Statistics with Ms Excel

- ✓ **Simple Statistics with Excel and Minitab**
- ✓ **Elementary Concepts in Statistics**
- ✓ **Multiple Regression**
- ✓ **ANOVA**

Elementary Concepts in Statistics

Overview of Elementary Concepts in Statistics. In this introduction, we will briefly discuss those elementary statistical concepts that provide the necessary foundations for more specialized expertise in any area of statistical data analysis. The selected topics illustrate the basic assumptions of most statistical methods and/or have been demonstrated in research to be necessary components of one's general understanding of the "quantitative nature" of reality (Nisbett, et al., 1987). Because of space limitations, we will focus mostly on the functional aspects of the concepts discussed and the presentation will be very short. Further information on each of those concepts can be found in the Introductory Overview and Examples sections of this manual and in statistical textbooks. Recommended introductory textbooks are: Kachigan (1986), and Runyon and Haber (1976); for a more advanced discussion of elementary theory and assumptions of statistics, see the classic books by Hays (1988), and Kendall and Stuart (1979).

Use of Excel for Statistical Analysis

Neil Cox, **Statistician, AgResearch Ruakura**
Private Bag 3123, Hamilton, New Zealand
16 May 2000

This article gives an assessment of the practical implications of deficiencies reported by McCullough and Wilson (1999) in Excel's statistical procedures. I outline what testing was done, discuss what deficiencies were found, assess the likely impact of the deficiencies, and give my opinion on the role of Excel in the analysis of data.
My overall assessment is that, while Excel uses algorithms that are not robust and can lead to errors in extreme cases, the errors are very unlikely to arise in typical scientific data analysis in AgResearch.

THE DEFICIENCIES OF EXCEL'S STATISTICAL ALGORITHMS

What Aspects Were Examined?

Excel's calculation of distributions (tail probabilities), mean and standard deviation calculations, analysis of variance, linear regression, non-linear regression (using Solver) and random numbers were scrutinised using data sets designed to reveal any shortcomings in the numerical procedures used in the calculations of statistics packages. The distributions were tested by Knusel (1998), the other aspects by McCullough and Wilson (1999). McCullough (1998, 1999) describes the methodology and the performance of SAS, SPSS and S-Plus.

How Did Excel Rate?

Generally Excel performed worse than the 3 statistics packages (SAS, SPSS, S-Plus) also examined, particularly in the non-linear regression problems. See below for more detail. The conclusion from these tests is that, in many cases, Excel uses naïve algorithms that are vulnerable to rounding and truncation errors and may produce very inaccurate results in extreme cases.

Distributions

Excel failed to give results for some discrete distributions; the failures occur when the number of cases is high and result from Excel producing, in its calculations, numbers too big to handle. The results are reliable when an answer is given. For the continuous distributions, such as the normal distribution, Excel's results for extreme tails beyond about 10^{-6} are poor; this is not normally an issue for significance testing.

Means, Standard Deviations, Analysis of Variance

Various data sets were used to check Excel's ability to get accurate results. The data sets are designed to discover whether the algorithms used are robust. For instance, the 2 data sets 90000001, 90000002, 90000003 and 1, 2, 3 have the same standard deviation (1) but Excel fails to get this answer in the first case. This is because it uses a naïve algorithm that results in subtracting two nearly equal very large numbers and the correct answer gets lost because computers store numbers with finite precision. Better algorithms avoid this problem. Excel failed to give satisfactory results in several of the more testing anova data sets but SAS (Anova) and SPSS did no better.

Linear Regression

Excel gave satisfactory results on all but one data set that had very high collinearity. SAS and SPSS report this problem and their inability to find a solution while Excel happily found a "solution" that is wrong.

Non-Linear Regression

Excel's "Solver" (an Excel add-in) was not able to give satisfactory results for several of the non-linear problems while the more sophisticated routines in the statistics packages gave satisfactory results for most of the problems. I have used Excel's Solver for a few problems and it has performed well (once I have parameterised the problem sensibly and found a reliable way of choosing starting values), giving results in close agreement to statistics packages. However, its performance for any particular application needs to be checked against a better package. And as you get no standard errors with the estimates, its usefulness is limited.

Random Number Generator

Excel's random number generator failed more of the tests of randomness than did the statistics packages examined. Hence bootstrap methods should not be used without further testing of the implications of the deficiencies in the generator. My own experience, using simulations to check difficult (for me) theoretical probability calculations, has been that the random number generator is very satisfactory.

THE IMPACT OF THESE DEFICIENCIES

In What Circumstances Will Excel be Unreliable?

- Standard deviations and statistics (eg t-tests) relying on standard deviation calculations where there are large numbers with low variation (eg see example below).
- Multiple regression with very high collinearity.
- Non-linear regression problems.
- Distribution tail areas beyond about 10^{-6}.
- Procedures (eg bootstrap) that rely on a good random number generator (I do not know whether or not the deficiencies of the generator are important here).

Will These Problems Affect You?

If you are using Excel for simple summaries, simple tests (t-tests, Chi-square, etc), regression analysis, it is most unlikely you will have any problems; Excel will give the right answers. The impact of the poorer algorithms used by Excel is less now that numbers are stored with about 15 significant digits than some years ago when numbers were often stored with only 7 significant digits. If you're dealing with very large numbers, scaling and/or re-centring your numbers can easily ensure you don't strike any rounding errors. Any serious statistics package will look after this for you; Excel does not.

CONCLUSIONS

What is Excel's Use in the Analysis of Scientific Data?

Excel is not a statistics package, more so for the very limited range of analysis tools available in it than for its naïve numerical algorithms. Nevertheless, it has a useful role in the analysis of data. Data analysis is much more than doing formal analyses and calculating P-values. When used effectively, Excel can be very useful in the exploratory analysis of data:

- viewing your data in graphs to detect errors, unusual values, trends and patterns
- summarising data with means and standard deviations

While some statistics packages have much more powerful exploratory graphing capability, Excel can often do all that is needed quite easily. Excel is of very limited use in the formal statistical analysis of data unless your experimental design is very simple. It is possible to write procedures in Excel to do more complex analyses and many people have produced statistical add-ins. Some producers of add-ins have used numerically sound procedures and have not relied on Excel's functions. However, the "Data Analysis Toolpack" provided with Excel is no easier to use than most statistics packages, has very limited capability, has known bugs and so, on the whole, is not worth bothering with. In AgResearch, we have a number of good statistics packages available and it is very easy to simply cut and paste your data into them to do formal statistical analyses. Any new statistical package (whether it be an Excel add-in or a stand-alone package) should be regarded with caution until it has been thoroughly checked out.

References

Knusel, L., (1998) On the accuracy of statistical distributions in Microsoft Excel 97. Computational Statistics and Data Analysis 26, 375-377

McCullough, B.D., (1998) Assessing the reliability of statistical software: Part I. The American Statistician 52, 358-366

McCullough, B.D., (1999) Assessing the reliability of statistical software: Part II. The American Statistician 53, 149-159

McCullough B.D. and Wilson B., (1999) On the accuracy of statistical procedures in Microsoft Excel 97. Computational Statistics and Data Analysis 31, 27-37

Simple Statistics with Excel and Minitab

http://www.bath.ac.uk/~ccsphc/excel.html

Contents

Introduction

The purpose of this article is to provide answers to some common questions about Microsoft Excel. My own interest is in the area of data presentation and analysis and I am going to concentrate on some simple statistical tests, namely regression and t-Test. I will also cover some plotting issues, particularly the summary of experimental results by means of a bar chart with error bars. These few topics cover a high proportion of recent questions at the BUCS help desk.
The article is based on Excel 5.0 and Minitab 10 for Windows, which are the versions currently available to students in the BUCS PC labs.
The first thing to say is that, in all these cases, Excel is not the best program to use. Excel is not a statistics package. We provide a good statistics packages in Minitab, which runs under Microsoft Windows (Minitab 10) and on the UNIX machines (Minitab 9.0). More powerful packages, such as Genstat, are also available. It has to be said that, for some types of graphs, Minitab can be hard work and Excel might be a better choice but for the simple types of common graphs, dealt with here, Minitab can produce the goods at the press of a button (more or less).

Installing the Analysis Toolpak.

It is possible to try some of these statistical tests using the raw functions provided by Excel, such as TTEST and LINEST. However, these functions are unfriendly and care must be taken to enter the arguments to the functions correctly. In addition, the output is uninformative. You can save yourself a lot of work by using some macros provided by Excel in the Analysis Toolpak. Look at the bottom of the Tools menu. If you do not have a Data Analysis section you need to install the Analysis Toolpak. In the student PC laboratories this is not installed by default.
In the Tools menu select Add-Ins and check the Analysis Toolpak option. Click OK and the macros will be installed. If the Analysis Toolpak does not appear as an option you will need to run setup again.
The Analysis Toolpak provides macros to perform linear regression, t-Tests, simple analysis of variance and histograms.

Regression

Regression with Excel using LINEST.

Linear Regression is fairly straightforward using the Analysis Toolpak. I will, however, describe how it is done using the LINEST function. since this will introduce the Excel array formula. The same method can be adapted to other functions which return arrays.

Entering an array formula

Many of the Excel functions return an array of output. These functions, known as array formulae, must be entered in a special way. Rather than entering the formula and pressing Enter, the formula is entered by pressing three keys at the same time; Ctrl+Shift+Enter.

If you click on the function wizard (*fx*) and choose the Statistical heading, you will see that Excel provides some 70 functions, about 7 of which concern regression. The function LINEST is the most useful for linear regression.

1. First enter your data into two columns the first of which should contain the x values. It is possible to enter the data in rows but this will make it difficult to paste the data into Minitab.
2. The LINEST function is entered as an array formula, that is it returns its output as an array the size of which depends on the number of fitted variables. For a single fitted variable an array of 2 columns by 5 rows must first be selected on the worksheet.
3. Next select a cell in which to enter the function and select the function wizard.
4. Choose the LINEST function from amongst the statistical functions and fill in the x-range by dragging the mouse over the column of x values in the worksheet. Enter the y-range in the same way. Enter the value 1 in the other two cells and click OK.
5. This function must now be entered as an array formula. Go to the function box where the full syntax of the function has been entered for you by the function wizard and press Ctrl+Shift+Enter (the control key, the shift key and Enter key at the same time). The output array will appear in the selected area of the worksheet. The output is arranged as follows (see the help entry for LINEST)

```
slope (b)                                intercept (a) standard
error of slope          standard error of intercept coefficient
of determination (r2)        standard error for y estimate (s)
Variance ratio (F)                       error degrees of
freedom Regression Sum of squares        error sum of squares
```

Plotting y against x with the fitted line.

1. Select the x and y columns (or rows) in the worksheet and, using the graph wizard, produce a scatter plot of y against x.
2. Now select the graph by double clicking on it.
3. Select the data set, by clicking on a data point.
4. From the insert menu choose the insert trend line. From the various options choose linear trend.

Linear Regression with Minitab.

Enter the data into columns, or paste the data in from Excel. Choose Regression from the statistics menu. select the columns and options and click OK. Choose the regression plot and residual plots to examine the regression fit and various residual plots. Notice the number of output options provided in Minitab.

Student's t-Test

In Excel use the Analysis Toolpak. Make sure you understand the difference between a paired t-Test and an unpaired t-Test and also decide whether you want a one or two-tailed test. If you do not have the Analysis Toolpak, for some reason, it is possible to use the TTEST function but be careful to set the correct options for the third and fourth arguments (2 and 2 for an unpaired two tailed t-Test). The only output you get is the probability.
In Minitab the t-Test is found under Basic Statistics. Various non-parametric equivalents, such as the Mann-Whitney U-test) can be found under the non-parametric section. Paired t-Tests in Minitab are carried out by subtracting the two columns and using the `TTEST` command. (or one sample t-Test from the Basic Statistics menu)

```
 MTB > let c3=c1-c2
 MTB > ttest c3

 TEST OF MU = 0.00 VS MU N.E. 0.00

              N      MEAN    STDEV   SE MEAN        T    P VALUE
 C3           5     -5.00    15.17      6.78    -0.74       0.50
```

This shows the Minitab output for a paired t-Test.

Plotting a Bar Chart with error bars.

A common way to summarise the results of an experiment with perhaps a control and several treatments, each of which has several replicates, is in a bar chart such as this one. The height of a bar represents the mean response for the several replicates for a particular treatment. The error bar in this case shows the 95% confidence limits for each mean. This chart can not be used to make any statistical decisions (unless the results are obvious, in which case you should decide not to carry out any further analysis) but it is a clear way to present the results. The appropriate statistical test is not discussed here but could be a oneway analysis of variance in some cases.

It is quite easy to produce such a bar chart with Excel, but including error bars is less straight forward. Produce the bar chart. Select the data set, as in the regression example, and choose insert error bars. Excel gives you 5 options the first four of which are of no use whatsoever. The only option which lets you put a separate error bar on each mean (what else would you want to do?) is the last one custom error bars. To use this you will need to calculate the confidence limits for each mean in a column beforehand and drag the mouse over these values to fill in the high and low boxes. Calculating the 95% confidence limits involves dividing the standard deviation by the square root n, where n is the number of measurements which the mean is based on. This is then multiplied by the appropriate value of Student's t. This is about 2 for sample sizes over 10 but rises rapidly for small samples.
Important... *Do not use Excel's CONFIDENCE function to calculate these limits. Excel always uses a value of 1.96 to calculate confidence limits. This is only valid if you know the variance of the population from which the sample is taken beforehand. This is almost never the case in practice, and will lead to serious errors for small samples.*

You can use the TINV function to look up the Student's-t for a given sample size. This particular type of plot could not be easier in Minitab. Enter the data in one column. A second column is used to index the treatment. Choose interval plot from the graph menu; and thats it. To be fair, a more complex bar chart in which bars are grouped which also required error bars would be hard work to produce in Minitab. This is a case where Excel would be quicker for most people.

Conclusion

Apart from a few simple Analysis of variance models, also provided by the Analysis Toolpak the above just about covers the full extent of Excel's statistical facilities. If you want to deal with more complex Analysis of variance models, non-parametric tests, multivariate techniques or chi-squared analysis of contingency tables you will have to use Minitab or Genstat anyway. The number of mistakes in the help files associated with Excel's statistical functions and macros and the often bizarre facilities provided in Excel make me wary of using Excel at all for statistics. In short I would strongly urge students to use Minitab instead.

http://www.statsoftinc.com/textbook/stathome.html

Multiple Regression

General Purpose

The general purpose of multiple regression (the term was first used by Pearson, 1908) is to learn more about the relationship between several independent or predictor variables and a dependent or criterion variable. For example, a real estate agent might record for each listing the size of the house (in square feet), the number of bedrooms, the average income in the respective neighborhood according to census data, and a subjective rating of appeal of the house. Once this information has been compiled for various houses it would be interesting to see whether and how these measures relate to the price for which a house is sold. For example, one might learn that the number of bedrooms is a better predictor of the price for which a house sells in a particular neighborhood than how "pretty" the house is (subjective rating). One may also detect "outliers," that is, houses that should really sell for more, given their location and characteristics.

Personnel professionals customarily use multiple regression procedures to determine equitable compensation. One can determine a number of factors or dimensions such as "amount of responsibility" (*Resp*) or "number of people to supervise" (*No_Super*) that one believes to contribute to the value of a job. The personnel analyst then usually conducts a salary survey among comparable companies in the market, recording the salaries and respective characteristics (i.e., values on dimensions) for different positions. This information can be used in a multiple regression analysis to build a regression equation of the form:

Salary = .5*Resp + .8*No_Super

Once this so-called regression line has been determined, the analyst can now easily construct a graph of the expected (predicted) salaries and the actual salaries of job incumbents in his or her company. Thus, the analyst is able to determine which position is underpaid (below the regression line) or overpaid (above the regression line), or paid equitably.

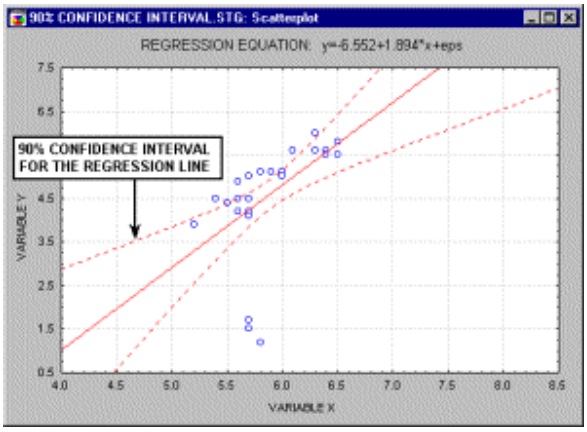
In the social and natural sciences multiple regression procedures are very widely used in research. In general, multiple regression allows the researcher to ask (and hopefully answer) the general question "what is the best predictor of ...". For example, educational researchers might want to learn what are the best predictors of success in high-school. Psychologists may want to determine which personality variable best predicts social adjustment. Sociologists may want to find out which of the multiple social indicators best predict whether or not a new immigrant group will adapt and be absorbed into society.

See also Exploratory Data Analysis and Data Mining Techniques, the *General Stepwise Regression* chapter, and the *General Linear Models* chapter.

To index

Computational Approach

The general computational problem that needs to be solved in multiple regression analysis is to fit a straight line to a number of points.



In the simplest case -- one dependent and one independent variable -- one can visualize this in a scatterplot.

- Least Squares
- The Regression Equation
- Unique Prediction and Partial Correlation
- Predicted and Residual Scores
- Residual Variance and R-square
- Interpreting the Correlation Coefficient R

See also Exploratory Data Analysis and Data Mining Techniques, the *General Stepwise Regression* chapter, and the *General Linear Models* chapter.

Least Squares. In the scatterplot, we have an independent or *X* variable, and a dependent or *Y* variable. These variables may, for example, represent IQ (intelligence as measured by a test) and school achievement (grade point average; GPA), respectively. Each point in the plot represents one student, that is, the respective student's IQ and GPA. The goal of linear regression procedures is to fit a line through the points. Specifically, the program will compute a line so that the squared deviations of the observed points from that line are minimized. Thus, this general procedure is sometimes also referred to as *least squares estimation*.

See also Exploratory Data Analysis and Data Mining Techniques, the *General Stepwise Regression* chapter, and the *General Linear Models* chapter.

The Regression Equation. A line in a two dimensional or two-variable space is defined by the equation *Y=a+b*X*; in full text: the *Y* variable can be expressed in terms of a constant (*a*) and a slope (*b*) times the *X* variable. The constant is also referred to as the *intercept*, and the slope as the *regression coefficient* or *B coefficient*. For example, GPA may best be predicted as *1+.02*IQ*. Thus, knowing that a student has an *IQ* of 130 would lead us to predict that her GPA would be 3.6 (since, 1+.02*130=3.6).

For example, the animation below shows a two dimensional regression equation plotted with three different confidence intervals (90%, 95% and 99%).

In the multivariate case, when there is more than one independent variable, the regression line cannot be visualized in the two dimensional space, but can be computed just as easily. For example, if in addition to *IQ* we had additional predictors of achievement (e.g., *Motivation, Self- discipline*) we could construct a linear equation containing all those variables. In general then, multiple regression procedures will estimate a linear equation of the form:

$$Y = a + b_1*X_1 + b_2*X_2 + ... + b_p*X_p$$

Unique Prediction and Partial Correlation. Note that in this equation, the regression coefficients (or *B* coefficients) represent the *independent* contributions of each independent variable to the prediction of the dependent variable. Another way to express this fact is to say that, for example, variable X_1 is correlated with the *Y* variable, after controlling for all other independent variables. This type of correlation is also referred to as a *partial correlation* (this term was first used by Yule, 1907). Perhaps the following example will clarify this issue. One would probably find a significant negative correlation between hair length and height in the population (i.e., short people have longer hair). At first this may seem odd; however, if we were to add the variable *Gender* into the multiple regression equation, this correlation would probably disappear. This is because women, on the average, have longer hair than men; they also are shorter on the average than men. Thus, after we remove this gender difference by entering *Gender* into the equation, the relationship between hair length and height disappears because hair length does *not* make any unique contribution to the prediction of height, above and beyond what it shares in the prediction with variable *Gender*. Put another way, after controlling for the variable *Gender*, the partial correlation between hair length and height is zero.

Predicted and Residual Scores. The regression line expresses the best prediction of the dependent variable (*Y*), given the independent variables (*X*). However, nature is rarely (if ever) perfectly predictable, and usually there is substantial variation of the observed points around the fitted regression line (as in the scatterplot shown earlier). The deviation of a particular point from the regression line (its predicted value) is called the *residual* value.

Residual Variance and R-square. The smaller the variability of the residual values around the regression line relative to the overall variability, the better is our prediction. For example, if there is no relationship between the *X* and *Y* variables, then the ratio of the residual variability of the *Y* variable to the original variance is equal to 1.0. If *X* and *Y* are perfectly related then there is no residual variance and the ratio of variance would be 0.0. In most cases, the ratio would fall somewhere between these extremes, that is, between 0.0 and 1.0. 1.0 minus this ratio is referred to as *R-square* or the *coefficient of determination*. This value is immediately interpretable in the following manner. If we have an *R-square* of 0.4 then we know that the variability of the *Y* values around the regression line is 1-0.4 times the original variance; in other words we have explained 40% of the original variability, and are left with 60% residual variability. Ideally, we would like to explain most if not all of the original variability. The *R-square* value is an indicator of how well the model fits the data (e.g., an *R-square* close to 1.0 indicates

that we have accounted for almost all of the variability with the variables specified in the model).

Interpreting the Correlation Coefficient R. Customarily, the degree to which two or more predictors (independent or *X* variables) are related to the dependent (*Y*) variable is expressed in the correlation coefficient *R*, which is the square root of *R-square*. In multiple regression, *R* can assume values between 0 and 1. To interpret the direction of the relationship between variables, one looks at the signs (plus or minus) of the regression or *B* coefficients. If a *B* coefficient is positive, then the relationship of this variable with the dependent variable is positive (e.g., the greater the IQ the better the grade point average); if the *B* coefficient is negative then the relationship is negative (e.g., the lower the class size the better the average test scores). Of course, if the *B* coefficient is equal to 0 then there is no relationship between the variables.

To index

Assumptions, Limitations, Practical Considerations

Assumption of Linearity. First of all, as is evident in the name multiple *linear* regression, it is assumed that the relationship between variables is linear. In practice this assumption can virtually never be confirmed; fortunately, multiple regression procedures are not greatly affected by minor deviations from this assumption. However, as a rule it is prudent to *always* look at bivariate scatterplot of the variables of interest. If curvature in the relationships is evident, one may consider either transforming the variables, or explicitly allowing for nonlinear components.

See also Exploratory Data Analysis and Data Mining Techniques, the *General Stepwise Regression* chapter, and the *General Linear Models* chapter.

Normality Assumption. It is assumed in multiple regression that the residuals (predicted minus observed values) are distributed normally (i.e., follow the normal distribution). Again, even though most tests (specifically the *F*-test) are quite robust with regard to violations of this assumption, it is *always* a good idea, before drawing final conclusions, to review the distributions of the major variables of interest. You can produce histograms for the residuals as well as normal probability plots, in order to inspect the distribution of the residual values.

Limitations. The major conceptual limitation of all regression techniques is that one can only ascertain *relationships*, but never be sure about underlying *causal* mechanism. For example, one would find a strong positive relationship (correlation) between the damage that a fire does and the number of firemen involved in fighting the blaze. Do we conclude that the firemen cause the damage? Of course, the most likely explanation of this correlation is that the size of the fire (an external variable that we forgot to include in

our study) caused the damage as well as the involvement of a certain number of firemen (i.e., the bigger the fire, the more firemen are called to fight the blaze). Even though this example is fairly obvious, in real correlation research, alternative causal explanations are often not considered.

Choice of the Number of Variables. Multiple regression is a seductive technique: "plug in" as many predictor variables as you can think of and usually at least a few of them will come out significant. This is because one is capitalizing on chance when simply including as many variables as one can think of as predictors of some other variable of interest. This problem is compounded when, in addition, the number of observations is relatively low. Intuitively, it is clear that one can hardly draw conclusions from an analysis of 100 questionnaire items based on 10 respondents. Most authors recommend that one should have at least 10 to 20 times as many observations (cases, respondents) as one has variables, otherwise the estimates of the regression line are probably very unstable and unlikely to replicate if one were to do the study over.

Multicollinearity and Matrix Ill-Conditioning. This is a common problem in many correlation analyses. Imagine that you have two predictors (*X* variables) of a person's height: (1) weight in pounds and (2) weight in ounces. Obviously, our two predictors are completely redundant; weight is one and the same variable, regardless of whether it is measured in pounds or ounces. Trying to decide which one of the two measures is a better predictor of height would be rather silly; however, this is exactly what one would try to do if one were to perform a multiple regression analysis with height as the dependent (*Y*) variable and the two measures of weight as the independent (X) variables. When there are very many variables involved, it is often not immediately apparent that this problem exists, and it may only manifest itself after several variables have already been entered into the regression equation. Nevertheless, when this problem occurs it means that at least one of the predictor variables is (practically) completely redundant with other predictors. There are many statistical indicators of this type of redundancy (tolerances, semi-partial *R*, etc., as well as some remedies (e.g., *Ridge regression*).

Fitting Centered Polynomial Models. The fitting of higher-order polynomials of an independent variable with a mean not equal to zero can create difficult multicollinearity problems. Specifically, the polynomials will be highly correlated due to the mean of the primary independent variable. With large numbers (e.g., Julian dates), this problem is very serious, and if proper protections are not put in place, can cause wrong results! The solution is to "center" the independent variable (sometimes, this procedures is referred to as "centered polynomials"), i.e., to subtract the mean, and then to compute the polynomials. See, for example, the classic text by Neter, Wasserman, & Kutner (1985, Chapter 9), for a detailed discussion of this issue (and analyses with polynomial models in general).

The Importance of Residual Analysis. Even though most assumptions of multiple regression cannot be tested explicitly, gross violations can be detected and should be dealt with appropriately. In particular outliers (i.e., extreme cases) can seriously bias the results by "pulling" or "pushing" the regression line in a particular direction (see the animation below), thereby leading to biased regression coefficients. Often, excluding just a single extreme case can yield a completely different set of results.

ANOVA/MANOVA

This chapter includes a general introduction to ANOVA and a discussion of the general topics in the analysis of variance techniques, including repeated measures designs, ANCOVA, MANOVA, unbalanced and incomplete designs, contrast effects, post-hoc comparisons, assumptions, etc. For related topics, see also Variance Components (topics related to estimation of variance components in mixed model designs), Experimental Design/DOE (topics related to specialized applications of ANOVA in industrial settings), and Repeatability and Reproducibility Analysis (topics related to specialized designs for evaluating the reliability and precision of measurement systems).
See also General Linear Models, General Regression Models; to analyze nonlinear models, see Generalized Linear Models.

Basic Ideas

The Purpose of Analysis of Variance

In general, the purpose of analysis of variance (ANOVA) is to test for significant differences between means. *Elementary Concepts* provides a brief introduction into the basics of statistical significance testing. If we are only comparing two means, then ANOVA will give the same results as the *t* test for independent samples (if we are comparing two different groups of cases or observations), or the *t* test for dependent samples (if we are comparing two variables in one set of cases or observations). If you are not familiar with those tests you may at this point want to "brush up" on your knowledge about those tests by reading *Basic Statistics and Tables*.

Why the name analysis of variance? It may seem odd to you that a procedure that compares means is called analysis of variance. However, this name is derived from the fact that in order to test for statistical significance between means, we are actually comparing (i.e., analyzing) variances.

- The Partitioning of Sums of Squares
- Multi-Factor ANOVA
- Interaction Effects

For more introductory topics, see the topic name.

- Complex Designs
- Analysis of Covariance (ANCOVA)
- Multivariate Designs: MANOVA/MANCOVA
- Contrast Analysis and Post hoc Tests
- Assumptions and Effects of Violating Assumptions

See also Methods for Analysis of Variance, Variance Components and Mixed Model ANOVA/ANCOVA, and Experimental Design (DOE).

The Partioning of Sums of Squares

At the heart of ANOVA is the fact that variances can be divided up, that is, partitioned. Remember that the variance is computed as the sum of squared deviations from the overall mean, divided by *n-1* (sample size minus one). Thus, given a certain n, the variance is a function of the sums of (deviation) squares, or *SS* for short. Partitioning of variance works as follows. Consider the following data set:

	Group 1	Group 2
Observation 1	2	6
Observation 2	3	7
Observation 3	1	5
Mean	2	6
Sums of Squares (SS)	2	2
Overall Mean	4	
Total Sums of Squares	28	

The means for the two groups are quite different (*2* and *6*, respectively). The sums of squares *within* each group are equal to *2*. Adding them together, we get *4*. If we now repeat these computations, ignoring group membership, that is, if we compute the total *SS* based on the overall mean, we get the number *28*. In other words, computing the variance (sums of squares) based on the within-group variability yields a much smaller estimate of variance than computing it based on the total variability (the overall mean).

The reason for this in the above example is of course that there is a large difference between means, and it is this difference that accounts for the difference in the *SS*. In fact, if we were to perform an ANOVA on the above data, we would get the following result:

	MAIN EFFECT				
	SS	**df**	**MS**	**F**	**p**
Effect	24.0	1	24.0	24.0	.008
Error	4.0	4	1.0		

As you can see, in the above table the total *SS (28)* was partitioned into the *SS* due to *within*-group variability (*2+2=4*) and variability due to differences between means (*28-(2+2)=24*).

SS Error and SS Effect. The within-group variability (*SS*) is usually referred to as *Error* variance. This term denotes the fact that we cannot readily explain or account for it in the current design. However, the *SS Effect* we *can* explain. Namely, it is due to the differences in means between the groups. Put another way, group membership *explains* this variability because we know that it is due to the differences in means.

Significance testing. The basic idea of statistical significance testing is discussed in *Elementary Concepts*. *Elementary Concepts* also explains why very many statistical test represent ratios of explained to unexplained variability. ANOVA is a good example of this. Here, we base this test on a comparison of the variance due to the between-groups variability (called *Mean Square Effect*, or MS_{effect}) with the within- group variability (called *Mean Square Error*, or Ms_{error}; this term was first used by Edgeworth, 1885). Under the null hypothesis (that there are no mean differences between groups in the population), we would still expect some minor random fluctuation in the means for the two groups when taking small samples (as in our example). Therefore, under the null hypothesis, the variance estimated based on within-group variability should be about the same as the variance due to between-groups variability. We can compare those two estimates of variance via the *F* test (see also F Distribution), which tests whether the ratio of the two variance estimates is significantly greater than 1. In our example above, that test is highly significant, and we would in fact conclude that the means for the two groups are significantly different from each other.

Summary of the basic logic of ANOVA. To summarize the discussion up to this point, the purpose of analysis of variance is to test differences in means (for groups or variables) for statistical significance. This is accomplished by analyzing the variance, that is, by partitioning the total variance into the component that is due to true random error (i.e., within- group *SS*) and the components that are due to differences between means. These latter variance components are then tested for statistical significance, and, if significant, we reject the null hypothesis of no differences between means, and accept the alternative hypothesis that the means (in the population) are different from each other.

Dependent and independent variables. The variables that are measured (e.g., a test score) are called *dependent* variables. The variables that are manipulated or controlled (e.g., a teaching method or some other criterion used to divide observations into groups

that are compared) are called *factors* or *independent* variables. For more information on this important distinction, refer to *Elementary Concepts*.

Multi-Factor ANOVA

In the simple example above, it may have occurred to you that we could have simply computed a *t* test for independent samples to arrive at the same conclusion. And, indeed, we would get the identical result if we were to compare the two groups using this test. However, ANOVA is a much more flexible and powerful technique that can be applied to much more complex research issues.

Multiple factors. The world is complex and multivariate in nature, and instances when a single variable completely explains a phenomenon are rare. For example, when trying to explore how to grow a bigger tomato, we would need to consider factors that have to do with the plants' genetic makeup, soil conditions, lighting, temperature, etc. Thus, in a typical experiment, many factors are taken into account. One important reason for using ANOVA methods rather than multiple two-group studies analyzed via *t* tests is that the former method is more *efficient*, and with fewer observations we can gain more information. Let us expand on this statement.

Controlling for factors. Suppose that in the above two-group example we introduce another grouping factor, for example, *Gender*. Imagine that in each group we have 3 males and 3 females. We could summarize this design in a 2 by 2 table:

	Experimental Group 1	Experimental Group 2
Males	2 3 1	6 7 5
Mean	2	6
Females	4 5 3	8 9 7
Mean	4	8

Before performing any computations, it appears that we can partition the total variance into at least 3 sources: (1) error (within-group) variability, (2) variability due to experimental group membership, and (3) variability due to gender. (Note that there is an additional source -- *interaction* -- that we will discuss shortly.) What would have happened had we not included *gender* as a factor in the study but rather computed a simple *t* test? If you compute the *SS* ignoring the *gender* factor (use the within-group means *ignoring* or *collapsing across gender*; the result is *SS*=10+10=20), you will see that the resulting within-group SS is larger than it is when we include gender (use the within- group, within-gender means to compute those *SS*; they will be equal to 2 in each group, thus the combined SS-within is equal to 2+2+2+2=8). This difference is due to the fact that the means for *males* are systematically lower than those for *females*, and this difference in means adds variability if we ignore this factor. Controlling for error variance increases the sensitivity (power) of a test. This example demonstrates another principal of ANOVA that makes it preferable over simple two-group t test studies: In ANOVA we can test each factor while controlling for all others; this is actually the

reason why ANOVA is more statistically powerful (i.e., we need fewer observations to find a significant effect) than the simple *t* test.

Interaction Effects

There is another advantage of ANOVA over simple *t*-tests: ANOVA allows us to detect *interaction* effects between variables, and, therefore, to test more complex hypotheses about reality. Let us consider another example to illustrate this point. (The term *interaction* was first used by Fisher, 1926.)

Main effects, two-way interaction. Imagine that we have a sample of highly achievement-oriented students and another of achievement "avoiders." We now create two random halves in each sample, and give one half of each sample a challenging test, the other an easy test. We measure how hard the students work on the test. The means of this (fictitious) study are as follows:

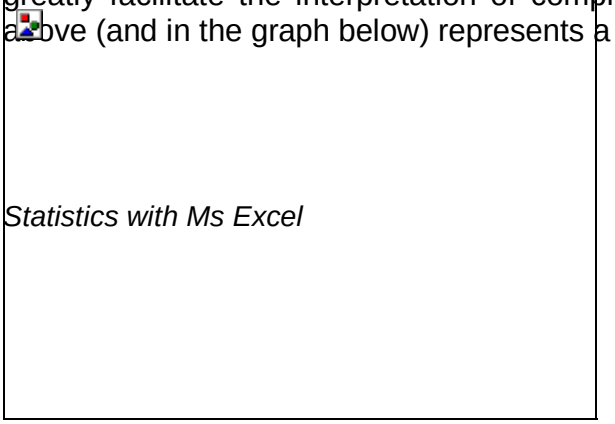
	Achievement-oriented	Achievement-avoiders
Challenging Test	10	5
Easy Test	5	10

How can we summarize these results? Is it appropriate to conclude that (1) challenging tests make students work harder, (2) achievement-oriented students work harder than achievement- avoiders? None of these statements captures the essence of this clearly systematic pattern of means. The appropriate way to summarize the result would be to say that challenging tests make only achievement-oriented students work harder, while easy tests make only achievement- avoiders work harder. In other words, the type of achievement orientation and test difficulty *interact* in their effect on effort; specifically, this is an example of a *two-way interaction* between achievement orientation and test difficulty. Note that statements 1 and 2 above describe so-called *main effects*.

Higher order interactions. While the previous two-way interaction can be put into words relatively easily, higher order interactions are increasingly difficult to verbalize. Imagine that we had included factor *Gender* in the achievement study above, and we had obtained the following pattern of means:

Females	Achievement-oriented	Achievement-avoiders
Challenging Test	10	5
Easy Test	5	10
Males	**Achievement-oriented**	**Achievement-avoiders**
Challenging Test	1	6
Easy Test	6	1

How could we now summarize the results of our study? Graphs of means for all effects greatly facilitate the interpretation of complex effects. The pattern shown in the table above (and in the graph below) represents a *three-way* interaction between factors.



Thus we may summarize this pattern by saying that for females there is a two-way interaction between achievement-orientation type and test difficulty: Achievement-oriented females work harder on challenging tests than on easy tests, achievement-avoiding females work harder on easy tests than on difficult tests. For males, this interaction is reversed. As you can see, the description of the interaction has become much more involved.

A general way to express interactions. A general way to express all interactions is to say that an effect is modified (qualified) by another effect. Let us try this with the two-way interaction above. The main effect for test difficulty is modified by achievement orientation. For the three-way interaction in the previous paragraph, we may summarize that the two-way interaction between test difficulty and achievement orientation is modified (qualified) by *gender*. If we have a four-way interaction, we may say that the three-way interaction is modified by the fourth variable, that is, that there are different types of interactions in the different levels of the fourth variable. As it turns out, in many areas of research five- or higher- way interactions are not that uncommon.

To index

Complex Designs

Let us review the basic "building blocks" of complex designs.

- Between-Groups and Repeated Measures
- Incomplete (Nested) Designs

For more introductory topics, click on the topic name.

- Basic Ideas
- Analysis of Covariance (ANCOVA)
- Multivariate Designs: MANOVA/MANCOVA
- Contrast Analysis and Post hoc Tests
- Assumptions and Effects of Violating Assumptions

See also Methods for Analysis of Variance, Variance Components and Mixed Model ANOVA/ANCOVA, and Experimental Design (DOE).

Between-Groups and Repeated Measures

When we want to compare two groups, we would use the *t* test for independent samples; when we want to compare two variables given the same subjects (observations), we would use the *t* test for dependent samples. This distinction -- dependent and independent samples -- is important for ANOVA as well. Basically, if we have repeated measurements of the same variable (under different conditions or at different points in time) *on the same subjects*, then the factor is a *repeated measures factor* (also called a *within-subjects factor*, because to estimate its significance we compute the within-subjects *SS*). If we compare different groups of subjects (e.g., males and females; three strains of bacteria, etc.) then we refer to the factor as a *between-groups factor*. The computations of significance tests are different for these different types of factors; however, the logic of computations and interpretations is the same.

Between-within designs. In many instances, experiments call for the inclusion of between-groups *and* repeated measures factors. For example, we may measure math skills in male and female students (*gender*, a between-groups factor) at the beginning and the end of the semester. The two measurements *on each student* would constitute a within-subjects (repeated measures) factor. The interpretation of main effects and interactions is not affected by whether a factor is between-groups or repeated measures, and both factors may obviously interact with each other (e.g., females improve over the semester while males deteriorate).

Incomplete (Nested) Designs

There are instances where we may decide to ignore interaction effects. This happens when (1) we know that in the population the interaction effect is negligible, or (2) when a complete *factorial* design (this term was first introduced by Fisher, 1935a) cannot be used for economic reasons. Imagine a study where we want to evaluate the effect of four fuel additives on gas mileage. For our test, our company has provided us with four cars and four drivers. A complete *factorial* experiment, that is, one in which each combination of driver, additive, and car appears at least once, would require 4 x 4 x 4 = 64 individual test conditions (groups). However, we may not have the resources (time) to run all of these conditions; moreover, it seems unlikely that the type of driver would interact with the fuel additive to an extent that would be of practical relevance. Given these considerations, one could actually run a so-called *Latin square* design and "get away" with only 16 individual groups (the four additives are denoted by letters A, B, C, and D):

	Car			
	1	**2**	**3**	**4**
Driver 1	A	B	C	D
Driver 2	B	C	D	A
Driver 3	C	D	A	B
Driver 4	D	A	B	C

Latin square designs (this term was first used by Euler, 1782) are described in most textbooks on experimental methods (e.g., Hays, 1988; Lindman, 1974; Milliken & Johnson, 1984; Winer, 1962), and we do not want to discuss here the details of how they are constructed. Suffice it to say that this design is *incomplete* insofar as not all combinations of factor levels occur in the design. For example, Driver 1 will only drive Car 1 with additive A, while Driver 3 will drive that car with additive C. In a sense, the levels of the *additives* factor (A, B, C, and D) are placed into the cells of the *car* by *driver* matrix like "eggs into a nest." This mnemonic device is sometimes useful for remembering the nature of *nested* designs.

Note that there are several other statistical procedures which may be used to analyze these types of designs; see the section on Methods for Analysis of Variance for details. In particular the methods discussed in the Variance Components and Mixed Model ANOVA/ANCOVA chapter are very efficient for analyzing designs with unbalanced nesting (when the nested factors have different numbers of levels within the levels of the factors in which they are nested), very large nested designs (e.g., with more than 200 levels overall), or hierarchically nested designs (with or without random factors).

To index

Analysis of Covariance (ANCOVA)

General Idea

The Basic Ideas section discussed briefly the idea of "controlling" for factors and how the inclusion of additional factors can reduce the error *SS* and increase the statistical power (sensitivity) of our design. This idea can be extended to continuous variables, and when such continuous variables are included as factors in the design they are called *covariates*.

- Fixed Covariates
- Changing Covariates

For more introductory topics, see the topic name.

- Basic Ideas
- Complex Designs
- Multivariate Designs: MANOVA/MANCOVA
- Contrast Analysis and Post hoc Tests
- Assumptions and Effects of Violating Assumptions

See also Methods for Analysis of Variance, Variance Components and Mixed Model ANOVA/ANCOVA, and Experimental Design (DOE).

Fixed Covariates

Suppose that we want to compare the math skills of students who were randomly assigned to one of two alternative textbooks. Imagine that we also have data about the general intelligence (IQ) for each student in the study. We would suspect that general intelligence is related to math skills, and we can use this information to make our test more sensitive. Specifically, imagine that in each one of the two groups we can compute the correlation coefficient (see *Basic Statistics and Tables*) between IQ and math skills. Remember that once we have computed the correlation coefficient we can estimate the amount of variance in math skills that is accounted for by IQ, and the amount of (residual) variance that we cannot explain with IQ (refer also to *Elementary Concepts* and *Basic Statistics and Tables*). We may use this residual variance in the ANOVA as an estimate of the true error *SS after* controlling for IQ. If the correlation between IQ and math skills is substantial, then a large reduction in the error SS may be achieved.

Effect of a covariate on the *F* test. In the *F* test (see also F Distribution), to evaluate the statistical significance of between-groups differences, we compute the ratio of the between- groups variance (MS_{effect}) over the error variance (MS_{error}). If MS_{error} becomes smaller, due to the explanatory power of IQ, then the overall *F* value will become larger.

Multiple covariates. The logic described above for the case of a single covariate (IQ) can easily be extended to the case of multiple covariates. For example, in addition to IQ, we might include measures of motivation, spatial reasoning, etc., and instead of a simple correlation, compute the multiple correlation coefficient (see *Multiple Regression*).

When the *F* value gets smaller. In some studies with covariates it happens that the *F* value actually becomes smaller (less significant) after including covariates in the design.

This is usually an indication that the covariates are not only correlated with the dependent variable (e.g., math skills), but also with the between-groups factors (e.g., the two different textbooks). For example, imagine that we measured IQ at the end of the semester, after the students in the different experimental groups had used the respective textbook for almost one year. It is possible that, even though students were initially randomly assigned to one of the two textbooks, the different books were so different that *both* math skills *and* IQ improved differentially in the two groups. In that case, the covariate will not only partition variance away from the error variance, but also from the variance due to the between- groups factor. Put another way, after controlling for the differences in IQ that were produced by the two textbooks, the math skills are not that different. Put in yet a third way, by "eliminating" the effects of IQ, we have inadvertently eliminated the true effect of the textbooks on students' math skills.

Adjusted means. When the latter case happens, that is, when the covariate is affected by the between-groups factor, then it is appropriate to compute so-called adjusted means. These are the means that one would get after removing all differences that can be accounted for by the covariate.

Interactions between covariates and factors. Just as we can test for interactions between factors, we can also test for the interactions between covariates and between-groups factors. Specifically, imagine that one of the textbooks is particularly suited for intelligent students, while the other actually bores those students but challenges the less intelligent ones. As a result, we may find a positive correlation in the first group (the more intelligent, the better the performance), but a zero or slightly negative correlation in the second group (the more intelligent the student, the less likely he or she is to acquire math skills from the particular textbook). In some older statistics textbooks this condition is discussed as a case where the assumptions for analysis of covariance are violated (see Assumptions and Effects of Violating Assumptions). However, because ANOVA/MANOVA uses a very general approach to analysis of covariance, you can specifically estimate the statistical significance of interactions between factors and covariates.

Changing Covariates

While fixed covariates are commonly discussed in textbooks on ANOVA, changing covariates are discussed less frequently. In general, when we have repeated measures, we are interested in testing the differences in repeated measurements on the same subjects. Thus we are actually interested in evaluating the significance of *changes*. If we have a covariate that is also measured at each point when the dependent variable is measured, then we can compute the correlation between the changes in the covariate and the changes in the dependent variable. For example, we could study math anxiety and math skills at the beginning and at the end of the semester. It would be interesting to see whether any changes in math anxiety over the semester correlate with changes in math skills.

To index

Multivariate Designs: MANOVA/MANCOVA

- Between-Groups Designs
- Repeated Measures Designs
- Sum Scores versus MANOVA

For more introductory topics, see the topic name.

- Basic Ideas
- Complex Designs
- Analysis of Covariance (ANCOVA)
- Contrast Analysis and Post hoc Tests
- Assumptions and Effects of Violating Assumptions

See also Methods for Analysis of Variance, Variance Components and Mixed Model ANOVA/ANCOVA, and Experimental Design (DOE).

Between-Groups Designs

All examples discussed so far have involved only one dependent variable. Even though the computations become increasingly complex, the *logic* and *nature* of the computations do not change when there is more than one dependent variable at a time. For example, we may conduct a study where we try two different textbooks, and we are interested in the students' improvements in math *and* physics. In that case, we have two dependent variables, and our hypothesis is that both together are affected by the difference in textbooks. We could now perform a multivariate analysis of variance (MANOVA) to test this hypothesis. Instead of a univariate *F* value, we would obtain a multivariate *F* value (Wilks' *lambda*) based on a comparison of the error variance/covariance matrix and the effect variance/covariance matrix. The "covariance" here is included because the two measures are probably correlated and we must take this correlation into account when performing the significance test. Obviously, if we were to take the *same* measure twice, then we would really not learn anything new. If we take a correlated measure, we gain *some* new information, but the new variable will also contain redundant information that is expressed in the covariance between the variables.

Interpreting results. If the overall multivariate test is significant, we conclude that the respective effect (e.g., textbook) is significant. However, our next question would of course be whether only math skills improved, only physics skills improved, or both. In fact, after obtaining a significant multivariate test for a particular main effect or interaction, customarily one would examine the univariate *F* tests (see also F Distribution) for each variable to interpret the respective effect. In other words, one would identify the specific dependent variables that contributed to the significant overall effect.

Repeated Measures Designs

If we were to measure math and physics skills at the beginning of the semester and the end of the semester, we would have a multivariate repeated measure. Again, the logic of significance testing in such designs is simply an extension of the univariate case. Note that MANOVA methods are also commonly used to test the significance of *univariate* repeated measures factors with more than two levels; this application will be discussed later in this section.

Sum Scores versus MANOVA

Even experienced users of ANOVA and MANOVA techniques are often puzzled by the differences in results that sometimes occur when performing a MANOVA on, for example, three variables as compared to a univariate ANOVA on the *sum* of the three variables. The logic underlying the *summing* of variables is that each variable contains some "true" value of the variable in question, as well as some random measurement error. Therefore, by summing up variables, the measurement error will sum to approximately 0 across all measurements, and the sum score will become more and more reliable (increasingly equal to the sum of true scores). In fact, under these circumstances, ANOVA on sums is appropriate and represents a very sensitive (powerful) method. However, if the dependent variable is truly multi- dimensional in nature, then summing is inappropriate. For example, suppose that my dependent measure consists of four indicators of success *in society*, and each indicator represents a completely independent way in which a person could "make it" in life (e.g., successful professional, successful entrepreneur, successful homemaker, etc.). Now, summing up the scores on those variables would be like adding apples to oranges, and the resulting sum score will not be a reliable indicator of a single underlying dimension. Thus, one should treat such data as multivariate indicators of success in a MANOVA.

To index

Contrast Analysis and Post hoc Tests

For more introductory topics, see the topic name.

See also Methods for Analysis of Variance, Variance Components and Mixed Model ANOVA/ANCOVA, and Experimental Design (DOE).

Why Compare Individual Sets of Means?

Usually, experimental hypotheses are stated in terms that are more specific than simply main effects or interactions. We may have the *specific* hypothesis that a particular textbook will improve math skills in males, but not in females, while another book would be about equally effective for both genders, but less effective overall for males. Now generally, we are predicting an interaction here: the effectiveness of the book is modified (qualified) by the student's gender. However, we have a particular prediction concerning the *nature* of the interaction: we expect a significant difference between genders for one book, but not the other. This type of specific prediction is usually tested via contrast analysis.

Contrast Analysis

Briefly, contrast analysis allows us to test the statistical significance of predicted specific differences in particular parts of our complex design. It is a major and indispensable component of the analysis of every complex ANOVA design.

Post hoc Comparisons

Sometimes we find effects in our experiment that were not expected. Even though in most cases a creative experimenter will be able to explain almost any pattern of means, it would not be appropriate to analyze and evaluate that pattern as if one had predicted it all along. The problem here is one of capitalizing on chance when performing multiple tests *post hoc*, that is, without *a priori* hypotheses. To illustrate this point, let us consider the following "experiment." Imagine we were to write down a number between 1 and 10 on 100 pieces of paper. We then put all of those pieces into a hat and draw 20 samples (of pieces of paper) of 5 observations each, and compute the means (from the numbers written on the pieces of paper) for each group. How likely do you think it is that we will find two sample means that are significantly different from each other? It is very likely! Selecting the extreme means obtained from 20 samples is very different from taking only 2 samples from the hat in the first place, which is what the test via the contrast analysis implies. Without going into further detail, there are several so-called *post hoc* tests that are explicitly based on the first scenario (taking the extremes from 20 samples), that is, they are based on the assumption that we have chosen for our comparison the most extreme (different) means out of *k* total means in the design. Those tests apply "corrections" that are designed to offset the advantage of *post hoc* selection of the most extreme comparisons.

To index

Assumptions and Effects of Violating Assumptions

- Deviation from Normal Distribution
- Homogeneity of Variances
- Homogeneity of Variances and Covariances
- Sphericity and Compound Symmetry

For more introductory topics, see the topic name.

- Basic Ideas
- Complex Designs
- Analysis of Covariance (ANCOVA)
- Multivariate Designs: MANOVA/MANCOVA
- Contrast Analysis and Post hoc Tests

See also Methods for Analysis of Variance, Variance Components and Mixed Model ANOVA/ANCOVA, and Experimental Design (DOE).

Deviation from Normal Distribution

Assumptions. It is assumed that the dependent variable is measured on at least an interval scale level (see *Elementary Concepts*). Moreover, the dependent variable should be normally distributed within groups.

Effects of violations. Overall, the *F* test (see also F Distribution) is remarkably robust to deviations from normality (see Lindman, 1974, for a summary). If the kurtosis (see *Basic Statistics and Tables*) is greater than 0, then the *F* tends to be too small and we cannot reject the null hypothesis even though it is incorrect. The opposite is the case when the kurtosis is less than 0. The skewness of the distribution usually does not have a sizable effect on the *F* statistic. If the *n* per cell is fairly large, then deviations from normality do not matter much at all because of the *central limit theorem*, according to which the sampling distribution of the mean approximates the normal distribution, regardless of the distribution of the variable in the population. A detailed discussion of the robustness of the *F* statistic can be found in Box and Anderson (1955), or Lindman (1974).

Homogeneity of Variances

Assumptions. It is assumed that the variances in the different groups of the design are identical; this assumption is called the *homogeneity of variances* assumption. Remember that at the beginning of this section we computed the error variance (*SS* error) by adding up the sums of squares within each group. If the variances in the two groups are different from each other, then adding the two together is not appropriate, and will not yield an estimate of the common within-group variance (since no common variance exists).

Effects of violations. Lindman (1974, p. 33) shows that the *F* statistic is quite robust against violations of this assumption (*heterogeneity* of variances; see also Box, 1954a, 1954b; Hsu, 1938).

Special case: correlated means and variances. However, one instance when the *F* statistic is *very misleading* is when the means are correlated with variances across cells of the design. A scatterplot of variances or standard deviations against the means will detect such correlations. The reason why this is a "dangerous" violation is the following: Imagine that you have 8 cells in the design, 7 with about equal means but one with a much higher mean. The *F* statistic may suggest to you a statistically significant effect. However, suppose that there also is a much larger variance in the cell with the highest mean, that is, the means and the variances are correlated across cells (the higher the mean the larger the variance). In that case, the high mean in the one cell is actually quite unreliable, as is indicated by the large variance. However, because the overall *F* statistic is based on a *pooled* within-cell variance estimate, the high mean is identified as significantly different from the others, when in fact it is not at all significantly different if one based the test on the within-cell variance in that cell alone.

This pattern -- a high mean and a large variance in one cell -- frequently occurs when there are *outliers* present in the data. One or two extreme cases in a cell with only 10 cases can greatly bias the mean, and will dramatically increase the variance.

Homogeneity of Variances and Covariances

Assumptions. In multivariate designs, with multiple dependent measures, the homogeneity of variances assumption described earlier also applies. However, since there are multiple dependent variables, it is also required that their intercorrelations (covariances) are homogeneous across the cells of the design. There are various specific tests of this assumption.

Effects of violations. The multivariate equivalent of the *F* test is Wilks' *lambda*. Not much is known about the robustness of Wilks' *lambda* to violations of this assumption. However, because the interpretation of MANOVA results usually rests on the interpretation of significant *univariate* effects (after the overall test is significant), the above discussion concerning univariate ANOVA basically applies, and important significant univariate effects should be carefully scrutinized.

Special case: ANCOVA. A special serious violation of the homogeneity of variances/covariances assumption may occur when covariates are involved in the design. Specifically, if the correlations of the covariates with the dependent measure(s) are very different in different cells of the design, gross misinterpretations of results may occur. Remember that in ANCOVA, we in essence perform a regression analysis within each cell to partition out the variance component due to the covariates. The homogeneity of variances/covariances assumption implies that we perform this regression analysis subject to the constraint that all regression equations (slopes) across the cells of the design are the same. If this is not the case, serious biases may occur. There are specific tests of this assumption, and it is advisable to look at those tests to ensure that the regression equations in different cells are approximately the same.

Sphericity and Compound Symmetry

Reasons for Using the Multivariate Approach to Repeated Measures ANOVA. In repeated measures ANOVA containing repeated measures factors with more than two levels, additional special assumptions enter the picture: The *compound symmetry* assumption and the assumption of *sphericity*. Because these assumptions rarely hold (see below), the MANOVA approach to repeated measures ANOVA has gained popularity in recent years (both tests are automatically computed in *ANOVA/MANOVA*). The *compound symmetry* assumption requires that the variances (pooled within-group) and covariances (across subjects) of the different repeated measures are homogeneous (identical). This is a *sufficient* condition for the univariate *F* test for repeated measures to be valid (i.e., for the reported *F* values to actually follow the *F* distribution). However, it is not a *necessary* condition. The *sphericity* assumption is a necessary and sufficient condition for the *F* test to be valid; it states that the *within-subject* "model" consists of independent (orthogonal) components. The nature of these assumptions, and the effects of violations are usually not well-described in ANOVA textbooks; in the following paragraphs we will try to clarify this matter and explain what it means when the results of the univariate approach differ from the multivariate approach to repeated measures ANOVA.

The necessity of independent hypotheses. One general way of looking at ANOVA is to consider it a *model fitting* procedure. In a sense we bring to our data a set of *a priori* hypotheses; we then partition the variance (test main effects, interactions) to test those hypotheses. Computationally, this approach translates into generating a set of contrasts (comparisons between means in the design) that specify the main effect and interaction hypotheses. However, if these contrasts are not independent of each other, then the partitioning of variances runs afoul. For example, if two contrasts *A* and *B* are identical to each other and we partition out their components from the total variance, then we take the same thing out twice. Intuitively, specifying the two (*not* independent)

hypotheses "the mean in Cell 1 is higher than the mean in Cell 2" *and* "the mean in Cell 1 is higher than the mean in Cell 2" is silly and simply makes no sense. Thus, hypotheses must be independent of each other, or *orthogonal* (the term *orthogonality* was first used by Yates, 1933).

Independent hypotheses in repeated measures. The general algorithm implemented will attempt to generate, for each effect, a set of independent (orthogonal) contrasts. In repeated measures ANOVA, these contrasts specify a set of hypotheses about *differences* between the levels of the repeated measures factor. However, if these differences are correlated across subjects, then the resulting contrasts are no longer independent. For example, in a study where we measured learning at three times during the experimental session, it may happen that the changes from time 1 to time 2 are negatively correlated with the changes from time 2 to time 3: subjects who learn most of the material between time 1 and time 2 improve less from time 2 to time 3. In fact, in most instances where a repeated measures ANOVA is used, one would probably suspect that the changes across levels are correlated across subjects. However, when this happens, the compound symmetry and sphericity assumptions have been violated, and independent contrasts cannot be computed.

Effects of violations and remedies. When the compound symmetry or sphericity assumptions have been violated, the univariate ANOVA table will give erroneous results. Before multivariate procedures were well understood, various approximations were introduced to compensate for the violations (e.g., Greenhouse & Geisser, 1959; Huynh & Feldt, 1970), and these techniques are still widely used.

MANOVA approach to repeated measures. To summarize, the problem of compound symmetry and sphericity pertains to the fact that multiple contrasts involved in testing repeated measures effects (with more than two levels) are not independent of each other. However, they do not need to be independent of each other if we use *multivariate* criteria to simultaneously test the statistical significance of the two or more repeated measures contrasts. This "insight" is the reason why MANOVA methods are increasingly applied to test the significance of univariate repeated measures factors with more than two levels. We wholeheartedly endorse this approach because it simply bypasses the assumption of compound symmetry and sphericity altogether.

Cases when the MANOVA approach cannot be used. There are instances (designs) when the MANOVA approach cannot be applied; specifically, when there are few subjects in the design and many levels on the repeated measures factor, there may not be enough degrees of freedom to perform the multivariate analysis. For example, if we have 12 subjects and $p = 4$ repeated measures factors, each at $k = 3$ levels, then the four-way interaction would "consume" $(k-1)^p = 2^4 = 16$ degrees of freedom. However, we have only 12 subjects, so in this instance the multivariate test cannot be performed.

Differences in univariate and multivariate results. Anyone whose research involves extensive repeated measures designs has seen cases when the univariate approach to repeated measures ANOVA gives clearly different results from the multivariate approach. To repeat the point, this means that the differences between the levels of the respective repeated measures factors are in some way correlated across subjects. Sometimes, this insight by itself is of considerable interest.

Methods for Analysis of Variance

Several chapters in this textbook discuss methods for performing analysis of variance. Although many of the available statistics overlap in the different chapters, each is best suited for particular applications.

General ANCOVA/MANCOVA: This chapter includes discussions of full factorial designs, repeated measures designs, mutivariate design (MANOVA), designs with balanced nesting (designs can be unbalanced, i.e., have unequal *n*), for evaluating planned and post-hoc comparisons, etc.

General Linear Models: This extremely comprehensive chapter discusses a complete implementation of the general linear model, and describes the sigma-restricted as well as the overparameterized approach. This chapter includes information on incomplete designs, complex analysis of covariance designs, nested designs (balanced or unbalanced), mixed model ANOVA designs (with random effects), and huge balanced ANOVA designs (efficiently). It also contains descriptions of six types of Sums of Squares.

General Regression Models: This chapter discusses the between subject designs and multivariate designs which are appropriate for stepwise regression as well as discussing how to perform stepwise and best-subset model building (for continuous as well as categorical predictors).

Mixed ANCOVA and Variance Components: This chapter includes discussions of experiments with random effects (mixed model ANOVA), estimating variance components for random effects, or large main effect designs (e.g., with factors with over 100 levels) with or without random effects, or large designs with many factors, when you do not need to estimate all interactions.

Experimental Design (DOE): This chapter includes discussions of standard experimental designs for industrial/manufacturing applications, including 2**(k-p) and 3**(k-p) designs, central composite and non-factorial designs, designs for mixtures, D and A optimal designs, and designs for arbitrarily constrained experimental regions.

Repeatability and Reproducibility Analysis (in the *Process Analysis* chapter): This section in the *Process Analysis* chapter includes a discussion of specialized designs for evaluating the reliability and precision of measurement systems; these designs usually include two or three random factors, and specialized statistics can be computed for evaluating the quality of a measurement system (typically in industrial/manufacturing applications).

Breakdown Tables (in the *Basic Statistics* chapter): This chapter includes discussions of experiments with only one factor (and many levels), or with multiple factors, when a complete ANOVA table is not required.