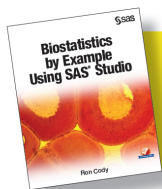


# **Biostatistics by Example Using SAS<sup>®</sup> Studio**



SAS University Edition

**Ron Cody**



From Biostatistics by Example Using  
SAS® Studio.  
Full book available for purchase [here](#).

# Contents

**About This Book vii**

**About the Author xi**

**Acknowledgments xiii**

**Chapter 1: What Is the SAS University Edition? ..... 1**

Introduction ..... 1

How to Download the SAS University Edition..... 2

Conclusions ..... 7

**Chapter 2: SAS Studio Tasks ..... 9**

Introduction ..... 9

Using the Built-in Tasks ..... 12

Taking a Tour of the Navigation Pane..... 13

Exploring the LIBRARIES Tab..... 14

Moving Columns ..... 20

Sorting Columns ..... 21

Filtering a Table (subsetting rows)..... 22

Conclusion ..... 25

**Chapter 3: Importing Data into SAS..... 27**

Introduction ..... 27

Exploring the Utilities Tab ..... 28

Importing Data from an Excel Workbook ..... 29

Listing the SAS Data Set ..... 35

Importing an Excel Workbook with Invalid SAS Variable Names ..... 37

Importing an Excel Workbook That Does Not Have Column Headings..... 38

Importing Data from a CSV File ..... 38

Shared Folders (Accessing Data from Anywhere on Your Hard Drive)..... 39

Demonstrating How to Read Data from a Shared Folder ..... 44

Conclusions ..... 45

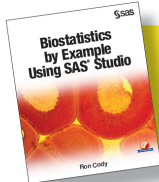
Problems ..... 45

<b>Chapter 4: Reading Data from Text Files .....</b>	<b>47</b>
Introduction .....	47
Understanding the Work Area .....	48
Some Basic Rules of SAS Programs.....	48
Writing a Program to Read a Text File Where Data Values Are Separated by Delimiters .....	49
Viewing Errors and Warnings .....	58
Reading CSV Files.....	59
Reading Text Files with Other Delimiters .....	60
Setting the Length of Character Variables .....	61
Reading Text Data in Fixed Columns.....	62
Conclusions .....	64
Problems.....	65
<b>Chapter 5: Descriptive Statistics – Univariate Analysis .....</b>	<b>67</b>
Introduction .....	67
Generating Descriptive Statistics for Continuous Variables .....	68
Investigating the Distribution for Systolic Blood Pressure.....	74
Adding a Classification Variable in the Summary Statistics Tab .....	76
Describing Categorical Variables .....	78
Editing the SAS Code Generated by the One-Way Frequencies Statistics Task .....	81
Conclusions .....	82
Problems .....	82
<b>Chapter 6: One-Sample Tests.....</b>	<b>85</b>
Introduction .....	85
Performing a One-Sample t Test.....	85
Nonparametric One-sample Tests .....	93
Conclusions .....	94
Problems .....	94
<b>Chapter 7: Two-Sample Tests .....</b>	<b>95</b>
Introduction .....	95
Unpaired t Test (t Test for Independent Groups).....	95
Nonparametric Two-sample Tests.....	101
Paired t Test.....	107
Conclusions .....	111
Problems .....	111

<b>Chapter 8: Comparing More Than Two Means (ANOVA)</b>	<b>113</b>
Introduction	113
Performing a One-Way Analysis of Variance	114
Performing a Nonparametric One-Way Tests	124
Conclusions	128
Problems	128
<b>Chapter 9: N-Way ANOVA</b>	<b>131</b>
Introduction	131
Performing a Two-Way Analysis of Variance	131
Selecting a Random Sample	131
Using the N-Way ANOVA Task	134
Interpreting the Two-Way ANOVA Results	141
Interpreting Models with Significant Interactions	143
Conclusions	145
Problems	145
<b>Chapter 10: Correlation</b>	<b>147</b>
Introduction	147
Creating a Permanent SAS Data Set	147
Reading the Exercise.xls Workbook and Creating a Permanent SAS Data Set	151
Using the Statistics Correlation Task	152
Generating Correlation and Scatter Plot Matrices	155
Interpreting Correlation Coefficients	160
Generating Spearman Non-Parametric Correlations	160
Conclusions	161
Problems	162
<b>Chapter 11: Simple and Multiple Regression</b>	<b>163</b>
Introduction	163
Describing Simple Linear Regression	164
Understanding the Diagnostic Plots	169
Demonstrating Multiple Regression	171
Demonstrating Stepwise Multiple Regression	176
Conclusions	181
Problems	182



<b>Chapter 12: Binary Logistic Regression.....</b>	<b>183</b>
Introduction .....	183
Preparing the Birth Weight Data Set for Logistic Regression.....	183
Selecting Reference Levels for Your Model.....	190
Conclusions .....	191
Problems .....	191
<b>Chapter 13: Analyzing Categorical Data .....</b>	<b>193</b>
Introduction .....	193
Describing the Heart_Attack Data Set.....	194
Computing One-Way Frequencies .....	195
Creating Formats .....	198
Producing One-Way Tables with Formats.....	200
Creating Two-Way Tables.....	201
Using Formats to Reorder the Rows and Columns of a Table.....	203
Computing Chi-Square from Frequency Data .....	206
Analyzing Tables with Low Expected Values .....	208
Conclusions .....	210
Problems .....	210
<b>Chapter 14: Computing Power and Sample Size .....</b>	<b>213</b>
Introduction .....	213
Computing Sample Size for a t Test .....	214
Calculating the Sample Size for a Test of Proportions .....	219
Computing Sample Size for a One-Way ANOVA Design.....	223
Conclusions .....	225
Problems .....	225
<b>Instructions for Problem Sets.....</b>	<b>227</b>
How to Use the Problem Set Data Files .....	227
How to Create a SAS Library .....	229
Using a SAS Data Set in the PROBLEMS Library .....	231
<b>Appendix: Solutions to the Odd-Numbered Problems.....</b>	<b>233</b>
<b>Index.....</b>	<b>241</b>



From Biostatistics by Example Using  
SAS® Studio.  
Full book available for purchase [here](#).

## Chapter 8: Comparing More Than Two Means (ANOVA)

<b>Introduction .....</b>	<b>113</b>
<b>Performing a One-Way Analysis of Variance .....</b>	<b>114</b>
<b>Performing a Nonparametric One-Way Tests .....</b>	<b>124</b>
<b>Conclusions .....</b>	<b>128</b>
<b>Problems .....</b>	<b>128</b>

---

### Introduction

When you want to compare means in a study where there are three or more groups, you cannot use multiple  $t$  tests. In the old days (even before my time!), if you had three groups (let's call them A, B, and C), you might perform  $t$  tests between each pair of means (A versus B, A versus C, and B versus C). With four groups, the situation gets more complicated; you would need six  $t$  tests (A versus B, A versus C, A versus D, B versus C, B versus D, and C versus D). Even though no one does multiple  $t$  tests anymore, it is important to understand the underlying reason why this is not statistically sound.

Suppose you are comparing four groups and performing six  $t$  tests. Also, suppose that the null hypothesis is true, and all the means come from populations with equal means. If you perform each  $t$  test with  $\alpha$  set at .05, there is a probability of .95 that you will make the correct decision—that is, to fail to reject the null hypothesis in each of the six tests. However, what is the probability that you will reject at least one of the six null hypotheses? To spare you the math, the answer is about .26 (or 26% if that is easier to think about). This is called an "experiment-wise" type I error. Remember, a type I error is when you reject the null hypothesis (claim the samples come from populations with different means—"the drug works")—when you shouldn't. So, instead of your chance of reporting a false positive result being .05, it is really .26.

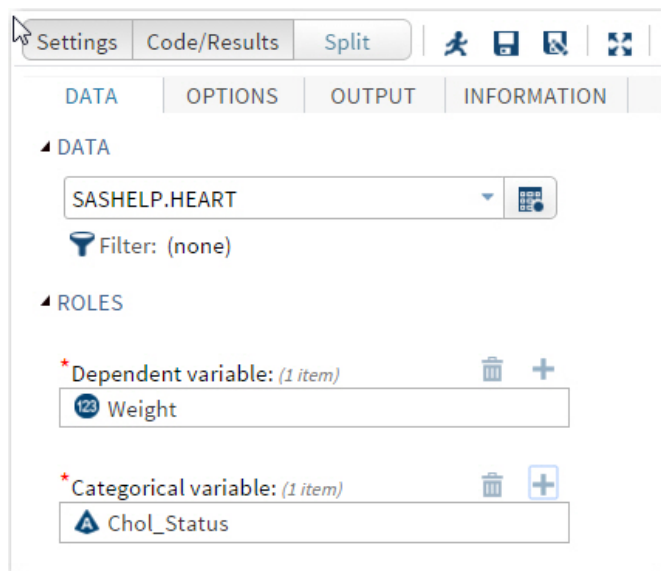
To prevent this problem, statisticians came up with a single test, called analysis of variance (abbreviated ANOVA). The null hypothesis is that all the means come from populations with the same mean; the alternative is that there is at least one pair of means that are different. You either reject or fail to reject the null hypothesis, and there is one  $p$ -value associated with the test. If you reject the null hypothesis, you can then investigate pairwise differences using methods that control the experiment-wise type I error.

## Performing a One-Way Analysis of Variance

Once again, let's start by using data from the SASHELP data set called Heart. This time you want to see if there are differences in the weight for each of the three levels of cholesterol (High, Borderline, and Desirable).

You start by choosing the task One-Way ANOVA from the statistics task list. This brings up the following screen:

**Figure 1: Data Tab for One-Way ANOVA**



The data set SASHELP.Heart was selected by clicking the icon to the right of the Data rectangle. The dependent and categorical variables (Weight and Chol\_Status, respectively) have also been selected. You may be more familiar with the term independent variable instead of categorical variable. In this context, they mean the same thing.

Once you have completed the Data screen, click the Options tab to see the following:

**Figure 2: Options for One-Way ANOVA (top portion)**

The screenshot shows the 'OPTIONS' tab in a SAS interface. It has four sub-tabs: DATA, OPTIONS, OUTPUT, and INFORMATION. The 'HOMOGENEITY OF VARIANCE' section is expanded, showing a 'Test:' dropdown menu set to 'Levene'. Below it is an unchecked checkbox for 'Welch's variance-weighted ANOVA'. The 'COMPARISONS' section is also expanded, showing a 'Comparisons method:' dropdown menu set to 'Tukey'. At the bottom, the 'Significance level:' is set to '0.05'.

One of the assumptions for performing an analysis of variance is that the variances in each of the groups are equal. The Levene test is one test that is used to determine if this assumption is reasonable. If this test is significant, you may choose to ignore it if the differences are not too large. (ANOVA is said to be robust to the assumption of equal variance, especially if the sample sizes are similar.) If you want to account for unequal variances, click the box for Welch's variance-weighted ANOVA.

Multiple comparisons are methods that we use in order to determine which pairs of means differ. There are several choices for these tests. The default is Tukey, a popular choice. Later in this chapter, you will see another multiple comparison test called SNK (Student-Newman-Keuls). You probably want to leave the significance level at .05.

Further down on the Options tab are plot options (Figure 3): You can accept the default plots or request all the plots as shown here. You also have a choice to display the diagnostic plots as a panel (several smaller graphs displayed in a grid) or as individual plots (the selection here). Finally, because the SASHELP.Heart data set has over 5,000 rows, you need to remove the 5,000-point default limit on plots to have them display correctly.

**Figure 3: Options for One-Way ANOVA (Bottom Portion)**

**▲ PLOTS**

Display plots:

Selected plots ▼

☒ Box plot

☒ Means plot

☒ LS-mean difference plot

☒ Diagnostics plot

Display as:

Individual plots ▼

Maximum number of plot points:

No limit ▼

It's time to run the procedure. Click the Run icon to produce the tables and graphs.

The first section of output displays class-level information. Don't ignore this! Make sure that the number of levels is what you expected (data errors can cause the program to believe there are more levels). Also, pay attention to the number of observations read and used. This is important because any missing values on either the dependent (Weight) or categorical (Chol\_Status) variable will result in that observation being omitted from the analysis. A large proportion of missing values in the analysis may lead to bias—subjects with missing values may be different in some way from subjects without missing values (i.e., missing values may not be random).

**Figure 4: Class-Level Information**

Class Level Information		
Class	Levels	Values
Chol_Status	3	Borderline Desirable High

Number of Observations Read	5209
Number of Observations Used	5051

You see three levels for Chol\_Status (as expected) and a relatively small number of subjects with missing values.



It's time to look at your ANOVA table (Figure 5 below):

**Figure 5: ANOVA Table**

Dependent Variable: Weight					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	42864.375	21432.188	25.90	<.0001
Error	5048	4176597.649	827.377		
Corrected Total	5050	4219462.024			

R-Square	Coeff Var	Root MSE	Weight Mean
0.010159	18.79164	28.76416	153.0689

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Chol_Status	2	42864.37515	21432.18758	25.90	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Chol_Status	2	42864.37515	21432.18758	25.90	<.0001

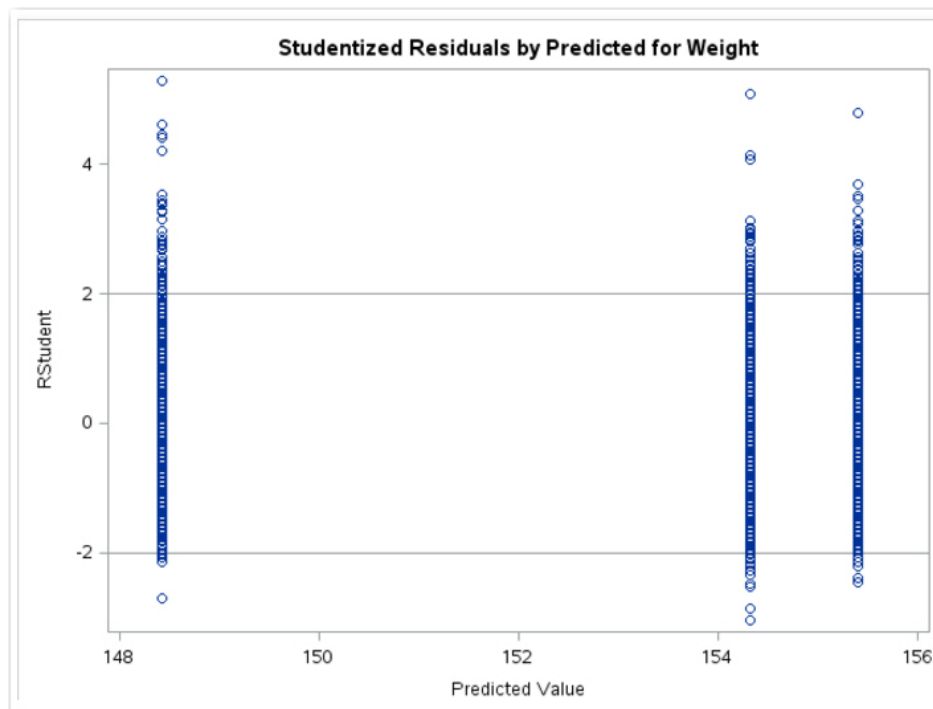
You can look at the F test and  $p$ -values in the ANOVA table, but you must remember that you also need to look at the several other parts of the output to determine if the assumptions for the test are satisfied. You will see in the diagnostic tests that follow that the ANOVA assumptions were satisfied, so let's go ahead and see what conclusions you can draw from the ANOVA table and the tables that follow.

Notice that the model has 2 degrees of freedom (because there were 3 levels of the independent variable). The mean squares for the model and error terms tell you the between-group variance and the within-group variance. The ratio of these two variances, the F value, is 25.90 with a corresponding  $p$ -value of less than .0001. A result such as this is often referred to as "highly significant." Remember, the term "significant" means that there is a low probability that one or more of the pairwise differences occurred by chance. It doesn't necessarily mean that the differences are significant in the common usage of the word, that is, important.

The next several plots are intended to help you decide if the ANOVA assumptions were satisfied and to graphically show you information about the 3 means and the distribution of scores in each of the 3 groups.

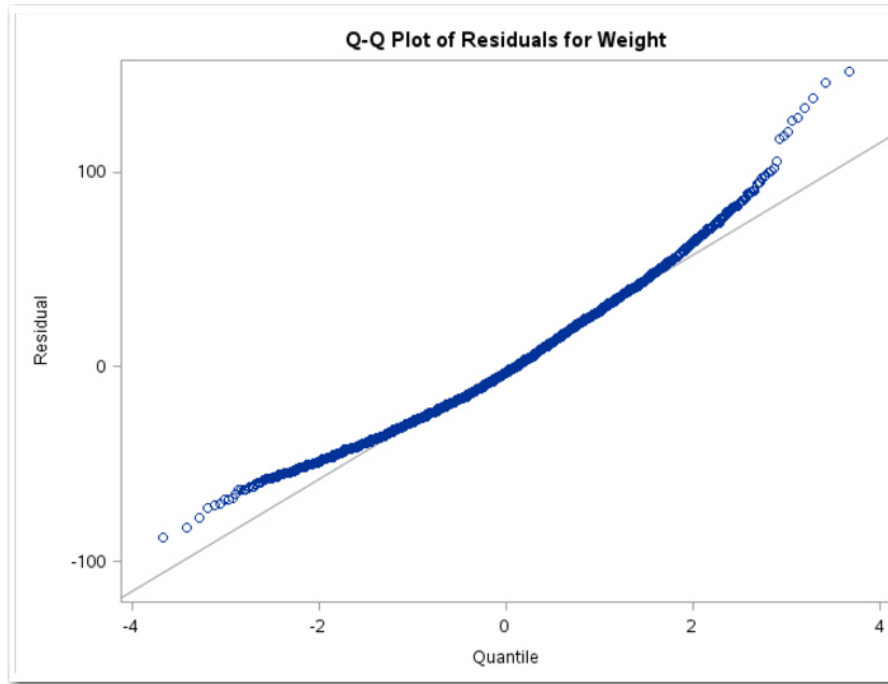
Note: The figures shown below were selected from a larger set of plots produced by the one-way ANOVA task.

The plot shown in Figure 6 shows the residuals (the differences between the mean of each group and each individual score) in that group. There are actually two residual plots produced by the one-way task. One (not shown) displays the residuals as actual scores (weights in this example). The one selected here displays the residuals as  $t$  scores (the number of standard deviations above or below the mean of the group). Both plots look very similar. You also see the predicted values (means of each group) shown on the x-axis.

**Figure 6: Residual Plot**

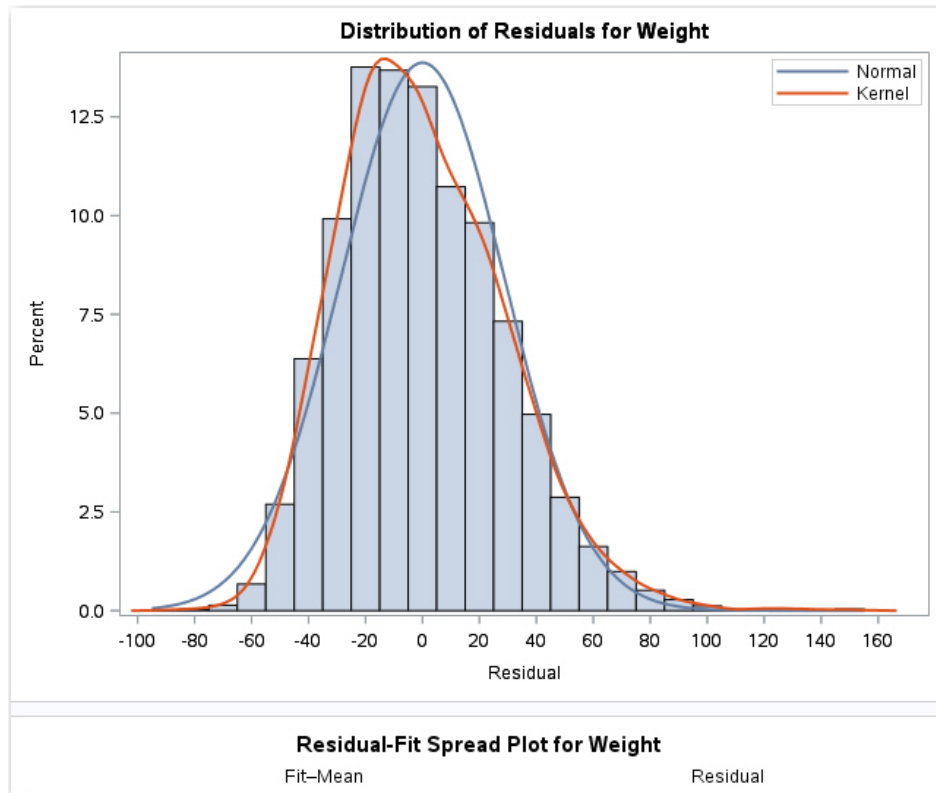
One of the assumptions for running a one-way ANOVA is that the errors (the residuals are estimates of these errors) are normally distributed. You have seen Q-Q plots earlier in this book, so you remember that data values that are normally distributed appear as a straight line on a Q-Q plot. The plot shown in Figure 7 shows small deviations from a straight line, but not enough to invalidate the analysis.

Figure 7: Q-Q Plot for Residuals



The residuals are also displayed as a histogram (see Figure 8):

**Figure 8: Histogram for Residuals**



To graphically display the distribution of weights in the 3 groups, the one-way ANOVA task produces a box plot (Figure 9). The line in the center of the box represents the median, and the small diamond represents the mean. Notice that the means, as well as the medians, of the three groups are not very different. Why then were the results so highly significant? The reason is the large (over 5,000) sample size. Large sample sizes give you high power to see even small differences.

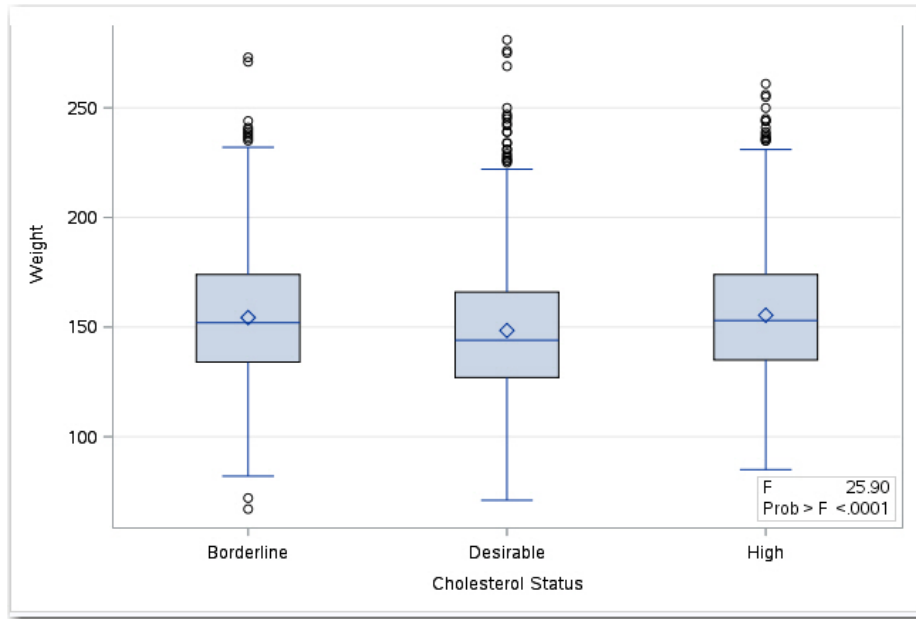
**Figure 9: Box Plot for Weight by Cholesterol Level**

Figure 10 shows the results for Levin's test of homogeneity of variance. Here, the null hypothesis is that the variances are equal. Because the  $p$ -value is .2194, you do not reject the null hypothesis of equal variance.

**Figure 10: Levin's Test for Homogeneity of Variance**

Levene's Test for Homogeneity of Weight Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Chol_Status	2	5399505	2699752	1.52	0.2194
Error	5048	8.9813E9	1779186		

Figure 11 show the means and standard deviations for the three groups.

**Figure 11: Group Means and Standard Deviations**

Level of Chol_Status	N	Weight	
		Mean	Std Dev
Borderline	1860	154.318280	28.5982126
Desirable	1403	148.431219	29.6364336
High	1788	155.408277	28.2367277



Because this is a one-way model, the least square means shown in Figure 12 are equal to the means in the previous figure. In unbalanced models with more than one factor, this may not be the case.

Below the table showing the three means, you see  $p$ -values for all of the pairwise differences. Each of the three cholesterol groups in the top table in the figure has what is labeled as the LSMEAN Number. In the table of  $p$ -values, the LSMEAN number is used to identify the groups. The intersection of any two groups displays the  $p$ -value for the difference. For example, group 1 (Borderline) and group 2 (Desirable) show a  $p$ -value of less than .0001. The  $p$ -value for the difference of Borderline (1) and High (3) is .4869 (not significant).

**Figure 12: Least Square Means**

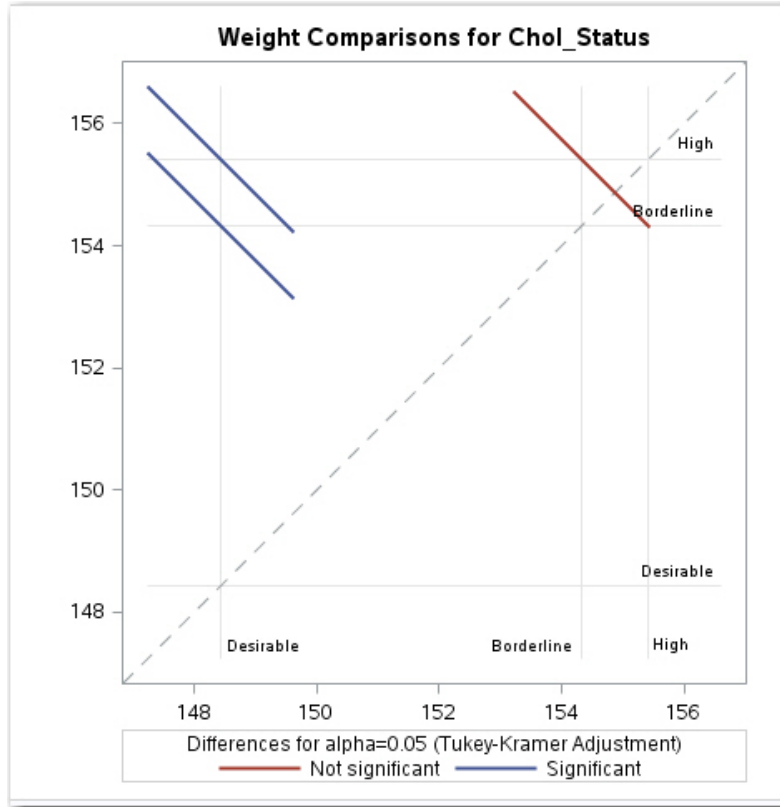
Least Squares Means		
Adjustment for Multiple Comparisons: Tukey-Kramer		
Chol_Status	Weight LSMEAN	LSMEAN Number
Borderline	154.318280	1
Desirable	148.431219	2
High	155.408277	3

Least Squares Means for effect Chol_Status			
Pr >  t  for H0: LSMean(i)=LSMean(j)			
Dependent Variable: Weight			
i/j	1	2	3
1		<.0001	0.4869
2	<.0001		<.0001
3	0.4869	<.0001	

Figure 13 shows a very clever way to display pairwise differences. At the intersection of any two groups, you see a diagonal line representing a 95% confidence interval for the difference between the two group means. If the interval crosses the main diagonal line (that represents no difference), the two group means are not significantly different at the .05 level. To make this clearer, significant differences are shown in blue and non-significant differences are shown in red.

Figure 13: Pairwise Comparison of Means



All of the previous figures were generated by the choices that you made in the Data and Options tabs. The last figure (below) shows an alternative method of determining pairwise differences, called the Student-Newman-Keuls test (also referred to in some texts as just Newman-Keuls). The SNK (the abbreviation for this test) test is similar to the Tukey test in that it shows group means and which pairs of means are different at the .05 level. The Tukey test has the advantage of computing  $p$ -values for each pair of means as well as a confidence interval for the differences. The SNK test can do neither of these two things but has a slightly higher power to detect differences. The SNK display shows the three means in order from highest to lowest. To the left of the means is a column labeled SNK Grouping. Any two means that have the same grouping letter are not significantly different. You can see here that the mean weights for the cholesterol groups High and Borderline are not significantly different (they both have As in the grouping column). The mean weight for the Desirable group is significantly different from the other two groups (it has a B in the grouping column).

**Figure 14: Student-Newman-Keuls Pairwise Comparisons**

Means with the same letter are not significantly different.			
SNK Grouping	Mean	N	Chol_Status
A	155.4083	1788	High
A			
A	154.3183	1860	Borderline
B	148.4312	1403	Desirable

## Performing a Nonparametric One-Way Tests

If you feel that the distribution assumptions are not satisfied by your data, another statistical task, Nonparametric One-Way analysis, provides a host of alternate tests. To demonstrate this, let's go back to the SASHELP data set called Fish and compare the weights of three species of fish.

This exercise also provides you with a demonstration of an alternate way of filtering data. Rather than creating the filter directly in the statistics task as you did in Chapter 7, you can use a Filter Data task under the list of Data tasks. To this end, let's add Bream to the weight comparison of Pike and Roach. You may find this method easier than having to write your own filter expression—you create a filter by choosing items in menus.

In the navigation pane, from the Task list, select Data ► Filter Data. This brings up the following:

**Figure 15: Creating a Filter with a Data Task**

▲ FILTER 1

\* Variable 1: (1 item)

▲ Species

Comparison:

Equal

Value type:

Enter a value

Enter a value

Select distinct value

Logical:

(none)

You selected Species as the first variable, Equal as the comparison, and Select a distinct value as the Value type. This brings up a list of all the species in the Fish data set. It looks like this:

**Figure 16: Selecting a Distinct Value for Species**

The screenshot shows a software interface for selecting a distinct value for Species. The interface includes the following elements:

- Variable 1:** (1 item) Species
- Comparison:** Equal
- Value type:** Select distinct value
- Value:** Bream
- Logical:**
  - (none)
  - (none)
  - AND
  - OR

A mouse cursor is pointing at the OR option in the Logical dropdown menu.

Because you want to add Roach and Pike to this list, select OR as your logical operator. This enables you to repeat the filtering process adding the other two species to the data set. Finally, on the tab labeled Output, select a name for your output data set (Three\_Fish was used in this example), and select which variables you want in the output data set (Species and Weight were selected here). Now, run the task.

This is certainly more tedious than simply writing a WHERE clause as you did in Chapter 7, but, by presenting you with lists of species, it helps avoid spelling or syntax errors.

It's time to run the Nonparametric One-Way Statistic task. The opening screen looks like this:

**Figure 17: Opening Screen of the Nonparametric One-Way Task**

The screenshot shows the 'DATA' tab of the Nonparametric One-Way Task interface. The 'DATA' section has a dropdown menu set to 'WORK.THREE\_FISH' and a 'Filter: (none)' option. The 'ROLES' section has two items: 'Dependent variable: (1 item)' with 'Weight' selected, and 'Classification variable: (1 item)' with 'Species' selected. There is a checkbox for 'Missing values are a valid level' which is currently unchecked. At the bottom, there is a section for 'ADDITIONAL ROLES'.

The data set Three\_Fish is selected, along with Weight as the Dependent variable and Species as the Classification variable. For this example, you are using all the default values except for a request for multiple comparisons that you decided to check (see Figure 18 below):

**Figure 18: Requesting a Multiple Comparison Test**

The screenshot shows the 'Additional Tests' section. It contains two checkboxes. The first checkbox, 'Empirical distribution function tests, including Kolmogorov-Smirnov and Cramer-von Mises tests, or the Kuiper test (for two-sample data)', is unchecked. The second checkbox, 'Pairwise multiple comparison analysis (asymptotic only)', is checked.



You are ready to run the analysis. Below are selected portions of the output:

**Figure 19: Wilcoxon Rank Sums and Kruskal-Wallis ANOVA Table**

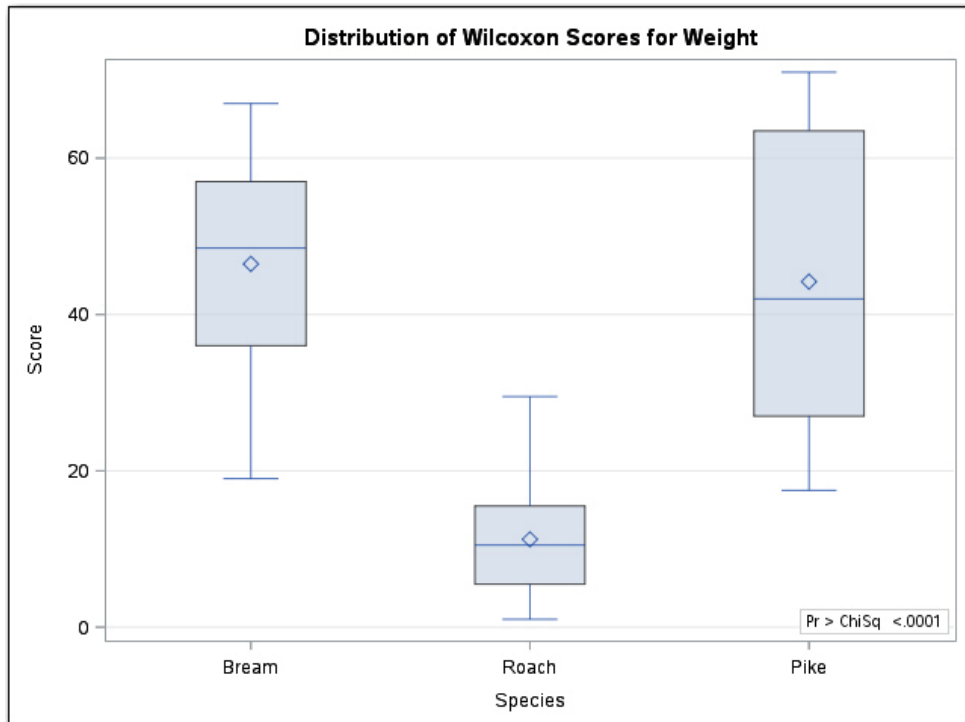
Wilcoxon Scores (Rank Sums) for Variable Weight Classified by Variable Species					
Species	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Bream	34	1580.00	1224.0	86.852273	46.470588
Roach	20	224.50	720.0	78.206158	11.225000
Pike	17	751.50	612.0	74.192876	44.205882
Average scores were used for ties.					

Kruskal-Wallis Test	
Chi-Square	40.2791
DF	2
Pr > Chi-Square	<.0001

Looking at the results of the Kruskal-Wallis test, you decide that the fish weights are not all equal ( $p < .0001$ ). Box plots are shown next:

**Figure 20: Box Plots for Fish Weights**



It looks like Roach are much lighter than either Bream or Pike. However, to determine which pairs of fish are unequal, look at the final piece of output (Figure 21) to see the  $p$ -values for each of the pairs. You see that the comparisons Bream versus Roach and Roach versus Pike are significantly different while the comparison of Bream versus Pike is not. Exactly what you would have guessed from the box plot.

**Figure 21: Pairwise Comparisons**

Pairwise Two-Sided Multiple Comparison Analysis			
Dwass, Steel, Critchlow-Fligner Method			
Variable: Weight			
Species	Wilcoxon Z	DSCF Value	Pr > DSCF
Bream vs. Roach	5.9671	8.4388	<.0001
Bream vs. Pike	0.4599	0.6504	0.8900
Roach vs. Pike	-4.9544	7.0066	<.0001

---

## Conclusions

You have seen how to conduct a one-way analysis of variance as well as a Kruskal-Wallis nonparametric test. You have also seen ways to determine if the two assumptions for a one-way ANOVA (normally distributed data and homogeneity of variance) are met. Finally, you saw an alternative way to filter data using the Filter Data task.

---

## Problems

8-1: Starting with the workbook Blood\_Pressure.xls, create a temporary SAS data set called BP. Use this data set to perform a one-way ANOVA, testing the three drugs' effects on SBP (systolic blood pressure). What is the overall  $p$ -value for the test? Using the Tukey (default) method of multiple comparisons, what do you conclude about the three drug levels (Placebo, Drug A, and Drug B)?

8-2: Repeat problem 8-1, except start with the SAS data set Blood\_Pressure.sas7bdat, which is located in the folder c:\SASUniversityEdition\myfolders\Problems. You may need to review the instructions describing the problem sets to see how to create a library.

8-3: Starting with the Diabetes.xls workbook, create a SAS data set called Diabetes. Test if there is a relationship between how often a person drinks diet drinks (variable Diet\_Drinks) and the glucose level. What is the overall  $p$ -value for the ANOVA; test if there are any pairwise differences. If so, what are they, and what are the  $p$ -values?

8-4: Repeat problem 8-3, except request the SNK (Student-Newman-Keuls) multiple comparison test. Because this test has a slightly high power to detect group differences, is the difference between the levels Rarely and Sometimes significant (at the .05 level)?

8-5: Using the SASHELP data set BMT, test if the T values are different for each of the three groups. What is the overall  $p$ -value, and which groups, if any, are significantly different at the .05 level?

8-6: You have measured the left ventricular ejection fraction (LVEF) on three groups of subjects with congestive heart failure (CHF). LVEF is the percentage of blood volume that is pumped from the left ventricle with each contraction. The three groups represent 1) Placebo, 2) Calcium channel blocker, and 3) Lasix. The experiment resulted in the following:

```
Placebo: 55 58 62 48 57 57 80 40 55 52
Calcium: 57 65 55 78 57 84 72 80 78 81
Lasix:   60 60 65 67 48 62 64 70 57 40
```

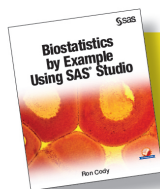
Run the program below to create the CHF data set. The variables in this data set are Subj, Group (Placebo, Calcium, or Lasix), and LVEF. There will be a short explanation following the program:

```
1. data CHF;
2.     do Group = 'Placebo','Calcium','Lasix';
3.         do Subj = 1 to 10;
4.             input LVEF @@;
5.             output;
6.         end;
7.     end;
8. datalines;
   55 58 62 48 57 57 80 40 55 52
   57 65 55 78 57 84 72 80 78 81
   60 60 65 67 48 62 64 70 57 40
;
```

The program starts with a DATA statement (1). Line 2 demonstrates a DO loop with character values. Group is first set to 'Placebo'. Then another DO loop creates a Subj variable with values from 1 to 10 (line 3). For each combination of Group and Subj, you read in a value for LVEF. The @@ on line 4 enables you to place several observations on a single line of data. Without the @@ on the INPUT statement, the program would go to a new line of data for each input. You finish each DO loop with an END statement. Finally, in line 8, you see a DATALINES statement. This enables you to enter the data value directly in the SAS program, avoiding the effort of first creating a text file and then using an INFILE statement to tell the program where to read the data values.

Run a one-way ANOVA comparing LVEF for each of the three groups. Include a test for Tukey multiple comparisons.

From *Biostatistics by Example Using SAS® Studio*, by Ron Cody. Copyright © 2016, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.



From *Biostatistics by Example Using SAS® Studio*.  
Full book available for purchase [here](#).



# Index

## A

- accessing data from hard drive 39–44
- Akaike's information criteria (AIC) 177, 188
- analysis of variance (ANOVA)
  - about 113
  - N-Way 131–144
  - performing nonparametric one-way tests 124–128
  - performing one-way 114–124
- analyzing
  - categorical data 193–210
  - tables with low expected values 208–210
- asterisk (\*) 52

## B

- backward slashes 29
- binary logistic regression
  - about 183
  - preparing birth weight data set for 183–190
  - selecting reference levels for models 190–191
- blank lines, in SAS statements 48
- blanks, in folder names 42
- BOOK library 14
- BOOKDATA library 148, 150–151, 164, 184, 195–196
- box plots, requesting 71–73
- built-in tasks 12–13

## C

- calculating
  - See also* power
  - See also* sample size
  - chi-square from frequency data 206–208
  - one-way frequencies 195–197
  - sample size for a test of proportions 219–223
  - sample size for one-way ANOVA design 223–225
- case sensitivity
  - in folder names 41
  - on Microsoft platforms 29
- categorical data
  - about 193
  - analyzing tables with low expected values 208–210

- computing chi-square from frequency data 206–208
- computing one-way frequencies 195–197
- creating formats 198–200
- creating one-way tables with formats 200
- creating two-way tables 201–203
- Heart\_Attack data set 194
- reordering table rows and columns using formats 203–206
- categorical variables 78–81
- CHANGE button 34
- character variables, setting length of 61–62
- chi-square, computing from frequency data 206–208
- CLASS statement 190
- classification variables, adding in Summary Statistics tab 76–78
- code, editing generated by One-Way Frequencies Statistics task 81–82
- CODE area 13, 48
- Cody, Ron
  - An Introduction to SAS University Edition* 1, 47, 64, 199
  - Learning SAS by Example* 47, 64, 199
- columns
  - See also* variables
  - importing Excel workbooks with no column headings 38
  - moving 20
  - resizing 19
  - sorting 21–22
- COMMENT statement 52, 55, 184
- comparing more than two means
  - See* analysis of variance (ANOVA)
- confidence interval 70–71, 99
- continuous variables, generating descriptive statistics for 68–73
- Cook's D 171
- correlation
  - about 147
  - creating permanent SAS data sets 147–152
  - creating Correlation and scatter plot matrices 155–160
  - creating Spearman non-parametric correlations 160–161



- interpreting correlation coefficients 160
- reading Exercise.xls workbook 151–152
- using Statistics Correlation task 152–154
- correlation coefficients, interpreting 160
- Correlation matrix 155–160
- CSV files
  - about 29
  - importing data from 38–39
  - reading 59–60
- CUMFREQPLOT 82

**D**

- data
  - accessing from hard drive 39–44
  - importing 27–45
  - importing Excel workbooks with invalid SAS variable names 37
  - importing from CSV files 38–39
  - listing SAS Data set 35–36
  - reading from shared folders 44–45
  - reading from text files 47–64
  - Utilities tab 28–29
  - values, separated by delimiters 49–58
- data sets 35–36
  - See also* tables
- DATA statement 55, 199
- DATA tab 68, 74, 76, 79
- DATASETS procedure 38
- delimiters
  - data values separated by 49–58
  - reading text files with other 60–61
- descriptive statistics
  - about 67
  - adding classification variables in Summary Statistics tab 76–78
  - categorical variables 78–81
  - distribution for systolic blood pressure 74–76
  - editing SAS code generated by One-Way Frequencies Statistics task 81–82
  - generating for continuous variables 68–73
- diagnostic plots 169–171
- distribution, for systolic blood pressure 74–76
- DLM= option 60–61
- dollar sign (\$) 55
- downloading SAS University Edition 2–6
- DSD option 60

**E**

- EDIT button 82
- errors, viewing 58–59
- Excel workbooks
  - about 29
  - importing data from 29–35

- importing with invalid SAS variable names 37
- importing with no column headings 38
- Exercise.xls workbook, reading 151–152

**F**

- F* test 99
- File Save icon 58
- files
  - case sensitivity and names of 29
  - CSV 29
  - selecting to import 31
  - XLS 29
- filtering
  - rows 23
  - tables 22–24
- filters, removing 23
- Fisher's Exact test 208–210
- fixed columns, reading text data in 62–64
- folders, shared 39–44
- FORMAT procedure 198–200
- FORMAT statement 199
- formats
  - creating 198–200
  - creating one-way tables with 200
  - reordering table rows/columns using 203–206
- FREQ procedure 81, 203–204
- FREQPLOT 82
- frequency data, computing chi-square from 206–208
- F*-value 139

**H**

- hard drive, accessing data from 39–44
- HEART library 15
- Heart\_Attack data set 194
- histograms 71–73, 98, 120
- horizontal scroll bars 18

**I**

- Import Utility 60, 86–87, 108
- importing
  - data 27–45, 38–39
  - data from CSV files 38–39
  - Excel workbooks with invalid SAS variable names 37
  - Excel workbooks with no column headings 38
- INFILE statement 55, 60–61, 64
- Information Center screen 12
- informats 63
- INPUT statement 55, 61–62, 63

interpreting  
     correlation coefficients 160  
     models with significant interactions 143–144  
     results of N-Way ANOVA 141–143  
*An Introduction to SAS University Edition* (Cody)  
     1, 47, 64, 199

## K

Kruskal-Wallis test 127  
 kurtosis 75

## L

*Learning SAS by Example* (Cody) 47, 64, 199  
 length, setting of character variables 61–62  
 LENGTH statement 61–62  
 Levin's test of homogeneity of variance 121  
 LIBNAME statement 148, 150  
 LIBRARIES tab 14–19, 35  
 linear regression 164–169  
 lines, in SAS statements 48  
 listing SAS Data set 35–36  
 LOG area 13, 48  
 logistic regression 183–191  
 LSMEANS 142–143

## M

Menu icon 53  
 models, selecting reference levels for 190–191  
 moving columns 20  
 multiple regression  
     about 163  
     demonstrating 171–175  
     stepwise 176–181  
 MYFMTS library 14

## N

names, in SAS 49  
 Navigation pane 13  
 NOCUM option 81–82  
 nonparametric one-sample tests 93  
 nonparametric one-way tests 124–128  
 nonparametric two-sample tests 101–107  
 N-Way ANOVA  
     about 131  
     interpreting models with significant  
       interactions 143–144  
     interpreting results of 141–143  
     performing two-way ANOVA 131–141  
     using 134–141

## O

one-sample tests  
     about 85

    nonparametric 93  
     performing one-sample *t* tests 85–93  
 one-way ANOVA  
     computing sample size for 223–225  
     performing 114–124  
 one-way frequencies 79, 81–82, 195–197  
 one-way tables 200  
 OPTIONS tab 33, 69–70, 74, 79  
 Oracle VM VirtualBox 1–2  
 ORDER= option 203–204

## P

PAD option 64  
 paired *t* test 107–111  
 Pearson correlation coefficient 153–154  
 permanent SAS data sets, creating 147–152  
 PLOT statement 224  
 PLOTS= option 81  
 plus (+) sign 23, 41  
 power 213  
 POWER procedure 223–225  
 Preferences menu 53  
 PRINT procedure 185, 199  
 Programmer Mode 13  
 programs rules 48–49  
*p*-value 98–99, 121–122, 139

## Q

Q-Q plot 74–75, 91–92, 98, 118–119, 170  
 QUIT statement 48

## R

random samples, selecting 131–134  
 reading  
     CSV files 59–60  
     data from shared folders 44–45  
     data from text files 47–64  
     Exercise.xls workbook 151–152  
     text data in fixed columns 62–64  
     text files 49–58  
     text files with other delimiters 60–61  
 reference levels, selecting for models 190–191  
 regression, multiple 171–175, 176–181  
 removing filters 23  
 reordering table rows/columns using formats 203–  
     206  
 resizing columns 19  
 Resources link 12  
 restarting virtual machines 43  
 RESULTS area 13, 48, 56

## rows

*See* observations

filtering 23

subsetting 22–24

R-square 173–174, 177

RUN icon 36, 56, 71, 103, 105

RUN statement 48, 55, 185, 199

**S**

## sample size

about 213

calculating for a test of proportions 219–223

computing for one-way ANOVA designs  
223–225

computing for *t* tests 214–218

## SAS Studio

*See also specific topics*

about 9–12

opening 12

using built-in tasks 12–13

## SAS University Edition

about 1–2

downloading 2–6

SASHELP library 15, 68, 148

SBC (Schwarz Bayesian Information Criterion)  
176–181

SC (Schwarz Criterion) 188–189

scatter plot matrices 155–160

Schwarz Bayesian Information Criterion (SBC)  
176–181

Schwarz Criterion (SC) 188–189

scroll bars 18

semicolon (;) 48

SET statement 184, 199

shared folders 39–45

## simple regression

about 163

diagnostic plots 169–171

linear 164–169

skewness 75

sorting columns 21–22

Spearman non-parametric correlations 160–161

Start SAS Studio button 12

Statistics Correlation task 153–154

Statistics tab 79

Statistics Task menu 68

stepwise multiple regression 176–181

Student-Newman-Keuls test 123

subsetting rows 22–24

Summary Statistics tab, adding classification  
variables in 76–78

systolic blood pressure, distribution for 74–76

**T***t* tests

computing sample size for 214–218

for independent groups 95–101

one-sample 85–93

## tables

analyzing with low expected values 208–210

filtering 22–24

reordering rows/columns using formats 203–  
206

TABLES statement 81

test of proportions, calculating sample size for a  
219–223

text data, reading in fixed columns 62–64

## text files

reading 49–58

reading data from 47–64

reading with other delimiters 60–61

TITLE statement 82

## two-sample tests

about 95

nonparametric 101–107

paired *t* test 107–111

unpaired *t* test 95–101

two-way ANOVA 131–141

two-way tables 201–203

**U**

underscore (\_) 49

## univariate analysis

*See* descriptive statistics

unpaired *t* test 95–101

Update icon 12

Utilities tab 28–29

**V**

VALUE statement 198–200

## values

analyzing tables with low expected 208–210

selecting 24

variable names, importing Excel workbooks with  
invalid 37

variance inflation factor (VIF) 173–175

vertical scroll bars 18

## virtual computer

defined 1–2

restarting 43

starting 9

VirtualBox 9–11

VMware Fusion 2

VMware Player 2

VMware Workstation Player 2

## **W**

warnings, viewing 58–59  
WHERE statement 102, 125, 184  
Wilcoxon Rank Sum Test 105  
Wilcoxon Signed Rank test 93  
work area 13, 48  
WORK library 14, 34, 36, 147–148, 176, 202  
workbooks, Excel 29

## **X**

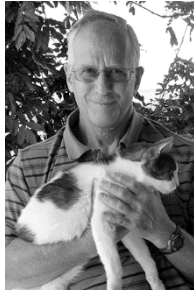
XLS files 29

## **Y**

Yesno format 199



## About The Author



Ron Cody, EdD, a retired professor from the Robert Wood Johnson Medical School now works as a private consultant and a national instructor for SAS Institute Inc. A SAS user since 1977, Ron's extensive knowledge and innovative style have made him a popular presenter at local, regional, and national SAS conferences. He has authored or co-authored numerous books, including *Learning SAS by Example: A Programmer's Guide*, and, *An Introduction to SAS University Edition*; as well as countless articles in medical and scientific journals.

*"Ron and Mickey" photo by Jan Cody*

Learn more about this author by visiting Ron Cody's author page at <http://support.sas.com/publishing/authors/cody.html>. There you can download free book excerpts, access example code and data, read the latest reviews, get updates, and more.



# About This Book

---

## Purpose

SAS University Edition and its user interface, SAS Studio, have become very popular. SAS Studio can also be used with standard versions of SAS, perhaps as an alternative to SAS Enterprise Guide. SAS Studio includes built-in tasks for importing data from external files (Excel, for example) and, best of all, it includes point-and-click statistical tasks for just about any statistical study. This book will benefit the reader by doing the following:

1. Providing step-by-step instructions, complete with screen shots, of how to use SAS University Edition and SAS Studio to perform most statistical queries.
2. Discussing the theory behind each statistical test, with emphasis on the assumptions that need to be satisfied before running each test.
3. Helping the reader negotiate some of the trickier aspects of running SAS University Edition. For example, this book explains how to access files on a local hard drive and make them available on the virtual machine where SAS is running.
4. Providing a detailed explanation of the output produced by each statistical procedure.
5. Presenting practice problems (with solutions to the odd-numbered problems).

---

## Is This Book for You?

The audience for this book consists mostly of students in a statistics or a biostatistics class. Although the book uses biostatistical examples, students in other classes such as educational or business statistics will also benefit. In addition, data analysts in the pharmaceutical industry will also find valuable information in this book.

---

## Prerequisites

There is NO prerequisite for readers of this book. It is written with the assumption that the reader has never used SAS before.

---

## Scope of This Book

The first section of this book explains how to install the SAS University Edition and the virtualization software needed to run it. Readers of this book may also be using SAS Studio with a standard edition of SAS (as opposed to the SAS University Edition).

Subsequent chapters describe how to import data from a variety of sources such as Excel workbooks and CSV files.



Following these chapters is a chapter on using the SAS Studio tasks to perform descriptive statistics, the first step in almost any data analysis project.

Most of the remaining chapters cover all the basic statistical tests commonly used in biostatistical analysis.

A final chapter is devoted to sample size and power calculations. This topic is not usually covered in a book of this type, even though it is a very important topic.

After reading this book, you will be able to understand temporary and permanent SAS data sets and how to create them from various data sources. The reader will also be able to use SAS Studio Statistics tasks to generate descriptive statistics for continuous and categorical data.

The inferential statistics portion of the book covers the following:

- Paired and unpaired  $t$  tests
- One-way analysis of variance
- N-Way ANOVA
- Correlation
- Simple and multiple regression
- Logistic regression
- Categorical data analysis
- Power and sample size calculations

---

## About the Examples

---

### Software Used to Develop the Book's Content

All of the statistical tasks described and demonstrated in this book are available to anyone using SAS Studio, either as part of SAS University Edition or as an interface to standard versions of SAS.

---

### SAS University Edition

If you are using SAS University Edition, you can use the code and data sets provided with this book. This helpful link will get you started:

[http://support.sas.com/publishing/import\\_ue.data.html](http://support.sas.com/publishing/import_ue.data.html).

---

### Output and Graphics Used in This Book

<If needed, add a description of the output and graphics used in this book.>

<Tell the reader how the output and graphics were generated.>

---

## Exercise Solutions

Each chapter, starting with Chapter 3, includes a set of problems for the reader to test his or her skills. Solutions to the odd-numbered problems are included in the book—solutions to the even-numbered problems are available from SAS Institute, on request.

---

## Additional Help

Although this book illustrates many analyses regularly performed in businesses across industries, questions specific to your aims and issues may arise. To fully support you, SAS Institute and SAS Press offer you the following help resources:

- For questions about topics covered in this book, contact the author through SAS Press:
  - Send questions by email to [saspress@sas.com](mailto:saspress@sas.com); include the book title in your correspondence.
  - Submit feedback on the author's page at [http://support.sas.com/publishing/bbu/companion\\_site/info.html](http://support.sas.com/publishing/bbu/companion_site/info.html).
- For questions about topics in or beyond the scope of this book, post queries to the relevant SAS Support Communities at <https://communities.sas.com/>.
- SAS Institute maintains a comprehensive website with up-to-date information. One page that is particularly useful to both the novice and the seasoned SAS user is its Knowledge Base. Search for relevant notes in the "Samples and SAS Notes" section of the Knowledge Base at <http://support.sas.com/resources>.
- Registered SAS users or their organizations can access SAS Customer Support at <http://support.sas.com>. Here you can pose specific questions to SAS Customer Support; under *Support*, click *Submit a Problem*. You will need to provide an email address to which replies can be sent, identify your organization, and provide a customer site number or license information. This information can be found in your SAS logs.

---

## Keep in Touch

We look forward to hearing from you. We invite questions, comments, and concerns. If you want to contact us about a specific book, please include the book title in your correspondence.

---

### Contact the Author through SAS Press

- By email: [saspress@sas.com](mailto:saspress@sas.com)
- Via the web: [http://support.sas.com/publishing/bbu/companion\\_site/info.html](http://support.sas.com/publishing/bbu/companion_site/info.html)

---

### Purchase SAS Books

For a complete list of books available through SAS, visit [sas.com/store/books](http://sas.com/store/books).

- Phone: 1-800-727-0025
- Email: [sasbook@sas.com](mailto:sasbook@sas.com)

---

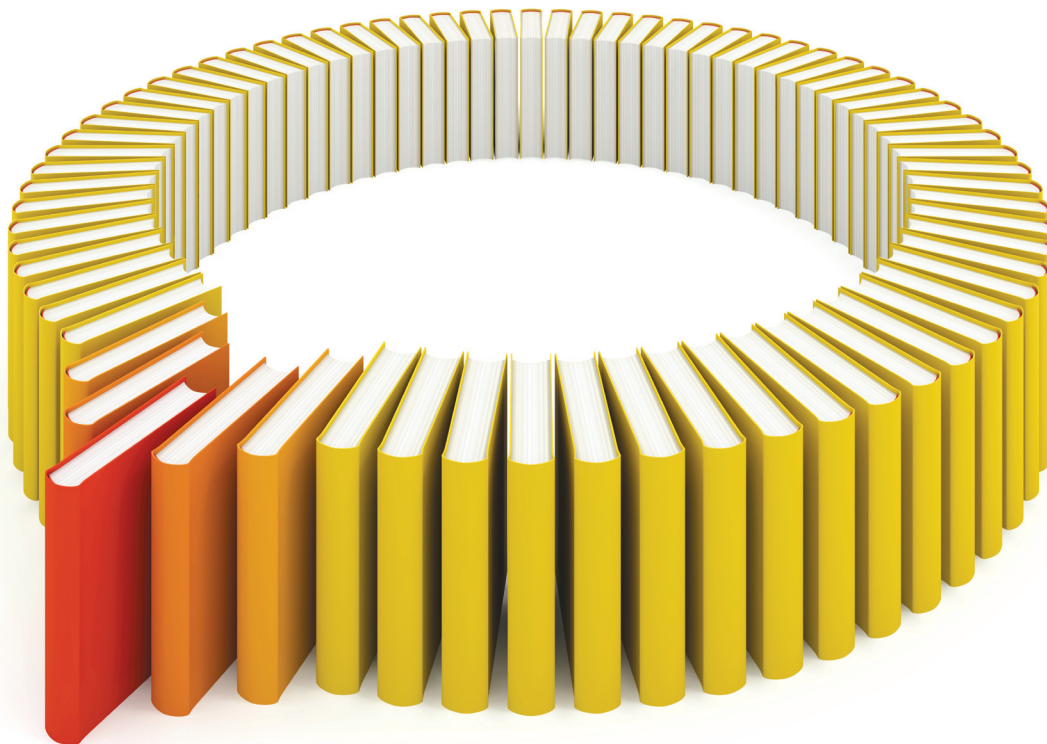
## **Subscribe to the SAS Learning Report**

Receive up-to-date information about SAS training, certification, and publications via email by subscribing to the SAS Learning Report monthly eNewsletter. Read the archives and subscribe today at <http://support.sas.com/community/newsletters/training!>

---

## **Publish with SAS**

SAS is recruiting authors! Are you interested in writing a book? Visit <http://support.sas.com/publishing/publish/index.html> for more information.



# Gain Greater Insight into Your SAS® Software with SAS Books.

Discover all that you need on your journey to knowledge and empowerment.

 [support.sas.com/bookstore](https://support.sas.com/bookstore)  
for additional books and resources.

 **sas**  
THE POWER TO KNOW®

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies. © 2013 SAS Institute Inc. All rights reserved. S107969US.0413