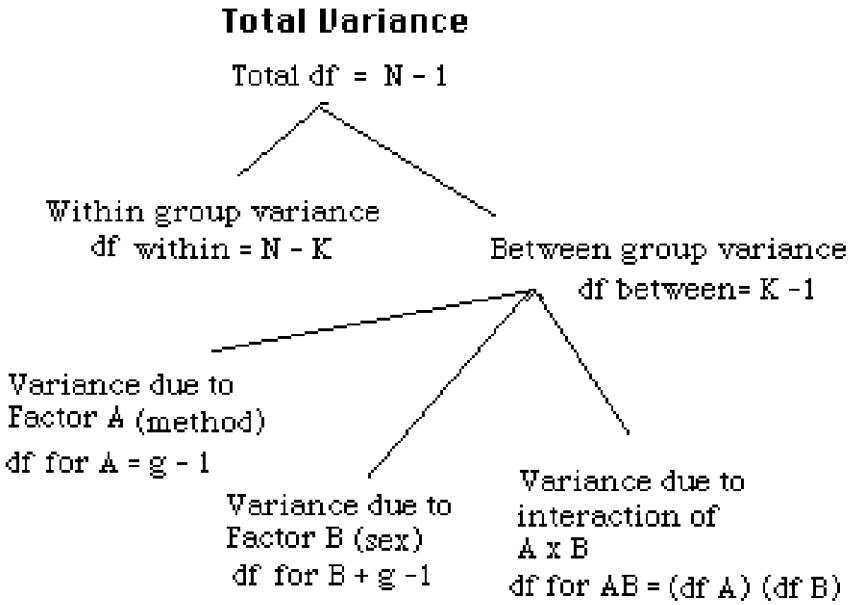

Fig. 3.11 Types of degrees of freedom

8. We find the F-ratio and complete the ANOVA test by dividing each sum of squares by its degrees of freedom to obtain the variance (MS) that can be attributed to each factor.

Source of variation		df	SS		MS (variance)
Between groups	(K-1)	3			
Method	(g-1)	1	61.25	61.25 / 1	61.25
Sex	(g-1)	1	11.25	11.25 / 1	11.25
Method X sex	(df A) (df B)	1	151.25	151.25 / 1	151.25
Within groups	(N-K)	16	80.00	80 / 16	5.00
Total	(N-1)	19	303.75		

We calculate the F-ratio for each factor (e.g. Factor A or Factor B) or the interaction of the factors by dividing the mean square (e.g. MS_a or MS_b) calculated for each factor by the mean square calculated for the variance within groups (MSW).

$$\text{F - ratio for Factor A (method)} = \frac{MSa}{MSW} = \frac{61.25}{5.00} = 12.25$$

$$\text{F - ratio for Factor B (sex)} = \frac{MSb}{MSW} = \frac{11.25}{5.00} = 2.25$$

$$\text{F - ratio for interaction} = \frac{MSab}{MSW} = \frac{151.25}{5.00} = 30.25$$

9. By referring to a statistical table, we find F for 1/16 df , which happens to be 4.49. This number indicates the critical value of the test.

10. In the end we summarize the ANOVA results within a table and indicate whether our null hypothesis is rejected or not:

Table 3. 2 ANOVA for gains in proficiency related to sex and method

Source of variation	SS Sum of squares	df Degrees of Freedom	MS Mean Square (Variance)	F-Ratio	F Critical Value
Between groups					

Method (A)	61.25	1	61.25	12.24**	From the given table. See Appendix C: Table C 4
Sex (B)	11.25	1	11.25	2.25	
A x B	151.25	1	151.25	30.25**	
Within groups	80.00	16	5,00		
Total	303.75	19			

Note: ** $p < .01$

As Brase & Brase (1995) recommend the results of the ANOVA results given in a table, they offer a chart for guidance:

B a s i c M o d e l

Source of variation	Sum of squares	Degrees of Freedom	Mean Square (Variance)	F-Ratio	F Critical Value	Test Decision
Between groups	SS_{BET}	df_{BET}	MS_{BET}	$\frac{MS_{BET}}{MS_W}$	From the given table	Reject H_0 or fail to reject H_0
Within groups	SS_W	df_W	MS_W			
Total	SS_T	N - 1				

(Brase & Brase, 1995, p. 790)

Interpretation of ANOVA

The effect of Factor A (method) exceeds the critical value. The difference in instruction method is important. The F value for sex does not exceed 4,49, so men and women made similar gains. If the interaction factor is significant, then, we must consider other claims that would contribute to the main effects of the variables that enter into that interaction. If we claim that the interaction is significant, this does not always mean that one method of instruction works better than the others. This may be due to the second factor, which indicates that males made strong gains in proficiency when instructed with the Cognitive-code Method. Women, however, made approximately the same gains regardless of method. We cannot make claims that Cognitive-code works better than audio-lingual even though there is a significant difference between methods. In this case, the significance is due to the interaction of two factors rather than one because the observed difference between methods is due to the better performance of men in Cognitive-code Method. Whenever we have a strong interaction between factors, we cannot always consider the main effects as important factors. The

significance of the difference can be attributed to the interaction of both factors rather than to one factor only.

Chi- square Test

Chi-square test, which is symbolized as X^2, is a nonparametric test of significance used to compare proportions actually observed with expected portions in order to see the significant difference. It is applied when the data are in the form of frequency counts grouped under two or more sets.

Some variables are nominal. In other words, the attribute either exists or does not exist. Bilingualism is a nominal variable. However, we can measure how much bilingualism a person has in ordinal or equal interval scale. Ordinal and interval variables can be changed to nominal scales such as beginning, intermediate, or advanced learners. When we measure nominal variables, we are not concerned with how much, but with how many and how often. Our data are in terms of frequency counts rather than scores. The Chi-square, a type of inferential statistics, is especially designed for nominal data.

	X_1	X_2	X_3
Y			

For example students at Çukurova University:

	Medicine	Engineer	Education
Students	543	437	52

The table displays the *observed* frequencies of the number of enrolled students grouped according to their major. Our task is to compare the observed frequencies with the *expected* frequencies; the frequencies that we would expect by chance. In such a case, we assume that the independent variable had no relationship to the distribution. If the area of research were not important, then the chances would be 1/3, 1/3, 1/3 . The question is whether there is a difference in the observed frequency and the expected frequency. Chi-square test (X^2) would give the answer to this question, and we pursue the following steps:

- We state the hypothesis (usually in the null form).
- We select the probability level for rejecting the null hypothesis (usually .05 or .01).
- We gather the data and display them in a frequency table. We can construct a table of expected frequency for each group (optional in the case of the questionnaire) by adding the obtained frequencies of all the groups and dividing the total by the number of groups involved. The X^2 test enables us to see if the difference between the obtained and the expected frequencies is large enough to allow us to reject the null hypothesis.

$$X^2 = \frac{\sum (\text{Observed} - \text{Expected})^2}{E}$$

E = (543+437+52) / 3(Agr., Med, Ed.)
E = 1032 /3

E = 342

	Observed f	Expected f	O - E	$(O-E)^2$	$(O-E)^2/E$
Agr.	543	342	+ 201	40401	118.13
Med.	437	342	+ 95	9025	26.39
Ed.	52	342	-290	84100	245.91
Total	1032				390.43

$\mathbf{X}^2 = 390.43$

The df will be based on the number of groups rather than the number of Ss. In our case, we have three groups.

df = number of columns - 1

- We calculate the df : 3 - 1 = 2 df
- We find the critical value from the provided lists on statistics (see Appendix C: Table C5). The critical value for 2 df is as follows:

  5.99 at .05
  9.21 at .01

As a result of calculation, we see that the difference between the obtained and the expected frequencies is large enough to reject the null hypothesis. If we have many levels for each of the two levels, the general procedure will not change. All that changes is the number of cells and the change in the degrees of freedom. For example, we want to find out whether children would choose material reinforcement or teacher praise if given a choice of reward for their work. We are also interested in knowing whether ethnic background, the independent variable, (X) has any relationship to the choice of reward, which is the dependent variable, (Y).

	X_1	X_2	X_3	
Y_1	1-1	1-2	1-3	n_1
Y_2	2-1	2-2	2-3	n_2
	n_1	n_2	n_3	N

Row totals:
n_1 = conditional distribution of X given Y_1
n_2 = conditional distribution of X given Y_2
N = marginal distribution of X

Column totals:
n_1 = conditional distribution of Y given X_1
n_2 = conditional distribution of Y given X_2
n_3 = conditional distribution of Y given X_3

	Turkish	American	Chinese	
Reward Type				Total
Material	61	36	21	(n_1 = 118)
Verbal	54	44	18	(n_2 = 116)
Total	(n_1 = 115)	(n_2 = 80)	(n_3 = 39)	N = 234

The first step to take is to determine what the expected frequencies would be if there were no special association between ethnic group membership and reward choice. For that reason, we need the marginal frequencies to help us find the expected cell frequencies (the frequencies if there were no connection between the levels of the ethnic membership variable and the choice of reward). The formula for the expected frequency for each of the six cells is

$$E_{ij} = \text{expected frequency} = \frac{\text{(raw total) (column total)}}{\text{sample size (N)}}$$

For Cell 1-1, the *observed* frequency is 61. According to the above formula, we multiply the two marginal frequencies in the row and column of this cell and divide the outcome by the total number of observations in the entire table.

$$E_{\text{cell 1-1}} = \frac{n_1 n_1}{N} = \frac{(115)\,(118)}{234} = 57.991$$

$$E_{\text{cell 1-1}} = \frac{n_2 n_1}{N} = \frac{(80)\,(118)}{234} = 40.34$$

$$E_{\text{cell 1-1}} = \frac{n_3 n_1}{N} = \frac{(39)\,(118)}{234} = 19.67$$

Expected Table

	X_1	X_2	X_3
Y_1	57.99	40.34	19.67
Y_2	57.01	39.66	19.33

Computation of Chi-square (X^2)

Row	Column	O	E	O - E	$(O - E)^2$	$(O-E)^2/E$
1	1	61	57.99	3.01	9.06	.16
1	2	36	40.34	-4.34	18.84	.47
1	3	21	19.67	1.33	1.77	.09
2	1	54	57.01	-3.01	9.06	.16
2	2	44	39.66	4.34	18.84	.47
2	3	18	19.33	-1.33	1.77	.09

$$X^2 = \sum(O-E)^2/E = \quad 1.44$$

To find the critical value of X^2 , we need to know the degrees of freedom. Therefore we apply the following formula:

df = (Rows - 1) (Columns -1)

We have 2 x 3 table (two rows and three columns), thus:

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

The critical value for 2 df is 5.99 at the .05 level of significance. The results of the Chi-square tell us that ethnic background does not influence choice of reward. Therefore, we do not reject our null hypothesis.

The following assumptions must be met for computing chi-square:

1. Each observation must fall in one and only one category.
2. The number of expected (not number of observed) frequencies in each cell must be at least five or we cannot legitimately use the test.
3. We need to use a correction factor for one-way (1x1) X^2 and two-way (2x2) X^2 where df is only 1.

Correction Factors

If we do a one way Chi-square (X^2) and have only 1 df, as is often the case with nominal variables of only two levels (e.g. male vs. female, or graduate vs. undergraduate), we will need to correct the estimate so that it fits the X^2 distribution because df is over 1. When the df is 1 and the design is one-way, we can correct it by adding or subtracting .5 from the observed values. If the observed value is larger than the expected value, we subtract .5 from the observed value. If the observed value is smaller than the expected value, we add .5 to the observed.

If we have a two-way (2 x 2)table and the df is 1: (2-1) (2-1) =1, the Yates correction factor allows us to adjust the data to fit the X^2 distribution :

Variable X

Variable Y			
	a	b	a + b
	c	d	c + d
	a + c	b + d	

Accordingly, we put the data into the corrected $\mathbf{X}^2$ formula:

$$\mathbf{X}^2 = \frac{N ([ad - bc] - N/2)^2}{(a + b) (c + d) (a + c) (b + d)}$$

Example: Suppose you asked a class of 50 students whether they liked using the computer lab (yes or no) and they are all either graduate or undergraduate students.

	Graduate	Undergraduate	Total
Yes	24	8	32
No	6	12	18
Total	30	20	

$$\mathbf{X}^2 = \frac{50 ([(24) (12) - (8) (6)] - 50/2)^2}{(32) (18) (30) (20)}$$

$$= \frac{50 ([288-48] -25)^2}{(576) (600)}$$

$$= \frac{50 (240-25)^2}{345600}$$

$$= \frac{2311250}{345600}$$

$$= \frac{2311250}{345600}$$

$$= 6.68$$

The critical value w/1 df is 3.84.

Application of Statistical Test on Data

The types of statistical analysis given here do not comprise all the available information. Brown (1990, p. 13) has adopted a table from Tuckman (1978, p. 255) categorizing the common types of analysis based on the number of variables, and the type of scales represented by the data. Using the adopted table three times, Brown has indicated, in boldface, the common analysis for different research methods. The information given in these tables will be cited in a list form:

Tests forDetermining Correlation

1. Simple regression, and Pearsonproduct-moment correlation coefficient r
   - one independent variable; data: interval scale
   - one dependent variable; data: interval scale

   e.g. The correlation between leadership and intelligence
   The correlation between intelligence scores of identical twins

2. Multiple regression, and Multiple R
   - more than one independent variable; data: interval scale
   - one dependent variable; data: interval scale
   or
   - one independent variable; data: interval scale
   - more than one dependent variable; data: interval scale

e.g. The prediction of ESL scores at the end of two years by examining the

1) overall proficiency scores in English,
2) language aptitude scores, and
3) score measure of individual motivation to learn a language.

3. Spearman rho, and Kendall's tau
   - one independent variable; data: ordinal scale
   - one dependent variable; data: ordinal scale

   e.g. To find the degree of similarity (in a small sample of less than 30) between acquisition rank orders as done by Krashen (1977).

4. Kendall's W
   - more than one independent variable; data: ordinal scale
   - zero dependent variable

e.g. To determine the tendency of agreement of four teachers in ranking students' ability to speak English

5. Biserial, and Point-biserial correlation
   - one independent variable; data: interval scale

- one dependent variable; data: nominal scale

e.g. To analyze the degree of relationship between being male or female (dichotomous nominal scale) and achievement in English

6. Phi coefficient, tetrachoric correlation
   - one independent variable; data: nominal scale
   - one dependent variable; data: nominal scale

e.g. To find out whether students who study first-semester Spanish in American colleges previously visited any Spanish speaking country (dichotomous for Phi coefficient)

To investigate the relationship between passing the Intensive English course and having been absent 10 or more times from class.

Common Statistics for Mean Comparison

1. z statistics (large sample), *t*-test (any sample)
   - one independent variable at levels; data: nominal scale
   - one dependent variable; data: interval scale

e.g. To interpret the test scores or observations from experimental and control groups.

2. One-way ANOVA
   - one independent variable at more than two levels; data: nominal scale
   - one dependent variable; data: interval scale

3. Two-way or three-way (etc.) ANOVA
   - more than one independent variable; data; nominal scale
   - one dependent variable; data: interval scale

4. Multivariate analyses
   - one independent variable; data: nominal scale
   - more than one dependent variable; data: interval scale
   or
   - more than one independent variable; data: nominal scale
   - more than one dependent variable; data: interval scale

e.g. To see the difference in test scores of graduate American students taking engineering as a major course or not, and Chinese graduate students taking engineering as a major course or not.

5. Kruskal-Wallis test
   - one independent variable; data: nominal scale
   - dependent variable; data: ordinal scale

e.g. To make comparisons in situations in which other powerful statistical methods do not work

6. Friedman's two-way ANOVA
   - more than one independent variable; data: nominal scale
   - one dependent variable; data: ordinal scale

e.g. To conduct multiple comparison by finding out that there is no effect of null hypothesis

Common Analyses for Comparing Frequencies

1. Phi coefficient, Tetrachoric correlation, and Fisher exact test
   - one independent variable; data: nominal scale
   - one dependent variable; data: nominal scale

2. Chi-square
   - one independent variable; data: nominal scale
   - one dependent variable; data: nominal scale
   or
   - more than one independent variable; data: nominal scale
   - zero dependent variable

e.g. To investigate the significant difference in frequency of signing up or not signing up for a language course at four different schools in different geographical regions.

If the data for independent variable represent ordinal scale, and the data for the dependent variable represent interval scale, the data cannot be assessed. To solve the problem, there are three alternatives:

1. Transform the ordinal variable into nominal scale, and apply the measures used to assess nominal independent variable and interval dependent variable.
2. Transform the interval variable into ordinal, and use the measurements to assess ordinal independent variable and ordinal dependent variable.
3. Transform both variables into nominal, and apply the measures for nominal scales.

If the data for independent variable represent interval scale and the data for dependent variable represent ordinal scale, the data cannot be assessed. To solve the problem, there are three alternatives:

1. Transform the ordinal variable into a nominal variable, and apply the measures for ordinal scales.
2. Transform the interval variable into ordinal, and use the measures for ordinal scale.
3. Transform the interval variable into a nominal variable, and use measures for nominal scale.

EXERCISES

A. Fill each of the blanks with a word that is most appropriate in statistical terms:

1. If 10 students are ranked from best to worst in terms of their scores in History 201, the ranks would be given on a(n) _____ scale.
2. Four brands of beer would be given on a(n) _____ scale.
3. If one value in a distribution is changed, it is certain that, by the raw score calculation, the _____ has changed.
4. In a scatterplot graph, if one of the points does not fall on the line of best fit, **r** cannot be _____.
5. The standard error of estimate is a kind of _____.
6. If high scores on one variable are associated, in general, with high scores on a second variable, their correlation is called _____.
7. One-way analysis of variance is ultimately concerned with hypotheses about _____.
8. The purpose of multiple comparison tests is to find out which groups differ after _____.
9. Homogeneity of variance refers to equality of _____.

B. Which statistical test is most appropriate to answer each of the following questions?

Test types:
Revision

a = z-statistics for one sample or for independent samples
b = z-statistics for dependent or correlated samples
c = *t*-test for one sample or for independent samples
d = *t*-test for dependent samples
e = Frequency test

1.__ Is the average length of adult male feet the same in Turkey as it was in 1882 ?
2.__ Is there any variation in the time of sunrise on July 10 over a 50-year period?
3.__ Is there any association between the age and weight of adults in Turkey?
4.__ Is the violin -playing performance of adult males more variable in quality than the performance of adult females?
5.__ Do equal number of male and female high school seniors possess valid driver's licenses?
6.__ Is the correlation between the intelligence and achievement test scores of teenagers greater than +.50?
7.__ Have freshmen learned more about using library facilities after they have completed a course on research design than they had before they signed up for the course?
8.__ Is the association between IQ at age 10 and adulthood similar for men and women?

C. Read the information/assumptions in the given box. Then try to solve the problem. accordingly.

Research Hypothesis : There is no difference in the reading achievement scores of your class(Class A) with Class B.

Significance level: .05
1 or 2-tailed: 2-tailed
Dependent variable: Reading achievement
Independent variable: Class
Number of students in each class: 25
Measurement: Nominal (class vs. population)
Statistical procedure: *t*-test

D. Suppose that the hypothesis of your research is as follows:

The mean score on a proficiency test will be higher for females than males when administered to the Preparatory Year students at English Language Teaching Departments at Çukurova University.

Again, suppose that after the administration of the proficiency test, you obtained the following data:

~X for females = 87.6
~X for males = 79.8
SEM = 3

a. State the null hypothesis.
b. Would you use a one-tailed or a two-tailed test of significance?
c. Check the table value at the 0.5 level significance. Would you reject or confirm the null hypothesis?
d. Would your result reject or confirm the hypothesis you stated initially?

E. The director of the Preparatory School wants to determine the differences between the two workshops conducted at the school:
1) the traditional inservice teacher workshop
2) the new inservice teacher workshop
A group of 20 instructors were randomly assigned into one of the two groups. *A posttest only control group design* was used. With the scores given below, conduct a statistical analysis to figure out which group has made a significant improvement.

New Inservice Workshop	Traditional Inservice Workshop
60	50
58	48
57	44
56	44
52	40
50	39
50	36
48	35
47	34
42	30

F. Consider a research study in which a group of Freshman at Middle East Technical University are asked to respond to the question using yes/no options. The aim is to find out these students' preference regarding English-medium instruction in their undergraduate studies.
Out of 220 students responding the questionnaire, 80 say no and 140 say yes.
Calculate the $\mathbf{X}^2$ score to determine whether their responses differ significantly from each other.