

# THE UNIVERSITY OF TEXAS AT AUSTIN

## SCHOOL OF INFORMATION

### MATHEMATICAL NOTES FOR LIS 397.1

#### INTRODUCTION TO RESEARCH IN

#### LIBRARY AND INFORMATION SCIENCE

Ronald E. Wyllys  
Last revised: 2003 Jan 15

## ANOVA: Analysis of Variance

### Introduction

The acronym, ANOVA (pronounced "a nova"), is the popular name for the statistical procedure whose full name is "analysis of variance." ANOVA carries out tests of hypotheses that say that some variable has the same mean value in two or more populations, i.e., in two or more situations. The procedure gets its name from the fact that it makes use of an elegant chain of reasoning about variances in order to reach a conclusion about whether the mean values of the variable are the same or not. The ANOVA technique was invented by a British mathematician, Sir Ronald Aylmer Fisher, in the 1920s.

In discussing the ANOVA procedure, we shall first take a look at the reasoning that underlies the procedure. After that, we shall explain how the arithmetic details are handled in terms of manual calculations. These calculational techniques were developed and polished in the 1920s and 1930s, when the only aids to calculation were mechanical and electromechanical calculating machines.

What is known as the ANOVA table was devised in that period, as a way of aiding the calculations through the display of intermediate steps along with the final results. Though no one would think of doing ANOVA these days except via a computer, the ANOVA table remains popular as a way of displaying the results of an ANOVA analysis. Computer programs for ANOVA regularly use the ANOVA table to report their results., and our review of the manual calculational techniques will help you understand the ANOVA table.

### The Reasoning Underlying the ANOVA Procedure

In this section we present an outline of the reasoning upon which ANOVA is based. To simplify things as much as possible, we shall illustrate the reasoning by showing how it works for a null hypothesis involving just two populations,

$$H_0: \mu_1 = \mu_2$$

where the populations are such that the hypothesis can be tested only via independent samples.

To test such a hypothesis, we need to take a sample from each population and, from each sample, we need to obtain the three key sample statistics that we use over and over again in inferential statistics. Let us suppose, then, that we have taken the two samples and have obtained the sample means,

$$\bar{X}_1, \bar{X}_2$$

the sample standard deviations,  $s_1, s_2$ , and the sample sizes,  $n_1, n_2$ . It happens that in the ANOVA procedure, it is often convenient to work with the sample variances,  $s_1^2, s_2^2$ , rather than the sample standard deviations. The variances, of course, embody the same information as the standard deviations, since we can move from one to the other by simply squaring or taking the square root.

To simplify matters even further, we shall suppose that the sizes of the two samples are the same. We shall use the symbol  $n_s$  to denote this shared sample size, in order to help avoid ambiguity later in the discussion.

## The Observed F-Ratio

What ANOVA does is to use the six key sample statistics from the two samples to form a ratio called the "observed F-ratio." The name F-ratio honors Sir Ronald Fisher, the inventor of ANOVA. (An American statistician, George Snedecor, is responsible for popularizing this name for the ratio.)

The F-ratio is ordinarily written as follows:

$$F_{obs} = \frac{\hat{\sigma}_B^2}{\hat{\sigma}_W^2}$$

This equation contains some special notation that, for historical reasons, is regularly used in the ANOVA procedure, and we need to discuss the notation briefly.

You can recognize the lower-case Greek sigmas, which you know are used to refer to the population standard deviation (with no exponent) and population variance (with an exponent of 2). The circumflex accents are usually read as "hat" in this context; e.g., the numerator of the F-ratio can be read as "sigma hat squared sub B". The circumflex accent is regularly used in inferential statistics to indicate that you are dealing with an *estimate* of the population parameter to which the accent is attached. Here the parameter being estimated is the population variance. The subscripts "B" and "W" are used to distinguish two different estimates of the population variance, called the "between groups" and "within groups" estimates, respectively. The difference between these two estimates is the key to the ANOVA procedure.

## Homoscedasticity

We need to note that one of the assumptions in the ANOVA procedure is that there is such a thing as *the* population variance. That is, a part of the procedure is the assumption that the two populations have equal variances, i.e., that  $\sigma_1^2 = \sigma_2^2$ . Since they are equal (by assumption), they have a common value, which is *the* population variance and which we can denote by  $\sigma_{pop}^2$ . That is, we assume that we have

$$\sigma_1^2 = \sigma_2^2 = \sigma_{pop}^2$$

This assumption, which is called the assumption of "homoscedasticity," is vital to the whole ANOVA procedure. (The word "homoscedasticity" has its roots in the Greek words *homos*, meaning "same," and *skedastikos*, meaning "able to scatter." Thus "homoscedasticity" conveys the notion of "same scatteredness," i.e., equal variability. The English word *scatter* is derived from the same Indo-European root as *skedastikos* and has preserved some of the sound.)

If the homoscedasticity assumption seems arbitrary to you, that is a natural and reasonable reaction. Fortunately, from a pragmatic standpoint, the assumption is well justified. One reason is that it can be shown theoretically that if the variances of populations in the null hypothesis are merely reasonably close though not identical, the ANOVA procedure still works very well. Another reason is this: Suppose you have several sets of numbers, each set being a different sample of observations of the same real-world phenomenon. It is a fact of nature that the variabilities (e.g., variances) of the different sets are more likely to be about equal than are the centers (e.g., the means) of the sets.

That is, variability tends to be a more stable property of sets of numbers that are samples of observations of a real-world phenomenon than does the location of the centers of the sets. A classic example is that a typical group of students taking a semester-long typing course will exhibit about the same amount of variation among their typing speeds in the 3rd-week test as in the 15th-week test, though the mean speed of the group will have increased considerably between the 3rd and 15th weeks. Another example is that a randomly chosen group of 100 women and a randomly chosen group of 100 men will exhibit about the same standard deviation in their two sets of height measurements, though the mean heights of the two groups will surely differ.

The foregoing discussion suggests that it is reasonable in the ANOVA procedure to assume that homoscedasticity prevails and that a single common population variance, *the* population variance, exists. What the observed F-ratio does is to compare two different ways of estimating this variance. We now discuss these two ways.

## The Two Estimates of the Population Variance

**The Within-Groups Estimate of Population Variance.** The denominator of the observed F-ratio contains

$$\hat{\sigma}_w^2$$

the within-groups estimate of the population variance. This estimate comes from the two observed sample variances,  $s_1^2$  and  $s_2^2$ , through the following reasoning.

Consider the fact that the two observed sample variances constitute two different sample values for the same population parameter: the population variance,

$$\hat{\sigma}_{pop}^2$$

The two sample variances represent two different pieces of evidence about the same thing; specifically, they are two estimates of the population variance. A moment's reflection will suggest to you that the sample variances are unlikely to be exactly equal to each other or to the true population variance; on the other hand, each of them is likely to be close to the latter. In this situation, we can combine the two pieces of evidence, thereby combining their information about the population variance, and thus we can obtain an improved estimate of that variance.

The combination is straightforward. We simply take the average of the two sample variances. (We are assuming, for simplicity, that the two samples are of equal size. If they were not, we would take a weighted average of them, with the weights being  $n_1 - 1$  and  $n_2 - 1$ , respectively.) Taking the average of the two sample variances, we let the within-groups estimate be defined as

$$\hat{\sigma}_w^2 = \frac{s_1^2 + s_2^2}{2}$$

Note that the information in this estimate involves only the variability *within* each sample, the variability around the mean of that sample, since that variability is what each of the two sample variances represents. This fact is the origin of the name of this estimate.

**The Between-Groups Estimate of Population Variance.** The numerator of the observed F-ratio contains

$$\hat{\sigma}_B^2$$

the between-groups estimate of the population variance. This estimate has a more subtle derivation than that of the within-groups estimate. The derivation begins with a fresh look at the standard error,  $s_{\bar{X}}$ . We remind you that, for the sake of simplicity, our illustration has just two populations, and that we have drawn a sample of size  $n_s$  from each of these populations.

You will recall that the standard error is a standard deviation: viz., the standard deviation of the set of means of all possible samples of a given size drawn from the population of interest. In the ANOVA procedure, *when the null hypothesis is true*, the two populations in the hypothesis have the same mean and we can consider their combination as being *the* population of interest. For convenience in the discussion that follows, we shall give this population the name "Population NHT," where "NHT" comes from "null hypothesis is true." Then the means of the two samples that we drew from the population of interest, Population NHT, can be considered to be elements of the set of means of all possible samples of size  $n_s$  that could be drawn from Population NHT.

It will be convenient to regard this set of means as a population in its own right. That is, the set of means of all possible samples of size  $n_s$  that could be drawn from Population NHT is itself going to be regarded as a population. To aid the discussion, we shall give this latter population the name "Population ASP," where "ASP" comes from "all samples possible." In the two sample means that have been obtained,

$$\bar{X}_1, \bar{X}_2$$

we have a sample of size 2 from Population ASP. This is a quite small sample, but nevertheless, as with any sample, we can calculate the sample mean and the sample standard deviation. We begin by calculating the mean of this sample:

Formula 1

$$\bar{X}_{ASP} = \frac{\bar{X}_1 + \bar{X}_2}{2}$$

We also can calculate the standard deviation of this sample. Recall that the definition of sample standard deviation is:

$$s = \sqrt{\frac{\sum C_X - \bar{X}h^2}{n-1}}$$

In working with the two observations from the population of all possible sample means, Population ASP, we will use the two observed sample means in the X term in this formula. The mean in the formula, the X with the bar over it, will be represented by the mean we calculated in Formula 1. We will call the resulting standard deviation  $s_{ASP}$ , and for it we have

Formula 2

$$s_{ASP} = \sqrt{\frac{\sum C_X - \bar{X}h^2}{n-1}} = \sqrt{\frac{C\bar{X}_1 - \bar{X}_{ASP}h^2 + C\bar{X}_2 - \bar{X}_{ASP}h^2}{2-1}}$$

At this point we have two standard deviations that we need to keep separate in our minds. We have just shown how to calculate  $s_{ASP}$ , the standard deviation of the sample of size 2 that we drew from the population of all possible sample means, Populations ASP. The other standard deviation that we are concerned with is the standard deviation of the sample that we drew from the elements of Population NHT. We shall call this sample standard deviation  $s_{NHT}$ , and we note that it is an estimate of  $\sigma_{NHT}$ , the standard deviation of the elements of Population NHT.

The population standard deviation of Population ASP is a very special standard deviation, the one that we call the (true) standard error of the mean and represent by the symbol  $\sigma_{\bar{X}}$ . Ordinarily we lack complete information about the population that is currently of interest, here Population NHT, and hence in place of the relation

Formula 3

$$\sigma_{\bar{X}} = \frac{\sigma_{NHT}}{\sqrt{n_s}}$$

we use the observed sample values in the corresponding relation

Formula 4

$$s_{\bar{X}} = \frac{s_{NHT}}{\sqrt{n_s}}$$

Since  $s_{\bar{X}}$  is based on information from the sample (rather than directly from the population), it is what we ordinarily have in mind when we talk about the standard error of the mean. Furthermore, since  $s_{\bar{X}}$  is the sample standard deviation of Population ASP, when our sample consists of two sample means,  $s_{\bar{X}}$  can be calculated directly, as shown in Formula 2 above. That is, we can write

$$s_{ASP} = \sqrt{\frac{\sum C_X - \bar{X}h^2}{n-1}} = \sqrt{\frac{C\bar{X}_1 - \bar{X}_{ASP}h^2 + C\bar{X}_2 - \bar{X}_{ASP}h^2}{2-1}} = s_{\bar{X}}$$

From this, by squaring the rightmost two elements, we have

$$\frac{\mathbf{C}\bar{X}_1 - \bar{X}_{ASP} \mathbf{h}^2 + \mathbf{C}\bar{X}_2 - \bar{X}_{ASP} \mathbf{h}^2}{2 - 1} = s_{\bar{X}}^2$$

Thus we can finally conclude that

$$s_{\bar{X}}^2 = \mathbf{C}\bar{X}_1 - \bar{X}_{ASP} \mathbf{h}^2 + \mathbf{C}\bar{X}_2 - \bar{X}_{ASP} \mathbf{h}^2$$

But also, from Formula 4 it follows that

$$s_{\bar{X}}^2 = \frac{s_{NHT}^2}{n_s}$$

and hence that

$$s_{NHT}^2 = n_s [\mathbf{C}\bar{X}_1 - \bar{X}_{ASP} \mathbf{h}^2 + \mathbf{C}\bar{X}_2 - \bar{X}_{ASP} \mathbf{h}^2]$$

Finally, note that the  $s_{NHT}^2$  in the preceding equation represents the sample variance of Population NHT, and that this sample variance is an estimate of the population variance of Population NHT. In other words, we have found that an estimate of this population variance is

$$s_{NHT}^2 = n_s [\mathbf{C}\bar{X}_1 - \bar{X}_{ASP} \mathbf{h}^2 + \mathbf{C}\bar{X}_2 - \bar{X}_{ASP} \mathbf{h}^2] = \hat{\sigma}_B^2$$

This estimate is what we call the between-groups estimate of population variance. The foregoing equation shows that  $\hat{\sigma}_B^2$  depends directly on the observed means of the two groups (i.e., samples) and, hence, on the difference between them. This is what led to the name of this estimate.

## Putting the Two Estimates of Population Variance Together

The observed F-ratio

$$F_{obs} = \frac{\hat{\sigma}_B^2}{\hat{\sigma}_W^2}$$

consists of the ratio of (a) the between-groups estimate of the variance of Population NHT to (b) the within-groups estimate of this variance. Here is how the F-ratio sheds light on the truth or falsity of the null hypothesis.

*When the null hypothesis is true*, the means of the two populations in the null hypothesis are equal. In this case the means of the two samples can be expected to be close together. That is, the distance between the two sample means is small, and hence the distance between each of them and their mean,

$$\bar{X}_{ASP}$$

will also be small. Therefore, the *between-groups estimate of population variance will be small*, since it depends directly on the terms

$$\mathbf{C}\bar{X}_1 - \bar{X}_{ASP} \mathbf{h}^2 \text{ and } \mathbf{C}\bar{X}_2 - \bar{X}_{ASP} \mathbf{h}^2$$

which involve the distances between each of the sample means and their joint mean.

Furthermore, *when the null hypothesis is true*, it turns out that the *between-groups estimate* and the *within-groups estimate* are usually about equal, and hence that the F-ratio will be *close to 1*.

When the null hypothesis is false, the means of the two populations in the null hypothesis are not equal; i.e., they are different and, hence, are some distance apart. It follows that the means of the two samples can also be expected to be some distance apart. In comparison with the situation when the null hypothesis is true, here the distances between the two sample means are larger, and hence the distance between each of them and their mean,

$$\bar{X}_{ASP}$$

will also be larger. Therefore, when the null hypothesis is false, the between-groups estimate of population variance will be larger.

What about the within-groups estimate of population variance in the situation when the null hypothesis is false? This estimate remains unchanged. The reason is that the within-groups estimate involves only information about how the observations within each group (i.e., sample) vary around the mean of that sample. Shifting the mean of the sample does not, in general, change the pattern of variability of the elements in that sample around the sample mean.

**The Gist of the F-Ratio.** Now you can see how the F-ratio behaves. When the null hypothesis is true, the numerator and denominator of the F-ratio can be expected to be about the same, and the F-ratio can be expected to be close to 1. But when the null hypothesis is false, the numerator of the F-ratio can be expected to be larger, while the denominator can be expected to remain unchanged. The result is that when the null hypothesis is false, the F-ratio can be expected to be larger than 1.

In short, when the F-ratio is close to 1, our decision with respect to the null hypothesis will be to continue to believe that it is true; i.e., our decision will be to accept  $H_0$ . When the F-ratio is sufficiently larger than 1, our decision will be to believe that the null hypothesis is false; i.e., our decision will be to reject  $H_0$ .

In any particular experiment or investigation, we must choose a boundary between values of the F-ratio that are "close enough to 1 that we will accept  $H_0$ " and values that are "so much larger than 1 that we will reject  $H_0$ ." That boundary will be chosen on the basis of three factors: the level of significance at which we want to work, the number of populations specified in the null hypothesis, and the numbers of observations made in the various samples. Tables of values of the F-distribution reflect these factors, and you will choose your boundary from an F-distribution table. The tabled boundary will serve as the threshold against which you will compare the observed value of the F-ratio.

## ANOVA Arithmetic and the ANOVA Table

The Analysis of Variance (ANOVA) procedure handles problems that can be reduced to the following question: Is it true that two or more populations have the same mean? An equivalent phrasing would be, does some variable have the same mean value in the two or more populations, i.e., in two or more different situations? In general, with  $k$  populations ( $k \geq 2$ ) in the problem, you write the null hypothesis as

$$H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$$

**Doing ANOVA by Manual Calculations.** The arithmetic details in this section are intended to help you understand what is in an ANOVA table, and why it is a useful way to display the results of applying the ANOVA procedure to observed data. Even though nowadays the ANOVA procedure is almost invariably handled by computer programs, you will gain a better understanding of the procedure by going through these manual details once.

The arithmetic steps necessary in manually carrying out an ANOVA procedure can be summarized as follows. You start by calculating what is called the Correction Factor (CF):

$$CF = \frac{\text{square of [the sum of all the observations]}}{\text{total number of observations}}$$

Next you calculate the Sum of Squares Total ( $SS_T$ ) (also called the "total sum of squares"):

$$SS_T = (\text{sum of the squares of the individual observations}) - CF$$

Next you calculate the Sum of Squares Between ( $SS_B$ ) (also called the "between-groups sum of squares"). In this calculation a major rôle is played by the "group totals." Each group total is the sum of the observations in a particular group, i.e., sample, drawn from one of the  $k$  populations. The calculation may be handled in one of two ways. If, as often happens the groups (samples) are not all the same size, then

$$SS_B = \frac{h_{\text{group 1 total}}^2}{\text{no. of obs. in group 1}} + \frac{h_{\text{group 2 total}}^2}{\text{no. of obs. in group 2}} + \dots + \frac{h_{\text{group k total}}^2}{\text{no. of obs. in group k}} - CF$$

If, on the other hand, the groups do all contain the same number of observations, then the denominators in all the fractions in the preceding equation will be the same. In this case the equation can be simplified down to

$$SS_b = \frac{\text{sum of [the squares of the group totals]}}{\text{no. of observations in any one group}} - CF$$

Finally, you enter the values of  $SS_T$  and  $SS_B$  into what is called the ANOVA table, and carry out the remaining calculations. Here is one way of setting up the ANOVA table.

	SS	DF	MS	F
T (total)	$SS_T$	(total number of observations)-1		
B (between)	$SS_B$	(no. of groups)-1=k-1	$\frac{SS_B}{DF_B} = MS_B = \hat{\sigma}_B^2$	$\frac{\hat{\sigma}_B^2}{\hat{\sigma}_W^2} = F_{obs}$
W (within)	$SS_T - SS_B$	$DF_T - DF_B$	$\frac{SS_W}{DF_W} = MS_W = \hat{\sigma}_W^2$	

In this table, row T contains the total sum of squares and the total number of degrees of freedom; row B contains the between-groups entries (and the final value calculated for F); and row W contains the within-groups entries.

Here is how to read the entries in this 3-row by 4-column table. You have already calculated  $SS_T$  and  $SS_B$ ; and from the samples that you have gathered, you know the number of groups (samples) involved and the total number of observations in all the samples put together. You enter  $SS_T$  and  $SS_B$  into the cells in the 1st and 2nd rows of the 1st column as shown; then you subtract  $SS_B$  from  $SS_T$  to get the entry in the cell in the 3rd row of the 1st column. This entry is  $SS_W$ , the Sum of Squares Within (i.e.,  $SS_W = SS_T - SS_B$ ).

We denote the cell containing  $SS_W$  by cell(W,SS), which you can read as "the cell in row W and column SS." We denote the other cells in similar fashion.

Into cell(T,DF) you enter the number that is one less than the total number of observations in all the samples; this entry is the number of degrees of freedom, df, associated with the total set of observations, and we shall call it  $DF_T$ , the Degrees of Freedom Total. Into cell(B,DF), you enter the number that is one less than the number of groups. This is the number of degrees of freedom associated with the Sum of Squares Between (or among) the groups; we shall call it the Degrees of Freedom Between and denote it by  $DF_B$ . Next, you subtract the contents of cell(B,DF) from the contents of cell(T,DF) to obtain the entry for cell(W,DF). This entry is the number of degrees of freedom, df, associated with the Sum of Squares Within the groups; we shall call it the Degrees of Freedom Within and denote it by  $DF_W$ .

Next, you obtain the entry in cell(B,MS), by dividing the contents of cell(B,SS) by the contents of cell(B,DF), as shown. This amounts to dividing the Sum of Squares Between by the Degrees of Freedom Between. Dividing a sum of squares by its associated number of degrees of freedom always yields a "mean square." (The process is almost exactly analogous to what you do when you obtain the standard deviation by dividing a sum of squares by  $n - 1$ . The variance that you obtain by this division, before you take the square root, is an adjusted mean square.) The mean square you have just calculated in the ANOVA table is called the Mean Square Between and denoted by  $MS_B$ .

Similarly, you divide the contents of cell(W,SS) by those of cell(W,DF) to obtain the entry in cell(W,MS). It is called the Mean Square Within and denoted by  $MS_W$ .

The two mean squares you now have in the MS column constitute the best information available from the sample concerning the two estimates of population variance, the "Between" (between-groups) estimate and the "Within" (within-groups) estimate. The point of these names (and the key to the ANOVA procedure) is that the between-groups estimate,  $MS_B$ , is based on information stemming from the *differences among* the sample means of the various groups, while the within-groups estimate,  $MS_W$ , is based on information stemming from the sample variances of the various groups, each of which variances is calculated using *only* information from *within* that particular group.

To emphasize that these are estimates, we use the standard statistical notation for estimates, viz., " $\hat{\sigma}$ ", the circumflex accent, over the symbol, " $\sigma$ ", for the population variance; the combination is usually pronounced "sigma hat." Also we use the subscript "B" to represent "between groups" and the subscript "W" to represent "within groups."

Finally, the observed value of the test-statistic,  $F_{obs}$ , is calculated by dividing the contents of cell(B,MS) by those of cell(W,MS). Thus  $F_{obs}$  is the ratio of the Mean Square Between to the Mean Square Within.

## Interpretation of the F-Ratio

Because of the nature of the two estimates (as pointed out earlier), the F-ratio will be close to 1 when the null hypothesis is true, i.e., when the population means are equal and, hence, when the sample means can be expected to be close together. If, on the other hand, the null hypothesis is false, then at least one of the population means is different from the others, and, hence, at least one of the sample means can be expected to be different from (i.e., to lie at a distance from) the others. This will make the numerator of the F-ratio larger than when the null hypothesis is true, so that when the null hypothesis is false, F will be bigger than 1.

If  $F_{obs} \leq 1$ , it is easy for you to decide that the null hypothesis is true. But when  $F_{obs} > 1$ , you have the problem of deciding whether  $F_{obs}$  is close enough to 1 for you to believe that the null hypothesis is true, or is enough bigger than 1 for you to believe that the null hypothesis is false. This decision depends on the degrees of freedom, i.e., ultimately, on the total number of observations and the number of groups, as well as on the information from the observations (viz., the sample means and the sample variances) that is bound up in the value of  $F_{obs}$ . Thresholds for making the decision are provided by tables of the F distribution.

Tables of the F distribution are set up in terms of the level of significance of the hypothesis test, and in terms of both the number of degrees of freedom in the numerator of the F-ratio and the number of degrees of freedom in the denominator of the F-ratio. Suppose you are using an F-distribution table as the source of a threshold value for an ANOVA test. Then the Degrees of Freedom Between,  $DF_B$ , will correspond to the degrees of freedom for the numerator in the table; and the Degrees of Freedom Within,  $DF_W$ , will correspond to the degrees of freedom for the denominator in the table. Usually, the numerator degrees of freedom will identify a column in the table, and the denominator degrees of freedom will identify a row (or a block of rows) in the table. Together, the column-wise degrees of freedom,  $DF_B$  (numerator), and the row-wise degrees of freedom,  $DF_W$  (denominator), will identify a specific cell within the F-distribution table. Within this cell will be at least two threshold values, which will be identified as corresponding to levels of significance of  $\alpha = .05$  and  $\alpha = .01$ , respectively.

## A Numerical Example of the ANOVA Table

Here is a numerical example of the use of the ANOVA table. To make the example short enough for you to be able to check the steps in the arithmetic easily, the example drastically over-simplifies matters and uses very small samples.

Suppose you wanted to study the question of whether graduates of your library school tended to earn higher salaries, on the average, than graduates of Brand X Library School. You gather two random samples, each of size 3, of salaries of graduates of the two schools, observed two years after graduation. Here are your data:

School	Salaries in 1000s of \$s		
UT-Austin	25	27	33
Brand X	21	22	28

A naive interpretation of these data would be that UT-Austin library-school graduates tend to earn higher salaries on the average, since the sample mean of their salaries is \$28,333 while the sample mean of the Brand X graduates'

salaries is only \$23,667. But you realize that this is indeed a naive interpretation and that a more rigorous analysis of the data is needed. Such an analysis is provided by the ANOVA procedure, which you decide to employ.

Using the sample data, you calculate as follows:

$$CF = \frac{(25 + 27 + 33 + 21 + 22 + 28)^2}{6} = \frac{156^2}{6} = \frac{24336}{6} = 4056$$

$$SS_T = (25^2 + 27^2 + 33^2 + 21^2 + 22^2 + 28^2) - 4056 = 4152 - 4056 = 96$$

$$SS_B = \frac{(25 + 27 + 33)^2}{3} + \frac{(21 + 22 + 28)^2}{3} - 4056 = \frac{85^2 + 71^2}{3} - 4056$$

$$= \frac{7225 + 5041}{3} - 4056 = 4088.6667 - 4056 = 32.6667$$

Next, you put these results into an ANOVA table, as follows:

	SS	DF	MS	F
T (total)	96	5		
B (between)	32.6667	1	$\frac{32.6667}{1} = 32.6667 = \hat{\sigma}_B^2$	$\frac{32.6667}{15.8333} = 2.0632 = F_{obs}$
W (within)	63.3333	4	$\frac{63.3333}{4} = 15.8333 = \hat{\sigma}_W^2$	

The threshold value of  $F_{obs}$  obtained from a table of the F distribution, with 1 and 4 degrees of freedom for the numerator and denominator respectively, is 7.71 for  $\alpha = .05$ . The observed value of the test statistic is less than the threshold, so your decision here is to accept the null hypothesis that the salaries of UT-Austin and Brand X library school graduates average the same. (Before you let yourself be too disheartened by this outcome, remember that these samples are extremely small.)

### ANOVA by Computers

As noted earlier, the ANOVA technique was invented in the 1920s. The ANOVA table, as presented in above, was developed early in the development of the overall ANOVA technique, as a way of simplifying the calculations--a major desideratum in those days when the best available computational aids were mere mechanical or electromechanical calculators. Because people have become accustomed to using the ANOVA table, and because it furnishes a concise way of displaying several related results, most computer programs that carry out ANOVA make use of the ANOVA table for output.

Some such programs rearrange the table. For example, here is how SPSS rearranges it.

ANOVA					
SALARY					
	Sum of	df	Mean	F	Sig.
	Squares		Square		
Between	32.667	1	32.667	2.063	.224
Groups					
Within	63.333	4	15.833		
Groups					
Total	96.000	5			

SPSS puts the between-groups row, the B row at the top of the table, directly underneath the column headings; the within-groups row, the W row, in the middle; and the totals row, the T row, on the bottom.

Furthermore, SPSS adds to the ANOVA table a new right-most column, labeled "Sig", in which it puts the numerical value of the "significance of the test": viz., the probability that the F-ratio will be as big as  $F_{obs}$  or bigger, by chance alone, when the null hypothesis is true. In other words, this is the probability--*when the null hypothesis is true*--of your having observed the value of the test statistic,  $F_{obs}$ , that you did observe.

Like SPSS, most computer programs that perform ANOVA include a column labeled something like "probability" or "significance," and they use this column to present the probability, calculated directly by the program, that the F-ratio will be as big as  $F_{obs}$  or bigger, by chance alone, when the null hypothesis is true. Many computer programs report similar probability results for other tests of statistical hypotheses.

The effect of knowing the probability--for the case when your null hypothesis *is* true--of your having observed the value of the test statistic that you did observe is that you can avoid going to a table of thresholds and can, instead, simply use the reported probability as the criterion by which you decide whether to accept or reject the null hypothesis.

The rule is: Accept  $H_0$  if the reported probability equals or exceeds your chosen level of significance; reject  $H_0$  only if the probability is less than your level of significance. For example, if you had decided to run an ANOVA test at the 5% = .05 level of significance, and if the entry in the "Prob" column turned out to be less than .05, you would decide to reject the null hypothesis.

The reason for this rule is that if the result you observed (viz., the observed value of the test statistic) has a large probability (scil., equal to or greater than  $\alpha$ ) of occurring when the null hypothesis is true, then you should conclude that your having observed what you did observe is something that is quite consistent with the idea that the null hypothesis is true. Since your predisposition is to believe that your null hypothesis is true, you conclude that the appropriate decision is to regard the observed result as leading you to continue to believe that the null hypothesis is true. On the other hand, if the result that you observed has a quite small probability (scil., less than  $\alpha$ ) of occurring when the null hypothesis is true, then that the fact you observed what you did observe is a good basis for deciding that the null hypothesis is false.

For example, for the data in the above numerical example of an ANOVA table, SPSS calculates a probability of .224 of observing, by chance alone, the observed F-ratio of 2.063 when the null hypothesis is true. This probability is bigger than .05, so by this method you reach the decision to accept the null hypothesis, just as you did by the method of comparing the observed F-ratio with the appropriate tabled threshold value.

As a further illustration, here is how Excel 97 displays the ANOVA results for the above example:

Anova: Single  
Factor

SUMMARY

Groups	Count	Sum	Average	Variance
Row 1	3	85	28.33333	17.33333
Row 2	3	71	23.66667	14.33333

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	32.66667	1	32.66667	2.063158	0.224248	7.70865
Within Groups	63.33333	4	15.83333			
Total		96				

You can see that Excel labels as "P-value" what SPSS calls "Sig" and that Excel also displays the threshold value (labeled "F crit", short for "critical F value") corresponding to the alpha of the test (here  $\alpha = .05$ ), calculated to 5 decimal places rather than the 2 decimal places of most tables of F. Excel also displays the mean and variance of each of the groups (here, the UT-Austin sample and the Brand X sample); this information is often useful, especially if the decision is to reject the null hypothesis that the population means are equal.