SAS® Press

# SAS® Statistics by Example

Ron Cody

sas

From *SAS® Statistics by Example*. Full book available for purchase here.

# Contents

List of Programs ..... ix
Acknowledgments ..... xv

**Chapter 1 An Introduction to SAS ..... 1**

Introduction ..... 2
What is SAS ..... 2
Statistical Tasks Performed by SAS ..... 3
The Structure of SAS Programs ..... 3
SAS Data Sets ..... 3
SAS Display Manager ..... 4
Excel Workbooks ..... 5
Variable Types in SAS Data Sets ..... 11
Temporary versus Permanent SAS Data Sets ..... 11
Creating a SAS Data Set from Raw Data ..... 12
Data Values Separated by Delimiters ..... 12
Reading CSV Files ..... 14
Data Values in Fixed Columns ..... 15
Excel Files with Invalid SAS Variable Names ..... 16
Other Sources of Data ..... 17
Conclusions ..... 17

**Chapter 2 Descriptive Statistics – Continuous Variables ..... 19**

Introduction ..... 19
Computing Descriptive Statistics Using PROC MEANS ..... 21
Descriptive Statistics Broken Down by a Classification Variable ..... 23
Computing a 95% Confidence Interval and the Standard Error ..... 25
Producing Descriptive Statistics, Histograms, and Probability Plots ..... 26

Changing the Midpoint Values on the Histogram ..... 32
Generating a Variety of Graphical Displays of Your Data ..... 33
Displaying Multiple Box Plots for Each Value of a Categorical Variable ..... 38
Conclusions ..... 39

**Chapter 3 Descriptive Statistics – Categorical Variables ..... 41**

Introduction ..... 41
Computing Frequency Counts and Percentages ..... 42
Computing Frequencies on a Continuous Variable ..... 44
Using Formats to Group Observations ..... 45
Histograms and Bar Charts ..... 48
Creating a Bar Chart Using PROC SGPLOT ..... 49
Using ODS to Send Output to Alternate Destinations ..... 50
Creating a Cross-Tabulation Table ..... 52
Changing the Order of Values in a Frequency Table ..... 53
Conclusions ..... 55

**Chapter 4 Descriptive Statistics – Bivariate Associations ..... 57**

Introduction ..... 57
Producing a Simple Scatter Plot Using PROG GPLOT ..... 58
Producing a Scatter Plot Using PROC SGPLOT ..... 61
Creating Multiple Scatter Plots on a Single Page Using PROC SGSCATTER ..... 63
Conclusions ..... 68

**Chapter 5 Inferential Statistics – One-Sample Tests ..... 69**

Introduction ..... 69
Conducting a One-Sample *t*-test Using PROC TTEST ..... 70
Running PROC TTEST with ODS Graphics Turned On ..... 71
Conducting a One-Sample *t*-test Using PROC UNIVARIATE ..... 74
Testing Whether a Distribution is Normally Distributed ..... 76

Tests for Other Distributions ..... 78
Conclusions ..... 78

**Chapter 6 Inferential Statistics – Two-Sample Tests ..... 79**

Introduction ..... 79
Conducting a Two-Sample *t*-test ..... 79
Testing the Assumptions for a *t*-test ..... 81
Customizing the Output from ODS Statistical Graphics ..... 83
Conducting a Paired *t*-test ..... 85
Assumption Violations ..... 87
Conclusions ..... 89

**Chapter 7 Inferential Statistics – Comparing More than Two Means ..... 91**

Introduction ..... 91
A Simple One-way Design ..... 92
Conducting Multiple Comparison Tests ..... 98
Using ODS Graphics to Produce a Diffogram ..... 101
Two-way Factorial Designs ..... 102
Analyzing Factorial Models with Significant Interactions ..... 106
Analyzing a Randomized Block Design ..... 108
Conclusions ..... 109

**Chapter 8 Correlation and Regression ..... 111**

Introduction ..... 111
Producing Pearson Correlations ..... 112
Generating a Correlation Matrix ..... 115
Creating HTML Output with Data Tips ..... 117
Generating Spearman Nonparametric Correlations ..... 119
Running a Simple Linear Regression Model ..... 120
Using ODS Statistical Graphics to Investigate Influential Observations ..... 125

Using the Regression Equation to Do Prediction ..... 129
A More Efficient Way to Compute Predicted Values ..... 132
Conclusions ..... 134

**Chapter 9 Multiple Regression ..... 135**

Introduction ..... 135
Fitting Multiple Regression Models ..... 136
Running All Possible Regressions with *n* Variables ..... 137
Producing Separate Plots Instead of a Panel ..... 141
Choosing the Best Model ($C_p$ and Hocking's Criteria) ..... 142
Forward, Backward, and Stepwise Selection Methods ..... 145
Forcing Selected Variables into a Model ..... 152
Creating Dummy (Design) Variables for Regression ..... 153
Detecting Collinearity ..... 155
Influential Observations in Multiple Regression Models ..... 157
Conclusions ..... 161

**Chapter 10 Categorical Data ..... 163**

Introduction ..... 163
Comparing Proportions ..... 164
Rearranging Rows and Columns in a Table ..... 166
Tables with Expected Values Less Than 5 (Fisher's Exact Test) ..... 169
Computing Chi-Square from Frequency Data ..... 172
Using a Chi-Square Macro ..... 173
A Short-Cut Method for Requesting Multiple Tables ..... 175
Computing Coefficient Kappa—A Test of Agreement ..... 175
Computing Tests for Trends ..... 178
Computing Chi-Square for One-Way Tables ..... 180
Conclusions ..... 182

**Chapter 11 Binary Logistic Regression ..... 183**

Introduction ..... 183
Running a Logistic Regression Model with One Categorical Predictor Variable ..... 184
Running a Logistic Regression Model with One Continuous Predictor Variable ..... 189
Using a Format to Create a Categorical Variable from a Continuous Variable ..... 191
Using a Combination of Categorical and Continuous Variables in a Logistic Regression Model ..... 193
Running a Logistic Regression with Interactions ..... 197
Conclusions ..... 203

**Chapter 12 Nonparametric Tests ..... 205**

Introduction ..... 205
Performing a Wilcoxon Rank-Sum Test ..... 206
Performing a Wilcoxon Signed-Rank Test (for Paired Data) ..... 209
Performing a Kruskal-Wallis One-Way ANOVA ..... 210
Comparing Spread: The Ansari-Bradley Test ..... 211
Converting Data Values into Ranks ..... 213
Using PROC RANK to Group Your Data Values ..... 216
Conclusions ..... 217

**Chapter 13 Power and Sample Size ..... 219**

Introduction ..... 219
Computing the Sample Size for an Unpaired *t*-Test ..... 220
Computing the Power of an Unpaired *t*-Test ..... 222
Computing Sample Size for an ANOVA Design ..... 225
Computing Sample Sizes (or Power) for a Difference in Two Proportions ..... 227
Using the SAS Power and Sample Size Interactive Application ..... 229
Conclusions ..... 234

**Chapter 14 Selecting Random Samples ..... 235**

Introduction ..... 235
Taking a Simple Random Sample ..... 236
Taking a Random Sample with Replacement ..... 237
Creating Replicate Samples using PROC SURVEYSELECT ..... 240
Conclusions ..... 241

**References ..... 243**

**Index ..... 245**

From *SAS® Statistics by Example* by Ron Cody. Copyright © 2013, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

From *SAS® Statistics by Example*. Full book available for purchase here.

# Descriptive Statistics – Categorical Variables

**Introduction.....41**
**Computing Frequency Counts and Percentages.....42**
**Computing Frequencies on a Continuous Variable.....44**
**Using Formats to Group Observations.....45**
**Histograms and Bar charts.....48**
**Creating a Bar Chart Using PROC SGPLOT.....49**
**Using ODS to Send Output to Alternate Destinations.....50**
**Creating a Cross-Tabulation Table.....52**
**Changing the Order of Values in a Frequency Table .....53**
**Conclusions.....55**

## Introduction

This chapter continues with methods of examining categorical variables. You will learn how to produce frequencies for single variables and then extend the process to create cross-tabulation tables. You will also learn several graphical approaches that are used with categorical variables. Finally, you will learn how to use SAS to group continuous variables into categories using a variety of techniques. Let's get started.

## Computing Frequency Counts and Percentages

You can use PROC FREQ to count frequencies and calculate percentages for categorical variables. This procedure can count unique values for either character or numeric variables. Let's start by computing frequencies for Gender and Drug in the `Blood_Pressure` data set used in the previous chapter.

### Program 3.1: Computing Frequencies and Percentages Using PROC FREQ

```
title "Computing Frequencies and Percentages Using PROC FREQ";
proc freq data=example.Blood_Pressure;
   tables Gender Drug;
run;
```

PROC FREQ uses a TABLES statement to identify which variables you want to process. This program selects Gender and Drug. Here is the output:

**Computing Frequencies and Percentages Using PROC FREQ**

The FREQ Procedure

| Gender | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| F | 28 | 48.28 | 28 | 48.28 |
| M | 30 | 51.72 | 58 | 100.00 |

Frequency Missing = 2

| Drug | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| Drug A | 20 | 33.33 | 20 | 33.33 |
| Drug B | 20 | 33.33 | 40 | 66.67 |
| Placebo | 20 | 33.33 | 60 | 100.00 |

By default, PROC FREQ computes frequencies, percentages, cumulative frequencies, and cumulative percentages. In addition, it reports the frequency of missing values. If you do not want all of these values, you can add options to the TABLES statement and specify what statistics you want or do not want. For example, if you want only frequencies and percentages, you can use the TABLES option NOCUM (no cumulative statistics) to remove them from the output, like this:

**Program 3.2: Demonstrating the NOCUM Tables Option**

```
title "Demonstrating the NOCUM Tables Option";
proc freq data=example.Blood_Pressure;
   tables Gender Drug / nocum;
run;
```

Because NOCUM is a statement option, it follows the usual SAS rule: it follows a slash. The following output shows the effect of the NOCUM option:

**Demonstrating the NOCUM Tables Option**

The FREQ Procedure

| Gender | Frequency | Percent |
|---|---|---|
| F | 28 | 48.28 |
| M | 30 | 51.72 |

Frequency Missing = 2

| Drug | Frequency | Percent |
|---|---|---|
| Drug A | 20 | 33.33 |
| Drug B | 20 | 33.33 |
| Placebo | 20 | 33.33 |

As you can see, the output now contains only frequencies and percents.

One TABLES option, MISSING, deserves special attention. This option tells PROC FREQ to treat missing values as a valid category and to include them in the body of the table. Program 3.3 shows the effect of including the MISSING option:

**Program 3.3: Demonstrating the Effect of the MISSING Option with PROC FREQ**

```
title "Demonstrating the effect of the MISSING Option";
proc freq data=example.Blood_Pressure;
   tables Gender Drug / nocum missing;
run;
```

Here is the output:

**Demonstrating the effect of the MISSING Option**

The FREQ Procedure

| Gender | Frequency | Percent |
|---|---|---|
| | 2 | 3.33 |
| F | 28 | 46.67 |
| M | 30 | 50.00 |

| Drug | Frequency | Percent |
|---|---|---|
| Drug A | 20 | 33.33 |
| Drug B | 20 | 33.33 |
| Placebo | 20 | 33.33 |

Notice that the two subjects with missing values for Gender are now included in the body of the table. Even more important, the percentages for females and males have changed. When you use the MISSING option, SAS treats missing values as a valid category and includes the missing values when it computes percentages. To summarize, without the MISSING option, percentages are computed as the percent of all nonmissing values; with the MISSING option, percentages are computed as the percent of all observations, missing and nonmissing.

## Computing Frequencies on a Continuous Variable

What happens if you compute frequencies on a continuous numeric variable such as SBP (systolic blood pressure)? Program 3.4 shows what happens when you try to compute frequencies on a continuous numeric variable:

### Program 3.4: Computing Frequencies on a Continuous Variable

```
title "Computing Frequencies on a Continuous Variable";
proc freq data=example.Blood_Pressure;
   tables SBP / nocum;
run;
```

This is the output:

**Computing Frequencies on a Continuous Variable**

The FREQ Procedure

| SBP | Frequency | Percent |
|---|---|---|
| 108 | 1 | 1.79 |
| 110 | 1 | 1.79 |
| 112 | 2 | 3.57 |
| 114 | 2 | 3.57 |
| 118 | 3 | 5.36 |
| 120 | 5 | 8.93 |
| 122 | 1 | 1.79 |
| 124 | 4 | 7.14 |
| 126 | 2 | 3.57 |
| 128 | 1 | 1.79 |
| 130 | 4 | 7.14 |
| 134 | 6 | 10.71 |
| 136 | 4 | 7.14 |
| 138 | 4 | 7.14 |
| 140 | 5 | 8.93 |
| 142 | 2 | 3.57 |
| 144 | 2 | 3.57 |
| 146 | 1 | 1.79 |
| 148 | 1 | 1.79 |
| 150 | 1 | 1.79 |
| 156 | 1 | 1.79 |

Frequency Missing = 4

Each unique value of SBP is considered a category. Now let's see how to group continuous values into categories.

## Using Formats to Group Observations

SAS can apply formats to character or numeric variables. What is a format? Suppose you have been using M for males and F for females but you want to see the labels Male and Female in your output. You can create a format that associates any text (Male, for

example) to one or more values. To demonstrate, let's start by making a format for Gender, SBP, and DBP, and using these formats with PROC FREQ.

**Program 3.5: Writing a Format for Gender, SBP, and DBP**

```
proc format;
   value $gender 'M' = 'Male'
                 'F' = 'Female';
   value sbpgroup low-140 = 'Normal'
         141-high         = 'High';
   value dbpgroup low-80 = 'Normal'
         81-high         = 'High';
run;

proc freq data=example.Blood_Pressure;
   tables Gender SBP DBP / nocum;
   format Gender $gender.
         SBP sbpgroup.
         DBP dbpgroup.;
run;
```

You use PROC FORMAT to create formats—labels associated with values. If you are planning to create formats for character variables, the format names must start with a dollar sign. Formats to be used with numeric variables cannot start with a dollar sign. In addition, format names cannot end with a number. All format names are limited to a maximum of 32 characters, including the initial dollar sign for character format names. Finally, format names can contain letters, digits, and the underscore character. You name each format on a VALUE statement. This statement lets you specify unique values, groups of values, or ranges of values on the left side of the equal sign, and labels that you want to associate with these values on the right side of the equal sign.

The first format in Program 3.5 is $gender. This name is a good choice because this format will be used later with the variable Gender (a character variable). All the format names are, however, completely arbitrary: you could have called this format $xyz if you wanted to. The $gender format associates the text "Male" with M and "Female" with F. You can use either single or double quotation marks when you create formats—just be sure to use double quotation marks if the format that you are creating contains an apostrophe (which is rendered as a single quotation mark).

The next two formats are to be used with the two variables SBP and DBP. For the SBPGROUP format, the range of values associated with the text "Normal" is from the lowest nonmissing value to 190. You can use the keywords LOW and HIGH when you are defining format ranges.

PROC FORMAT creates formats, but it does not associate any of these formats with SAS variables (even if you are clever and name them so that it is clear which format will go with which variable). To associate a format with one or more SAS variables, you use a FORMAT statement. You can place this statement in either a DATA step or a PROC step. If you place a FORMAT statement in a PROC step (as in Program 3.5), the format will be associated with the variables only for the duration of that procedure. If you place a FORMAT statement in a DATA step, the formats will be permanently assigned to the variables.

In a FORMAT statement, you start with the keyword FORMAT, followed by one or more variables names, followed by the format you want to associate with the variables you listed. On a FORMAT statement, you *must* follow each format name with a period. If you omit the period, SAS will think that you are writing a variable name and not a format. It is slightly confusing—when you create the format with a VALUE statement, you do not end the name with a period (SAS knows this is a format name). When you write a FORMAT statement, you must end the format name with a period.

Let's see what happens when you run Program 3.5—here is the output:

**Computing Frequencies on a Continuous Variable**

The FREQ Procedure

| Gender | Frequency | Percent |
|---|---|---|
| Female | 28 | 48.28 |
| Male | 30 | 51.72 |

Frequency Missing = 2

| SBP | Frequency | Percent |
|---|---|---|
| Normal | 48 | 85.71 |
| High | 8 | 14.29 |

Frequency Missing = 4

| DBP | Frequency | Percent |
|---|---|---|
| Normal | 24 | 42.86 |
| High | 32 | 57.14 |

Frequency Missing = 4

Instead of F's and M's you now see Female and Male. Instead of frequencies for individual values of SBP and DBP, you see only two categories, Normal and High.

## Histograms and Bar Charts

Sometimes it is useful to show frequencies in a graphical display. With SAS, you have several options: First, there is an older SAS procedure called GCHART, which is part of the SAS/GRAPH collection of procedures. A newer procedure, PROC SGPLOT, can produce a wide variety of plots and charts.

The first example of a bar chart uses PROC GCHART to display the frequencies of a variable called Region (region of the country) from a data set called `store`. You can skip this section and go right to the next section, which shows you how to create a bar chart using PROC SGPLOT. However, at some point you might need to run or modify an older SAS program that uses PROC GCHART. Here is the code:

**Program 3.6: Generating a Bar Chart Using PROC GCHART**

```
goptions reset=all;
pattern value = solid color = blue;
title "Generating a Bar Chart - Using PROC GCHART";
proc gchart data=store;
   vbar Region;
run;
quit;
```

The first statement (GOPTIONS, which stands for graphic options), is not mandatory, but if you have been using any SAS/GRAPH procedures during your SAS session, it is a good idea to reset all the options to their default values. Why? Because when you set any graphic option such as color or plotting symbol, these options remain in effect until you change them. This behavior is similar to TITLE statements, which persist unless you change them or omit the titles completely.

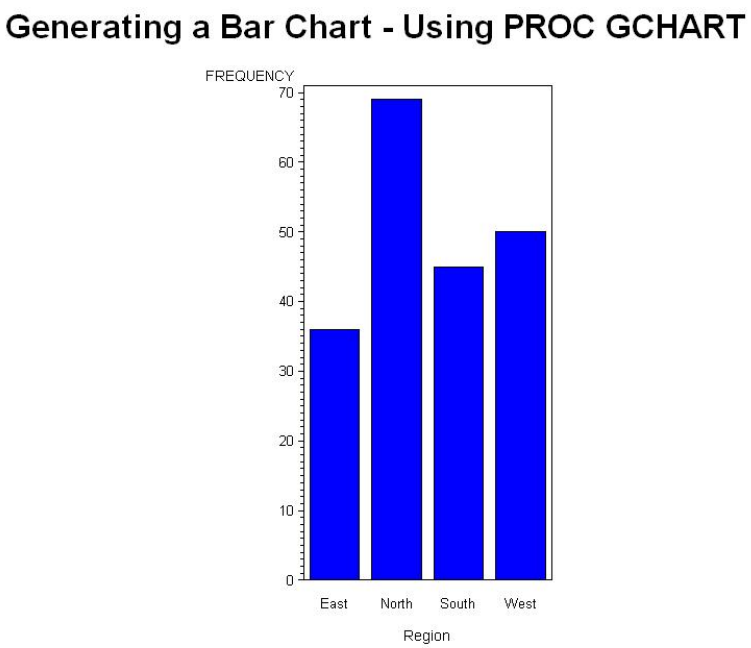
The PATTERN statement enables you to select the type of bar (SOLID in this case) and the color of the bars. A useful hint is to set VALUE=EMPTY if you are sending the output to a dot matrix printer. The EMPTY option displays only the outline of the box and keeps you from rushing out to the office supply store to buy more ink cartridges.

The VBAR (vertical bar) statement lets you list the variables for which you want to generate bar charts. If you prefer a horizontal bar chart, use the HBAR statement instead.

Notice the QUIT statement in this program. Certain procedures in SAS such as PROC GCHART have something called RUN-group processing. This kind of processing keeps

the procedure in memory, even after it encounters a RUN statement. Because the procedure is still in memory, you can request additional charts or, in the case of other procedures, new models, etc. The QUIT statement ends the procedure. If you omit a QUIT statement, the procedure ends when the next DATA or PROC step executes.

Here is the output from Program 3.6:




## Creating a Bar Chart Using PROC SGPLOT

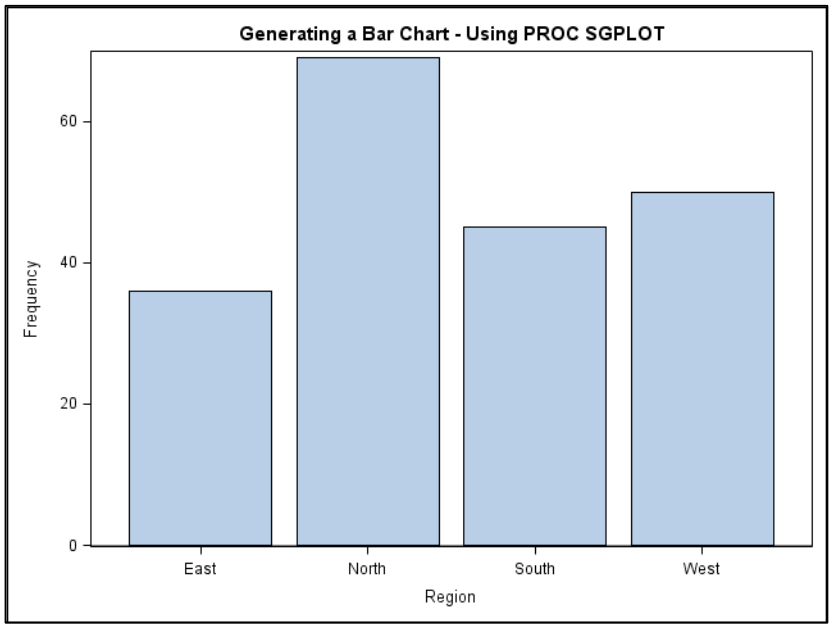
You can create a similar bar chart using PROC SGPLOT. A number of built-in styles make it very easy to customize your output. For example, a style called JOURNAL produces black and white output, suitable for publication in a journal. A style called STATISTICAL gives you output that is designed for statistical purposes.

Program 3.7 shows how to produce a chart similar to the one produced in Program 3.6:

**Program 3.7: Generating a Bar Chart Using PROC SGPLOT**

```
title "Generating a Bar Chart - Using PROC SGPLOT";
proc sgplot data=store;
   vbar Region;
run;
```

The syntax is almost identical to PROC GCHART. You enter the keyword VBAR, followed by one or more variables for which you want to create a bar chart. Here is the output:




## Using ODS to Send Output to Alternate Destinations

To demonstrate the flexibility of the SGPLOT procedure, the next example shows you how to use a built-in style to send the same chart to a PDF file.

**Program 3.8: Using ODS to Create PDF Output**

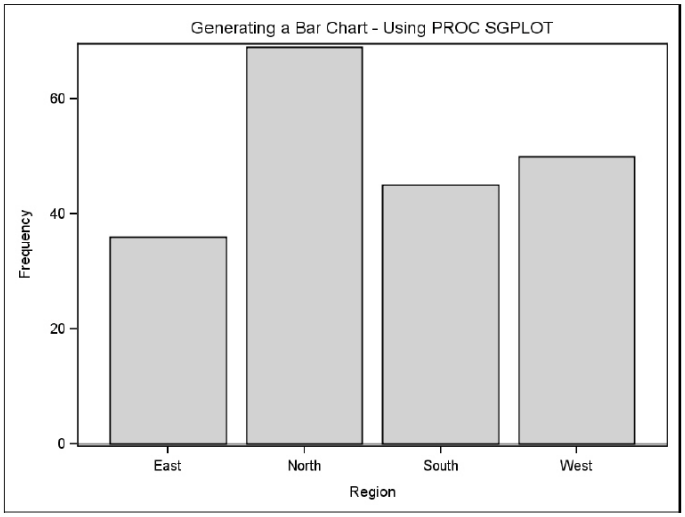
```
ods listing close;
ods pdf file='c:\books\statistics by example\bar.pdf'
        style=journal;
title "Generating a Bar Chart - Using PROC SGPLOT";
proc sgplot data=store;
   vbar Region;
run;
quit;
ods pdf close;
ods listing;
```

This program is identical to the previous one, except you place an ODS statement before the procedure that tells SAS two things: 1) you want to produce a PDF file and 2) you want to use the built-in style called JOURNAL. Following the procedure, you close the destination using another ODS statement.

You should close all your ODS destinations before you exit your SAS session. It is also a good idea to include the ODS LISTING CLOSE statement before the procedure so that you don't get two outputs—one sent to the PDF file and the other sent to the normal SAS output location. Remember that you need to reopen the listing file using the ODS LISTING statement following the procedure.

The PDF file that was created by Program 3.8 looks like this:

***Generating a Bar Chart - Using PROC SGPLOT***



This PDF file can be read by Adobe and used in any application that can work with PDF files.

## Creating a Cross-Tabulation Table

You can use PROC FREQ to create a cross-tabulation table. You start out with the keyword TABLES. Following this, you specify the two variables of interest, separated by an asterisk. For example, the `store` data set contains the variables Region and Gender. If you want to see the distribution of Gender across all values of Region, you proceed with Program 3.9:

### Program 3.9: Creating a Cross-Tabulation Table Using PROC FREQ

```
title "Demonstrating a Cross-Tabulation Table using PROC FREQ";
proc freq data=store;
   tables Gender * Region;
run;
```

This program requests a table of Gender by Region. In this example, Gender will form the rows of the table, and Region will form the columns.

The general form of a cross-tabulation request is:

```
tables  row-variable * column-variable;
```

Here is the output from Program 3.9:

**Demonstrating a Cross-Tabulation Table using PROC FREQ**

The FREQ Procedure

| Frequency<br>Percent<br>Row Pct<br>Col Pct | Table of Gender by Region | | | | |
|---|---|---|---|---|---|
| | Region | | | | |
| Gender | East | North | South | West | Total |
| Female | 22 | 39 | 23 | 26 | 110 |
| | 11.00 | 19.50 | 11.50 | 13.00 | 55.00 |
| | 20.00 | 35.45 | 20.91 | 23.64 | |
| | 61.11 | 56.52 | 51.11 | 52.00 | |
| Male | 14 | 30 | 22 | 24 | 90 |
| | 7.00 | 15.00 | 11.00 | 12.00 | 45.00 |
| | 15.56 | 33.33 | 24.44 | 26.67 | |
| | 38.89 | 43.48 | 48.89 | 48.00 | |
| Total | 36 | 69 | 45 | 50 | 200 |
| | 18.00 | 34.50 | 22.50 | 25.00 | 100.00 |

Each box in the table contains four values; the meaning of these values is found in the key in the upper-left corner of the table. As you can see, the top number in each box is the number of observations. For example, there are 22 females in the Eastern region. The next number is a percentage. In this example, 11% of all observations are females in the Eastern region. The third number in each box is a row percentage—20% of the females were in the Eastern region. Finally, the fourth number in each box is a column percentage; 61.11% of the observations from the Eastern region are female.

Notice the order of the rows and columns in the output. By default, SAS orders the rows and columns in a table (or for a single variable) by the internal value of the variable—alphabetically for character variables and numerically for numeric variables. This is why the rows in the previous table were ordered Female→Male and the order of the columns was East→North→South→West.

## Changing the Order of Values in a Frequency Table

Whether you have a one-way or a two-way table, you might want to control the order that SAS uses for the rows, the columns, or both. In the previous example, maybe you want the regions to be ordered North→East→South→West. Or you might be computing an odds ratio in a 2x2 table and want the first column to be labeled Yes and the second column to be labeled No.

You can accomplish these goals in several ways. One is to create a new variable from the existing variable, where the internal values are in the desired order. Another, easier, method is to associate formats that are in the target order and associate that format with your variable. You can then use a PROC FREQ option called ORDER=FORMATTED to tell SAS to order the rows, columns, or both by their formatted values, rather than by their internal values.

The example that follows uses this method to force the order of the regions to be North, East, South, and West. First the program, then the explanation.

**Program 3.10: Changing the Order of Values in a PROC FREQ Table By Using Formats**

```
proc format;
   value $region 'North' = '1 North'
                 'East'  = '2 East'
                 'South' = '3 South'
                 'West'  = '4 West';
run;

title "Change the Order in a PROC FREQ Output";
proc freq data=store order=formatted;
   tables Gender * Region;
   format Region region.;
run;
```

The four formatted values created by the $region format are in the desired order alphabetically. (Note that the digits 1, 2, 3, and 4 are part of the format labels, and 1 comes before 2 alphabetically, etc.) Including a FORMAT statement in PROC FREQ associates the $region format with the variable called Region. (Remember that the association is made by the FORMAT statement, not because the name of the format is similar to the name of the variable.) Finally, to tell SAS to order the table by the formatted values rather than by the internal values, you must include the PROC FREQ option ORDER=FORMATTED.

Here is the output from Program 3.10:

**Change the Order in a PROC FREQ Output**

The FREQ Procedure

| Frequency<br>Percent<br>Row Pct<br>Col Pct | Table of Gender by Region | | | | |
|---|---|---|---|---|---|
| | Region | | | | |
| Gender | 1 North | 2 East | 3 South | 4 West | Total |
| Female | 39 | 22 | 23 | 26 | 110 |
| | 19.50 | 11.00 | 11.50 | 13.00 | 55.00 |
| | 35.45 | 20.00 | 20.91 | 23.64 | |
| | 56.52 | 61.11 | 51.11 | 52.00 | |
| Male | 30 | 14 | 22 | 24 | 90 |
| | 15.00 | 7.00 | 11.00 | 12.00 | 45.00 |
| | 33.33 | 15.56 | 24.44 | 26.67 | |
| | 43.48 | 38.89 | 48.89 | 48.00 | |
| Total | 69 | 36 | 45 | 50 | 200 |
| | 34.50 | 18.00 | 22.50 | 25.00 | 100.00 |

The regions are now ordered 1 North, 2 East, 3 South, and 4 West. Because female comes before male alphabetically, the order is Female, then Male.

## Conclusions

In this chapter, you learned how to display values of categorical variables, both in tabular and graphical form. Although you can use several methods to change the order of rows and columns in a table, using formats might be the simplest.

You also learned how to use the built-in styles to create attractive output with a minimum of effort. And you saw how to use ODS to send this output to a variety of destinations, such as HTML, PDF, and RTF.

The next chapter finishes up our discussion of descriptive statistics by showing you how to produce numerical and graphical displays for bivariate relationships.

From *SAS® Statistics by Example* by Ron Cody. Copyright © 2013, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

From *SAS® Statistics by Example*. Full book available for purchase here.

# Index


## A

AB option, NPAR1WAY procedure 212
ADJUST= option, LSMEANS statement (GLM) 98, 100
adjusted R-square, generating plots of 142, 144
AGREE option, TABLES statement (FREQ) 166, 175–177
AIC (Akaike Information Criterion) 187, 199–200
Akaike Information Criterion (AIC) 187, 199–200
ALL option, TABLES statement (FREQ) 166
ALPHA= option, TTEST procedure 70–71
analysis of variance
  *See* ANOVA (analysis of variance)
ANOVA (analysis of variance)
  analyzing randomized block designs 108–109
  computing sample size for 225–227
  Kruskal-Wallis ANOVA 210–211
  multiple comparison tests 97–101
  one-way design 92–97, 210–211
  producing diffograms 101–102
  two-way factorial designs 102–108
Ansari-Bradley test 211–213
at sign (@) 15, 212
automatic selection methods 145, 198

## B

backward selection method 145–152, 198
bar charts, creating 49–50
BEST= option, MODEL statement (REG) 139, 145
binary logistic regression
  *See* logistic regression models
bivariate associations
  producing scatter plots using GPLOT 58–60
  producing scatter plots using SGPLOT 61–62
  producing scatter plots using SGSCATTER 63–68
block designs, randomized 108–109
blood_pressure sample data set 20
box plots
  displaying for categorical variables 38–39
  displaying outliers 37
  GLM procedure and 97
  horizontal 35–36
  labeling outliers on 38
  producing 33
  vertical 36

## C

CALL MISSING routine 153–154
case sensitivity in variables 6
categorical variables
  and continuous variables in regression models 193–196
  bar charts and 48–50
  chi-square macros 173–175
  comparing proportions 164–166
  computer percentages 42–44
  computing chi-square for one-way tables 180–182
  computing chi-square from frequency data 172–173
  computing frequency counts 42–44
  computing Kappa coefficient 175–177
  computing tests for trends 178–180
  creating dummy variables for 153–155
  creating from continuous variables 191–193
  cross-tabulation tables and 52–53
  displaying box plots for 38–39
  Fisher's exact test 169–171
  frequency tables and 53–55
  grouping observations with formats 45–48

categorical variables (*continued*)
histograms and 48–49
ODS considerations 50–52
rearranging rows and columns in tables 166–169
requesting multiple tables 175
running logistic regression models with 184–189
CATEGORY= option
HBOX statement (SGPLOT) 38–39
VBOX statement (SGPLOT) 38–39
character variable types 11
chi-square
Cochran-Mantel-Haenszel 178
computing for one-way tables 180–182
computing from frequency data 172–173
computing tests for trends 178–180
Pearson's 228
chi-square macros 173–175
CHISQ option, TABLES statement (FREQ) 165–166
CLASS statement
GLM procedure 92–93, 97
LOGISTIC procedure 185, 194
MEANS procedure 23–25
NPAR1WAY procedure 207
TTEST procedure 80
classification variables, breaking down descriptive statistics by 23–25
CLM option, MEANS procedure 22, 25
CLODDS= option, MODEL statement (LOGISTIC) 185, 189
CMH option, TABLES statement (FREQ) 166
Cochran-Armitage test for trends 178–180
Cochran-Mantel-Haenszel test for trends 178–180
coefficient of agreement 175–177
coefficient of variation (CV) 28
collinearity, detecting 155–156
colon (:) 176
column pointer (@ sign) 15
columns
*See also* variables
changing order of values in 53–55
rearranging in tables 166–169
COMPARE statement, SGSCATTER procedure 63–66
confidence intervals, computing 25–26
continuous variables
and categorical variables in regression models 193–196
breaking down descriptive statistics 23–25
computing confidence intervals 25–26
computing descriptive statistics 21–23
computing frequencies on 44–45
computing standard error 25–26
creating categorical variables from 191–193
generating graphical displays 33–39
producing descriptive statistics 26–32
producing histograms 26–33
producing probability plots 26–32
running logistic regression models with 189–191
converting
data values into ranks 213–216
Excel files into data sets 6–9
Cook's D statistics 126–128, 157
CORR procedure
ID statement 118, 125
NOPRINT option 133
NOSIMPLE option 112
OUTEST= option 133
Pearson correlations 112
PEARSON option 119
PLOTS= option 113–115, 118, 125, 128
RANK option 112
SPEARMAN option 119
Spearman rank correlations 119–120
VAR statement 112, 115
WITH statement 112
correlation matrix, generating 115–117
cross-tabulation tables 52–53
CSV files
about 5
converting into data sets 6–9

reading 14–15
CV (coefficient of variation) 28
CV option, MEANS procedure 22

## D

data descriptors, displaying 9
DATA= option
MEANS procedure 22
SURVEYSELECT procedure 236
data sets
*See also* specific sample data sets
about 3
converting Excel files into 6–9
creating from raw data 12–16
data descriptors and 3, 9
data values in fixed columns 15–16
data values separated by delimiters 12–14
metadata and 3
observations and 6
permanent 11–12, 21
reading CSV files 14–15
taking random sample from 236
temporary 11–12, 21
variable types in 11
DATA step
about 13
Ansari-Bradley test 211–212
creating categorical variable from continuous variable 191
DATALINES statement 130
DO loops 212
ELSE-IF statement 37
Fisher's exact test example 170
FORMAT statement 47, 176
IF-THEN statement 37
INFILE statement 13–14, 130
INFORMAT statement 176
INPUT statement 13, 15–16, 130, 170, 176
LENGTH statement 13
OUTPUT statement 212
RUN statement 13
SET statement 37
Data Tips 117–119
data values
converting into ranks 213–216
grouping with RANK procedure 216–217
in fixed columns 15–16
separated by delimiters 12–14
database systems 17
DATALABEL= option, HBOX statement (SGPLOT) 38
DATALINES statement 130
delimiters, data values separated by 12–14
descriptive statistics
about 19–21
bar charts and 48–50
broken down by classification variables 23–25
computer percentages 42–44
computing confidence intervals 25–26
computing frequency counts 42–44
computing standard error 25–26
computing with MEANS procedure 21–23
cross-tabulation tables and 52–53
frequency tables and 53–55
generating graphical displays 33–38
grouping observations with formats 45–48
histograms and 26–33, 48–49
ODS considerations 50–52
probability plots and 26–32
producing 26–32
scatter plots and 58–68
design variables 153–155
DIAGONAL= option, MATRIX statement (SGSCATTER) 67
diffograms 101–102
Display Manager
about 4
Log window 4
Output window 4
Program Editor 4
Submit icon 8
DO loops 212
dollar sign ($) 176
DSD option, INFILE statement 14
dummy variables 153–155

## E

ELSE-IF statement 37
EST keyword 27
EVENT= option, MODEL statement (LOGISTIC) 185
EXACT statement, NPAR1WAY procedure 208–209, 211–212, 216
Excel files
  about 5
  converting into data sets 6–9
  with invalid variable names 16
exercise sample data set 70, 112, 136

## F

factorial designs
  GLM procedure and 93
  two-way 102–108
false-positive rates 197
FILE= option, ODS HTML statement 118
FISHER option, TABLES statement (FREQ) 166
Fisher's exact test 169–171, 228
fixed columns, data values in 15–16
folded F-test 211
FORMAT procedure
  about 46–48
  VALUE statement 46–48
FORMAT statement
  about 47, 176
  FREQ procedure 167
  LOGISTIC procedure 192
formats
  associating variables with 176
  creating 46–47
  creating categorical variables from continuous variables 191–193
  grouping observations with 45–48
forward selection method 145–152
FREQ procedure
  *See also* TABLES statement, FREQ procedure
  about 42–44
  Fisher's exact test example 169
  FORMAT statement 167
  ORDER=FORMATTED option 53–54, 167
  TEST statement 177
  WEIGHT statement 173, 181
frequency tables
  changing order of values 53–55
  computing chi-square from 172–173

## G

GCHART procedure
  about 48–49
  HBAR statement 48
  VBAR statement 48
GLM procedure
  CLASS statement 92–93, 97
  LSMEANS statement 98, 100, 107
  MEANS statement 93, 95, 97–98
  MODEL statement 92–93, 102, 108
  multiple comparison tests 97–101
  one-way ANOVA and 92–97
  pairwise differences and 101
  PLOTS= option 92, 96, 101
  RUN-group processing and 120
GLMPOWER procedure 225
global statements 9
  *See also* specific statements
GOPTIONS statement 48, 59
GPATH= option, ODS HTML statement 118
GPLOT procedure
  about 58–60
  PLOT statement 58
graphical output 33–38
  *See also* ODS Graphics
  *See also* SAS/GRAPH software
GROUP= option
  COMPARE statement (SGSCATTER) 65–66
  SCATTER statement (SGPLOT) 62
grouping observations with formats 45–48
GROUPMEANS= option, POWER procedure 220, 223
GROUPS= option, RANK procedure 216

## H

HBAR statement, GCHART procedure 48
HBOX statement, SGPLOT procedure
about 35–36
CATEGORY= option 38–39
DATALABEL= option 38
HIGH keyword 46
HISTOGRAM statement
SGPLOT procedure 33–35
UNIVARIATE procedure 27, 30–33
histograms
changing midpoint values 32–33
distribution of income example 206
ODS Graphics example 73, 82, 123
producing with CORR procedure 115
producing with GCHART procedure 48–49
producing with SGPLOT procedure 34–35
producing with SGSCATTER procedure 67
producing with UNIVARIATE procedure 26–32
Hocking's criteria 142–145
horizontal box plots 35–36
HOVTEST option, MEANS statement (GLM) 93
HTML output, creating with Data Tips 117–119

## I

ID statement
CORR procedure 118, 125
UNIVARIATE procedure 26, 30
IF-THEN statement 37
IMAGEMAP= option, ODS GRAPHICS statement 118
Import Wizard
converting Excel files into data sets 6
CSV file and 15
sources supported 17
IN operator 153
INCLUDE= option, MODEL statement (REG) 152
inferential statistics
comparing more than two means 91–109
one-sample *t*-tests 69–78
two-sample tests 79–89
INFILE statement
about 13
DSD option 14
regression equation example 130
INFLUENCE option, MODEL statement (REG) 125–126, 157
influential observations
in multiple regression models 157–161
investigating 120–125
INFORMAT statement 176
informats
associating variables with 176
defined 15
INPUT statement
@ symbol and 212
@@ symbol and 130, 170
about 13
colon in 176
data values in fixed columns 15–16

## J

JOURNAL style 49, 51

## K

Kappa coefficient 175–177
KEEP= data set option 37
Kruskal-Wallis one-way ANOVA 210–211
kurtosis 28

## L

labeling outliers on box plots 38
least squares regression equation 129–132
LENGTH statement 13
leptokurtic distribution 28
LIBNAME statement 21
likelihood ratio chi-square test 228
linear regression, running with REG procedure 120–125
Log window (Display Manager) 4

LOGISTIC procedure
CLASS statement 185, 192, 194
FORMAT statement 192
logistic regression models example 187
MODEL statement 184–185, 189, 198
ODDSRATIO statement 198
PLOTS= option 198
running logistic regression with interactions 197–203
UNITS statement 190–191, 194
logistic regression models
creating categorical variables from continuous variables 191–193
running with interactions 197–203
running with one categorical variable 184–189
running with one continuous variable 189–191
selection methods and 198
using categorical and continuous variable combinations in 193–196
LOW keyword 46
LSMEANS statement, GLM procedure
ADJUST= option 98, 100
PDIFF option 98, 100
SLICE option 107

## M

macros, chi-square 173–175
Mallows' $C_p$ 138, 142–145
Mann-Whitney U test 207
MATRIX statement, SGSCATTER procedure 63, 67–68
MAX option, MEANS procedure 22
MAXDEC= option, MEANS procedure 22
MEAN option, MEANS procedure 22
means-means plot 101–102
MEANS procedure
CLASS statement 23–25
CLM option 22, 25
computing descriptive statistics 21–23
CV option 22
DATA= option 22
MAX option 22
MAXDEC= option 22
MEAN option 22
MEDIAN option 22
MIN option 22
N option 22, 24
NMISS option 22, 24
PRINTALLTYPES option 24
STD option 22
STDERR option 22, 25
VAR statement 22
MEANS statement, GLM procedure
HOVTEST option 93
multiple comparison tests and 97–98
WELCH option 95
MEASURES option, TABLES statement (FREQ) 166
MEDIAN option, MEANS procedure 22
metadata, defined 3
METHOD= option, SURVEYSELECT procedure 236
MIDPOINT option, HISTOGRAM statement (UNIVARIATE) 32–33
MIN option, MEANS procedure 22
MISSING option, TABLES statement (FREQ) 43–44
MODEL statement, GLM procedure
one-way designs 92–93
randomized block designs 108
short-cut notation 198
two-way factorial designs 102
MODEL statement, LOGISTIC procedure
CLODDS= option 185, 189
EVENT= option 185
logistic regression with one categorical variable 184
logistic regression with one continuous variable 189
short-cut notation 198
SLSTAY option 198–199
MODEL statement, REG procedure
BEST= option 139, 145
INCLUDE= option 152

INFLUENCE option 125–126, 157
P option 131
R option 125–126
SELECTION= option 138, 145
SLENTRY= option 145–152
SLSTAY= option 145–152
VIF option 156
MU= keyword 27
multiple comparison tests
conducting 97–101
diffograms and 101–102
multiple regression models
automatic selection methods 145–152
creating dummy variables for 153–155
detecting collinearity 155–156
fitting 136–137
forcing selected variables into 152
Hocking's criteria 142–145
influential observations in 157–161
Mallows' $C_p$ 142–145
producing separate plots instead of panels 141–142
running all possible regressions 137–141
running with REG procedure 120
multiple tables, requesting 175

## N

n. informat 15
$n. informat 15
N option, MEANS procedure 22, 24
NEXTROBS= option, UNIVARIATE procedure 30
NMISS option, MEANS procedure 22, 24
NOCUM option, TABLES statement (FREQ) 42–43
nonparametric tests
converting data values into ranks 213–216
grouping data values with RANK procedure 216–217
Kruskal-Wallis one-way ANOVA 210–211
Wilcoxon rank-sum test 206–209
Wilcoxon signed-rank test 29, 209–210
NOPRINT option, CORR procedure 133
normal distribution
skewness and kurtosis 28
testing for 76–78
NORMAL option
PROBPLOT statement (UNIVARIATE) 27
UNIVARIATE procedure 76
NOSIMPLE option, CORR procedure 112
NPAR1WAY procedure
AB option 212
CLASS statement 207
EXACT statement 208–209, 211–212, 216
VAR statement 207
WILCOXON option 207
NPERGROUP= option, POWER procedure 221, 225, 228
numerical variable types 11

## O

observations
defined 6
groups with formats 45–48
in multiple regression models 157–161
investigating 120–125
odds ratios 188, 201
ODDSRATIO statement, LOGISTIC procedure 198
ODS Graphics
analyzing factorial models 107
CORR procedure and 112–113
creating HTML output with Data Tips 117–119
customizing output from 83–85
investigating influential observations 125–129
odds ratios 197
one-way ANOVA and 92
producing diffograms 101–102
REG procedure and 121
ROC curve 197
TTEST procedure and 71–74, 81–83
ODS GRAPHICS statement
IMAGEMAP= option 118

ODS GRAPHICS statement (*continued*)
OFF option 71, 81
ON option 71, 81, 115
ODS HTML statement
FILE= option 118
GPATH= option 118
PATH= option 118
ODS LISTING statement
about 51
CLOSE option 51
ODS SELECT statement 209
ODS Statistical Graphics
*See* ODS Graphics
OFF option, ODS GRAPHICS statement 71, 81
ON option, ODS GRAPHICS statement 71, 81, 115
one-sample *t*-tests
conducting with TTEST procedure 70–74
conducting with UNIVARIATE procedure 74–78
examples 29
ODS Graphics example 72
one-way ANOVA 92–97, 210–211
one-way tables 180–182
ONEWAYANOVA option, POWER procedure 225
ONLY keyword 142
ORDER=FORMATTED option, FREQ procedure 53–54, 167
OUT= option
SCORE procedure 134
SURVEYSELECT procedure 236
OUTEST= option, CORR procedure 133
OUTHITS option, SURVEYSELECT procedure 237–239
outliers
displaying in box plots 37
labeling on box plots 38
OUTPUT statement 212
Output window (Display Manager) 4

## P

P option, MODEL statement (REG) 131
PAIRED statement, TTEST procedure 85–87
pairwise comparisons
conducting 98–101
diffograms and 101–102
PARAM= option, CLASS statement (LOGISTIC) 194
PATH= option, ODS HTML statement 118
PATTERN statement
about 48
VALUE=EMPTY option 48
PDF files 50–52
PDIFF option, LSMEANS statement (GLM) 98, 100
Pearson correlations 112–116
PEARSON option, CORR procedure 119
Pearson's chi-square 228
period (.) 11
permanent data sets 11–12, 21
platykurtic distribution 28
PLOT statement
GPLOT procedure 58
POWER procedure 221–223
SGSCATTER procedure 63–64
PLOT statement, POWER procedure 221
PLOTS= option
CORR procedure 113–115, 118, 125, 128
GLM procedure 92, 96, 101
LOGISTIC procedure 198
REG procedure 141–142, 157
TTEST procedure 81, 83–84
plus sign (+) 58
power
computing for difference in proportions 227–229
computing for unpaired *t*-test 222–224
SAS Power and Sample Size program 229–234
Power and Sample Size program 229–234
POWER= option, POWER procedure 223
POWER procedure
GROUPMEANS= option 220, 223
NPERGROUP= option 221, 225, 228
ONEWAYANOVA option 225

PLOT statement 221–223
POWER= option 223
STDDEV= option 220
TEST= option 228
TWOSAMPLEMEANS option 220
predicting values
efficiently 132–134
with least squares regression equation 129–132
PRINT procedure 13
PRINTALLTYPES option, MEANS procedure 24
probability plots
distribution of income example 206
producing 26–32
PROBPLOT statement, UNIVARIATE procedure
about 27, 30–31
NORMAL option 27
testing distribution with 78
Program Editor (Display Manager) 4
proportions
comparing 164–166
computing power for difference in 227–229
computing sample size for difference in 227–229

## Q

Q-Q plots
linear regression example 123
ODS Graphics example 74, 82
one-way designs 97
testing *t*-test assumptions 87–88
QQPLOT statement, UNIVARIATE procedure 78
quantiles
commonly used 30
ODS Graphics example 74
QUIT statement 48–49, 93, 120
quotation marks 21

## R

R option, MODEL statement (REG) 125–126
R-square
computing 156
generating plots of 142–143
random samples
creating replicate samples 240–241
selecting 235–241
taking 236–237
taking with replacement 237–240
randomized block designs 108–109
RANK option, CORR procedure 112
RANK procedure
about 213–214
grouping data values with 216–217
GROUPS= option 216
RANKS statement 214
VAR statement 214–215
ranks, converting data values into 213–216
RANKS statement, RANK procedure 214
reading CSV files 14–15
REG procedure
*See also* MODEL statement, REG procedure
PLOTS= option 141–142, 157
predicting values 131–134
running regression models 120–125, 136, 154
relative kurtosis 28
RELRISK option, TABLES statement (FREQ) 166–167
REPS= option, SURVEYSELECT procedure 240
risk sample data set 164, 184
ROC curve 189, 197–198, 203
root MSE 122
rows
*See also* observations
changing order of values in 53–55
rearranging in tables 166–169
RSQUARE selection method 137–138

RUN-group processing
about 120
GCHART procedure and 48–49
GLM procedure and 93, 120
REG procedure and 120
RUN statement 13

## S

sample size
computing for ANOVA design 225–227
computing for difference in proportions 227–229
computing for unpaired *t*-test 220–222
SAS Power and Sample Size program 229–234
SAMPRATE option, SURVEYSELECT procedure 236, 240
SAMPSIZE= option, SURVEYSELECT procedure 236, 240
SAS
about 2
sources of data 17
statistical tasks performed by 3
SAS/GRAPH software
GOPTIONS statement and 48
graphical output and 26
producing scatter plots 58–62
SAS macros 173–175
SAS Power and Sample Size program 229–234
SAS programs, structure of 3
SAS/STAT software 219, 235
SC (Schwartz Criteria) 187, 199
scatter plots
producing with CORR procedure 112
producing with GPLOT procedure 58–60
producing with SGPLOT procedure 33
SCATTER statement, SGPLOT procedure 61–62
Schwartz Criteria (SC) 187, 199
SCORE= option, SCORE procedure 134
SCORE procedure
computing predicted values 132–134
OUT= option 134
SCORE= option 134
VAR statement 134
SEED= option, SURVEYSELECT procedure 236
selection methods 145–152, 198
SELECTION= option
MODEL statement (LOGISTIC) 198
MODEL statement (REG) 138, 145
semicolon (;) 8
sensitivity of models 197
SET statement 37
SGPLOT procedure
about 33
creating bar charts 49–50
HBOX statement 35–36, 38–39
HISTOGRAM statement 33–35
producing scatter plots 61–62
SCATTER statement 61–62
VBOX statement 36, 38–39
SGSCATTER procedure
about 33, 63, 114
COMPARE statement 63–66
MATRIX statement 63, 67–68
PLOT statement 63–64
SIDES= option, TTEST procedure 70
SIGMA= keyword 27
sign test 29
skewness 28
SLENTRY= option, MODEL statement (REG) 145–152
SLICE option, LSMEANS statement (GLM) 107
SLSTAY= option
MODEL statement (LOGISTIC) 198–199
MODEL statement (REG) 145–152
SNK (Student-Newman-Keuls) test 98–99
Spearman nonparametric correlations 119–120
SPEARMAN option, CORR procedure 119
spread, comparing 211–213
SS (sum of squares) 93
standard error, computing 25–26
statistical analysis
*See* descriptive statistics

STATISTICAL style 49
STD option, MEANS procedure 22
STDDEV= option, POWER procedure 220
STDERR option, MEANS procedure 22, 25
stepwise selection method 145–152
store sample data set 58, 92
Student-Newman-Keuls (SNK) test 98–99
Submit icon (Display Manager) 8
sum of squares (SS) 93
SURVEYSELECT procedure
about 235
DATA= option 236
METHOD= option 236
OUT= option 236
OUTHITS option 237–239
REPS= option 240
SAMPRATE option 236, 240
SAMPSIZE= option 236, 240
SEED= option 236
taking simple random sample 236–237
SYMBOL statement 58, 60–61

## T

*t*-tests
*See also* one-sample *t*-tests
*See also* two-sample *t*-tests
computing power for 222–224
computing sample size for 220–222
tables
cross-tabulation 52–53
Fisher's exact test and 169–171
frequency 53–55, 172–173
generating 164
one-way 180–182
rearranging columns in 166–169
rearranging rows in 166–169
requesting multiple 175
two-way 164
TABLES statement, FREQ procedure
about 42, 52–53, 164
AGREE option 166, 175–177
ALL option 166
CHISQ option 165–166
CMH option 166
FISHER option 166
MEASURES option 166
MISSING option 43–44
NOCUM option 42–43
RELRISK option 166–167
requesting multiple tables 175
TESTF= option 180–181
TESTP= option 180–181
TREND option 166
temporary data sets 11–12, 21
TEST= option, POWER procedure 228
TEST statement, FREQ procedure 177
TESTF= option, TABLES statement (FREQ) 180–181
testing
for normal distribution 76–78
nonparametric tests 29, 205–217
sign test 29
*t*-test assumptions 87–89
tests for trends 178–180
TESTP= option, TABLES statement (FREQ) 180–181
TITLE statement
about 9, 21
scatter plots example 60
TREND option, TABLES statement (FREQ) 166
TTEST procedure
ALPHA= option 70–71
CLASS statement 80
conducting one-sample *t*-test 70–71
conducting two-sample *t*-test 80–81
NPAR1WAY procedure and 207
ODS Graphics and 71–74, 81–83
PAIRED statement 85–87
PLOTS= option 81, 83–84
running with ODS Graphics 71–74
SIDES= option 70
VAR statement 80
Tukey test 98–99

two-sample *t*-tests
assumption violations 87–89
conducting 79–81
conducting paired *t*-test 85–87
converting data values into ranks example 215
testing assumptions for 81–83
two-way factorial designs
about 102–105
analyzing 106–108
two-way tables 164
TWOSAMPLEMEANS option, POWER procedure 220

## U

underscore (_) 5, 16
UNITS statement, LOGISTIC procedure 190–191, 194
UNIVARIATE procedure
about 26–32
computing quantiles 30
conducting one-sample *t*-test 74–76
HISTOGRAM statement 27, 30–31
ID statement 26, 30
NEXTROBS= option 30
NORMAL option 76
plotting distribution of income 206
PROBPLOT statement 27, 30–31, 78
QQPLOT statement 78
VAR statement 27, 74
WEIGHT option 29
Wilcoxon signed-rank test 209
unpaired *t*-test
computing power for 222–224
computing sample size for 220–222
SAS Power and Sample Size program 229–234
URS method 237

## V

VALIDVARNAMES system option 16
VALUE statement, FORMAT procedure 46–48
VALUE=EMPTY option, PATTERN statement 48
VAR statement
CORR procedure 112, 115
MEANS procedure 22
NPAR1WAY procedure 207
RANK procedure 214–215
SCORE procedure 134
TTEST procedure 80
UNIVARIATE procedure 27, 74
variables
*See also* categorical variables
*See also* continuous variables
associating formats with 176
associating informats with 176
case sensitivity 6
classification 23–25
design 153–155
dummy 153–155
forcing into regression models 152
invalid 16
variance inflation factor (VIF) 156
VBAR statement, GCHART procedure 48
VBOX statement, SGPLOT procedure
about 36
CATEGORY= option 38–39
generating bar charts 50
vertical box plots 36
VIF (variance inflation factor) 156
VIF option, MODEL statement (REG) 156

## W

Wald test 188
WEIGHT option, UNIVARIATE procedure 29
WEIGHT statement, FREQ procedure 173, 181
WELCH option, MEANS statement (GLM) 95
WILCOXON option, NPAR1WAY procedure 207
Wilcoxon rank-sum test 206–209
Wilcoxon signed-rank test 29, 209–210
WITH statement, CORR procedure 112

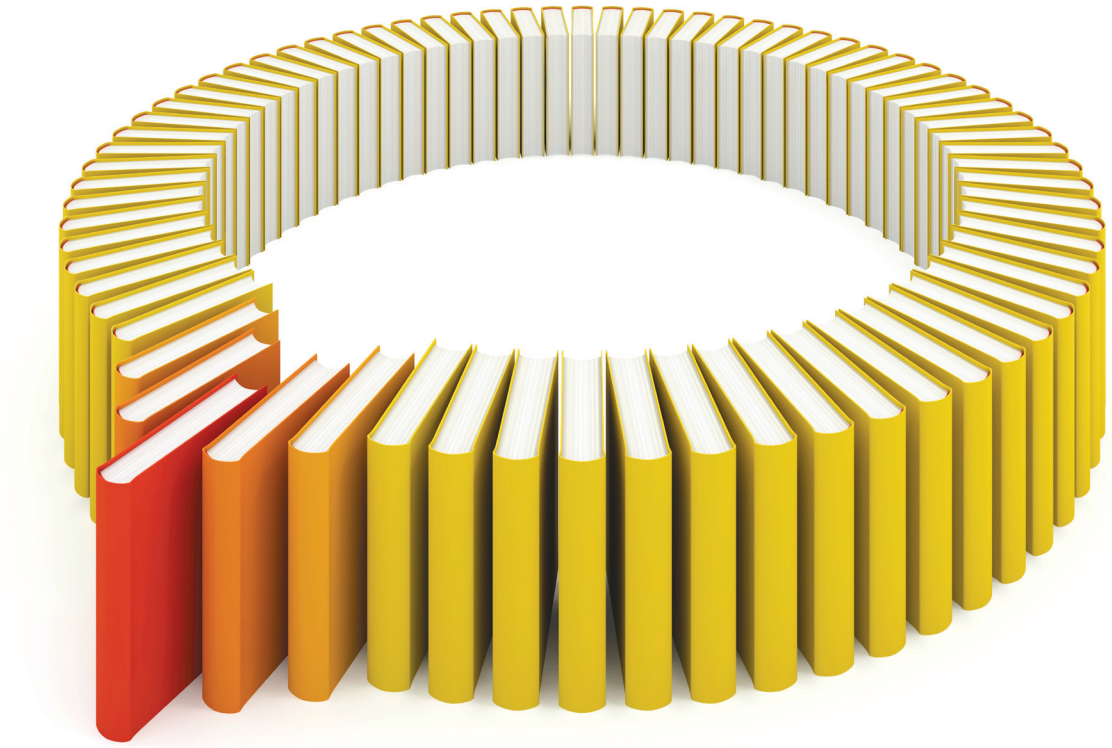
## Y

Yates' corrected value 171

## Symbols and Numbers

@ (at sign) 15, 212
$ (dollar sign) 176
+ (plus sign) 58
_ (underscore) 5, 16
: (colon) 176
. (period) 11
; (semicolon) 8
$1. informat 15
3. informat 15
$3. informat 15
95% confidence interval, computing 25–26
@@ symbol 130, 170

From *SAS® Statistics by Example by Ron Cody*. Copyright © 2013, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.



# 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
*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