STATISTICS FOR NON-STATISTICIANS

REVIEW OF KEY CONCEPTS

Kirk Hallahan

Descriptive Statistics

Statistics Describing Relationships between Two or More Variables

Statistics Comparing Differences Between Two or More Groups

Note to Students: This handout is intended as a refresher for students who might have taken or more elementary statistics courses. It recaps basic concepts and thus provides only a cursory review of some of the most basic statistical concepts that students might use in theses or read about in journal articles.

A classic overview of introductory statistical reasoning is found in Frederick Williams, Reasoning With Statistics. Simplified Examples in Communications Research. New York: Holt, Rinehart and Winston, 1968.

Various more advanced tests might be appropriate for particular research problems. Consult your adviser.

February 2009

STATISTICS FOR NON-STATISTICIANS
PART I

Some Basic Notations

=	*Equal*	<	*Less than*	≤	*Less than/equal to*
≠	*Not equal to*	>	*More than*	≥	*More than/equal to*
				Σ	*Sum (Sigma)*

Descriptive Statistics

(*Statistics* applies to samples; *parameters* applies to populations)

Measures of Central Tendency

Frequency -- number (count) of occurrences
Mean (M) -- arithmetic average, computed by summing values of all scores, divided by number of all scores
Median (Md) -- midpoint; 50% of scores are higher; 50% of scores are lower (useful in splitting and comparing groups)
Mode (Mo) -- most frequent value or score (can be unimodal, bimodal, trimodal or multi-modal).

Mean, median and mode are same only in the case of all scores being the same, which negates the value of most research (because there is no variation to study).

Measures of Dispersion

Range -- lowest to the highest value or scores, an indication of the amount of variation that can be observed.
Percentiles, deciles, quartiles -- divisions of the scores by hundredths, tenths, quarters, respectively. Often useful for same purpose as a median split (comparison of groups).
Variance -- measure of dispersion away from a mean

Variance in a population: sigma-squared: σ^2
Variance in a sample: s-squared s^2

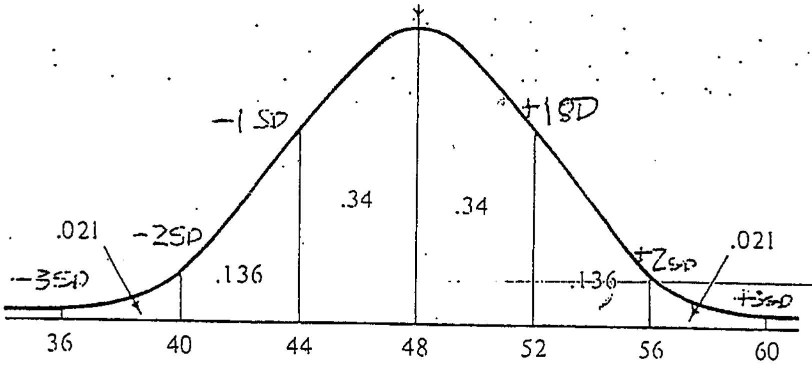
Standard deviation (SD or s.d.) -- square root of the variation.

Standard deviation in a population: sigma: σ
Standard deviation in a sample: s

This is a useful method for standardizing the variance found in any population or sample. In a normal distribution:

± 1 standard deviation = 68.26% of cases (34.13% each direction)
± 2 standard deviations = 95.46% of cases (additional 13.59%)
± 3 standard deviations = 99.87% of cases (additional .0214%)
± 3.25 standard deviations = 99.94% of cases

The following chart from Williams (1979) shows a hypothetical normal distribution, with a population mean (Mu), and population standard deviation.



The example assumes Mu=48 and the standard deviation of the population=4.0 (the variance would be 16). One standard deviation says that 68% of all scores fall between 44 and 52; two standard deviations says that 95% of all scores fall between 40 and 56.

Inferential Statistics

While we want to draw conclusions about populations in research, the data we have usually comes from a sample. Implicit in this fact is the idea that there will be *measurement error*, i.e. a particular sample will not necessarily be an accurate measure of the population mean and variance (which we don't know usually).

The statistical validity concern to researchers is the *probability* that any particular sample does not accurately reflect the population it is intended to measure. Researchers are willing to provide for a chance finding once in every 20 samples. We denote that with a probability statement:

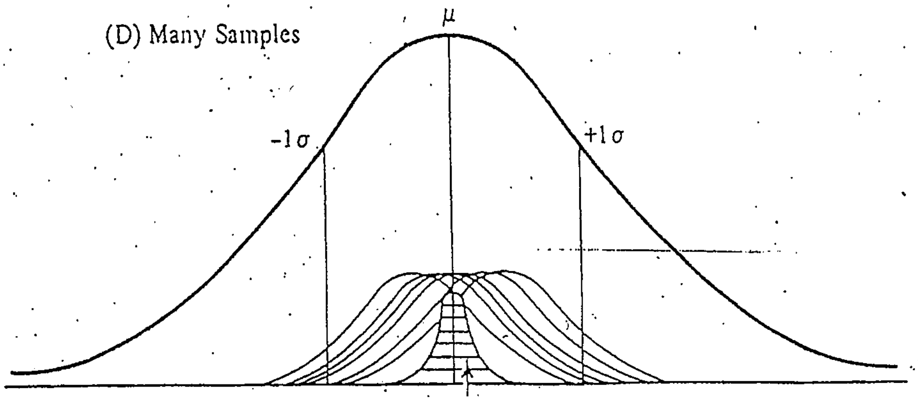
p≤.05: There is a 5% or less chance that the results obtained were purely by chance (95% confidence level).
p≤.01: The probability is reduced to one-in-100 that the results are by chance (99% confidence).
p≤.001: The probability is only one-in-1,000 that the results are by chance.

Note: Probability should <u>not</u> be construed to mean that there is a 95% chance that the results are <u>accurate</u>.

In hypothesis testing, behavioral researchers technically test for the null hypothesis (H_0), i.e. there is no significant difference that couldn't be accounted for by chance.

Some important concepts:

Sampling distribution -- The patterns of normal distributions for samples taken, compared to the hypothetical population distribution. The idea of sampling distribution suggests that over repeated attempts, samples drawn will be normally distributed.



Sampling distribution of sample means

Sampling error -- An estimate of how statistics can be expected to deviate from parameters when sampling randomly from a given population. This is calculated by first computing the *standard error:* the standard deviation (square root of variance), divided by the square root of the number of observations.

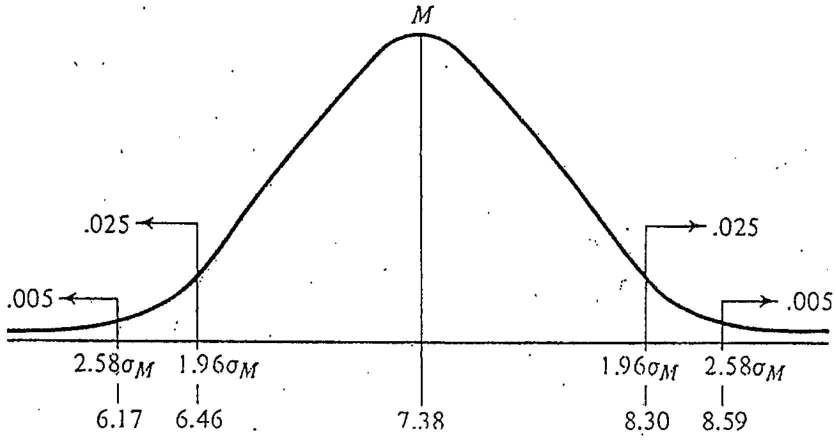
Then, depending on probability level selected, it possible to identify the *sampling error:*

For p≤=.05, multiple the standard error by 1.96 -- the standardized z-score for 95.00 compared to 95.46% (2.00) using standard deviation units.

For p≤=.01, multiple the standard error by 2.58 -- the standardized z-score for 99.00 compared to 99.87% (3.00) using standard deviation units.

Important: The size of the standard error is a direct function of the size of the sample since its square root is used in the denominator: The smaller the sample; the larger error term. To narrow the chance of error, increase sample size!

Consider this example from Williams (1979):



The example above assumes that the mean of the population is 7.38 and that the standard error (σm)=.47. The graph shows values needed to calculate the sampling error at the 95% and 99% levels:

For p≤=.05, .47 x 1.96= ±.92 sampling error

For p≤=.01, .47 x 2.58= ±1.21 sampling error

Confidence Interval. The *values* of the range where the mean is believed to fall, based on the mean obtained and the upper and lower values of the sampling error

For p≤=.05 (95% confidence interval): 7.38 ± .92 = 6.46-8.30

For p≤=.01 (99% confidence interval): 7.38 ± 1.21 = 6.17-8.59

STATISTICS FOR NON-STATISTICIANS
PART II

Some Basic Notations

0x0	*Numbers of treatments*
χ^2	*Chi-square*
ϕ	*Phi*
r	*Correlation coefficient*
r	*Coefficient of determination*

Statistics Describing Relationships Between Two or More Variables

Although describing the central tendency and variance for one variable in a population or a sample is important, more interesting questions often involve examining two or more variables at one time. This involves bivariate statistics (involving how two or more variables vary together) versus univariate statistics (the description of one variable).

General Procedures

Bivariate statistical tests involve knowing the descriptive statistics for each variable (mean, variance or standard deviation, sample size), then considering how the two vary together. You also must determine *a priori* the acceptable level of probability that any results are due to chance (p value: $p \leq .05$ says there is only one chance out of 20; $p \leq .01$ says there is only one chance out of 100; $p \leq .001$ says there is only one chance out of 1,000).

In each case, the basic procedure is to:

1) Calculate the statistic

2) Compare the result to a <u>critical value</u> that can be found in a table already compiled by statisticians. Different critical values are calculated for each probability level. Modern statistical computer programs automatically compare the results and report the probability levels for each statistic.

   Note: In some cases, it will be necessary to know that number of categories or the number of observations. This involves determining the number of *degrees of freedom* that might be associated with a particular statistic. Degrees of freedom refer to the number of scores that are free to vary once any one value is known--and usually is 1 less than the number of categories or number of observations (depending on the statistic).

Statistical Significance. If the result of the statistic calculation exceeds the critical value~~, results~~ are said to be *statistically significant* at the probability level specified (i.e. $p \leq .05$, $p \leq .01$ or $p \leq .001$).

Statistical significance suggests that results are not the result of chance but does not necessarily connote the results are accurate or true, or that a change in one variable causes change in another. However, statistical relationships between variables are necessary conditions to conclude causation.

Although there are other tables for highly specialized situations, most statistical tests involve use of one of four tables of critical values. These are tables of the critical values or distributions of:

chi-square (categorical data),
binomial distribution (probabilities of two alternative results occurring, such as coin-flipping).
t-test distribution (comparison of two means in a normal distribution), and
F-test distribution (comparison of three or more means).

Trivia: t-test is a special case of the F-test. The critical value of t is equivalent to the square root of an F statistic.

Specific Statistics

Nominal (Categorical) and Ordinal Data

Description: Contingency Tables or Cross-Tabulations -- This technique is used to compare two or more sets of categorical data. The simplest example is a 2x2 contingency table, which can be presented using frequencies or percentages.

Frequency Distribution
of Managers by Gender

	Males	Females	
Executives	160	40	**200**
Supervisors	240	60	**300**
	400	**100**	**500**

Cross-tabs can also be used to analyze combinations of categorical and ordinal measures. For example:

Frequency Distribution of
Manager Gender by Education

	Males	Females	
Graduate School	20	5	**25**
Bachelor's	180	45	**225**
Some College	180	45	**225**
High School	20	5	**25**
	400	**100**	**500**

Statistical Tests for Nominal (Categorical) or Ordinal Data. A special set of statistical tests can be performed that are predicated on the notion of comparing the frequencies of each cell to those that would be expected if no variation existed. In the 2x2 contingency above, the values of the cell are exactly what would be expected based upon the marginal frequencies of the rows and columns. (To calculate the expected value of a particular cell, multiple the row total by the column total and divide by the overall total: 200 x 400 = 80,000, divided by 500 = 160).

Chi-square (χ^2) is a test of *statistical independence* ("goodness of fit") of categorical data. To compute chi-square, the researcher calculates the differences between the *obtained value* and *expected value* in each cell, squares the result and then divides it by the expected value for that cell. The resulting values for each cell are then summed.

Using a table of χ^2 critical values, a researcher can compare results obtained in the calculation to determine whether the value obtain exceeds the critical value necessary. To obtain the critical value of chi-square, it is necessary to know the desired alpha level and the degrees of freedom (categories less 1).

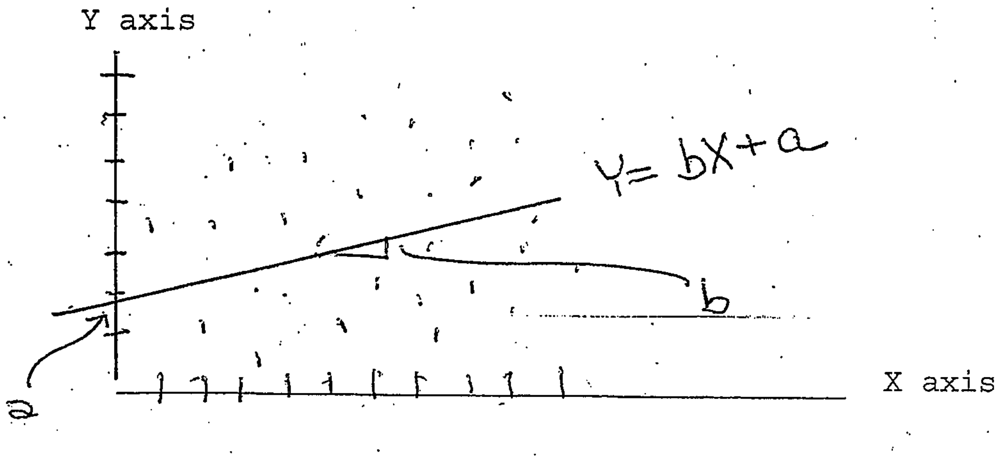
Phi (ϕ) is a test of association using the Chi-square statistic that is specifically for 2x2 contingency tables. In term of a formula, Phi is the square root of the chi-square statistic divided by the number of observations. *Cramer's V* is a more generalized chi-square based statistic that can be used for categorical situations larger than 2x2 and is based on Phi (ϕ).

Ordinal Data. Several nonparametric measures can be used to analyze ordinal or ranked data: These include: *Kendall's Tau and Tau, Gamma, Spearman's rho* (tests ordering).

Interval and Ratio Data

Most research involving interval data involves more than 30 observations. The Central Limits Theorem posits it is therefore possible to assume normality of distribution. As such, it is possible to use a series of statistical specially designed to test the relationship between two interval measures.

Description of Interval and Ratio Data: Scatter-plots are used to track to two sets of interval data, such as grade point average and college entrance test scores. For an example of such a plot.



Linear regression can be thought of a measure of the combined values of both measures. A regression line can be drawn through the data that provides "best fit" explanation of the data pattern. Given the score on one variable (X), regression allows prediction of the score or value of the second (Y). A regression line can be drawn by knowing where the line crosses the Y-axis (a) and knowing the slope of the line (b).

The general formula is: Y= bX + a, where:
Y= value you want to predict
X= value upon which you will predict Y
b= the slope of the line on the y access compared to the x axis (run versus rise: as x increases/ decreases, y increases/decreases by some proportion).
a= value where the regression line crosses the y-axis (y-intercept)

Two important variations of this concept are:

Multiple regression -- Uses two or more variables to determine which the best predictors of the value of y (or the criterion variable).

Curvilinear regression -- Assumes other than a straight line (linear) relationship, such as a U-shape or inverted U-shape.

Statistical Tests for Interval and Ratio Data: Correlation
Correlation can be thought of a measure of the association of two interval scores, based on calculating the variance. The notion is to calculate how a change in one variable affects a change in another.

The most common measure is the *Pearson product-moment correlation coefficient (Pearson r)*, which ranges in value from .00 (no relationship) to 1.00 (perfect correlation). Note: two variables can be *positively related*, i.e. an increase in one leads to an increase in the other. Or, two variables can be *negatively related*, an increase in one leads to a decrease in the other. It provides a single-figure measure of relationship.

It is rare that a perfect correlation is found. In social science, Pearson r correlations can be interpreted as follows:

less than .20:	slight (possible relationship)
.20-39:	low (some relationship)
.40-69:	moderate (substantial relationship)
.60-89:	high (marked relationship)
more than .90:	strong (definitely related)

Pearson r is calculated using the raw scores on interval data. For each observation, the scores are simply squared by themselves, also multiplied together. The sum of each set of scores, the sums of the two set of squared scores, and the sum of the multiplied scores are then added. All these results are then combined in a formula that is a variation of the formula used to calculate variance.

Again, the number of observations is critical. With a very large number of observations, somewhat small correlations can be statistically significant. With a small number of observations, however, statistical significance (a conclusion that the result is other than by chance) requires relatively high correlation numbers.

Coefficient of determination: This potentially useful statistic allows a researcher to estimate the total amount of variance in scores that this explained by knowing the strength of relationship between two interval-level variables. The coefficient of determination is calculated by squaring the correlation coefficient. For example, if r=.40, then r^2=.16 (.40x.40=.16). This suggests that 16% of the total variance in scores can be explained; the remainder of the variance must be explained by other factors, such as the influence of other factors or simple random error.

Two other ideas related to correlation:

Partial correlation -- It is possible to show more clearly the relationship between two variables by taking into account the possible influence of another variable, thus "controlling" for the influence variable. Another way to think about partial correlation is in terms of accounting for variance that might be explained by a third variable. This often results in a high correlations (and higher coefficients of determination).

Factor analysis -- A technique based on correlation that involves reducing a large number of similar items, such as items used in a scale, to distill a fewer number of underlying constructs or dimensions that are represented. These items are believed to be related because they vary together.

STATISTICS FOR NON-STATISTICIANS
PART III

Some Basic Notations

p≤	*probability less than or equal to*	*M*	*Mean (M_1, M_2 etc.)*	*k*	*any number*
		n	*no. observations*	*1Q*	*one-tailed test*
d.f.	*degrees of freedom*	σ_{diff}	*standard error of difference*	*2Q*	*two-tailed test*
t	*t statistic*			*v1*	*column d.f*
F	*F statistic*			*v2*	*row d.f.*

Statistics Comparing Differences Between Two or More Groups

Much research involves determining the means and variances for different groups of observations, and then comparing the results to see if the scores can be attributed to causes other than chance.

Such comparisons are common in survey research, where it is valuable to compare scores between demographic categories (e.g. gender). Such comparisons also are the foundation for most experiments, which involve exposing subjects to two or more levels of an independent (categorical) variable and then comparing responses on a dependent variable (usually an interval measure).

Comparing Two Means: *t*-test

The simplest differences test involves two means, and is called a Student's *t*-test. Researchers need to know: a) the means obtained, b) the variance or standard deviation for each group, or the combined variance for all groups, and c) the number of subjects in each group.

$$t = \frac{M_1 - M_2}{\sigma_{diff}}$$

where M_1 and M_2 represent the two means obtained, and σ_{diff} represents the standard error of the difference (which is calculated by knowing the variance and sample size for each sample).

Consider this example:

	Group A	**Group B**	
Means	57	52	
n	5	5	
σ_{diff}	2	2	(equal)

The $t_{obtained}$ is calculated based on the formula:

$$t = \frac{57-52}{2} = \frac{5}{2} = 2.50$$

(Note: In this example the sample sizes and standard errors for each group are the same. If this were not the case, it is necessary to compute the standard error using a weighting formula.)

To determine whether the means are different statistically, the researcher compares the $t_{obtained}=2.50$ to a critical value of *t* that can be found in a table included in most statistics books.

A *t*-table is organized based on p-values (columns) and the number of observations (rows). The latter is expressed as *degrees of freedom* (d.f. or df). To read the table:

First, choose a desired p-value: p=.05, p=.01 or p=.001.
Second, if you can predict in advance the *direction* of the difference (i.e. which numbers are higher and lower), you can choose to use a one-tailed test (which improves your chances of finding a significant difference). Otherwise, rely on the two-tailed option.
Third, determine the degrees of freedom (d.f) which apply, based on the number of observations:

$$d.f. = n_1 + n_2 - 2$$

where n_1 is the number of subjects in group 1 and n_2 is the number of subjects in group 2.
Then, go down the column (showing p-values and one- or two-tailed direction, 1Q versus 2Q) to the row showing the number of degrees of freedom (or the number that most closely approximates it). The value at the intersection is the critical value of *t*.

If the $t_{obtained}$ is larger than the $t_{critical}$, the difference is significant statistically. If less, it is probable that the result obtained was merely by chance at the p-level selected.

In our example:

From calculation:	$t_{obtained}$	=	2.50
From t table:	$t_{(8).05,2t}$	=	2.306

Because $t_{obtained}=2.50 > t_{critical}=2.306$, we can conclude the 5-point difference in means is statistically significant (only 1 in 20 odds that it was by chance at the $p \leq .05$ level using a two-tailed test).

Analysis of Variance: Fundamentals

The *t*-test is appropriate only when researchers analyze differences between two groups or two measurements of subjects within the same group. A *t*-test is a streamlined version of a more general procedure for comparing differences, *analysis of variance*.

In ANOVA, the sources of all possible combinations of variation are analyzed at the same time. The aim is to determine whether the proportion of variance accounted for by any particular variable, or combination of variables (called an *interaction*), is substantial compared to all of the remaining unexplained variance. Think of ANOVA in terms of a partitioning process: 100% of the variance can be put together, then partitioned and repartitioned (sliced and diced) in combinations.

To compute ANOVAs:

1) The variance for each score is calculated, then squared (to eliminate effects of negative and positive values). Each variance is then summed; the result is the *sum of squares*.

The SS for the *variance explained* and the SS for the *variance unexplained* (residual) add up to *total variance*.

2) Each sum of squares is divided by the applicable degrees of freedom (based on the total number of treatments related to the variable under analysis, less one). This results in a *mean square*.

3) The *F ratio* is computed by dividing the resulting mean square for each variable or combination of variables by the mean square computed for the *residual*, representing all the remaining variance.

As with the *t*-test, statistical tables can be found in statistics books that provide the critical values of *F*. However

1) Separate *F* tables exist for each probability level or p-value adopted: Upper 5 Percent Points (p=.05), Upper 1 Percent Points (p=.01), etc.
2) Two different degrees of freedom are reported and used to determine the critical value. Most *F* statistics are reported as follows:

$$F_{(1,26).05}=3.76$$

The first number in the parenthesis (v1) represents the number of treatments analyzed, calculated by multiplying the number of treatments for variable 1 less one, times the number of treatments for variable 2 less one:

Treatments	1x3 = 2 d.f.	2x2 = 1 d.f.	3x3 = 4 d.f.
	1x4 = 3 d.f.	2x3 = 2 d.f.	3x4 = 6 d.f.
	1x5 = 4 d.f.	2x4 = 3 d.f.	3x5 = 8 d.f.

The corresponding critical value is found in the applicable columns of the *F*-table.

The second number in the parenthesis (v2) represents the total number of subjects, less one, and is found in the rows of the *F*-table.

For 10 subjects: d.f. = 9
For 20 subjects: d.f. = 19 etc.

Given an F-statistic, the same general procedure is used to determine whether it is significant statistically: Compare the the $F_{obtained}$ to the $F_{critical}$ found in the table. If the value exceeds the tabled critical value, the difference is significant.

Note: The t distributions and F distributions are related: A t-statistic is the equivalent of the square root of F in the case of an ANOVA with only two means being compared ($F_{1,k}$). Thus, in the example above, $t_{(8).05,2t}=2.306$ is the same as $F_{(1,8).05,2t}=5.517$. The first d.f. quoted in an F statistic is understood to be 1 in the case of a t-test. Thus, in a t-table, v1 does not need to be specified and only v2 is used.

There are two basic types of ANOVAs: one-way and factorial.

Comparing Three or More Means For One Variable: One-Way ANOVA

The One-Way ANOVA allows a researcher to compare three or more treatments based upon one variable and to determine if they are statistically different. Operationally, a One-Way ANOVA is computed by dividing

$$F \text{ value} = \frac{MS_{bet}}{MS_{with}} \qquad \frac{\text{Mean Square Between}}{\text{Mean Square Within}}$$

where the mean square between is the variance explained related to the variable being investigated, compared to all random (unexplained) variance found within the subjects.

A One-Way ANOVA differs from the t-test because a) more than two means can be compared, but 2) only the fact that a difference exists can be determined. To determine which mean(s) is (are) different from the other(s) requires use of *multiple comparison* procedures. Multiple comparisons (often referred to as *a priori* or *post-hoc* comparisons) operate like multiple, simultaneous t-tests, but are conducted within a desired probability level.

Factorial Designs: General ANOVA Model

Analysis of variance is useful because it allows multiple variables to be analyzed simultaneously. In addition, it permits exploration of *interactions* between variables or *factors*. Consider this simple example from Hallahan's study comparing news and advertising:

Believability Scale

(Means based on 7-point scale: 1=not believable, 7=highly believable)

	News	**Ads**	**All**
Females	4.84	4.43	4.64
Males	5.06	4.76	4.91
Totals	4.96	4.62	4.79

Instead of a single ratio, multiple F ratios are computed, and are presented in a single table such as this:

F Table

	Sum of Squares	D.F.	Mean Square	F	p Signif- icance
Main Effects					
Gender	24.658	1	24.658	17.390	.000
Content Class	38.885	1	38.885	21.284	.000
Interaction					
Gender X Class	.965	1	.965	.528	.468
Explained	64.507	3	21.502	11.700	.000
Residual (Unexplained)	2387.803	1301	1.827		
Total	2441.337	1304	1.872		

Fs were computed as follows:

Gender (24.658 divided by 1.827) = $F_{(1,1301).05}=17.390$, $p\leq=.000$
Class (38.885 divided by 1.827) = $F_{(1,1301).05}=38.885$, $p\leq=.000$
Interaction (.965 divided by 1.827) = $F<1$, n.s.

Note: 1304 d.f. is based on 1,316 observations, less 11 incomplete cases, minus 1 degree of freedom.

A critical issue for researchers relates to the correct calculation of the error term (residual) used as the denominator in the F computation. Slightly different error terms are used depending on whether a *between-subjects* or *within-subjects* design is used. Most computer programs require the researcher to specify the type of design.

Some related procedures:

ANCOVA -- Analysis of covariance follows generally the same procedure, but first adjusts the scores based on some other (spurious) variable that is suspected of affecting the results. The effect of this covariate is netted out before the regular ANOVA is calculated. The effects of the covariate are examined by computing a mean square that is divided by the mean square of the residual in the same way as a regular ANOVA. The same kind of F ratio and p-values are calculated.

MANOVA -- Multivariate analysis of variance treats several variables as a single dependent measure, upon which the analysis of variance is performed. Most often MANOVA is used when the several dependent measures are very highly correlated, suggesting that the effects should be the same on each of the underlying variables. Example: In consumer behavior research, attitude toward the advertisement, attitude toward the brand, and purchase intent are inter-related ideas. All of the dependent variables are analyzed frequently as if they were the same, which eliminates the need to duplicate similar analyzes on each. However, additional v1 degrees of freedom are used when dependent variables are combined in this procedure.

MANCOVA -- Combines ANCOVA and MANOVA, allowing the researcher to control for one or more extraneous or spurious variables before conducting an analysis of variance procedure on multiple dependent measures.