Acknowledgements

The writer would like to thank Pat Stewart of Pennsylvania State University, Marie Rowland of the City of San Luis Obispo, California, and numerous students enrolled in marketing research courses for their useful feedback on previous editions of this guide.
## Contents

Acknowledgements ..................................................................................................... ii

To the Instructor ........................................................................................................... v

To the Student .............................................................................................................. ix

What Can SPSS/ PC+ Do for Me? ............................................................................. 1

Introduction ................................................................................................................... 9

Step 1: Coding Your Data ............................................................................................ 11

Step 2: Writing An SPSS/ PC+ Program ...................................................................... 20

Step 3: Entering The Program And Data ................................................................... 30
  Exercise 1 ............................................................................................................. 33

Step 4: Checking Your Data And Program For Errors ........................................... 35

Step 5: Using Statistical Procedures And Computations ....................................... 38

  Frequencies ............................................................................................................. 38
  Creating New Variables: Compute ......................................................................... 39
  Recoding Variables ............................................................................................... 39
  Crosstabulation ..................................................................................................... 41
  Product Moment (Pearson) Correlation .................................................................. 45
  Multiple Linear Regression ................................................................................. 47
  Discriminant Analysis ......................................................................................... 48
  "Count: "Counting" on how many variables a criterion is met ........ 49
  Descriptives: A summarized version of Frequencies ..................................... 49
  Npar tests: Non-parametric statistical tests ....................................................... 50
  Means: Providing a breakdown of population means by subgroup .................. 51
  Oneway and ANOVA: Analysis of Variance ...................................................... 52
  Plot: Turning bivariante data into a scatterplot ................................................. 52
To the Instructor

This text is intended to meet the needs of students enrolled in a one term course which is only partially devoted to data analysis. In such a situation, the instructor cannot afford to take out a major part of the class to discuss the theoretical aspects of data analysis and processing. In fact, this book resulted from the frustration I experienced in not being able to find a textbook suitable for a one quarter marketing research methods course of which only the last three weeks were devoted to data analysis.

My emphasis has been on getting the student started on doing his or her own analysis as quickly as possible. Rather than discussing the concepts and commands in a highly theoretical manner, I have presented some very readable and illustrative examples. A discussion of the concepts and mechanics behind these examples is intended to enable the student to generalize these illustrations to their own projects.

I have taken several steps aimed at enhancing the student's understanding of the analytical process. For example, I have chosen to consistently place data in the same file as the SPSS/PC+ program and have confined any treatment of the topic of system files to an appendix. While this may at first glance seem to be a wasteful approach which requires the user to run the same program over and over, I believe it helps the student more clearly understand where the data comes from. For the same pedagogical reasons, only a fixed data format is used. Not only does a fixed data column arrangement promote a better appreciation of how each variable is associated with every case; it also avoids inevitable sequence displacement problems that occur when the free format is used.

I have deliberately omitted discussion of the menu driven command system and the optional data entry module. While some users find these features useful, I believe that a clearer understanding is reached by typing in one's own commands. The menus can't supply such information as the variable and value labels, which account for the majority of the program anyway. What is more, once one gets into trouble with the menus, it is often
very difficult to get out. As for typing in the data directly, rather than through the data entry module, it is my experience that such practice helps the user appreciate the fixed format of the data entry. Also, this method of data entry is more efficient for large data sets since the "key puncher" is not required to enter a carriage return after entering each variable.

These approaches, along with my avoidance of alphanumeric variables, further allow for a better generalization to other statistical packages and formats.

Students today are increasingly computer literate, and it is not unusual for students to be familiar with such software programs as Lotus 123 and dBase. I have included appendices that illustrate similarities and differences between these programs in order to help students who already know the programs easily understand comparable procedures and operations of SPSS/PC+. An appendix also indicates how data can be imported from such sources as Lotus 123.

The other side of the coin is that many people still feel uneasy when it comes to dealing with computers. Many official software manuals, and even a number of secondary texts, tend to present the material in a relatively abstract and "sterile" manner. At the loss of a slight degree of generality, I have instead chosen to present examples that will help the user "fill in the blanks" on his or her own programs. That is, I frequently use "real" variable names rather than referring to some abstract notion such as "varlist." In addition, I have used a great deal of humor and anthropomorphism to put the "reluctant" computer user off guard. On that topic, I see no reason why humorous examples cannot be as informative and educational as boring ones. The purpose of examples is to show the student how data can be analyzed, and while "real World" projects may be less engrossing, the skills learned in a humorous situation can be generalized to a routine, or perhaps even boring, situation.

The ease with which computers can perform a large number of computations within a few moments creates the potential for a great deal of abuse. It is almost certain that something "significant" will show up if one performs enough tests. Thus, I have strongly emphasized the practice of
making a limited number of hypotheses before running the statistical procedures. The program named "SIGNIF.EXE," which is included on the distribution diskette for this manual, allows the user to compute the probability of making at least one Type I error given \( n \) significance tests. Having noted students' tendency to attempt a very large number of analyses, it is strongly recommended that the instructor stress the concept of accumulating error levels in class.

Since this is intended as an introductory textbook, it only discusses those statistical techniques that would likely be encountered in an introductory course. However, as students complete this book and the exercises contained herein, they should be well prepared to consult the official manuals issued by SPSS, Inc. These manuals also provide an excellent discussion of the theory behind the statistical procedures involved.

Whenever one attempts to write a textbook for even a mildly diverse group of readers, the question of how much background should be assumed from the reader arises. In the present case, such a concern is particularly relevant when it comes to deciding how statistical output should be interpreted. My choice, for better or for worse, has been to leave any extensive discussion of the statistical principles involved to the instructor and/or any other textbooks and reference materials that may have been made available to the student. I have only touched lightly on such topics as statistical significance, although examples and brief discussions of the techniques available might suggest applications appropriate for specific statistical procedures. In order to accommodate users with highly diverse statistical backgrounds, I have included, as an appendix, an expanded glossary that explains many of the terms that the student may find unfamiliar.

With SPSS being available in so many forms, one may wonder about the wisdom about using the personal computer version as opposed to mainframe versions such as SPSS-X available on most campuses. After all, SPSS/PC+ may be installed only on a few computers on campus while mainframe terminals are readily available to students, some of whom may even be able to dial up the university mainframe by modem from home. I think there are several reasons why the PC version is preferable. First, many
students have already had experience on the IBM PC and thus need much less introduction. Secondly, should the student wish to include part of the SPSS output directly in a report, taking it from the "SPSS.LIS" output file is much easier than downloading an output file from the mainframe. Finally, those students who will end up using SPSS on the job are much more likely to find a SPSS/PC+ than the mainframe version in industry.

In that same vein, a question arises as to whether one should use the complete SPSS/PC+ program or SPSS/PC+ Studentware. While I feel that using the "real thing" will be a better preparation for practical industry applications, I don't think that students who only have the Studentware edition available will be seriously shortchanged in an introductory course.
To the Student

Much has changed since I was first exposed to the Statistical Package for the Social Sciences (SPSS) in the early 1980s. As an industry standard, SPSS now exists in versions for many different computers. At the time, however, I was confined to a mainframe version which did not seem terribly user friendly.

I didn't find the textbook assigned to be appreciably more user friendly than the program. After several chapters referring to such abstract terminology as a "field," we finally got to write some programs, but the coverage was still rather abstract. I have chosen a different, and hopefully more readable approach, to writing about SPSS/PC+.

Many textbooks take the approach of presenting the theory fully before providing any examples. I don't find that a good way to learn. Few of us learned to talk by studying a dictionary. Most of us instead learned by listening to others and then adapting the language to suit our own needs.

I have chosen a similar approach in writing this book. First, we will look at some SPSS/PC+ programs and explore what they do. Looking at these examples, we will discuss how you can generalize the techniques presented to meet your own needs.

This book is an introduction to SPSS/PC+, and as such, it covers only a fraction of the options available. If you go on to use SPSS/PC+ extensively, you will probably find the official manuals published by SPSS/PC+ to be invaluable. Not only do these manuals provide a great source of reference for the SPSS/PC+ procedures; they also provide an excellent and very readable discussion of the statistical techniques available.

I have included some relatively humorous examples in this text. Aside from my own exaggerated sense of humor (which has a tendency to get me into trouble on many occasions), I think a witty approach may help to (1) motivate you to keep up with the reading, and (2) help those people
who feel uncomfortable with computers to feel more at ease with the subject. You should not feel guilty about enjoying the reading, however. While you may risk getting some dirty looks in the library if you laugh out loud, the examples, although often far fetched, illustrate real research issues and are just informative as boring examples. Why shouldn't enlightened poodle breeders commission marketing research just like the manufacturers of laundry detergent? Those people who split their investment funds between an inventory of poodles and a controlling interest in Proctor & Gamble will be just concerned about increasing levels of prejudice against small dogs as about consumer trends toward buying generic household products.

A few final cautions. The computer has today made it very easy to perform statistical calculations that could literally have taken a person months to perform in past decades. With this potential, however, comes an opportunity for serious abuse. This can take two forms. First, anyone can do complex statistical calculations in SPSS/PC+, but the output may not be at all meaningful. In class, you may have discussed the distinction between nominal, ordinal, interval, and ratio scales of measurement. However, the computer doesn't know where your data comes from and will gladly comply with your request to include nominal data in a procedure that really requires interval, or even ratio, level data. Therefore, be sure you understand the assumptions behind a statistical technique before running it.

The second potential for abuse results from our ability to indiscriminately perform a great number of analyses at the same time. Intuitively, it makes sense that if we try long enough to find something that is "very unlikely," we will. Suppose, for example, that you decide to make a telephone survey to find out the birthdays of the respondent and his or her spouse. You call up a thousand married people at random from the phonebook (i.e., you terminate the interview once you find out that someone is not married). You then find out that three couples have their birthdays on the same date. A ny great discovery? Well, in each trial (call), the chance of "success" is approximately 1/365=0.00273. Multiplying our one thousand trials times that 0.273% chance, we would expect about 2.73 couples in our sample to have the same birthday. Similarly, if you perform fifty significance tests at an $\alpha=0.05$ level of significance, you would expect 2.5
tests to come out significant by chance alone.
What Can SPSS/PC+ Do for Me?

SPSS/PC+ is an incredibly beautiful program. If you like computer software, you might think of SPSS/PC+ as being as beautiful as programs like WordPerfect and Lotus 123. Unless you are a real nature enthusiast, you will probably find SPSS/PC+ even more impressive than a beautiful river or mountain range. SPSS/PC+ will probably compare favorably with the grandest piece of literature or greatest work of art you have ever seen. And, depending on how humanistic you are, you may also find SPSS/PC+ more beautiful than the one you love. It's not surprising if, at first, you find this statement difficult to believe, so let's get right into the features of SPSS/PC+. You be the judge! (This chapter is intended to show you the variety of statistical procedures available within SPSS/PC+, and may cover statistical methods that you have not yet studied. The purpose is only to show you what is available, and consequently, the chapter is not intended for detailed study.)

At the most basic level, you might want to tabulate some data you have collected in a survey or through other means. Later on in the book, we will meet a dog breeder who is very interested in whether people own dogs or not and what kind they prefer. Suppose he has asked you to do a survey. After you have entered the data, you can ask SPSS/PC+ for a table that indicates how many people gave each of the possible answers to a question:

<table>
<thead>
<tr>
<th>DOGOWN</th>
<th>Ownership of Dog</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Valid</td>
</tr>
<tr>
<td></td>
<td>Value Label</td>
</tr>
<tr>
<td>Yes</td>
<td>1 82</td>
</tr>
<tr>
<td>No</td>
<td>2 82</td>
</tr>
<tr>
<td>Not sure</td>
<td>3</td>
</tr>
<tr>
<td>Missing</td>
<td>9</td>
</tr>
<tr>
<td>TOTAL</td>
<td>175</td>
</tr>
</tbody>
</table>

Mode 1.000 Minimum 1.000 Maximum 3.000

For example, the SPSS/PC+ output indicates that eighty-two people claimed to have a dog, eighty-two claimed not to have one, three people...
were not sure if they owned a dog or not, and nine people supplied answers that were not interpretable or supplied no answer at all; hence, their answers are "missing."

Although SPSS/PC+ does not provide good graphics capabilities except in an optional graphics module that many institutions do not have, you can use a spreadsheet or graphics package like Lotus 123, Excel, or Quattro to create a graph to depict the responses given. It might look something like this:¹

OK, so this saved you some time and provided an output that was somewhat neater and more organized than what you would have obtained if you had done the calculations by hand. Is that all we have to be excited about?

Of course not! We are just beginning. SPSS/PC+ allows us to do more involved things as well. For example, we might be interested in assessing the relationship between two or more variables. One popular feature allowing us to do this is the "crosstabs" table. Let’s suppose that you have been hired by a major airline that wants to diversify into the hospitality industry at its destination sites, offering consumers an integrated vacation package. In order to establish the kinds of restaurants that will appeal most to vacationers at each location, the airline would like to know if there is any relationship between food preference and favorite vacation destination. After consulting your marketing textbook, you decide to do a crosstabulation. You think that people who prefer the Orient would be more likely to prefer Chinese food; those people who prefer Europe would like

¹There is no need to cut and paste! Lotus graphics (*.pic files) can be imported directly into WordPerfect 5.0 or 5.1.
Italian and French food; and those people preferring the Continental U.S. would prefer Western type food such as steaks, hamburgers, and fries. You are not quite sure about those who prefer to visit Hawaii. SPSS/PC+ allows you to test your hypotheses:

<table>
<thead>
<tr>
<th>FOODPREF</th>
<th>U.S. Mainland</th>
<th>Orient</th>
<th>Europe</th>
<th>Hawaii</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>20  33  59  29  5  150</td>
<td>24  33  59  29  5  150</td>
<td>24  33  59  29  5  150</td>
<td>24  33  59  29  5  150</td>
</tr>
<tr>
<td>Row Pct</td>
<td>66.7 3.3 16.7 10.0 3.3 20.0</td>
<td>83.3 3.0 8.5 10.3 20.0</td>
<td>83.3 3.0 8.5 10.3 20.0</td>
<td>83.3 3.0 8.5 10.3 20.0</td>
</tr>
<tr>
<td>Vertical</td>
<td>1  1  1  1  2  16</td>
<td>1  1  1  1  2  16</td>
<td>1  1  1  1  2  16</td>
<td>1  1  1  1  2  16</td>
</tr>
<tr>
<td>Col Pct</td>
<td>6.3 6.3 68.8 6.3 12.5 10.7</td>
<td>4.2 3.0 18.6 3.4 40.0</td>
<td>4.2 3.0 18.6 3.4 40.0</td>
<td>4.2 3.0 18.6 3.4 40.0</td>
</tr>
<tr>
<td>Row Pct</td>
<td>3.6 49.1 7.3 40.0 36.7</td>
<td>8.3 81.8 6.8 75.9</td>
<td>8.3 81.8 6.8 75.9</td>
<td>8.3 81.8 6.8 75.9</td>
</tr>
<tr>
<td>Column</td>
<td>24  33  59  29  5  150</td>
<td>24  33  59  29  5  150</td>
<td>24  33  59  29  5  150</td>
<td>24  33  59  29  5  150</td>
</tr>
<tr>
<td>Total</td>
<td>16.0 22.0 39.3 19.3 3.3 100.0</td>
<td>16.0 22.0 39.3 19.3 3.3 100.0</td>
<td>16.0 22.0 39.3 19.3 3.3 100.0</td>
<td>16.0 22.0 39.3 19.3 3.3 100.0</td>
</tr>
</tbody>
</table>

Chi-Square: 156.76735  D.F.: 12  Significance: .0000  Min E.F.: .533  Cells with E.F.<5: 80 (40.0%)  Cramer's V: .59023

If you are familiar with the Chi square statistic, you can see that there is strong evidence to reject the null hypothesis that food and vacation preference are "independent." As a matter of fact, the Cramer's V statistic

2 Normally, you should define hypotheses more specifically before testing them. For now, we are just testing whether the two variables in question (food preference and favorite vacation destination) are dependent.
even suggests a modest relationship.

But, you are of course not limited to non-parametric statistics, or procedures that only use ordinal measures, in SPSS/PC+. You can also do a Pearson correlation analysis if you have interval data or "better." Let's suppose that the airline is interested in tourist travel and would like you to conduct a study of how to best predict how much money a person spends on vacation(s) every year. You decide to correlate amount spent on vacations against various other variables.

Correlations: FOOD AUTO ENTERT DINING PHONE

VACATION .4304 .5754 .5728 .5620 .3498
(263) (263) (263) (263) (263)
P = .000 P = .000 P = .000 P = .000 P = .000

Once you have done a correlation analysis, you might feel that the proper next step is to do a multiple regression analysis to see if you can improve your ability to predict based on the introduction of additional variables. Unlike Lotus 123, SPSS/PC+ gives you several choices as to which method you want to use (forward inclusion, backward deletion, stepwise consideration, or "forced" entry). If a traditional method doesn't suit your needs, you can introduce non-linear or log-linear models. Let's try to "predict" a person's telephone bill from his or her expenditures on other items and other demographic information.

---

Actually, a correlation analysis is in practice applied many times even when only ordinal level data is available. This is not sanctioned as correct by most statisticians, but the this approach can sometimes still yield meaningful results. When one or both of the variables depart seriously from the assumption of interval properties, the true relationship between the variables may be greatly underestimated. On the other hand, a correlation will rarely provide "false positives" or suggest a relationship that does not exist.
Variable(s) Entered on Step Number
 3. COMPUTER ownership of computer

Multiple R .67310
R Square .45306
Adjusted R Square .44673
Standard Error 84.09635

Analysis of Variance

<table>
<thead>
<tr>
<th></th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F</th>
<th>Signif F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1517312.98023</td>
<td>505770.99341</td>
<td>71.51541</td>
<td>.0000</td>
</tr>
<tr>
<td>Residual</td>
<td>1831698.85247</td>
<td>7072.19634</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

F = 71.51541 Signif F = .0000

Equation Number 1 Dependent Variable PHONE total household phone bill

--------------- Variables in the Equation ---------------

<table>
<thead>
<tr>
<th>Variable</th>
<th>B</th>
<th>SE B</th>
<th>Beta</th>
<th>T</th>
<th>Sig T</th>
</tr>
</thead>
<tbody>
<tr>
<td>DINING</td>
<td>.09375</td>
<td>.01064</td>
<td>.58737</td>
<td>8.811</td>
<td>.0000</td>
</tr>
<tr>
<td>AUTO</td>
<td>.22249</td>
<td>.10231</td>
<td>.14281</td>
<td>2.175</td>
<td>.0306</td>
</tr>
<tr>
<td>COMPUTER</td>
<td>-36.34246</td>
<td>16.83655</td>
<td>-.10385</td>
<td>-2.159</td>
<td>.0318</td>
</tr>
<tr>
<td>(Constant)</td>
<td>408.56124</td>
<td>31.01232</td>
<td></td>
<td>13.174</td>
<td>.0000</td>
</tr>
</tbody>
</table>

--------------- Variables not in the Equation ---------------

<table>
<thead>
<tr>
<th>Variable</th>
<th>Beta In Partial Min Toler</th>
<th>T</th>
<th>Sig T</th>
</tr>
</thead>
<tbody>
<tr>
<td>FOOD</td>
<td>-.05499 - .07431 .47474 -1.197 .2324</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENTERT</td>
<td>.02691 .01453 .13203 .233 .8156</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HOME</td>
<td>-.02361 .02499 .36798 .401 .6884</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VCR</td>
<td>-.02999 -.03802 .45710 -.611 .5417</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

End Block Number 1 PIN = .050 Limits reached.

Not all statistical analyses have to be this involved. Maybe by now you are getting tired of doing research for someone else. Suppose that, for some reason, you have a good cause to believe that there is a killing to be made in retailing tall people's clothing. Not having a preference for either men's or women's clothing, but realizing that you need to specialize to compete, you toss a coin and end up in the women's apparel business. Naturally, you will want to find a geographic location where there are a lot of tall women. Since you can't afford to research the heights of women all over the nation, you decide to focus on Texas in the belief that Texas cities will be the most likely spots for success. You now proceed to collect data on the heights of adult women (in inches) in Dallas and Houston. One way to
test for such differences would be to employ a t-test (two-tailed since you have no preconceptions), but a more direct way might be to "break down" information by city. Using the SPSS/PC+ procedure Means you get the following result:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Value</th>
<th>Label</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>For Entire Population</td>
<td>66.6960</td>
<td>Height of subject (inches)</td>
<td>1.7183</td>
<td>200</td>
<td></td>
</tr>
<tr>
<td>CITY</td>
<td>1</td>
<td>Dallas</td>
<td>68.1830</td>
<td>.9038</td>
<td>100</td>
</tr>
<tr>
<td>CITY</td>
<td>2</td>
<td>Houston</td>
<td>65.2090</td>
<td>.8069</td>
<td>100</td>
</tr>
</tbody>
</table>

Total Cases = 230
Missing Cases = 30 OR 13.0 PCT.

Now you can optionally calculate a confidence interval for the means of heights of the women on in each. Tentatively, it looks like Dallas might be the best bet. (An added benefit is that by choosing this location, you will be closer to the Oil Barons' Club).

Of course, there is always the possibility that you decided to split the cost of doing the survey with a classmate who believes that there is more money to be made in tall men's clothing. In that case, of course, you would want to find out about the heights of the men in the different cities as well. However, you would want to keep track of the heights of the men and women separately, both because the city that has the tallest men might not have the tallest women and because the great between-sex height differences would greatly inflate the estimate of within-gender variability. "Means" allows you to produce this table:
Now that you have been working with SPSS/PC+ for quite some time, you are really getting good at marketing research, and you feel that you can handle almost anything—even the unexpected. One day, you receive a distressed phone call from Rudolph the Redneck Reindeer. Rudolph is hysterical because his employer, an elderly man who likes to wear a red suit during the winter months, has warned his long time sleigh puller that he may have to lay him off because people are beginning to demand greater sophistication from reindeer. You agree to do a survey for Rudolph to find out how important that aspect really is to consumers. However, having studied marketing research for some time, you realize that one question or "item" will not give you a result that is reliable enough to give you an answer that is dependable. You therefore decide to create a scale of "Appreciation of Sophistication in Reindeer," where subjects will be asked to indicate their level of agreement or disagreement with Likert type questions on a scale ranging from "strongly agree" (1) "to neither agree nor disagree" (4) "to strongly disagree" (7). Now you want to test whether the average score on the questions will be reliable enough to be meaningful. You and Rudolph hope that people will score as low as possible on that scale, suggesting that his employer's concern is unwarranted. After you "reverse score" item #4 (which is worded in the opposite direction of the other questions), you are ready to generate the following estimate of internal consistency:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Value Label</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>CITY</td>
<td>1 Dallas</td>
<td>70.6840</td>
<td>2.6261</td>
<td>100</td>
</tr>
<tr>
<td>SEX</td>
<td>1 Male</td>
<td>73.1160</td>
<td>.8728</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>2 Female</td>
<td>68.2520</td>
<td>1.0492</td>
<td>50</td>
</tr>
<tr>
<td>CITY</td>
<td>2 Houston</td>
<td>67.2800</td>
<td>2.2614</td>
<td>100</td>
</tr>
<tr>
<td>SEX</td>
<td>1 Male</td>
<td>69.3700</td>
<td>.8929</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>2 Female</td>
<td>65.1900</td>
<td>.7875</td>
<td>50</td>
</tr>
</tbody>
</table>

Total Cases = 230
Missing Cases = 30 OR 13.0 PCT.
RELIABILITY ANALYSIS - SCALE (SOPHIST)

1. SOPHIS1  Good reindeer are educated
2. SOPHIS2  Reindeer should have good manners
3. SOPHIS3  A reindeer should have a good cultural background
4. SOPHIS4  Redneck reindeer are OK
   (Reverse scored)
5. SOPHIS5  Reindeer should use good grammar
6. SOPHIS6  Reindeer should be graceful
7. SOPHIS7  A reindeer's style is more important than the color of his nose

RELIABILITY COEFFICIENTS

N OF CASES = 100.0   N OF ITEMS = 7

ALPHA = .9218

Assessing internal consistency is an advanced topic, and you may not appreciate this capability yet, but the time will come! Please be patient until the end of the quarter.

SPSS/PC+ allows you to do many other beautiful things such as discriminant analysis, one-way analysis of variance, multivariate analysis of variance, factor analysis, cluster analysis, and various non-parametric statistics. However, by now you ought to have seen enough to make an informed judgment.

What's your verdict? I think you will agree that SPSS/PC+ is as beautiful as WordPerfect and various other software programs. At least it is more beautiful than a bouquet of roses or a beautiful waterfall. Now, how does it compare to that special person?
Introduction

Today, the data analysis involved in a marketing research project of any real size is almost universally done by computer. Most statistical procedures involve a great number of arithmetic calculations which are not really "difficult" to do by hand, but require a tremendous amount of repetitious work. Not only is this work boring and time consuming, but it also provides a great deal of opportunities for little mistakes which can seriously distort your actual results.

Today, many statistical software packages are available to help the researcher avoid the repetitious work involved in number crunching. Not only does such software allow us to save time, but the programs will also allow one to do analyses which simply wouldn't have been practical to perform in past years. While a researcher normally can't afford to literally spend four weeks doing the calculations involved in a regression analysis of two hundred subjects with, say, ten independent variables, the computer will do this analysis for him or her in seconds--that is, once the data set has been entered, or typed, into the computer.

Although the micro computer version of the Statistical Package for the Social Sciences (SPSS/PC+) is one of the leading statistical software packages on the market, it is by no means the only useful one available. The group of other powerful statistical packages includes the mainframe version (SPSS-X and SPSS v. 9.0) and such programs as the Statistical Analysis System (SAS--available in both micro and mainframe versions) and the Biomedical Statistical Package (BMDP). Generally, you will find that the skills you learn while using SPSS can easily be transferred to these other software with minor only modifications. Other programs, such as Minitab, are slightly easier to learn but not nearly as powerful and flexible.

SPSS has many features, of which you will probably only be using a few. It is important not to lose sight of the forest for the trees (or, in more modern terms, not to lose sight of the computer for the chips). This manual contains descriptions of many simple procedures to run, with more information being available in various manuals put out by SPSS, Inc. and
third party sources. (If you are unsure about particular statistical procedures, these manuals also function very well as statistical texts since they contain very good, real life illustrations of the statistical techniques discussed). The attempt of this book is not to teach you all the details of SPSS/PC+, but simply to allow you to adapt sample programs to your needs.

Please don't be intimidated by the reference to "programming" within SPSS/PC+. All this involves is putting together a few instructions for the computer telling it information about your data and how you would like to analyze it. As a reasonably "user friendly" program, SPSS/PC+ accepts relatively "English-like" commands that make a great deal of sense even those people who don't spend most of their lives reading computer books and magazines. Questions that SPSS/PC+ would like to have answered relate to issues such as:

- How many variables are in the data?
- What do the different values mean?
- What happens if a person didn't answer a question?

Considering the amount of time SPSS/PC+ saves us in doing the calculations, I think it is fair enough that SPSS/PC+ expects to get that kind of information from us as a sort of "retainer." (In any event, it is not worth the effort to try to bargain with SPSS/PC+ since we need it more than it needs us).

As we go through the writing of a program for a hypothetical questionnaire, you will be able to modify the statements of information for that program to write one that fits your data.
Step 1: Coding Your Data

Assuming that you have already collected your data, the statistical analysis generally involves five stages:

1. Data coding, variable naming, and classification;
2. Statistical program writing;
3. Data Entry;
4. Error checking; and
5. Data analysis

We will go through each stage separately. In this chapter, we will discuss data coding.

In Figure 1-1, you will find a sample questionnaire, commissioned by a poodle breeder concerned about possible increasing prejudice against small dogs, for which we will code and write an SPSS/PC+ program. Please understand that this is not supposed to be an example of a good questionnaire. As a matter of fact, if you plan to use one like it in your marketing class, you should probably be prepared for a relatively low grade. The purpose of the questionnaire, instead, is to demonstrate how to code a number of different questions. Later in this book, we will get to the touching story of the poodle breeder who has hired you to do a study of, among other things, prejudice against poodles.
DOG PREFERENCE QUESTIONNAIRE

We are interested in your opinion about dogs and dog care. Please take a few minutes to respond to the questions below.

1. Do you own a dog? ___ Yes ___ No ___ Not sure
   (If no, please skip to question 6)

2. How many dogs do you own? _______

3. What is the breed of your favorite dog? _____

4. Please rank the following foods in the order you prefer to feed them to your dog:
   ____ generic dry dog food      ____ Lucky Dog food
   ____ generic canned dog food   ____ Kit 'n' Caboodle
   ____ Kit 'n' Caboodle

5. How much do you spend on feeding your dog per week? $_______

6. Please write next to each of the questions below the number from the following scale which most closely matches your level of agreement or disagreement:

   1 2 3 4 5 6 7

   Strongly agree          Strongly disagree
   ___ 1. Poodles are fragile
   ___ 2. Poodles are stupid
   ___ 3. Poodles are self-centered
   ___ 4. Poodles are cute
   ___ 5. Poodles are over-priced

7. Age: ____ Sex (Please circle): Male Female
       Annual household income: $______________

Figure 1: The Poodle Breeder’s Questionnaire
Whenever you plan to enter the contents of several questionnaires into the computer, it is always a good habit to number each questionnaire. Let's do that in the top right corner.

The need to number the questionnaires introduces an important data coding concept. Since we will ultimately be entering the data as numbers on one or more lines of text, we will want to determine the maximum number of digits that a variable may take up. Suppose we administered three hundred questionnaires and numbered them, starting with one. In order to be able to express all these ID numbers, we would need a minimum of three digits for the ID field (since the numbers between 100 and 300 each take up three digits. For a questionnaire administered to less than one hundred people, we would only need two digits for the ID.) If we assume that we have three hundred surveys, we will eventually enter the ID number this way:

```
001
002
003
...> more data here <....
...
300
```

It is, of course, acceptable to allocate more digits than needed to a variable.

Now we get to actually code the questions. This includes giving the question (or variable) a one word name, assigning a number to each possible answer on the questionnaire, determining how many digits are needed for the question, determining in which columns the data will be put and, SPSS/PC+ actually allows the use of alphanumeric characters, that is, letters of the alphabet, as data. However, the use of alphanumeric data will often cause problems which are difficult to solve and it may be a poor practice since many other statistical programs will not allow such data.
optionally, assigning the question and answers "labels," i.e. short phrases describing their meaning. Although this process may sound overwhelming, it is quite simple once we go through it.

The name of the variable can simply refer to the number of the question (e.g. "Q1," "ITEM 18"), or it can be descriptive of the meaning of the variable (e.g. "AGE," "INCOME."). The rules for naming SPSS/PC+ variables are very similar to those for naming DOS files, i.e.

- A variable must begin with a letter of the alphabet (unlike DOS, a number is not acceptable) and may be up to eight characters long.
- There may be no spaces in the name (thus, we say "ITEM 1" and not "ITEM 1").
- A variable name may not be the same as certain "reserved" words, i.e. words that SPSS/PC+ uses for internal and programming purposes. These reserved words are very few and far between, but you cannot name your variables "AND," "OR," or "IF" as well as a few other obscure names.
- SPSS/PC+ does not distinguish between upper and lower case letters. "QUEST1," "Quest1," and "quest1" all refer to the same variable and can be used interchangeably.
- Unlike DOS file names, a variable name cannot have an extension.

The following variable names are valid:

Q1
QUEST1
QUEST1A
QUESTA1

However, the following names are not acceptable:

1st Starts with a number
Quest 1 Contains a space
Marketshare Consists of more than eight characters
1st question Violates all of the above
Let's look at question number 1. Our first task is to name it. Either you can call it something like Q1, to keep it simple, or you can name it something more descriptive like "DOGOWN." When you write your own program, you have a choice; for now, we will call it DOGOWN. Next, notice that there are three possible answers. (Your client insisted on including the "Not sure" option since the questionnaire would be administered in the neighborhood of a major university, making it likely that a number of absent minded professors would be asked to respond.) We now have to assign a number to each. Let's assign a "1" to "Yes," a "2" to "No," and a "3" to "Not sure." Now, are these response categories enough?

Not really. Two things could happen. First, the respondent might accidentally overlook or refuse to answer the question (a common situation when you ask about such emotionally charged and private topics as income or extra marital activities). The next several questions illustrate the situation that occurs when not everybody is supposed to answer a question, and we will discuss how to code such instances when we talk about those questions. For this question, missing data can only arise when a person either omits the question or provides one that is not useful. This could happen if someone wrote a sarcastic comment instead of answering or simply overlooked the question. Whenever a person fails to answer a question that he or she should have answered, we will assign a numeric value. For this question we will code it as a "9." When the question is not applicable, just leave the space blank and the computer will assign the response as a "system missing value." (Notice that using "9" as the missing value for this question will result in all the numbers in between four and eight, inclusive, not being used).

The next question, which we will call DOGCOUNT, asks how many dogs the respondent owns. We will assume that the respondents are reasonably normal and do not own more than 98 dogs; hence, we will reserve only two digits for that variable. Note that here we may encounter the kind of missing data that occurs when people legitimately omit questions since the questions between sections two and five should only be answered by those people who own dogs. When people "legitimately" skip questions, we will put blank spaces in the columns designated for the
variable. For those people who indicated that they own at least one dog by answering "Yes" to the first question but failed to answer this question, we will put in the missing value of "99."

The third question, which we will call FAVORITE, is somewhat more complicated. This is what is called an "open ended" question; in other words, the subject is asked to write an answer and is not given a list of "acceptable" answers from which to choose. Therefore, we have to try to match each written answer with some kind of code that is general enough to be meaningful. Since more than nine breeds are likely to be mentioned, we will reserve two digits for this question. Notice that in order to occupy two columns, the numbers zero through nine must either be preceded by a blank or a zero. Thus, "01" might be "Poodle;" 02, "Fox Terrier;" 03, "Yorkshire Terrier;" and so forth. Other options such as "10" as "German Shepherd" are possible. (Several other breeds are listed in the program that we will get to). Again, a correctly omitted answer will result in two blank spaces and other cases of missing data will result in a code of "99."

Note that open ended questions have a great potential for missing data. Suppose that someone misunderstood the question and thought he or she was asked about his or her favorite pet. If he or she answered "Polar bear," you would most likely classify the answer as missing. You should be prepared for certain other kinds of "missing answers." Perhaps a respondent unsympathetic to the objectives of our research might scribble in something like "I hate all dogs!"

To keep organized, we might consider giving very small dogs low numbers, say, below 29; medium sized dogs intermediate numbers between 30 and 59, and big dogs the bigger numbers between 60 and 98. Depending on our research objective, we might also like to "cluster" the numbers on the basis of other variables such as price, friendliness/ viciousness, durability, guarding ability, or lifespan. Or, if we cared about none of those issues, we could simply list the dogs in alphabetical order.

The next questions require very little discussion since subjects will be responding directly with numbers. Thus, for section 4, all we have to do is to assign one digit to each variable (i.e. each dog food that we ask the
respondent to rate) and reserve "9" as a missing value. Skipping slightly ahead, the same holds for the Likert scale questions of Section 6.

In section 5, we will assume that no one spends more than $99.98 per week on dog costs, and we will thus assign five digits and a missing value of "99.99." (SPSS/ PC+ allows you to designate a variable as a dollar amount rather than as plain number. However, not all programs have that capability, so let's not get into that.) Also notice that we have reserved space for the period. We could arrange to use four digits instead, but why be so stingy with the space?

For age, we will assign two digits and a missing value of "99." For sex, "male" will be assigned "1" and "female" "2."

For the question of annual household income, we will assume that the figure does not exceed $999,998 for any respondent since we are located in a university community. We will thus reserve six digits since SPSS/ PC+ would not appreciate the comma (unlike the period allowed in the question on weekly spending). Incidentally, anyone earning over a million dollars a year would, in statistical jargon, probably be considered an "outlier"--a sort of maverick who would probably be excluded from our analysis anyway.

The table below summarizes the lengths, missing value indicators, and column positions of each variable. A "case," or collection of answers from one respondent (who owns a dog), could thus look like this:

```
001 1 02 02 52314 03.50 66727 29 2 026000
```

while a case from a person who does not own a dog, and thus was not asked to answer some of the questions, could look like this:

```
002 2 22252 31 1 032150
```

Remember, we need $999,999 (i.e. 999999) for missing data.
<table>
<thead>
<tr>
<th>Variable</th>
<th>Missing Value</th>
<th>Length</th>
<th>Columns</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>None</td>
<td>3</td>
<td>1-3</td>
<td>Identification number</td>
</tr>
<tr>
<td>OWNDOG</td>
<td>9</td>
<td>1</td>
<td>5</td>
<td>Ownership of dog</td>
</tr>
<tr>
<td>DOGCOUNT</td>
<td>99</td>
<td>2</td>
<td>7-8</td>
<td>Number of dogs owned</td>
</tr>
<tr>
<td>FAVORITE</td>
<td>99</td>
<td>2</td>
<td>10-11</td>
<td>Breed of favorite dog</td>
</tr>
<tr>
<td>FOOD1</td>
<td>9</td>
<td>1</td>
<td>13</td>
<td>Ranking of generic dry</td>
</tr>
<tr>
<td>FOOD2</td>
<td>9</td>
<td>1</td>
<td>14</td>
<td>Ranking of generic canned</td>
</tr>
<tr>
<td>FOOD3</td>
<td>9</td>
<td>1</td>
<td>15</td>
<td>Ranking of Mighty Dog</td>
</tr>
<tr>
<td>FOOD4</td>
<td>9</td>
<td>1</td>
<td>16</td>
<td>Ranking of Lucky Dog</td>
</tr>
<tr>
<td>FOOD5</td>
<td>9</td>
<td>1</td>
<td>17</td>
<td>Ranking of Kit 'n' Caboodle</td>
</tr>
<tr>
<td>SPEND</td>
<td>99.99</td>
<td>5</td>
<td>19-23</td>
<td>Weekly expense of dog food</td>
</tr>
<tr>
<td>LIKERT1</td>
<td>9</td>
<td>1</td>
<td>25</td>
<td>Poodles are fragile</td>
</tr>
<tr>
<td>LIKERT2</td>
<td>9</td>
<td>1</td>
<td>26</td>
<td>Poodles are stupid</td>
</tr>
<tr>
<td>LIKERT3</td>
<td>9</td>
<td>1</td>
<td>27</td>
<td>Poodles are self-centered</td>
</tr>
<tr>
<td>LIKERT4</td>
<td>9</td>
<td>1</td>
<td>28</td>
<td>Poodles are cute</td>
</tr>
<tr>
<td>LIKERT5</td>
<td>9</td>
<td>1</td>
<td>29</td>
<td>Poodles are over-priced</td>
</tr>
<tr>
<td>AGE</td>
<td>99</td>
<td>2</td>
<td>31-32</td>
<td>Age</td>
</tr>
<tr>
<td>SEX</td>
<td>9</td>
<td>1</td>
<td>34</td>
<td>Sex</td>
</tr>
<tr>
<td>INCOME</td>
<td>999999</td>
<td>6</td>
<td>37-42</td>
<td>Annual household income</td>
</tr>
</tbody>
</table>

A look at the column assignments and the two example cases will show that we have put spaces between some variables and not between others. SPSS/PC+ leaves it entirely up to us whether we want to leave such spaces or not. One of the first commands of the program simply tells it the column positions of each variable. Thus, we can to put spaces where they improve our ability to type in and read the data.

For each case, a variable will be in the same column position(s). This can be a great help to ensure that we are on target when we type in the data. Consider, for example, the following two cases:

003 1 02 02 52314 03 50 66727 29 2 026000 004 1 01 03 32512 05 50 67615 21 2 018500

Not only do we know that each case should end in the same column; we also know that many of the blank spaces should be in the same places. In this case, we have reserved blank spaces between most of the variables, but none within the variables in the Likert scale and rank-order sections. Putting in spaces there would make the data seem more confusing.
You should try to be as consistent as possible when coding different questions within the same questionnaire. If for one question you use the code "1" for yes and "2" for no, you should try to keep that practice throughout the questionnaire.

When several people enter data, be sure that everyone understands and agrees upon a coding system. There are numerous ways the same variable can be coded. The variable SEX above, for example, could have the codes "1" for male and "2" for female (as we did); "1" for female and "2" for male; and "0" for female and "1" for male. None is inherently better or more correct than the other (although the first one is the most commonly used approach), but if several individuals use different codes in the same file, a big editing job could result.

Sometimes, the best way to code a variable does not coincide with the coding that may be "suggested" by the questionnaire. Consider this example:

How many persons over the age of 55 currently live in your household?
___ a. None
___ b. One
___ c. Two
___ d. Three
___ e. Four or more

In this case, it might be tempting to start off at the beginning and number the choices from 1 through 5. This, however, would be a bad approach, since the code "2" would now correspond to the answer "one person." Instead, you would be better off starting with zero and advancing. That way, the code will correspond to the number of
people actually in the household.
Step 2: Writing An SPSS/PC+ Program

Now we are ready to pursue the bottom line of this text, that is, the writing of the SPSS/PC+ program. The next chapter will discuss how to enter the data and commands into SPSS/PC+; here, we will just talk about what to enter. Here is a program for the questionnaire we have been discussing. Please don't be intimidated if it looks overwhelming at first. We will go through it line by line.

```
TITLE "Dog Preference Study".
DATA LIST / id 1-3 owndog 5 dogcount 7-8
   favorite 10-11 food1 to food5 13-17
   spend 19-23 (2) likert1 to likert5 25-29 age 31-32 sex 34
   income 37-42.

VARIABLE LABELS
   owndog "Ownership of dog"
   dogcount "Number of dogs owned"
   favorite "Breed of favorite dog"
   food1 "Rating of generic dry dog food"
   food2 "Rating of generic canned dog food"
   food3 "Rating of Mighty Dog"
   food4 "Rating of Lucky Dog"
   food5 "Rating of Kit 'n' Caboodle"
   spend "Weekly spending on dog food"
   likert1 "Poodles are fragile"
   likert2 "Poodles are stupid"
   likert3 "Poodles are self-centered"
   likert4 "Poodles are cute"
   likert5 "Poodles are over-priced"
   income "Annual household income".

VALUE LABELS
   owndog 1 "Yes" 2 "No" 3 "Not sure"/
   favorite 1 "Poodle" 2 "Fox Terrier" 3 "Yorkshire Terrier"
   4 "Dashund" /
   10 "German Shepherd" 11 "Collie" 12 "Saint Bernard"
   13 "Pit Bull" 14 "Malamute" 15 "Afghan" 16 "Cocker Spaniel" /
   17 "Dobemmand" 18 "Golden Retriever" 19 "Rotweiler"/
   food1 to food5 1 "Generic dry dog food" 2 "Generic dry cat food"
   3 "Generic canned dog food" 4 "Kit 'n' Caboodle" 5 "Mighty Dog"/
   sex 1 "Male" 2 "Female"/
   likert1 to likert5 1 "Strongly agree" 7 "Strongly disagree".
MISSING VALUE owndog likert1 to likert5 (9)/
dogcount favorite age (99)/ spend (99.99)/ income 999999.
BEGIN DATA.
001 1 02 02 52314 03.50 66727 29 2 026000
002 2                   22252 31 1 032150
003 2                   45445 26 2 135000
004 1 01 03 32512 05.50 67615 21 2 018500
---> MORE DATA HERE <-----
999 2                   77627 45 2 053000
END DATA.
FREQUENCIES VARIABLES=owndog to income/ STATISTICS=all.
```
Let's take a look at what the lines in this program look like. The program consists of several commands which are intended to instruct the computer about the data and what to do with it.

First, note that some lines are indented while others are not. In general, indented lines are continuations of commands were that started immediately at the left margin on some line above it. SPSS/PC+ really doesn't care if you indent or not, but it will make your program more readable. As you can see, each command eventually ends with a period, which tells the computer to take in the next line as a new command. If you leave out the period, the computer will not understand your commands and will give you an error message. Fortunately, such errors are easy to detect and correct, so if you leave out a few periods, it only means that you will have to do some editing after you first try to run the program. Even experienced SPSS/PC+ users often have problems in their first attempts at any program, but the more experience you get with the program, the easier it gets to correct the problems.

Now, let's start from the beginning.

TITLE "Dog Preference Study".

This first command is actually optional. Eventually, SPSS/PC+ will give pages of output consisting of statistical computations, and putting in the "title" command will put some heading that you choose on top of every page. In terms of syntax, the word "title" is followed by the title you wish to assign in quotes. Notice that the period, contrary to what your junior high school English teacher taught you, goes outside the quotes.

<table>
<thead>
<tr>
<th>On The Use of Capitals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unlike an English teacher, SPSS/PC+ doesn't care if you use upper or lower case letters. Thus, you can mix and match as you find it convenient.</td>
</tr>
</tbody>
</table>
DATA LIST /id 1-3 owndog 5 dogcount 7-8
favorite 10-11 food 1 to food 5 13-17
spend 19-23 (2) likert1 to likert5 25-29 age 31-32 sex 34
income 37-42.

This command is the most crucial part of the program. It may also be the one that can cause you the most frustration.

The "data list" command tells the computer about the positions of the variables on the data lines. If you have made a table detailing this information, you already have all the information needed. Otherwise, you will have to do some arithmetic now to calculate the beginning and ending columns of each variable.

After typing in the command "data list," we will type in a slash to indicate that we are starting the definition of a line of data. The command will not be recognized if you leave this slash out. After that, we will list each variable followed by the column(s) it represents. Thus, we state that ID covers the columns one through three:

DATA LIST /id 1-3...

The variable OWN DOG only takes up one column, so no dash is needed to indicate a range.

For a long questionnaire, we don't quite have to type in the information for each variable individually. Notice that we are able to consolidate the information for the Likert scale section into one declaration:

likert1 to likert5 25-29

Since each variable takes up the same number of columns (in this case one), SPSS/PC+ will recognize that the above declaration is identical to the following more elaborate statement:

likert1 25 likert2 26 likert3 27 likert4 28 likert5 29
This statement only works because each variable takes up the same number of columns. If some of the listed variables took up two columns and some just one, SPSS/PC+ would not have enough information to determine how much each one got.

We have alluded to the fact that SPSS/PC+ knows that

\texttt{item1 to item6}

is the same as

\texttt{item1 item2 item3 item4 item5 item6}.

However, SPSS/PC+ does not allow you to create five variables in the variable list by saying

\texttt{item1a to item1e}.

You would thus have to type in:

\texttt{item1a item1b item1c item1d item1e}.
### Technical Note

Once you have defined a range of variables such as `item1a item1b item1c item1d item1e`

you can refer to the variables as `item1a to item1e`

in subsequent procedures since SPSS/PC+ now knows the sequence in which they occur in the program.

This is no different from the fact that, in ensuing procedures, you will, using the example of our questionnaire, be able to refer to `owndog to likert5`

which would include the variables OWNDOG, DOGCOUNT, FOOD1 to FOOD5, SPEND, and LIKERT1 to LIKERT5.
Handling Long Questionnaires

SPSS/PC+ only allows each line of text to be eighty characters long. That means that when all information from one questionnaire can’t fit into eighty columns, you have to use more than one line for each case.

To tell SPSS/PC+ that you are continuing a case on a new line, put a slash (/) before the first variable to go on the following line. If you have the ID1 number and ITEM1 to ITEM50 on the first line and ID2 and ITEM51 to ITEM100 on the second line, your data list might look like this:

```
data list /id1 1-3 item1 to item50 5-54
    /id2 1-3 item51 to item100 5-54.
```

If you have a very long questionnaire requiring more than three lines of text, see Appendix C for some strategies.
VARIABLE LABELS
owndog "Ownership of dog"
dogcount "Number of dogs owned"
favorite "Breed of favorite dog"
food1 "Rating of generic dry dog food"
food2 "Rating of generic canned dog food"
food3 "Rating of Mighty Dog"
food4 "Rating of Lucky Dog"
food5 "Rating of Kit 'n' Caboodle"
spend "Weekly spending on dog food"

The "variable labels" command is rather long, so only the first few lines appear above. Here, we are assigning a "label," or descriptive phrase, to each variable or question. Note that the variable name is first listed and its "label" is then put in quotes. Each variable label may be up to forty characters long. The main purpose of the variable label is to remind you or the reader of your report what the question is about. The computer really doesn't care what label you give a variable but just feeds it back to you on certain statistical procedures. Be sure to put in the period at the end of the last label! (Note that there is no period at the end of each individual label).

VALUE LABELS
owndog 1 "Yes" 2 "No" 9 "Not sure"
favorite 1 "Poodle" 2 "Fox Terrier" 3 "Yorkshire Terrier"
4 "Daschund"
10 "German Shepherd" 11 "Collie" 12 "Saint Bernard"
13 "Pit Bull" 14 "Malamute" 15 "Afghan" 16 "Cocker Spaniel"
17 "Doberman" 18 "Golden Retriever" 19 "Rotweiler"
food1 to food5 1 "Generic dry dog food" 2 "Generic dry cat food"
3 "Generic canned dog food" 4 "Kit 'n' Caboodle"
5 "Mighty Dog"
sex 1 "Male" 2 "Female"
likert1 to likert5 1 "Strongly agree" 7 "Strongly disagree"

The value labels command assigns names or labels to the values, or "answers," associated with each variable. First, the variable name is stated. Then each value is listed, followed by its description in quotes. Note that there is a slash at the end of each question. Also note that if a set of labels applies to more than one question, they can be listed simultaneously. For example, if "1" means yes and "2" means no for Q1 to Q17 and Q28, you could say

VALUE LABELS q1 to q17 q28 1 'yes' 2 'no'.

Note that some questions don't have value labels. It is not meaningful, for example, to assign labels to the values associated with
AGE and INCOME since these are self-explanatory. This is the case for most interval and ratio scaled variables; the quantity expressed usually carries its own meaning. (In some cases, you will probably want to express the unit of measurement, such as pounds, inches, years, or dollars, in the variable label).

Both the "variable labels" and "value labels" commands are optional and are available solely for your convenience. Putting them in will, however, tend to improve the readability of your output.

MISSING VALUE own dog likert 1 to likert 5 (9)/
dogcount favorite age (99)/spend (99.99)/income 999999.

We discussed the meaning of missing values in the chapter on coding. Note that SPSS/PC+ will automatically interpret blanks as missing values; thus you only put in those missing values you have defined as referring to a non-legitimate missing answer. From the above, you can see that the syntax is the variable name(s) followed by the missing value in parentheses, separating each range by a slash.

BEGIN DATA.
END DATA.

These two commands, both followed by a period, tell the computer that the data will now begin and stop, respectively. In between, you can then type in the data as it is defined in the data list.
Sending Output
To the Printer

By default, SPSS/PC+ displays the statistical output on the screen. If, instead, you would like to send the output to the printer, simply put the following two lines in your program:

```
set more off.
set printer on.
```

Whatever SPSS/PC+ produces after these commands are encountered in the program will be sent to the printer. Thus, if you don't want to send your whole program to the printer, you can put the commands immediately before the statistical procedures.

The "set more off" command frees you from having to press <RETURN> or <SPACE> at the end of each screenful of data. This means that the output may "pass you by" before you have a chance to read it. If this happens, you may either want to leave out this command or wait to read the output until you have it printed on paper.

Some printers, particularly laser printers and other printers that user single sheets instead of continuous "tractor" fed paper, cause sometimes cause particular problems. See Appendix M for more details.
When typing in the data, be sure to check that you are "on target" with respect to the columns. Generally, all the lines should be equally long. Also, be sure to check that, when you have typed in a complete entry, you are one column farther out than the last position listed in the data list.

FREQUENCIES VARIABLES=all/STATISTICS=all.

The "frequencies" command is an example of a statistical procedure or command--SPSS/PC+'s raison d'être. Once we have looked through the data for errors, we will go over other statistical commands, which normally go here in the program, after any recoding and computations, which we will also discuss.
Step 3: Entering The Program And Data

There are several ways you can enter the program and data into SPSS/PC+. Since SPSS/PC+ uses an ASCII, ("plain text") file to handle the input data, you can either use REVIEW, the editor supplied with SPSS/PC+, or a word processor such as WordStar, WordPerfect, or Microsoft Word. When using a word processor, be sure to set the margins so that the lines can be long enough. The default margins for most word processors will normally allow only about sixty-five characters on a line. Be sure to save the file as an ASCII file, i.e. not in word processing format.

LOTUS 123 provides a nice way of entering data for small questionnaires. See Appendix E for details.

To use REVIEW, start SPSS/PC+. How you will do this may depend on the setting of the computer you will be using. In some cases, there will be a "menu" on the computer, and all you have to do is to enter the number or letter that corresponds to SPSS/PC+. On other computers, you may have to start from the DOS prompt. If SPSS/PC+ is found on the "\SPSS" directory of the hard drive, you might type in the underlined part of the following:

```
C: \> CD \SPSS
C: \SPSS >SPSSPC
```

If neither of these approaches work, you will have to find "local" instructions for what to do. Fortunately, most of the rest of this manual's approaches will be more universally applicable.

---

6 Use the text-in/ text-out feature (<CONTROL> <F5>) to create an ASCII file.

7 Also notice that some word processors set margins in terms of length rather than characters. In the newer versions of WordPerfect or Microsoft Word, margins are by default set in terms of inches. In such programs, you may wish to switch to a smaller font instead of adjusting the margins.
Once you are in SPSS/PC+, a logo will first flash and you will then be presented with a menu. Press <ALT>-M, then <F3> <RETURN>. Now specify the name you want to give the file that will contain your program and data and press <RETURN>. For example, if you were entering the questionnaire about dogs, you might call it "A:POODLE.SPS." (It is traditional to use the file name extension ".SPS," but if you like to be different, this convention is not required).

You are now in REVIEW, the editor associated with SPSS/PC+. To move around in the text you create, use the arrows to move up and down and in the right and left directions. When you are ready to save, move to the top of the file (press <CONTROL>-<HOME>), press <F9>, and press <RETURN> twice.

If your file already exists, you will probably be brought into the document at the bottom. To go to the top, press <CONTROL>-<HOME>. (As you might expect, <CONTROL>-<END> will bring you to the bottom of the document). You can use the arrows on the keyboard to move one space or line at a time. The <INS> key will toggle between insert and write-over (that is, whether the computer will type new text on top of existing text or move the old text over to make room for the new). (The default is "on"; however, you may want to turn it off if you are editing data and you want to overwrite some incorrect contents).

One tricky situation involves the insertion of a new line on the very top of the file. (Suppose you want to insert "SET PRINTER ON" above the DATA LIST statement.) To do this, go to the top and inserting a blank space at the beginning of the line. You can now press <RETURN> and the blank line will be inserted at the top of the file.

For some reason, REVIEW will occasionally save only what is

---

³If your floppy disk is in Drive A:, the filename should start with "A:;", e.g. "A:poodle.sps" in our case.
below the cursor. Therefore, you should always be sure to go to the top of the file before saving it. The complete sequence to save, including this first step, is:

<CONTROL>-HOME>
<RETURN> filename <RETURN>

where "filename" represents a new filename and specification you may optionally give the file. If you want to keep the "suggested" filename, just press <RETURN>.

Remember that SPSS/PC+ does not automatically save your data. Therefore, it is not a bad idea to save every fifteen to twenty minutes to guard yourself against a power surge or other interruption which might destroy your data. Also, it is a good practice to have at least two diskettes with the one being used as a backup.

You may not have time to type in all of your program and your entire set of data in one sitting. You can leave REVIEW at any time and resume at a later point.
Exercise

You are now ready to enter data into SPSS/PC+ and analyze it. Your first exercise is simply to type in the following program and run it. From the menu, call up SPSS/PC+, then press <ALT>-M <F3> <RETURN>, followed by the filename "A:exerc1.sps" (make sure you have a formatted floppy disk in Drive A:), and press <RETURN> again.

Now type in the following program, substituting only your name in the "TITLE" line. Make sure that you get the punctuation right--SPSS/PC+ is even pickier than your junior high school English teacher!

set printer on.
set screen off.
title "YOUR NAME".
data list / id 1-3 age 5-6 educ 8-9 sex 11 income 13-15 vacation 17-20.
variable labels
  age 'age of respondent'
  educ 'years of formal education completed'
  sex 'sex of respondent'
  income 'annual household income in hundreds of dollars'
  vacation 'amount spent by household on vacations last year'.
value labels
  sex 1 'male' 2 'female'.
missing value
  age, educ(99)/ sex(9)/ income(999)/ vacation(9999).
begin data.
  001 52 13 1 232 3270
  2 50 13 1 235 4279
  3 57 12 1 274 3643
  4 30 10 1 217 3504
  5 74 17 1 379 4247
  6 63 15 1 353 3715
  7 53 15 2 277 4059
  8 57 16 1 294 4154
  9 35 10 2 266 2908
10 38 11 2 234 3602
11 29 14 2 243 3492
12 24 11 1 185 3391
13 61 15 2 316 3637
14 47 12 2 225 3422
15 49 13 1 259 4289
end data.
frequencies variables=age to vacation/ statistics=all.

You are now ready to run the program. Press <CONTROL> <HOME> to go to the top, <F9> <RETURN> <RETURN> to save, and <F10> <RETURN> to run. If you have made a syntactic mistake, SPSS/PC+ may point it out to you and you will have to fix it and rerun.

Analysis

As is evident, this survey contains demographic information about surveyed individuals as well as information on how much they spent on vacations.

1. Using the FREQUENCIES output you will receive, find the mean, median, mode, and standard deviation for each variable. Do those seem to be representative of the population at large?
Step 4: Checking Your Data And Program For Errors

So you thought that you were finally done with the program and data entry? Well, not quite yet! Your data is probably good enough to be published in a tabloid magazine as it is, but there is one additional step that a conscientious researcher must take.

When you type in a large amount of data, there is a significant chance that you might make a typographical error. Not all errors can be caught and some errors won't make that much difference, but some are relatively easy to catch and should be eliminated.

Now that you have all your data entered, make sure that there is nothing below the "end data" line. (In our example, you would delete the "frequencies" line).

Now press <CONTROL> <HOME> to go to the top of the file, press <F9> followed by <RETURN> <RETURN> to save, and <F10> to run. Unless you have turned the printer option on, you will get one screen at a time. When a complete screen has been displayed, you will hear an obnoxious beep and you will be prompted with the message "MORE." Press <RETURN> to see the next screen.

The program may point out some errors in your program. Such errors are often caused by (1) omission of a period, quote, or slash, (2) the misspelling of a command or variable name, or (3) other "typographical" error. The computer will beep and stop after each screen of information has been displayed. Note down any errors and press <RETURN> to continue.

Once the program has come to an end, press <F3> <RETURN>, followed by your filename and <RETURN>. You are now back in editing mode and you can now fix any problems you have been able to identify. Continue running the program this way until all errors have been fixed.

Note that taking care of one error may fix several other complaints that SPSS/PC+ had in the previous run. If, for example,
you leave out a period or slash, SPSS/ PC+ may encounter several subsequent "errors"--expressions that are not allowed in the given context. In other words, if you left out some punctuation, SPSS/ PC+ may expect something that is not forthcoming and will continue to complain.

Also note the way SPSS/ PC+ chooses to describe your error may not be very informative. The reference to "an unrecognized expression," for example, can mean almost anything. Instead, focus on where error occurs. Should there have been a period immediately before? Did you misspell a command?

When you are satisfied that the errors have been removed from the program, add the line

```
frequencies variables=all.
```

to the bottom of the file, press <CONTROL><HOME> to go to the top, <F9><RETURN><RETURN> to save, and then <F10><RETURN> to run the program. After going through the data definition part of the program, the computer will display the frequency counts of each variable. You should now look for "illegitimate" values for each variable. Let's take a look at the below example:

<table>
<thead>
<tr>
<th>Value</th>
<th>Label</th>
<th>Valid</th>
<th>Cum</th>
<th>Value</th>
<th>Frequency</th>
<th>Percent</th>
<th>Percent</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>1</td>
<td>82</td>
<td>46.9</td>
<td>49.1</td>
<td>49.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>2</td>
<td>81</td>
<td>46.3</td>
<td>48.5</td>
<td>97.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not sure</td>
<td>3</td>
<td>3</td>
<td>1.7</td>
<td>1.8</td>
<td>99.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>1.6</td>
<td>.6</td>
<td>100.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing</td>
<td>9</td>
<td>8</td>
<td>4.6</td>
<td>4.6</td>
<td>MISSING</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TOTAL 175 100.0 100.0

Valid Cases 167  Missing Cases 8

In this case, it is quite evident that an error has been made since there is no such legitimate value as "4" for this question. That is, you either own a dog, don't own a dog, don't know if you own a dog, or
refuse to answer the question. Therefore, the code "4" cannot represent a acceptable answer. We now know that something went wrong, and we will want to track down the error. Also note that we really would have no way of detecting if the value of "3" had been entered one time too many (at the expense of some other code) since that would not show up as an illegitimate value. Be sure to note down all unacceptable values. (In this case, there is only this one "objectionable" value). Note that the period indicates a "system missing value" (or a blank) and that "9" is our defined missing value to be used when the given answer is not usable).

When you have found all the illegitimate values in the program, run the program again, this time putting in the following two lines at the bottom:

```
PROCESS IF (dogown EQ 4).
FREQUENCIES VARIABLES=id.
```

In the above example, you would modify the part in parentheses to meet the condition relevant to your case. On the left side of the "EQ" put the name of the variable that gave you an offending value and on the right side, put the value in question. Now the computer will select only the case that has given you the problem. The next line will give you the case number of the problem variable. When you run the program, you will identify the offending case and you can make appropriate corrections.

Once you have weeded out the incorrect values, you may want to run the frequencies check again to see if you got them all or if new ones have come about as a result of editing.
Step 5: Using Statistical Procedures And Computations

Statistical commands normally go at the bottom of the file, after any computation and recoding commands.

Frequencies

You have already been exposed to the Frequencies command, which provides a frequency count answers to each variable specified. The Frequencies command can give you more information, however. By saying

\texttt{FREQUENCIES VARIABLES=all/STATISTICS=all.}

you will get the mean, standard deviation, median, mode, and various other statistics associated with the distribution.

<table>
<thead>
<tr>
<th>DOGS</th>
<th>Number of dogs owned or leased</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Valid</td>
</tr>
<tr>
<td>Value</td>
<td>Label</td>
</tr>
<tr>
<td>.00</td>
<td></td>
</tr>
<tr>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>2.00</td>
<td></td>
</tr>
<tr>
<td>3.00</td>
<td></td>
</tr>
<tr>
<td>4.00</td>
<td></td>
</tr>
<tr>
<td>6.00</td>
<td></td>
</tr>
<tr>
<td>41.00</td>
<td></td>
</tr>
<tr>
<td>99.00</td>
<td></td>
</tr>
<tr>
<td>TOTAL</td>
<td></td>
</tr>
</tbody>
</table>

| Mean    | 1.190 | Std Err | .656 | Median | .000 |
| Mode    | .000  | Std Dev | 5.208 | Variance | 27.124 |
| Kurtosis| 57.522| SE Kurt | .595 | Skewness | 7.448 |
| S E Skew| .302  | Range   | 41.000 | Minimum  | .000 |
| Maximum | 41.000| Sum     | 75.000 |

Valid Cases 63  Missing Cases 3

For example, this table indicates that 42 of the people own or lease no dogs, 13 people have one, four people have two, and three, four, six, and forty-one dogs are possessed each by one person. We have missing data for three people, and the mean number of dogs owned is 1.19 (although the median is 0.00). It looks as though our
average has been brought up quite a bit by the person who owns forty-one.

**Creating New Variables: Compute**

Computations can be very useful. Suppose you have collected data on how much people spend during Christmas for presents (GIFTS), comestibles (FOOD), travel (TRAVEL), and additional Christmas related expenses (OTHER). If the variable names are the ones given in the parentheses, you can calculate total Christmas expenses (EXPENSES) by

\[
\text{COMPUTE expenses=gifts+food+travel+other.}
\]

If for some reason you wanted to find the average of those figures, you would say

\[
\text{COMPUTE avgexp=(gifts+food+travel+other)/4.}
\]

**Recoding Variables**

Recoding can sometimes be useful when doing crosstabs and other nominal statistics where you would like to "collapse" the data to make it more interpretable. In our questionnaire, we might want to collapse the AGE variable:

\[
\text{RECODE age (0 THRU 25=1) (26 THRU 40=2) (41 THRU 65=3) (66 THRU 98=4).}
\]

As is evident from the example above, you first state the name of the variable. Each value range is then specified in parentheses, followed by an equal sign, and the desired recoded value.

Details on specific statistical procedures are found in the SPSS/PC+ manuals, however, the syntax for a few procedures is contained below.
Reverse Scoring

When using Likert scales and other measures of opinion or attitude, it is sometimes desirable to word questions in the opposite direction of what one is looking for. There are two reasons for this approach. First, wording the question one way may be clearer or more natural than wording it the other way. Secondly, it may be desirable to reverse the polarity of the question to prevent respondents from simply checking the same answer for each question.

For example, when we attempted to measure prejudice against poodles in our questionnaire, we included the item “Poodles are cute.” Since strong agreement with this question (Likert4) could be expected to signify lack of prejudice, we should “reverse score” it before we add it to the others to compute the scale of prejudice against poodles.

In reverse scoring, we will essentially turn the scale upside down. That is, we will convert the highest value to the lowest, the lowest value to the highest, and so forth. In this example, since we have a seven point scale, the command would look like this:

```
RECODE likert4 (7=1) (6=2) (5=3) (3=5) (2=6) (1=7).
```

(Notice that "(4=4)" is superfluous. That is, if the person neither agrees nor disagrees, that fact is not going to change with the polarity of the question).
Crosstabulation

A crosstabulation allows us to explore the relationship between two variables by tabulating one against the other. Consider this example:

### Crosstabulation

<table>
<thead>
<tr>
<th>ANIMAL &gt;</th>
<th>Giraffe</th>
<th>Elephant</th>
<th>Bear</th>
<th>Cow</th>
<th>Pig</th>
<th>Dog</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Short)</td>
<td>6</td>
<td>10</td>
<td>14</td>
<td>5</td>
<td>4</td>
<td>60</td>
<td>33.3</td>
</tr>
<tr>
<td>2 (Medium)</td>
<td>4</td>
<td>16</td>
<td>5</td>
<td>19</td>
<td>6</td>
<td>60</td>
<td>33.3</td>
</tr>
<tr>
<td>3 (Tall)</td>
<td>10</td>
<td>17</td>
<td>5</td>
<td>15</td>
<td>7</td>
<td>60</td>
<td>33.3</td>
</tr>
</tbody>
</table>

Column: 20 26 47 15 55 17 180 Total: 11.1 14.4 26.1 8.3 30.6 9.4 100.0

Chi-Square: D.F. | Significance | Min E.F. | Cells with E.F.<5
-----------------|--------------|----------|-------------------
6.17035          | 10           | .8008    | 5.000             | None With ANIMAL

As you will note, the values of one variable are listed horizontally, and the values of the other are listed vertically. In the cells, you see the number of subjects falling into the "intersection" of the two. The table shows, for example, that fifteen people are both tall and prefer pigs. You will notice that some statistics are also provided. (We will soon get to why anyone would care about the relationship between these two variables).

The general syntax for crosstabs table is:

```
CROSSTABS TABLES=firstvar by secondv/STATISTICS=all.
```

where "firstvar" and "secondv" are the two variables you want to tabulate against each other. Notice that the optional statistics take up
a lot of room. If you only want Chi square ($\chi^2$), you can reduce the output by substituting "STATISTICS=1" for "STATISTICS=all."

Or, you could select the statistics available from this list:

<table>
<thead>
<tr>
<th>Statistics Available in Crosstabs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Chi square</td>
</tr>
<tr>
<td>2 Phi or Cramer's V, depending on the number of variables</td>
</tr>
<tr>
<td>3 Contingency coefficient</td>
</tr>
<tr>
<td>4 Lambda</td>
</tr>
<tr>
<td>5 Uncertainty coefficient</td>
</tr>
<tr>
<td>6 Kendall's Tau-b</td>
</tr>
<tr>
<td>7 Kendall's Tau-c</td>
</tr>
<tr>
<td>9 Somers' d</td>
</tr>
<tr>
<td>10 Eta</td>
</tr>
<tr>
<td>11 Pearson's r</td>
</tr>
</tbody>
</table>

Ordinarily, you probably wouldn't want to look at anything more than statistics numbers one and two. However, calculating the others won't take the computer very long at all, so there is little penalty in saying "STATISTICS=all."

If you want more detail, you can get row and column percentages, i.e. the percentage of the row and column that each cell contributes, by putting in the "/ OPTIONS=3,4" parameter. Thus, if you were tabulating "firstvar" with "secondvar" and wanted these features as well as Chi square and Cramer's V, the command would be:

```
CROSSTABS TABLES=firstvar BY secondvar
/STATISTICS=1,2
/OPTIONS=3,4.
```
Note that the period goes at the very end of all the subcommands. You would not place a period after "firstvar BY secondvar" if you included additional subcommands as we did in this case.

If you want to create multiple tables with one command, you can specify a variable list on each side of the "BY" part of the command. For example, you could say:

```
CROSSTABS TABLES=var1 to var5 by var6 to var10/
STATISTICS=1.
```

Notice, however, that this would create $5 \times 5 = 25$ tables! It is easy to write a statement, without realizing it, that would create hundreds of tables. This would most likely provide you with severe information overload and would make any sort of meaningful interpretation impossible. For example, if you have thirty questions and you want to see how "everything relates to everything," you might think about saying "tables=var1 to var30 by var1 to var30." However, this would create 420 non-redundant tables! You would be sitting by the printer for a long time and would probably not have time to interpret all of them.

To find out in advance how many tables you would get by trying a list of variables against the same list, use the formula

$$T = \frac{n^2 - n}{2}$$

where $T$ is the resultant number of tables and $n$ is the number of items in the list.

To find out how many tables would result from running two different lists against each other, multiply the number of variables in each list by each other. For example,

```
CROSSTABS TABLES=var1 to var10 by var15 to var20
```

would result in $(10 \times 6) = 60$ tables.
Let’s suppose that you have been hired to do a marketing study for a manufacturer of stuffed animals. The manufacturer wants to start a poster media campaign to promote his stuffed animals to the public. A media consultant (of questionable reputation) that he has hired believes that advertising will be most effective if it is placed at eye level. In order to enable the manufacturer to target customers of different height, the manufacturer has asked you to find out whether there is a relationship between a person’s height and his or her preferred stuffed animal species. You collect data and run the following crosstabulation:

<table>
<thead>
<tr>
<th>Count</th>
<th>ANIMAL &gt;</th>
<th>Giraffe</th>
<th>Elephant</th>
<th>Bear</th>
<th>Cow</th>
<th>Pig</th>
<th>Dog</th>
<th>Row</th>
</tr>
</thead>
<tbody>
<tr>
<td>HEIGHT</td>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>Total</td>
</tr>
<tr>
<td>1</td>
<td>Short</td>
<td>6</td>
<td>10</td>
<td>14</td>
<td>5</td>
<td>21</td>
<td>4</td>
<td>60</td>
</tr>
<tr>
<td>2</td>
<td>Medium</td>
<td>4</td>
<td>10</td>
<td>16</td>
<td>5</td>
<td>19</td>
<td>6</td>
<td>60</td>
</tr>
<tr>
<td>3</td>
<td>Tall</td>
<td>10</td>
<td>6</td>
<td>17</td>
<td>5</td>
<td>15</td>
<td>7</td>
<td>60</td>
</tr>
<tr>
<td>Column</td>
<td></td>
<td>20</td>
<td>26</td>
<td>47</td>
<td>15</td>
<td>55</td>
<td>17</td>
<td>180</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>11.1</td>
<td>14.4</td>
<td>26.1</td>
<td>8.3</td>
<td>30.6</td>
<td>9.4</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Chi-Square D.F. Significance Min E.F. Cells with E.F.<5
---------- ---- ------------ -------- ------------------
6.17035 10 .8008 5.000 None With ANIMAL

Statistic Symmetric Dependent Dependent
-------- --------- ------------- ------------

Lambda .04898 .08333 .01600
Uncertainty Coefficient .01260 .01585 .01045
Somers’ D -.02035 -.01877 -.02222
Eta .15250 .05750

Statistic Value Significance
-------- ----- ------------
Cramer’s V .13092
Contingency Coefficient .18065
Kendall’s Tau B -.02042 .3730
Kendall’s Tau C -.02222 .3730
Pearson’s R -.03093 .3401
Gamma -.02806

Number of Missing Observations = 0
Looking at the intersection of the two variables in the table, we can see six people are both short and prefer giraffes, 16 medium sized people prefer bears, etc. That is a large amount of information not easily interpretable without any kind of statistical summary.

Is there a relationship between the two variables? It is difficult to say just from looking at the table. However, the Chi square ($\chi^2$) statistic will test the null hypothesis that the two variables are "independent," i.e. that knowing information about the one does not tell us anything about the likelihood of the other. Normally, we require the significance level to be less than 0.05. Since we did not come anywhere near that this time around, we conclude that there is not enough evidence to support the height hypothesis, and we recommend to the manufacturer that he find another method of segmenting his advertising.

**Product Moment (Pearson) Correlation**

The general syntax for Pearson Correlation in SPSS/PC+ is:

```
CORRELATION varlist WITH varlist
/OPTIONS=options
```

In practical terms, if we wanted to generate the below matrix of the (boring) variables X60 to X62 correlated against X63 to x65, we would write:

```
CORRELATION x60 to x62 WITH x63 to x65/OPTIONS=5.
```
Correlations: X63  X64  X65

<table>
<thead>
<tr>
<th></th>
<th>X60</th>
<th>X61</th>
<th>X62</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.0267</td>
<td>.0350</td>
<td>.1223</td>
</tr>
<tr>
<td></td>
<td>(-.0900)</td>
<td>.1456</td>
<td>-.2251</td>
</tr>
<tr>
<td></td>
<td>.0448</td>
<td>-.0413</td>
<td>.1382</td>
</tr>
<tr>
<td></td>
<td>(72)</td>
<td>(72)</td>
<td>(72)</td>
</tr>
<tr>
<td>P</td>
<td>.412</td>
<td>.385</td>
<td>.153</td>
</tr>
<tr>
<td></td>
<td>(.226)</td>
<td>(.111)</td>
<td>(.029)</td>
</tr>
<tr>
<td></td>
<td>(.354)</td>
<td>(.365)</td>
<td>(.123)</td>
</tr>
</tbody>
</table>

(Coefficient / (Cases) / 1-tailed Significance)

". " is printed if a coefficient cannot be computed

The "/ OPTIONS=5" statement provides for the printing of the number of cases and the one-tailed significance levels associated with the respective correlation coefficient.

If you are not sure whether to expect a positive or negative relationship between the two variables, you may want a two-tailed probability associated with the correlation coefficient. You would then add option 3, making the above statement:

```
CORRELATION x60 to x62 WITH x63 to x65
/OPTIONS=3,5.
```

If you want to create a square matrix of variables correlated against each other, you can simply say

```
CORRELATION varlist /OPTIONS=options.
```

Thus,

```
CORRELATION x60 to x65 /OPTIONS=5.
```

would create a matrix of x60 to x65 against each other. Notice that close to half of the correlations would be redundant (i.e. X62 against X63).

\[\text{Note that, unless you change the options, this significance level is one-tailed.}\]
X63 is the same as X63 against X62) and the coefficients of the correlations on the diagonal would be all ones since we are correlating each variable with itself. (That is, a variable perfectly "predicts" itself).

**Multiple Linear Regression**

There are three main methods of doing linear regression analysis within SPSS/PC+\(^{10}\): **forward inclusion**, **backward elimination**, and **stepwise**. To run an analysis, you must make a decision as to which one to use. Theoretical and philosophical considerations will enter your decision as to which method to select, but if you wanted to keep things simple, you would most likely use forward inclusion.

A regression analysis contains exactly one **dependent variable** and one or more **independent variables**. The syntax for the command is:

```
REGRESSION VARIABLES=varlist/DEPENDENT=depvar
/METHOD=method.
```

Thus, to do a forward inclusion regression analysis trying to "predict" X38 from X3,X4,X5,X6, and X7, you would write:

```
REGRESSION VARIABLES=x38,x3 to x7/DEPENDENT=x38
/METHOD=forward.
```

If you wanted to do a stepwise or backward elimination procedure, you would substitute the words "STEPWISE" and "BACKWARD," respectively, for the term "FORWARD."

Please note that, when doing a regression analysis, you are not guaranteed any result. If no predictor variables are significant at the

\(^{10}\) Some more esoteric varieties (such as hierchical regression) are also available but will not be discussed.
first step, the process will simply tell you that the "PIN limit" of .05 (or whatever level you may have specified if you chose to override the default) has been reached and terminate. This is a frequent outcome, reflecting an empirical reality, and does not indicate that an error has been made.

Also beware that the computer can calculate a meaningless regression equation if the underlying assumptions of the model are not met.

**Discriminant Analysis**

Discriminant analysis is a relatively complex topic. In order to perform this procedure appropriately, you will need a solid background in statistics. It is expected that you will have obtained this background from a statistical text or other appropriate source prior to reading this manual.

The procedure DSCRIMINANT (note the lack of "i" in the spelling) is found in the Advanced Module of SPSS/PC+. This means that you may not have access to this feature in industry even if your company has the Base version of SPSS/PC+.

The syntax for the procedure is:

```
DSCRIMINANT GROUPS=vargroup (val1,val2)/
   VARIABLES=predictor variables
   /STATISTICS=statistics.
```

In the GROUPS specification, you indicate the values for each of the two relevant groups within the appropriate grouping variable. The variable specification parameter allows you to specify

---

11 Technically, it is possible to specify more than two groups. If, for example, you were specified "GROUPS=CLEVEL(2,3,4)," you could involve sophomores, juniors, and seniors in your analysis.
which variables you wish to attempt to use as predictors. Thus, if you were trying to predict purchase of your product (purchase=1) vs. non-purchase of your product (i.e. no purchase or a purchase of your competitor's product) (purchase=2), using income, sex, age, education, and various other demographic variables (DEMO1 to DEMO15), your procedure might look like this:

```
DSCRIMINANT GROUPS=purchase(1,2)/
VARIABLES=income,sex,age,educ,demo1 to demo10/STATISTICS=1,2.
```

Since the specification of "STATISTICS=ALL" will normally result in a very voluminous output, it is recommended that you confine yourself to "STATISTICS=1,2," which will give you the mean and standard deviations for each potential predictor variable broken down by the two groups.

For more information, see the SPSS/PC+ Advanced Statistics Manual.

**Count: "Counting" on how many variables a criterion is met**

Sometimes, you may be interested in "counting" the number of variables on which subjects meet a certain criterion. In less abstract language, you might be interested in knowing how many of a list of products have been bought by a particular consumer. For example, if we wanted to know how many major brands of soft drinks a person had bought and we had each of the major brands coded as an indicator variable (i.e. 1=purchase, 0=non-purchase), we might make a count as follows:

```
COUNT sdrink=drink1 TO drink10(1).
```

In the above example, we are interested in those variables among "drink1" to "drink10" to which people gave the answer 1 (purchase).
Descriptives: A summarized version of Frequencies.

Descriptives (known as Condescriptive in the mainframe version and in certain other packages) provides a summary of the statistics which appear in the frequencies table, such as the mean, median, and mode, without the explicit listing of the frequency of each response. This may be useful under certain circumstances, particularly when you are using interval data with many different values. For example, if you were to measure annual household income with a precision of $100, you might expect to get a large number of discrete answers if from a survey of several hundred people. Thus, you might be better off with a set of summary statistics including the mean, median, mode, and standard deviation, at least for inclusion in your report. (Note: You may still want to do a complete frequencies printout to look for outliers).

Using the income example, the following command might create the below table:

```
DESCRIPTIVES VARIABLES=income
/STATISTICS=1,5,13,10,11.
```

```
Number of Valid Observations (Listwise) = 49.00
Variable      Mean    Std Dev   Minimum   Maximum      N  Label
INCOME      270.43      46.10       209       356     49  annual household inc
```

It goes without saying, but is worth reiterating, that "descriptives" would not generate a meaningful output for variables in an nominal type scale.

For those technically interested, Descriptives can also be used to create z-scores, i.e. standardized scores measuring relative deviation from the mean rather than absolute magnitude. (This is useful when you want to run a regression analysis on data that has an odd distribution).
Npar tests: Non-parametric statistical tests.

SPSS/PC+ allows you to perform various non-parametric tests such as Wilcoxon, Cochran, Friedman, Mann-Whitney, and various others. See the SPSS/PC+ Base Manual for details.

Means: Providing a breakdown of population means by subgroup.

Again, the title sounds a little esoteric, but the basic concept is quite simple. Sometimes, a single population mean may be deceptive because it is heavily influenced by some underlying variable. Suppose, for example, that we know that the average height of a group is 69 inches (5'9"). We might find it useful to break down this figure by men and women, realizing that a lot of the variability is due to between group (sex) differences:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Value</th>
<th>Label</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>For Entire Population</td>
<td></td>
<td></td>
<td>58.9350</td>
<td>5.0615</td>
<td>200</td>
</tr>
<tr>
<td>SEX</td>
<td>1</td>
<td>male</td>
<td>62.3200</td>
<td>3.5587</td>
<td>100</td>
</tr>
<tr>
<td>SEX</td>
<td>2</td>
<td>female</td>
<td>55.5500</td>
<td>3.9603</td>
<td>100</td>
</tr>
<tr>
<td>Total Cases =</td>
<td></td>
<td></td>
<td>200</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

But we can do better than this. Suppose that we also know that half of the people are basketball players (bball=1) and half are not (bball=0). We can now do a breakdown at two levels. Notice that sex is still the outermost criterion for distinction since this is considered the stronger source of influence. (That is, we expect the average male non-basketball player to be taller than the average female player).
Summaries of HEIGHT
By levels of SEX
BBALL basketball player or not

Variable Value Label        Mean   Std Dev  Cases
For Entire Population       58.9350  5.0615    200
SEX  1 male                   62.3200  3.5587    100
BBALL 0 no                   60.0800  2.6942     50
BBALL 1 yes                  64.5600  2.8440     50
SEX  2 female                55.5500  3.9603    100
BBALL 0 no                   53.1600  3.0463     50
BBALL 1 yes                  57.9400  3.2789     50

Total Cases = 200

ONEWAY and ANOVA: Analysis of Variance

SPSS/PC+ allows for one way as well as more complex methods of analysis of variance. For ONEWAY the syntax is:

ONEWAY VARIABLES=dependents BY independent(min,max).

where dependents refers to the list of dependent variables to be tested and independent refers to the criterion variable on which subjects are to be distinguished. "Min" and "max" refer to the minimum and maximum values of the independent variable on which subjects are to be distinguished. Thus, for example, if we have the following occupational groups:

1=construction
2=other blue collar
3=professional
4=other white collar
5=other

and we are interested in testing for income differences between the first four groups, our statement would look as follows:

ONEWAY VARIABLES=income BY occup(1,4).
The ANOVA program is somewhat more complex and should be attempted only by individuals well versed in statistics. Details are in the SPSS/PC+ Base Manual.

Plot: Turning bivariate data into a scatterplot.

When you do a correlation analysis, you may come up with a result like $r=0.33$. From this, you may be able to deduce that the one variable accounts for something like $0.33^{0.5} = 11.1\%$ of the variability of the other. However, perhaps a plot of the two variables against each other might be more illustrative and help in your interpretation. To create a plot, use the PLOT command. The syntax for a plot suitable for the above situation is:

```
PLOT FORMAT=REGRESSION/PLOT=var1 with var2.
```

where "var1" and "var2" are the names of the two variables you want to plot.

For example, this plot displays the relationship between expenditures on chocolate and spending at service stations among a group of hypothetical consumers. The correlation is approximately 0.33.
(The numbers indicate how many people fall approximately at that point in the intersection of the two variables).

Plot allows a number of options which are described in the manual.

**Reliability: Finding coefficient Alpha and other measures of reliability in a scale**

If you ask a question or about a concept in many several different ways or versions, you may be interested in attempting to infer the reliability of the measure from the repeated measurements. Of particular interest may be coefficient alpha. Suppose you have collected a survey containing a scale consisting of variables PREJUD16 to PREJUD25, all measuring prejudice against poodles. To find coefficient alpha, the command would be:

```
RELIABILITY VARIABLES=prejud16 TO prejud25
/SCALE=prejud16 TO prejud25
/MODEL=ALPHA.
```

(It is not a mistake that the variable specification, "prejud16 TO prejud25," appears identically twice in the command).
Other models, such as split half and parallel forms, are described in the Advanced Statistics manual.

**t-test: Testing for differences in two population means.**

There are two kinds of t-test commonly used. The one is used to test for differences between groups, as if we were going to test for differences in height between males (sex=1) and females (sex=2). The syntax for this kind of t-test is

```
T-TEST GROUPS=sex(1,2)/VARIABLES=height.
```

Another example: Is there a difference in the number of soft drinks consumed (sdrinks) by males and females?

```
T-TEST GROUPS=sex(1,2)/VARIABLES=sdrinks.
```

Independent samples of SEX

<table>
<thead>
<tr>
<th>Group</th>
<th>SEX</th>
<th>Number of Cases</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>1</td>
<td>100</td>
<td>28.2300</td>
<td>10.744</td>
<td>1.074</td>
</tr>
<tr>
<td>Group 2</td>
<td>2</td>
<td>100</td>
<td>30.4100</td>
<td>10.181</td>
<td>1.018</td>
</tr>
</tbody>
</table>

Pooled Variance Estimate

<table>
<thead>
<tr>
<th>Degrees of 2-Tail Value</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>F 2-Tail Value Prob.</td>
<td>1.11</td>
</tr>
<tr>
<td>t Degrees of 2-Tail</td>
<td>-.47</td>
</tr>
<tr>
<td>t Degrees of 2-Tail Value Freedom Prob.</td>
<td>.142</td>
</tr>
<tr>
<td>t Degrees of 2-Tail Value Freedom Prob.</td>
<td>.142</td>
</tr>
</tbody>
</table>

The other kind is used to test for significant differences in repeated measures or same-subject variables. Suppose we are interested in finding out if people changed their consumption patterns of soft drinks after an advertising campaign. If we collected data on consumption both prior to (predrink) and after (postdink) the
campaign, the syntax might look as follows:

\[
\text{t-test pairs=predrnk with postdrnk.}
\]

Paired samples t-test: PREDRNK beverages consumed prior to campaign
POSTDRNK beverages consumed after campaign

<table>
<thead>
<tr>
<th>Variable</th>
<th>Number</th>
<th>Standard Mean</th>
<th>Standard Deviation</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>PREDRNK</td>
<td>200</td>
<td>29.3200</td>
<td>10.497</td>
<td>.742</td>
</tr>
<tr>
<td>POSTDRNK</td>
<td>200</td>
<td>30.7800</td>
<td>9.887</td>
<td>.699</td>
</tr>
</tbody>
</table>

-1.4600 | 13.797 | .976 | .085 | .233 | -1.50 | 199 | .136 |

Since the significance level associated with this test depends on whether the test is one or two tailed, it is particularly critical that you make hypotheses prior to running this test. For example, it might be reasonable to hypothesize that soft drink consumption would increase after the campaign, thus avoiding the need to allow for the left tail (that is, the probability that sales could also decline as a result of the campaign). In any event, we must be careful to notice that even if there is a significant change, we cannot positively say that it was caused by the advertising campaign. We hope, however, that this is the best explanation available.

**Advanced features: Factor, Cluster, Hiloglinear, and MANOVA.**

The SPSS/PC+ Advanced Statistics Manual contains information about these more complex procedures. In general, these should be attempted only by individuals very knowledgeable about the particular procedures since quite meaningless results can be generated and misinterpreted by an inexperienced user.
Using Only Selected Cases

For some analyses, you may want to include only selected subjects in your calculations. For example, if you want to figure out the average age of only those respondents to your survey who are dog owners, you might enter the following command:

```
PROCESS IF (dogown EQ 1).
DESCRIPTIVES VARIABLES=age.
```

In the parentheses, we indicate a logical condition. In this case, the logical condition is that the value on the dog ownership variable is 1.

Other logical "operators" are:

- **GT**: Greater than
- **GE**: Greater than or equal
- **LT**: Less than
- **LE**: Less than or equal
- **NE**: Not equal

Thus, if we wanted a Frequencies table of the number of dogs owned by people whose age is less than 35, we would say

```
PROCESS IF (age LT 35).
FREQUENCIES VARIABLES=dogown.
```
Appendix A
Common Questions About SPSS/PC+

Q.: I use SPSS/PC+ on a computer that is connected to a dot matrix printer. Is there any way I can print the output from another computer that has a better printer attached?

A.: Yes. SPSS/PC+ output is sent to the file C:\ SPSS\SPSS.LIS. You can import this file into a word-processor and then move the file over to the other computer (See Appendix I for information on how to import SPSS/PC+ output into WordPerfect). Caution: Beware that SPSS/PC+ output is often voluminous--printing all of it could be time consuming.

Q.: Sometimes, the arrow keys don't work while I am in SPSS/PC+. Did I do anything wrong?

A.: No. There is a bug that will often "invalidate" the arrow keys after an SPSS/PC+ program has been run. The best way to get around this problem is to save your data and reboot the system.

Q.: I have some LOTUS data I would like to analyze with SPSS/PC+. Do I have to retype it all?

A.: No. Either print it to an ASCII file (using PrintFile) or use the IMPORT command in SPSS/PC+.

Q.: Can I do nice graphics in SPSS/PC+?

A.: Although a graphics module is available, it has not been purchased by all institutions. Therefore, you are better off
creating graphs with a program such as LOTUS or Harvard Graphics. (For example, by typing in the information from a frequencies run, you could create a nice pie-chart).

Q.: I am really sold on SPSS/PC+ and I would love to have it at home. How do I go about getting it?

A.: Since SPSS/PC+ is a commercial product, it is copyrighted, and you will have to buy your own copy. The Base version lists for about $800 and contains the basic modules for such procedures as frequencies, crosstabs, t-test, ANOVA, regression, descriptives, and means. The Advanced supplement costs about the same and contains the modules for reliability, factor analysis, and other advanced functions. You will need eight to ten megabytes on your hard drive. You can, however, buy a more limited student version for about $40.

Q.: Sometimes when I press <ALT>-M <F3>, the computer beeps and says that the selection is not available in this mode. What went wrong?

A.: The <ALT>-M is a so called "toggle" switch which changes between the menu and "type-in" modes. Apparently, you were already in type-in mode before you started the command. Simply repeat the procedure (<ALT>-M <F3>), and you should be able to continue.

Q.: I have noticed that some people use the menu system in SPSS/PC+ to write programs. Why isn't it covered in this text?

A.: It's true that the menus free you from typing in certain commands in their entirety, but in the author's view, they create more problems than it solves.
Q.: What about the Data Entry module? Doesn't that save time?

A.: Some people find it easier to use this module, but since it has to be purchased as a separate product, it may not be available to you in industry, and it is therefore not a good idea to become dependent on it. Also, if you are entering a long questionnaire, it is inefficient and frustrating to have to press <RETURN> after entering each variable.

Q.: I just have a short survey and all I want is some averages and perhaps some standard deviations. Do I really have to use SPSS/PC+ for this?

A.: No. Lotus 123 will allow you to compute the mean, standard deviation, maximum, and minimum for a data range. In addition, you can do a regression analysis (DataRegression).
Appendix B

**Working With SPSS/PC+ Output**

Everything should be made as simple as possible, but no simpler than that. --Albert Einstein

SPSS/PC+ allows you to do very complex analyses, including some whose interpretation would be challenging even for individuals holding advanced degrees in statistics and related fields. Therefore, we cannot cover the majority of analyses available, but in this section, we will attempt to look at the output from some of the procedures typically used in a beginning research methods class. Specifically, we will look at how to prepare output from the frequencies, crosstabs, and Pearson correlation procedures for analysis and interpretation.

**Frequencies**

Recall that the fundamental purpose of statistics is to make inferences about some underlying population. Unless the sample size employed is very large or unless the variability of the data is very low, statistics based on a sample will rarely ever be exactly equal to the population parameters. In some cases, however, statistics can be good estimates of the underlying population parameters, and we can use statistical principles to evaluate the likely accuracy of our estimate.

One approach to the assessment of the accuracy of a statistical estimate is the use of a confidence interval. A confidence interval refers to the range within which a given proportion of statistics will occur in repeated sampling of the same size from the respective population.

---

12 Figures based upon information about the entire population is known as a parameter. By definition, there are no confidence intervals associated with parameters.
population. In less abstract terminology, if obtain, from sample data, a 95% confidence interval of $34,598 to $38,710 for the mean household income in a neighborhood in which we are considering opening a convenience store, we can be ninety-five percent certain that the true population mean is within this range. There is, however, a 2.5% (5%/2) chance that the income is higher and a corresponding chance that they are lower.
Depending on the situation involved, you might be interested in one of two different kinds of confidence intervals, i.e. those for either

Warning!

Missing data can severely distort your statistics since SPSS/PC+ can only base its calculations on "valid" cases—that is, the cases that have a meaningful value for the variable in question. This is problem is particularly evident when we calculate the mean of a variable.

When you encounter a variable that has many cases missing, examine whether the missing cases may be systematically different from the valid cases. For example, perhaps those people who did not respond to a question of how many library books they checked out last month don't care that much about libraries and checked out fewer books than those who were conscientious enough to answer the question.

Also, what is your population of interest? If you are interested to find out how much the average person spends on dog food, then those people who do not own dogs should not be counted as missing—instead, we would count those that appear to have relatively normal eating habits as spending $0.00. If we counted only those people who have dogs, we would greatly overestimate the market.
Confidence interval for a mean. We illustrated the idea that a sample mean can be used to generate a confidence interval for a population mean above. Now, let's look at some SPSS/PC+ output to see how generate the data:

In the above survey, the respondents claimed to have a mean level of education of 13.97 years (i.e. being 0.03 years short of completing two years of college). The sample standard deviation is 2.71 years. Using the Lotus 123 spreadsheet MEANTINT.WK1 on the distribution disk, get the result.

13 For this example, we will use information from a Frequencies printout, but you could also get the needed information from the procedure Means.

14 For formulas, see an introductory statistics or research methods text.
From the above, we can conclude with 95% certainty that the "true" population mean is somewhere between 13.64 and 14.30 years. If we wanted to be more certain than that, the confidence interval would be wider.

Confidence intervals for a proportion. As you know, an important step in business strategy formation is competitor analysis. Tallon Termite & Pest Control, a California company that uses liquid nitrogen to kill termites with "coldness instead of chemicals," is concerned about aardvarks, the only other natural enemy that termites have. The aardvark is commonly known as the ant-eater, but that's a misnomer since, according to unverified company intelligence, two out of three aardvarks prefer termites over ants. Although aardvarks reportedly chew with their mouths open, which is not a pretty sight, they are could nevertheless represent a serious competitive threat.

Let's suppose that you have been hired by Tallon Termite & Pest Control, Inc., to assess the extent to which aardvarks constitute serious competition nitrogen program. You are directed to find out snack preferences of the rivals by sampling one hundred aardvarks and asking if they prefer ants or termites. The results are as follow:

<table>
<thead>
<tr>
<th>PREF</th>
<th>Snack preference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>Label</td>
</tr>
<tr>
<td>Value</td>
<td>Frequency</td>
</tr>
<tr>
<td>Ants</td>
<td>1</td>
</tr>
<tr>
<td>Termites</td>
<td>2</td>
</tr>
<tr>
<td>TOTAL</td>
<td>100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Sample</th>
<th>Lower</th>
<th>Upper</th>
<th>size(n) limit</th>
<th>limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>Std</td>
<td>size(n)</td>
<td>limit</td>
<td>limit</td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td>13.97</td>
<td>2.71</td>
<td>263</td>
<td>13.64</td>
<td>14.30</td>
</tr>
</tbody>
</table>

From the above, we can conclude with 95% certainty that the "true" population mean is somewhere between 13.64 and 14.30 years. If we wanted to be more certain than that, the confidence interval would be wider.

Confidence intervals for a proportion. As you know, an important step in business strategy formation is competitor analysis. Tallon Termite & Pest Control, a California company that uses liquid nitrogen to kill termites with "coldness instead of chemicals," is concerned about aardvarks, the only other natural enemy that termites have. The aardvark is commonly known as the ant-eater, but that's a misnomer since, according to unverified company intelligence, two out of three aardvarks prefer termites over ants. Although aardvarks reportedly chew with their mouths open, which is not a pretty sight, they are could nevertheless represent a serious competitive threat.

Let's suppose that you have been hired by Tallon Termite & Pest Control, Inc., to assess the extent to which aardvarks constitute serious competition nitrogen program. You are directed to find out snack preferences of the rivals by sampling one hundred aardvarks and asking if they prefer ants or termites. The results are as follow:

<table>
<thead>
<tr>
<th>PREF</th>
<th>Snack preference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>Label</td>
</tr>
<tr>
<td>Value</td>
<td>Frequency</td>
</tr>
<tr>
<td>Ants</td>
<td>1</td>
</tr>
<tr>
<td>Termites</td>
<td>2</td>
</tr>
<tr>
<td>TOTAL</td>
<td>100</td>
</tr>
</tbody>
</table>
Using the Lotus template PROPINT.WK1 on the distribution disk, we get the following results, where \( p \) is the proportion of aardvarks preferring termites and \( q = 1-p \), we obtain

<table>
<thead>
<tr>
<th>Variable</th>
<th>( p )</th>
<th>( q )</th>
<th>Sample size (n)</th>
<th>Lower limit</th>
<th>Upper limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Snack preference</td>
<td>0.74</td>
<td>0.26</td>
<td>100</td>
<td>0.65</td>
<td>0.83</td>
</tr>
</tbody>
</table>

Since we can conclude aardvarks are potentially a serious threat (we are 95% sure that between 65% and 83% of aardvarks prefer termites over ants), your next task is to find out what percentage of home owners in the company's trade area considers aardvarks' table manners offensive. Because the methodology is the same, the task is left to you. (If funding is available, you might explore whether attitudes are influenced by whether a home owner has children of an age where they are particularly susceptible to influence).

**Crosstabs**

The crosstabulation feature is one of the more popular programs in SPSS/PC+ since it allows the user to assess a relationship between two variables when one or both of them are nominally scaled. Note that in this example, the assignment of numbers to various categories of animals is nominal or arbitrary while the scaling for height is ordinal:

---

\(^{15}\) \( q \) is **not** always the equivalent of the proportion preferring ants. In some cases, there may be several ways in which something can be "not \( p \)" without qualifying as \( q \). For example, suppose that the aardvarks were also given the choice "other." In that case, \( q \) would be equal to the proportion preferring ants plus the proportion preferring other snacks.
Crosstabulation: \[ \text{HEIGHT} \quad \text{Height of customer} \]
By \[ \text{ANIMAL} \quad \text{Species of preferred stuffed animal} \]

<table>
<thead>
<tr>
<th>ANIMAL</th>
<th>Giraffe</th>
<th>Elephant</th>
<th>Bear</th>
<th>Cow</th>
<th>Pig</th>
<th>Dog</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>HEIGHT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Short</td>
<td>1</td>
<td>6</td>
<td>10</td>
<td>14</td>
<td>5</td>
<td>21</td>
<td>40</td>
</tr>
<tr>
<td>Medium</td>
<td>2</td>
<td>4</td>
<td>16</td>
<td>19</td>
<td>6</td>
<td>60</td>
<td>33.3</td>
</tr>
<tr>
<td>Tall</td>
<td>3</td>
<td>10</td>
<td>17</td>
<td>15</td>
<td>7</td>
<td>60</td>
<td>33.3</td>
</tr>
<tr>
<td>Column</td>
<td>20</td>
<td>26</td>
<td>47</td>
<td>15</td>
<td>55</td>
<td>17</td>
<td>180</td>
</tr>
<tr>
<td>Total</td>
<td>11.1</td>
<td>14.4</td>
<td>26.1</td>
<td>8.3</td>
<td>30.6</td>
<td>9.4</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Chi-Square | D.F. | Significance | Min E.F. | Cells with E.F.< 5 |
-----------|------|--------------|----------|---------------------|
6.17035   | 10   | .8008        | 5.000    | None With ANIMAL    |

Statistic | Symmetric | Dependent |
-----------|-----------|-----------|
Lambda     | .04898    | .08333    | .01600   |
Uncertainty Coefficient | .01260 | .01585 | .01045 |
Somers' D  | -.02035   | -.01877   | -.02222 |
Eta        | .15250    | .05750    |

Statistic | Value | Significance |
-----------|-------|--------------|
Cramer's V | .13092|
Contingency Coefficient | .16205|
Kendall's Tau B  | -.02042| .3730 |
Kendall's Tau C  | -.02222| .3730 |
Pearson's R    | -.03093| .3401 |
Gamma       | -.02806|

Number of Missing Observations = 0

Crosstabs provides a number of statistical tests. One of the more well known of these is the Chi square (\( \chi^2 \)) test of independence, which assesses the probability that one could obtain a relationship in the sample as strong as the one obtained if the sample were taken from a population in which no relationship actually existed. The formula compares the ratio of the actual count in each cell to that expected under no relationship, given the distributions of each variable:
where \( k \) is the number of "cells" (i.e. different combinations of the two variables), and \( O_i \) and \( E_i \) are the observed and "expected" number of cases in the respective cell. SPSS/PC+ calculates the expected number of cases in each cell by multiplying the proportion of cases meeting each condition and then multiplying by the total number of cases. For example, if we have one hundred and eighty cases (\( n=180 \)) and sixty people are tall (\( n_T=60 \)) and forty-seven people prefer bears (\( n_B \)), then, if there is no relationship between the two variables, we would expect a total of

\[
\frac{60}{180} \times \frac{47}{180} \times 180 = 15.67
\]

people to fall into the category of being both tall and preferring bears. This figure compares to the actual, or observed, count of seventeen people falling into that cell.

To each Chi square value corresponds a figure for degrees of freedom and a significance level. In our case, the significance level is 0.8008, giving us insufficient evidence to reject the null hypothesis that the two variables are independent. (Even if the Chi square value had been significant, we would not be able to conclude anything more than that some kind of relationship probably exists between the variables).

On the negative side, Chi square has some serious drawbacks:

1. It is not a powerful test, and we can only count on confirming a
relationship if it is strong. If you have interval data or better for both variables, you would be better off using correlation or some other more powerful technique. Also, you are unlikely to pick up a significant relationship with a small sample size.

2. It is not likely to yield significant results when there are many "levels," or possible values, associated with each variable. Variables having many levels should probably collapsed (e.g. on an age scale, you might collapse age 35-50 to one value).

3. When a large number of cells are empty, the test may not be valid.

By specifying "/ STATISTICS=ALL" when running Crosstabs, you get several other nominal level statistics. More details are in the SPSS/PC+ Base Manual.

Pearson Correlation

The Pearson correlation coefficient is used to assess the extent to which two variables can be used to "predict" each other. It is theoretically assumed that the two variables are both at least at the interval level; however, in practice, the Pearson correlation coefficient is often used when one or both of the variables in question are only ordinal. When assumptions are violated in this manner, any true relationship between the two variables may be underestimated or entirely overlooked. Using a variable that does not achieve interval or ratio scaling could also lead to an overestimate of the true relationship, but that is less likely.

In practice, many researchers include Likert scale items, which are ordinally scaled variables, in correlation analysis.
Also note that correlation is intended to explore any linear relationship between two variables. That is, it is assumed that the change in one variable as a result of the change in the other will be constant regardless of the magnitude of either variables. For example, if we correlated grocery spending against income, it would be assumed that a $1,000 increase in income would result in the same increase in grocery spending whether the original income were $18,000 or $249,500. If this assumption is not viable, you may want to transform one or both of the variables. This topic is beyond the scope of this book, but is generally covered in methodology texts.

Also note that correlation is intended to explore any linear relationship between two variables. That is, it is assumed that the change in one variable as a result of the change in the other will be constant regardless of the magnitude of either variables. For example, if we correlated grocery spending against income, it would be assumed that a $1,000 increase in income would result in the same increase in grocery spending whether the original income were $18,000 or $249,500. If this assumption is not viable, you may want to transform one or both of the variables. This topic is beyond the scope of this book, but is generally covered in methodology texts.

Assuming you specified "Options=5" when you ran the analysis, your output will consist of three parts. Let's take a look at a fictitious example between income and vacation spending:

Correlations: VACATION

INCOME .5636
( 263)
P = .000
The first figure is the sample correlation coefficient. This ranges between -1.0, when there is a perfect negative correlation between the two variables, and 1.0, when there is a perfect positive correlation between the variables.

On the second line, the number of cases on which the correlation is based appears in parentheses. Sometimes, this may be a lower figure than your total number of subjects because there may be missing data among some of your subjects\(^ {16}\). Some people, for example, might have felt that one or more of these questions was too personal to answer.

Finally, on the last line is the significance level associated with the correlation coefficient. As we noted above, the figure of .5636 is the sample correlation coefficient and may not be the actual population coefficient.\(^ {17}\) The significance level is the probability that the correlation coefficient is zero, not the probability that it differs from the figure shown (which it probably will). It cannot be concluded from the above information (without making further calculations) that the population correlation coefficient is greater than, say, 0.35. However, the sample coefficient is our best available estimate.

Now, what is the meaning of a correlation coefficient? In general, the rule says that the square of the value of the coefficient represents the proportion of the variability of the one variable "explained" by the other. For example, in the above example,

\[ .5636^2 = 0.3176. \]

suggesting that income "explains" about 31.76% of the variation in vacation spending between people.

\(^ {16}\) Note that a case will be missing in correlation if the only one of the two variables is missing.

\(^ {17}\) Because we are taking a finite sample from a population, we expect our sample results differ somewhat from those of the actual population.
Correlation coefficients are not very intuitive. Sometimes, a visual representation is more informative. Using the Plot command, we get the following illustration:

```
<table>
<thead>
<tr>
<th></th>
<th>100</th>
<th>200</th>
<th>300</th>
<th>400</th>
</tr>
</thead>
<tbody>
<tr>
<td>amount spent by household on vacations</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2800</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3600</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4400</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4800</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

(The numbers in the box represent the count of cases falling at that coordinate. In a perfect correlation, all the numbers would have been on a diagonal; with a complete lack of correlation, the numbers would have been scattered around the box without pattern).
Appendix C

Working With Large Data Sets

In Chapter 2, we discussed how a case can take up more than one line. We noted that in the "data list" command, we mark the beginning of a new line by a slash. Thus, if we have ID1 and questions 1 through 50 on line 1 and ID2 and questions 51 through 100 on line 2, our data list might look like this:

```
data list /id1 1-3 q1 to q50 5-54
/id2 1-3 q51 to q100 5-54.
```

Two cases might look like this:

```
001 8005395460366893284890676150103753760134568443947962
001 230737418052740284674032009273491437943283227406462
002 375233138065194175558179849761573065614784546728337
002 190235863323536560428504071766165022113656862678986
```

Now suppose that instead of just two lines per case, we have a data set that requires five lines per subject. Two cases, using that system, might look like this:

```
data list /id1 1-3 q1 to q50 5-54
/id2 1-3 q51 to q100 5-54.
```

Two cases might look like this:

```
001 518810421387925828815224790024613226620946180660912
001 1736096447695181568034889642259507514245342577273341
001 9072540086525965683452094258943757649786874275834
001 858725673569373816576796816641488407924430157401661
001 93770572368521277158218564000020348140603127470916776
002 2619282693275359149696660713293518078070153532387853052
002 467192347053286242781463574656343764730236589590917
002 565662632732205780016862471949627692765870597258624634
002 92326111451257030399355546291915734015614726564885112
002 8518487544497481966211345279596745459258907987424135
```

That format is not very readable. If you accidentally leave out one line from a case or put it in twice, the computer would warn you that you that the number of lines is not evenly divisible by the number of lines per case, but it might be difficult, or at least laborious, to track down where the problem occurred. Therefore, we might want to take two preventive steps:

1. First, assign a variable to each line to identify its number. Thus,
you might have a data list beginning like this:

data list /id1 1-3 line1 4 q1 to q50 6-55
   /id2 1-3 line2 4 q51 to q100 6-55

and going on like that until five lines have been created. For every case, then "LINE1" would be equal to "1," "LINE2," to 2, etc.

2. At the end of the five lines, create a data line that only contains the one "variable" called "BOGUS." This will create a blank line that makes it more obvious where each case ends. Note that you must assign the columns that "BOGUS" could take up even though the variable will be "system missing" for all cases. We will just say that it occupies column 1 of line 6. The complete data list from the above might look like this:

   data list /id1 1-3 line1 4 q1 to q50 6-55
      /id2 1-3 line2 4 q51 to q100 6-55
      /id3 1-3 line3 4 q101 to q150 6-55
      /id4 1-3 line4 4 q151 to q200 6-55
      /id5 1-3 line5 4 q201 to q250 6-55
      /bogus 1.

Two cases from this data set might look like this:

   0011 51881042413879258528815244790246113226620946180660912
   0012 173609604476951815680348896422595507514245342577273341
   0013 9027254080865592565683450084248943756499768779375834
   0014 8587725673569373631657679816664188404897224430157401661
   0015 9377057236851277718218564000020348310063127470916776
   0021 28192629327531949569666071329351801787015328783052
   0022 4671923347050326842478146357465634376473026358959917
   0023 565682632732205780901682471949627692765870597258624634
   0024 9232621145125703059935554629191573401564726564885112
   0025 851848754449748196622113452795996745459258090787424135

If your data set becomes so big that it takes a long time to run, you might consider making a "system file," which will increase the execution speed considerably. See Appendix D for more information.
Appendix D

Using System Files

So far, we have had to run our program each time we wanted to execute a series of one or more statistical procedures. This is generally the easiest way to work when your data set is small. However, as you noticed, running the program and having the computer "read" the data each time takes a great deal of time. The bigger the data set, the more time it takes to run the program.

Creating a "system file," which stores the data in a way that enables the computer to access it more quickly, may help you make the reading of the data go faster. There are a few costs associated with this, however:

- It takes time to create the system file
- The system file takes up a great deal of space on the hard disk. If the computer is found in a college computer lab, there may not be space enough on the hard disk for everyone to store his or her data as a system file, and you may not be able to fit all your data on a floppy disk.
- Although variable and value labels will show up in your output as usual, you no longer have much of the information you used to create your program in your file. If you forget a variable name, you may either have to refer to a printout or "get" the file to find this information.

To create a system file, first type in the SPSS/PC+ program and the data as you normally would. When finished, you may want to run a Frequencies procedure to find out if there are any errors in the data. You can also do any data modifications (such as Recode and Compute statements) now. When you are satisfied with the shape of your data, add a "save" statement to the very end of your program:

```
SAVE /OUTFILE="filename.SYS".
```
where "filename" refers to the name you want to give the system file you are creating. If your data file is named "SURVEY.SPS," for example, you might want to name the system file "SURVEY.SYS." (On a floppy disk in Drive A:, this would be "A:SURVEY.SYS." Note, however, that there might not be enough space on a floppy disk to accommodate a system file containing a large data set.)

Now you are ready to access the system file for further analysis. To do that, create a new file (say, "FASTSURV.SPS"). Put in a "get" statement as the very first line in this file:

```
GET /FILE="filename.SYS".
```

where "filename" refers to the name of your system file.

After the "get" statement, you can add any statistical procedures and run the program as normal. Thus, in our example, a run involving a system file might involve the following:

```
GET /FILE="fastsurv.sps".
CORRELATION age WITH income.
CROSSTABS TABLES=dogpref BY sex.
```
Advanced Tip

Your data file might contain some redundant and unnecessary information that you don't want to include in your system file. For example, if you have five lines of data, you might have five lines of data (e.g. ID1, ID2, ID3, ID4, and ID5) and five line identification numbers (e.g. LINE1, LINE2, LINE3, LINE4, and LINE5). To exclude these variables from your system file, use the "drop" subcommand of the Save command. If we want to get rid of the above variables and our file and we want to name our system file "SURVEY.SYS," the command would look like this:

```
SAVE /OUTFILE="survey.sys" /DROP=id1,
    id2,id3,id4,id5,line1,line2,
    line3,line4,line5.
```

**WARNING:** In this case, "ID1 to ID5" would not be equivalent to "ID1,ID2,ID3,ID4,ID5" since there are a lot of variables that come in between the identification numbers. Thus, if you said "/DROP=ID1 to ID5," you would be excluding all but the last line of text!
Appendix E

**Importing Data From Lotus 123**

Sometimes, you may want to analyze some data you already have typed into LOTUS 123 in SPSS/PC+. There are at least two situations in which this could happen. First, suppose started an assignment in LOTUS 123. Later, you realize that you wish to do some statistical analyses not available in LOTUS. The second possibility is that you may want to do some of your data input on a computer that has LOUTS, but not SPSS/PC+, installed.

The following will illustrate how you can import sets of data that consist of at most one line per case into SPSS/PC+. (If you have more than one line of data, you will most likely have to use an SPSS/PC+ system file. See Appendix D and the section on Import in the SPSS/PC+ Base Manual.) Since you already have the data entered into LOTUS, it is assumed that you are reasonably familiar with the program.
First, get into LOTUS 123 and bring up the file that contains the data you wish to use in SPSS/PC+. Make sure that your data is aligned properly in the columns by issuing the "/Range Format Fixed" sequence of commands.

Now make sure that the data does not take up more than eighty columns or spaces. If it does, you might try to reduce the length by first reducing excess columns assigned to any one variable. For example, while LOTUS assigns nine digits to each column by default, you don't need that many for a variable like age. (For reasons beyond the scope of this book, you need one more column than the number of digits of the greatest number.) To reduce the number of columns reserved for age to three, get into the column containing the age and issue the directive "Worksheet Column Set width" and then specify 3. If your data still does not fit, you might think about only taking some of the variables into SPSS/PC+. You can do that by moving only those variables you want into another area of the spreadsheet.

Warning!

Lotus is not very strict about whether you put your cases across rows or columns. That is, when you put the first variable from a case in a cell, you have a choice between putting the next variables to the right of this variable or below. For example, if you put VAR1 in A1, you can either put variables 2 through 5 in A2..A5 or B1..E1.

SPSS/PC+ is a little more choosy. In SPSS/PC+, variables associated with the same case must occur in rows. This means that if you put the first variable in A1, the next one should go in A2. Then variable 1 of the next case might then go in B1.
To take the data to SPSS/PC+, we will now "print" the variables to an ASCII (i.e. plain text) file. You may be familiar with the "/Print Printer" command. Now instead say "/ Print File." The computer will prompt you for the name of the output file. Let's call it "A:LOTUSDAT.SPS." (Make sure you have a formatted disk in drive A:.) Press "Range" and specify the range you want to take into SPSS/PC+. Now specify "Options" in the print menu and choose "Margins Left" and specify "0." Finally, press "Align Go." When the computer has finished "printing" to the disk, that is, when the menu comes back, choose "Quit," then "/ Quit" to exit LOTUS 123.

Now get into SPSS/PC+ and get into the file "A:LOTUSDAT.SPS" (i.e. by <ALT>-M <F3>).
Warning!

Even when you "print" to a file, LOTUS still thinks in terms of a "sheet" of paper. Since a "sheet" in LOTUS cannot be set to be longer than 100 lines (with the default being 66 lines), you will have some empty lines when each "page" ends in the file. Once you get into SPSS/PC+, be sure to delete these empty lines.

Once you are in SPSS/PC+, it is time to put in the SPSS/PC+ commands. You can get the information for the data list command by cursoring over the beginning and end of each variable and looking at the position indicator in the bottom right part of the screen. Then insert the needed statements such as Data list, Begin data, End data, and any Variable labels, Value labels, and Missing value definitions that you would like.

The program will now run just like any other program.
Appendix F

**Similarities Between SPSS/PC+ and Lotus 123**

There are some similarities between some SPSS/PC+ functions and those in Lotus 123 that you may already know. Let’s suppose you want to add up a person’s total utility expenditures from his or her phone, water, natural gas, and power bills. In Lotus we might have the data entered as follows:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>33.99</td>
<td>11.44</td>
<td>23.44</td>
<td>29.95</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>144.54</td>
<td>12.34</td>
<td>53.60</td>
<td>44.45</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>67.41</td>
<td>9.50</td>
<td>22.34</td>
<td>34.35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>34.45</td>
<td>10.51</td>
<td>34.45</td>
<td>29.45</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

To add up the first person’s total utility expenditures, you might type in the formula

@SUM(A3..D3)

in cell E3. Alternatively, you could use the expression

\[+a3+b3+c3+d3\]

After you completed the calculation for the first case, you would have to copy the formula from E3 to E4 through E6:

/E3 <CR> E4..E6 <CR>

Now look at how similar the procedure is in SPSS/PC+! Assuming that we already have created the variables PHONE, WATER, GAS, and POWER, we can use the latter approach from Lotus
and say:

```
COMPUTE total=phone+water+gas+power.
```

(Unfortunately, there is no function comparable to @SUM in SPSS/PC+). For practical purposes, SPSS/PC+ will automatically do the "cell copying" for us! (Remember, for statistical purposes we generally work with the entire data set without specific interest in individual cases).

SPSS/PC+ provides other functions as well. These include both the traditional ones such as the absolute value, e.g.

```
COMPUTE differnc=abs(score1-score2).
```

and more esoteric ones such as logarithmic and trigonometric functions\(^\text{18}\), which are also available in Lotus (See p. C-24 of the SPSS/PC+ v. 2.0 Base Manual for more information).

You can also use arithmetic expressions in calculations. While in Lotus you might multiply by putting the expression

```
+A12*A13
```

in cell A14, you might say

```
COMPUTE interest=prncp*rate.
```

in SPSS/PC+.

\(^{18}\) For those interested in econometric and forecasting methodology, it is also possible to use lagged data. Details are found in the manual.
Important Note!

In Lotus, you are used to the automatic updating which takes place when you change the values in a cell on which a formula is based. In SPSS/PC+, such updating will generally not take place. Consider this example of a command sequence in SPSS/PC+:

\begin{verbatim}
COMPUTE discret=income-expenses.
COMPUTE income=income*(1-taxrate).
\end{verbatim}

SPSS/PC+ will not adjust discretionary income for the effect of taxes in this example! To accomplish this task, the second COMPUTE statement would have to precede the first. In other words, the effect of the change in the income calculation is not "retroactive" as it would have been in Lotus if you had gone back to change the formula in the cell in which (pre-tax) income was computed. It will, however, be in effect any time after the change has been performed.
Appendix G

Similarities Between SPSS/PC+ and dBase III+

Readers familiar with dBase III and similar database programs may notice certain conceptual similarities in SPSS/PC+. Although understanding how database software functions is not necessary to understanding how SPSS/PC+ works, it may help you understand some concepts more easily.

First of all, both SPSS/PC+ and dBase clearly divide the data into individual cases or "records." In dBase, the data associated with each entity is kept within a single record. Thus, if we have a database of student grades, a dBase record might look like this:

<table>
<thead>
<tr>
<th>Record 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
</tr>
<tr>
<td>MIDTERM</td>
</tr>
<tr>
<td>PAPER</td>
</tr>
<tr>
<td>FINAL</td>
</tr>
</tbody>
</table>

In SPSS/PC+, each case is assigned a specific number of lines. Thus, our SPSS/PC+ case might look like this:

5551515555 084095098

Each record or case is further subdivided. In dBase, we call each variable a "field." For example, the field "MIDTERM" contains the student's midterm score of 84. In SPSS/PC+, the variables are
assigned specific column positions. If we look immediately after the blank space following the ID number (probably the student’s social security number), we can see the entry of "084," signifying the score of 84. Notice that the separation of the variables is completely up to us; we defined a blank space between the ID number and the scores, but none between the scores themselves.

DBase allows us to compute several statistical functions. Thus, we could use the "average()" and "std()" functions to get much the same information we could get with "descriptives" command in SPSS/PC+. However, DBase does not give us a direct way to do anything as complicated as crosstabs, correlation, regression, or even frequencies.

Sometimes, we are only concerned with part of our sample. Thus, in DBase, if we wanted to find the average on the final for those people who scored below 70 on the midterm, we could say

\[
\text{AVERAGE(final) FOR midterm<70}
\]

while in SPSS/PC+, we might issue the commands

\[
\text{PROCESS IF (midterm LT 70).}
\text{DESCRIPTIVES VARIABLES=final.}
\]
Appendix H

**SPSS-X: The Mainframe Version**

As mentioned in the introduction, a mainframe version, known as SPSS-X, is also available on many university campuses. There are certain circumstances under which it may be desirable to use the mainframe. Some features are not readily available on the PC or may take too long to execute. This will normally not be a problem except you wish to undertake very advanced analyses.

Another advantage of using a mainframe computer is that many university computer centers have a "high speed" printer connected to their mainframes. These printers can often print out a hundred pages in less than one minute. Although you normally have to wait for the printouts to be distributed to an "out" box at some central location, you at least save the time you would have to sit in front of a slower "ordinary" printer.

SPSS-X commands are very similar the SPSS/PC+ commands, the main exception being that there are normally no periods after the end of each command. You may also have to include some "file handle" commands under certain circumstances and, of course, you will have to deal with some computer commands very different from those of the IBM PC.

Extensive documentation on the use of SPSS-X is available from SPSS, Inc. You may also have to get some handouts on your mainframe's implementation of SPSS-X.
Appendix I

Iincorporating SPSS/PC + Output Into
WordPerfect Reports

Sometimes, you might like to include the results of a statistical analysis in a report. While it is possible to manually type the results from your computerized analysis into a word processing document, it can be cumbersome and the complexity involved can lead to a number of mistakes. Fortunately, there is an easier way.

SPSS/PC+ creates an output file to which everything you see on the screen or on printer output is sent. This file, called "\SPSS\SPSS.LIS," is overwritten every time you restart SPSS/PC+, so you will have to "rescue" the material immediately if you need it.19

Although this manual will illustrate the use of WordPerfect 5.0 or 5.1 to "grab" important output, most word processors will provide the same capability.20

Immediately after you exit SPSS/PC+, select WordPerfect from the menu. If you have already created the document in which you would like to include SPSS/PC+ output, call it up now. Regardless of whether you take that step or not, now switch to document #2 (<SHIFT>-<F3>). Use the "text-in" function to retrieve the file "\SPSS\SPSS.LIS":

<CONTROL>-<F5> 1 2 C:\SPSS\SPSS.LIS <RETURN>

Find the part of the output you would like to include in your

19 Alternatively, can use the "SET RESULTS" command to direct your output to a permanent file (see manual).

20 SPSS.LIS is an ASCII file and should be edited and read in accordingly.
document and use the block function (<ALT>-<F4>) to mark the appropriate text. To copy the text, use <CONTROL>-<F4> to copy and press <SHIFT>-<F3> to move back to document #1. Use the cursors to move where you want the text inserted, then press <RETURN>.

If you want more text from the output, you can repeat the process.

**NOTE:** In some cases, the lines of SPSS/PC+ tables are more than 65 columns long. If you encounter such a table, you will want to be sure to use a font smaller than the default Courier, 10 pitch. Use the base font option (<CONTROL>-<F8>-4) to select. (The SPSS/PC+ output included in this text was printed in Letter Gothic, 16.67 pitch).

Also, beware that you should never use a proportionally spaced font for tables. (Helvetica and Times Roman are two of the most common proportionally spaced fonts). While text printed in a proportionally spaced font looks very beautiful, this scheme tends to distort tables that depend on proper alignment of variables.
Appendix J

Dealing With Printer Problems

Unless you specify otherwise, SPSS/PC+ assumes that you are printing to printer that uses continuous sheets of paper, each of which is sixty-six lines long. This is fine when you print to dot matrix printers whose paper is "tractor fed."

However, you may run into some problems when you try to print to a laser printer or other kind printer that prints on stacks of single sheets of paper. This happens because, although there is room for sixty-six lines of text on a laser printer, not all of the space on the paper is available for the printer's use. Thus, the pages may become disalligned; that is, the page breaks may not occur at the right places.

To deal with this problem, enter the following series of commands where you would otherwise place the "set printer on" command:

SET SCREEN ON.
SET MORE OFF.
SET LENGTH=59.
SET PRINTER ON.
SET EJECT ON.

Be sure to enter the commands in the suggested order.

NOTE: When SPSS/PC+ is done printing, you may need to press the <FORM FEED> key on the printer to get the last page out.
When studying a subject such as statistics, it is easy to lose sight of the forest for the trees (or, in more contemporary terms, to lose sight of the computer for the chips). In order to get the "big picture" in statistics, it is important that you remember the fact that statistics is all about making inferences about an underlying population from a limited sample. If you do a study of grocery shoppers, you really don't care in particular about the people you are interviewing. A sample made up of 100 people is in and of itself irrelevant (it is a very small part of the total market). You are really interested in generalizing the responses of the sample to the population as a whole. This is not to say that the selection of your sample is not important. In order to make sure that the sample is representative of the population, the researcher should use a sampling scheme that ensures proper selection. Depending on the circumstances, one may want to use either a random or systematic sample or some variation of these.

As you know from introductory statistics, a sample mean will rarely ever equal the "true" population mean. You may recall the Central Limit Theorem, which suggests that the larger the sample size, the closer its mean will be to the actual population mean. In general, the principle that a larger sample size will allow you to make more precise statistical inferences applies to other statistical procedures, too.

Every profession has its peculiar jargon, and statistics is no exception. One of the common terms in this field is that of statistical "significance." Despite what the word suggests, statistical significance has nothing to do with importance. As ironic as it sounds, something can be very significant but not at all important. Consider this fact: There is a statistically "significant" relationship between the number of people at a Southern California beach and the number of leaves on the surrounding eucalyptus trees. (This is, in terms of another piece of
statistical jargon, a spurious correlation in the sense that both of these phenomena are "caused" by a common factor [the season of the year] although there is no intercausal relationship between the two).

In the broadest possible terms, statistical significance refers to the probability that one would observe a sample statistic as extreme as the one obtained if, in fact, there were no such relationship in the population.

You will recall from your statistics classes the use of the null and alternative (or "research") hypotheses \( (H_0 \text{ and } H_a) \). You assume the null hypothesis, which states that there is no relationship, to be true until you find overwhelming evidence that it is not—in that case, you "accept" the alternative hypothesis, knowing that there is a certain chance (e.g. 5%) that the relationship does not exist.

Consider this fictitious example: A researcher samples five athletes and five non-athletes to see how many soft drinks they each consumed last week. The results are as follows:

<table>
<thead>
<tr>
<th></th>
<th>Soft drinks consumed</th>
<th>Standard deviation</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Athletes</td>
<td>19.1</td>
<td>7.7</td>
<td>5</td>
</tr>
<tr>
<td>Non-athletes</td>
<td>14.9</td>
<td>6.5</td>
<td>5</td>
</tr>
</tbody>
</table>

Intuitively, would you be confident in concluding that athletes drink more soft drinks than non-athletes? Although you could do a t-test to be sure, you would probably conclude without doing so that you don't have enough evidence to conclude that there is a difference. (This would particularly be the case if you had had not made a hypothesis in advance that athletes drink more soft drinks than non-athletes). The fact is that the variance within each group is so high compared to the difference in the two sample means suggests that no confident conclusion is possible. Now let's suppose, however, that we sampled ten thousand athletes and ten thousand non-athletes and got the following results:
In this case, the variance is much smaller relative to the differences in the sample means, and although you might not feel confident intuitively concluding that there is a difference, the evidence now seems impressive enough to attempt a formal t-test.

As previously indicated, it is possible to generalize the notion of significance to other statistical tests. Consider the example of correlation. Recall that the Pearson correlation coefficient ranges from -1, a perfect negative relationship, to 1, a perfect positive relationship. When the population coefficient $r=0$, there is no relationship. Let us suppose that there is in fact no relationship between GPA and the number of soft drinks consumed the previous week. However, a group of students survey of fifty individuals and find a sample correlation of $r=0.05$. Does this mean that there is a small but real relationship between the two variables? No! The correlation is so close to zero that it may just have happened that the number came out this way. If a "competing" group of students wishing to prove that the relationship was negative attempted another survey, they might find $r=-0.08$. Another group might find $r=0.01$. Thus, the only thing we seem to be able to conclude is that, if any relationship exists in the population, it is probably trivial.

This suggests that a certain amount of "error" variance takes place. In a limited sample size, the correlation coefficient obtained may not be the true population coefficient, even if the sample is a perfect random sample. For this reason, you will find a significance value below the correlation coefficient and sample size if you use the parameters in the correlation statement described in this text; that is, for example:

```
CORRELATION var1 WITH var2
/OPTIONS=5.
```

If you choose a 95% level of confidence, for example, you would
say that there is a relationship only when the significance level is less than 0.05. This significance level only implies that there is either a positive or negative relationship, not how big it is. (Beware that these are one-tailed levels of significance. This means that you would have had to make a directional hypothesis (e.g. "Soft drink consumption is positively associated with GPA," not "Soft drink consumption is associated (either positively or negatively) with GPA).

When running the correlation procedure, you will tend to find that, for a small sample size, a relationship will have to be strong to show up as significant. Also beware that even if a relationship shows up as significant, the true correlation coefficient may be greatly different from what was obtained. For example, you might observe a sample coefficient of 0.41. However, the 95% confidence interval for the true coefficient may be something like 0.25-0.57--quite a bound!

To reiterate, it is useful, in general, to look upon significance as the probability that a relationship as extreme as what was observed would have happened in a sample of a given size if the null hypothesis were true; that is, if no relationship actually existed. You will find significance levels in virtually all statistical tests including, but not limited to, regression, chi square, discriminant analysis, t-tests, and correlation. Some analyses such as regression may contain several significance tests. There will be significance levels at each step (as each additional variable is entered into or removed from the equation). Further, each step will have a separate significance test for each variable, measuring the confidence we can have that the particular variable should in fact be in the "true" regression variable. The overall regression significance test examines whether we in fact have an equation which significantly "explains" a portion of the variance in the population dependent variable.

A related topic is that of simultaneous significance testing. This is a relatively difficult subject to grasp, but one that is nevertheless very important.

Remember that, if we choose the 0.05 level of significance, each
time we run a correlation analysis, there is a five percent chance of committing what is called a "Type I" error, i.e. concluding that there is a significant relationship when there in fact isn't. Now, let's suppose that we run twenty analyses and find that one is significant. Should we conclude with confidence that a relationship exists?

Doing so may be risky. We ran twenty analyses, in each of which we ran a 5% risk of being wrong. Multiplying twenty (the number of analyses we ran) by 0.05 (the probability of each one being wrong), we find that the "expected" number of "false alarms" is 1. Thus, the relationship may not be truly significant.

We can also look at this issue another way. If we make one significance test at $\alpha=0.05$, the probability of a Type I error is 0.05. If testing all our hypotheses requires us to make ten such tests, what is the probability of making at least one Type I error? This probability is not $10(0.05)=0.50$, although that may seem an intuitive answer. Instead, we must use the Bernoulli Formula. When you encountered this formula in finite mathematics, you most likely used it to calculate the something like the probability that there would be two or more left-handed people in a class of thirty.

The Bernoulli Formula involves what in finite mathematics is called a "combination," or the number different ways one can select a sample of a desired size from a population of a given size.\(^{21}\) The formula involves only basic arithmetic, but lots of it, so the distribution disk includes a program called "SIGNIF.EXE" which will calculate the probability of getting at different numbers of significant outcomes under a given number of tests at the desired significance level. (With the distribution disk in Drive A:, type in "A:SIGNIF.EXE" at the DOS prompt to run the program).

\(^{21}\) The theory behind Bernoulli trials and combinations is beyond the scope of this text but is generally discussed in introductory finite mathematics texts.
Statisticians have never satisfactorily resolved the issues and problems that surround simultaneous significance testing. However, there are some ways of limiting its potential dangers:

• Be sure to make your hypotheses before running your tests and do not run any more tests than are necessary to answer the questions raised by your hypotheses. Whenever possible, hypotheses should be one-tailed, i.e. predicting the direction of the relationship.

• Confine your tests to those for which you have an adequate sample size. If you have only thirty subjects, for example, it is unlikely that you will find a truly significant correlational relationship, so running that procedure is not appropriate.

• Realize that sporadic significant tests are perhaps significant by chance. (If, for example, you obtained three significant independent tests out of forty, you could use the Bernouli Formula to figure out the probability of obtaining that many "successes" by chance.

Significance is not an easy topic, and you may have to do a lot of thinking in your project to find out how the concept relates to your particular findings.
Appendix L

Expanded Glossary

Statistics

**alpha (α):** (1) the maximum probability of a Type I error allowed in a statistical test. The level of α=0.05 is frequently used in most tests. (2) The level of internal consistency in an alpha reliability analysis; α is a conservative estimate of the test-retest correlation between a scale and a hypothetical parallel version. (See also reliability.)

**alternative (research) hypothesis (H_a):** The hypothesis that the researcher is trying to support. For example, if you are trying to prove that athletes on the average drink more soft drinks than non-athletes, the alternative hypothesis, H_a, might be \( \mu_a > \mu_n \), where \( \mu_a \) is the mean number of soft drinks consumed by athletes and \( \mu_n \) is the mean number of soft drinks consumed by non-athletes. Note that the alternative hypothesis can either be one tailed, where we hypothesize that one mean is greater than the other, or two-tailed, where we merely hypothesize that the two means are different.

**analysis of variance (ANOVA):** A statistical method that enables the user to assess the contribution of one or more independent variables to the variability of the dependent variable. For example, a researcher may wish to test the hypothesis that the number of soft drinks consumed by a student is influenced by both major (1=business, 2=liberal arts, 3=engineering, 4=other) and athletic status (1=athlete, 2=non-athlete). ANOVA is a very flexible statistical procedure that allows the user to create sophisticated statistical models. (See also ONEWAY.)

**case:** One entity or unit of data on one or more variables; usually one subject, individual, or other unit (such as a firm). For example,
when you administer a questionnaire, each person is a case, but if you have both a pre-test and a post-test, the two tests for the same subject together constitute a case.

Central Limit Theorem: The theorem which holds that, regardless of the distribution of a particular variable, the means of any large sample from the given distribution will follow a normal distribution. Thus, we are really talking about "means of means," or in less formal terms, averages of averages. Knowing this property allows us to perform certain tests without assuming that a variable involved is normally distributed. The distribution of the means is dependent on both the sample size and the variability of the population. Taking this information into consideration, we can calculate the standard error of the mean, a concept comparable to the standard deviation of a variable. The standard error of the mean is appreciably smaller than the standard deviation of the variable; the formula is

\[ \text{Standard Error of the Mean} = \frac{s}{\sqrt{n}} \]

One application of the Central Limit Theorem is the calculation for the mean of a "large" sample (i.e. a sample that has an \( n \) greater than 30).

Chi square (\( \chi^2 \)): A statistical test that determines the probability that two nominal variables, given the sample data, are "independent," i.e. whether knowing one variable will help "predict" the other. The formula is

\[ \chi^2 = \sum \frac{(O - E)^2}{E} \]

Chi square may be used on ordinal, interval, or ratio scaled variables, but more efficient tests are available for such variables.
where \( k \) is the number of "cells" (i.e. different combinations of the two variables), and \( O_i \) and \( E_i \) are the observed and expected number of cases in the respective cell. (See also crosstabulation.)

**confidence interval**: In context of the mean of a variable, the interval that is estimated with a given certainty (e.g. 95%) to contain the true mean. The Central Limit Theorem allows us to estimate confidence interval for the mean of a variable can be calculated by using the standard error of the mean \( (s/\sqrt{n}) \). The formula is:

\[
\text{Confidence Interval} = \bar{x} \pm z_{\alpha/2} \left( \frac{s}{\sqrt{n}} \right)
\]

where \( \bar{x} \) is the sample mean, \( s \) is the sample standard deviation, \( n \) is the sample size, and \( z_{\alpha/2} \) is the two-tailed z-value for the given level of significance.

**coding**: The process of translating research data (such as questionnaire responses) into consistent numerical codes for input into the computer. For example, to the question of "Did you vote in the last election?" we might assign "1" to "yes," "2" to "no" and "9" for missing data.

**correlation**: A measure of how well two variables "predict" each other. Correlation can either take the form of the Pearson Product-Moment Correlation, which assumes interval data, or Spearman Rank-Order Correlation, which assumes only ordinal data. (See also regression.)

**crosstabulation**: A table showing the relationship between two variables. For example, the below table shows a hypothetical relationship between a person's height and stuffed animal preference. Of the respondents, six people were both short and preferred giraffes.
Crosstabulation:  

<table>
<thead>
<tr>
<th>ANIMAL</th>
<th>Giraffe</th>
<th>Elephant</th>
<th>Bear</th>
<th>Cow</th>
<th>Pig</th>
<th>Dog</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short</td>
<td>6</td>
<td>10</td>
<td>14</td>
<td>5</td>
<td>60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>4</td>
<td>10</td>
<td>16</td>
<td>5</td>
<td>60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tall</td>
<td>3</td>
<td>10</td>
<td>17</td>
<td>5</td>
<td>60</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Height of customer</th>
<th>Count</th>
<th>Giraffe</th>
<th>Elephant</th>
<th>Bear</th>
<th>Cow</th>
<th>Pig</th>
<th>Dog</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short</td>
<td>1</td>
<td>6</td>
<td>10</td>
<td>14</td>
<td>5</td>
<td>60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>2</td>
<td>4</td>
<td>10</td>
<td>16</td>
<td>5</td>
<td>60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tall</td>
<td>3</td>
<td>10</td>
<td>17</td>
<td>5</td>
<td>60</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ANIMAL</th>
<th>Giraffe</th>
<th>Elephant</th>
<th>Bear</th>
<th>Cow</th>
<th>Pig</th>
<th>Dog</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short</td>
<td>6</td>
<td>10</td>
<td>14</td>
<td>5</td>
<td>60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>4</td>
<td>10</td>
<td>16</td>
<td>5</td>
<td>60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tall</td>
<td>3</td>
<td>10</td>
<td>17</td>
<td>5</td>
<td>60</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Chi-Square:  

<table>
<thead>
<tr>
<th>Chi-Square</th>
<th>D.F.</th>
<th>Significance</th>
<th>Min E.F.</th>
<th>Cells with E.F.&lt;5</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.17035</td>
<td>10</td>
<td>.8008</td>
<td>5.000</td>
<td>None</td>
</tr>
</tbody>
</table>

The SPSS/PC+ crosstabs procedure allows you to calculate several statistical tests and functions such as Chi square ($\chi^2$). (See also Chi square.)

data list: The command in SPSS/PC+ which informs the computer of the names and positions of the variables in the data. For example, the line

data list / id 1-3 age 5-6 sex 7.

means that for every case, the ID number is contained in columns one through three; age in columns five through six; and sex in column seven. The data list is a required command.

data set: The total collection of cases and variables included in a file. For example, if you administer a questionnaire to one hundred people and enter the answers to all the questions into the computer, this would be your data set.

dependent variable: The variable we are trying to "predict" from one or more independent variables. For example, if we wish to test whether athletes on the average consume more soft drinks than
non-athletes, the number of soft drinks consumed is the dependent variable and athletic status is the independent variable. See also independent variable.

**descriptives**: a procedure that calculates the mean and various optional statistics associated with a variable.

"**dummy**" variable: see indicator variable.

**frequencies**: A procedure that indicates the number of cases associated with each value of a variable. For example, the below table indicates the number of people who gave each of various answers to the question of whether they owned a dog. Various statistics, such as the mean, median, mode, standard deviation, and standard error of the mean, optionally are available with this procedure.

<table>
<thead>
<tr>
<th>Value Label</th>
<th>Value</th>
<th>Frequency</th>
<th>Percent</th>
<th>Percent</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>1</td>
<td>82</td>
<td>46.9</td>
<td>49.1</td>
<td>49.1</td>
</tr>
<tr>
<td>No</td>
<td>2</td>
<td>82</td>
<td>46.9</td>
<td>98.2</td>
<td></td>
</tr>
<tr>
<td>Not sure</td>
<td>3</td>
<td>3</td>
<td>1.7</td>
<td>1.8</td>
<td>100.0</td>
</tr>
<tr>
<td>Missing</td>
<td>9</td>
<td>8</td>
<td>4.6</td>
<td></td>
<td>MISSING</td>
</tr>
<tr>
<td>TOTAL</td>
<td>175</td>
<td>100.0</td>
<td>100.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**independent variable**: A variable used to "predict" a dependent variable. See dependent variable.

**indicator variable**: A binary variable used to designate whether a case falls into a category or not. Usually, "1" denotes that the subject has the characteristic and "0" that it does not. Indicator variables are most commonly used in correlation or regression analyses where they can be used to represent such dichotomies as sex (1=Female, 0=Male) or residential status (1=Urban, 0=Other).

**internal level of measurement**: See levels of measurement.
**Levels of Measurement**: Variables can be divided into four categories or levels of measurement.

At the "lowest" level are nominal scales, where the numbers are arbitrary. Thus, if, for a given variable, the code "1" corresponds to "Democrat" and "2" corresponds to "Republican," the scale is not meant to imply that Republicans are exactly twice as much of anything as Democrats. In fact, we could reverse the two codes without making any real difference.

In ordinal (or monotonic) scaling, the measure increases or decreases with the number. Thus, if we assign a "7" to "strongly agree", "5" to "slightly agree," and a "4" to "neither agree nor disagree," the person who responds "7" agrees more than the other two. We cannot say, however, that the person who chooses "7" agrees 1.75 times (7/4) as much as the person who is neutral.

In an interval scale, the units of increment are equal, but there is no absolute zero. A classic example of this is a temperature scale, where 0°F or 0°C are not absolute minima.

A ratio scale has an absolute zero, and increments are equal. Using a scale of income, a person who earns $40,000 earns exactly twice as much as the person who earns $20,000. (Note, however, that the scale is no longer a ratio scale if it has been collapsed [1=$0-10,000, 2=$10,001-20,000, etc.])

**Likert Scale**: A scale which asks the respondent to rate his or her opinion on a bi-polar scale. For example, a respondent might be asked to rate his or her level of agreement with a statement using a scale such as:

<table>
<thead>
<tr>
<th>Strongly agree</th>
<th></th>
<th>Strongly disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Other dimensions might include, but are not limited to, "Extremely satisfied/Not at all satisfied," and "Extremely important/Not at all important."

**mean:** The quotient of sum of all the cases of a variable to the sample size; i.e., for the sample mean,

\[
\text{mean} = \frac{\text{sum of all the cases of a variable}}{\text{sample size}}
\]

This quantity is also known as the arithmetic mean or, in less precise language, the average.

**median:** The middle value of a variable, i.e. the value of a variable such that an equal number of variables are less than and greater than or equal to this quantity. If \( n \) is even, the median is the mean of the two middle values. For example, in the set \( \{1, 3, 5, 7, 8\} \), the median is five. In the set \( \{1, 3, 4, 5, 7, 8\} \), the median is \((4+5)/2=4.5\). Note that even if we rearrange the order of the set to \( \{3, 8, 7, 1, 2, 5\} \), the median is still 4.5 since the result is based on the ranked values.

**missing data:** The condition that arises when data is not available or not applicable on one or more variables for one or more cases. Missing data can take two forms.

The first form occurs when a subject fails to respond to a question that he or she should have answered. (This omission may be deliberate or accidental). We often designate a missing value consisting of one more "9s" for this purpose, although the exact value designated is arbitrary. For example, we might designate the missing value of "99" for the variable "age" since
we normally assign two digits for this variable.

The other kind of missing data occurs when a person is not supposed to respond to a question. For example, if we have already determined that a person does not own a dog, it does not make sense to ask him or her to respond to a question about his or her expenditures on dog food (except, of course, if we are surveying people who tend to be on unusual diets). We frequently leave the data columns assigned to the variable blank in such cases.

**mode**: The value that occurs most frequently on a given variable. For example, in the set \{1, 2, 3, 3, 3, 3, 4, 4\}, the mode is 3.

**nominal level of measurement**: See levels of measurement.

**multiple r**: See \(R^2\), the square of this quantity, which is a more interpretable statistic.

**non-parametric statistics**: Statistical procedures that do not make assumptions about the distribution or level of measurement of the data. For example, Chi square (\(\chi^2\)) is a non-parametric statistic.

**null hypothesis (H\(_0\))**: The hypothesis that no statistical relationship or difference exists between two variables in the population. For example, if we were interested in testing the hypothesis that athletes drink more soft drinks than non-athletes, the null hypothesis would be \(\mu_a = \mu_n\). See also alternative hypothesis.

**O NEWAY Analysis of Variance**: A method of testing for differences in group means on a variable. For example, we might use the O NEWAY procedure to test the hypothesis that there is a difference in the average number of soft drinks consumed by students, professors, and clerical workers. While it would have been possible to make pairwise comparisons between these groups using t-tests, error levels would accumulate since an
alpha probability of a Type I error is associated with each test.

**ordinal level of measurement**: see *levels of measurement*.

**parametric statistics**: Statistical procedures that assume an interval or ratio level scale of measurement. For example, Pearson product-moment correlation and multiple regression are parametric procedures.

**population**: The complete group about which we want to make generalizations from our statistical data. Specifying a population is important but can be difficult. For example, if we are interested in the soft drink consumption habits of athletes, is our population "all professional athletes in the United States," "all professional, amateur, and recreational athletes in the World," or some other group?

Technical note: In defining a population, statisticians generally refer to the actual units of comparison rather than the subjects with which they are associated. Thus, to a statistician, the population in the above example would be the soft drink consumption counts scores and *not* the athletes. However, in the social sciences and in this text, the term is more loosely used to refer to the subjects, that is, in this case, the athletes themselves.

**random sample**: A sample selected in a manner such that each element has the same probability of being picked from the population. Random samples are typically selected using a random number table or through random pickings by a computer.

**ratio level of measurement**: See *levels of measurement*.

**recoding**: The process of systematically changing values of a variable into others. In SPSS/PC+, recoding can take several forms. First, we may want to "collapse" a variable. For example,
supposed we asked a person to name his or her favorite kind of food. We might have a separate code for several kinds of Asian food (E.g. Chinese=11; Japanese=12; Korean=13; Thai=14) and several kinds of Italian food (pizza=21; pasta=22). Part of our recode statement might thus look like this:

\[
\text{recode food (11 thru 19=10) (21 thru 29=20).}
\]

Secondly, we may want to "reverse" score a variable. For example, a scale of prejudice against poodles might include both of these Likert scale items:

\[
\begin{align*}
\text{___: Poodles are stupid} \\
\text{___: Poodles are loving dogs.}
\end{align*}
\]

Note that the two variables go in the opposite direction of each other; agreeing with the first (poodle1) would imply a negative attitude toward poodles while agreeing with the second (poodle2) one would imply a positive attitude. Thus, we should recode the second variable so that a high score will become a low score and vice versa. If we asked the subject to respond based on a seven point Likert scale, our recode command might look like this:

\[
\text{recode poodle2 (1=7) (2=6) (3=5) (5=3) (6=2) (7=1).}
\]

**regression**: A statistical procedure which attempts to predict a dependent variable from one or more independent variables. Regression requires interval or ratio data, although, in practice, it is often used with ordinal data. In such cases, interpretation is somewhat open to question. Regression differs from correlation in that more than one variable can be used to predict the dependent variable.

**reliability**: The extent to which a variable consistently measures some phenomenon. Reliability can be assessed in various ways. If we have two "parallel" or similar measures of the same variables, such as two comparable versions of the same test, we can correlate these. Another method is to administer the same
"test" or questionnaire twice and correlate the scores against each other. A third method is the measurement of "internal consistency" of a scale consisting of several variables through coefficient alpha. See also alpha.

\[ r^2 \]: The proportion of the total variance in a dependent variable explained by a regression equation. \( r^2 \) is the square of \( r \) in a simple correlation or multiple \( r \) in multiple regression analysis.

**reverse scoring**: see recoding.

**sample**: A subset of the population selected for analysis. Often, it is not possible to survey or examine the entire population, making it necessary to generalize from a sample. For example, it would be logistically impossible, let alone cost effective, to contact all athletes to find out how many soft drinks they consume. Thus, we might ask only a limited number, such as 200, and try to generalize to the entire population with some level of confidence.

**significance**: The probability that some statistical phenomenon would happen by chance in a sample of a given size if no difference or relationship actually existed in the population. For example, in a t-test, if \( p=0.03 \), differences as large those observed would be expected only in three percent of samples if no actual difference exists in the population.

**simultaneous significance testing**: The condition that arises when several significance tests are made at the same time. Since a separate alpha (\( \alpha \)) value, or probability of a Type I error, is associated with each test, error levels accumulate. Thus, for example, if we perform six independent tests at the same time, there is a higher probability (approximately, but not exactly, .30) that at least one test will be significant at the 0.05 level even if no relationship or difference exists in the population.
**standard deviation**: A measure of variability within a variable. The standard deviation of a sample is:

$$\text{standard deviation} = \sqrt{\frac{\sum (x_i - \mu)^2}{n}}$$

When calculating by hand, the following formula is more efficient to use:

$$\text{standard deviation} = \sqrt{\frac{\sum x_i^2}{n} - \left(\frac{\sum x_i}{n}\right)^2}$$

**standard error of the mean**: An estimate of the variability of the mean of a sample from a given population. The greater the variability (i.e. standard deviation) of the variable in the sample, the less precisely we can estimate the true mean of the population for a given sample size. See also Central Limit Theorem.

**t-test**: A statistical method used to test for differences in the means of either (1) the same variable between two different groups or (2) between two different variables relating to the same subject.

The first form is illustrated by the testing of our hypothesis that athletes drink more soft drinks than non-athletes.

The second form is frequently used to test for differences between a pre-test and post-test. That is, we first give the subjects a test, we then make some "intervention," and then test again to see if it appears that the intervention has had an effect. For example, suppose we first determine how many soft drinks
of our brand that each member of a random sample of people in two different regions consume. We then conduct a promotional campaign in the one region and none in the other region (which will serve as our control group). (Assume we have selected regions that are relatively similar in terms of climate and social stratification.) After the promotional campaign is over, we again determine how many soft drinks are consumed by people in each sample. If there is a significant difference between consumption before and after for those people in the region where the campaign was conducted, we might conclude that the campaign had an effect on consumption of our brand. However, if there were also significant differences in the consumption in the control group (which received no promotional campaign), we should be alert to the possibility that the increase in sales has actually been caused by some other effect (such as changes in the weather or a change in consumer preferences). In any event, we can never be absolutely sure that increase in consumption has been caused by our intervention, but if all other reasonable causes have been eliminated, we can make a good case for that assumption.

**type I error**: The error of incorrectly rejecting the Null Hypothesis (H₀) when it is in fact true. For example, our sample might lead us to conclude with, say, ninety-five percent confidence, that athletes drink more soft drinks than non-athletes when, on the average, the two groups actually consume the same number. The probability of committing a type I error is denoted by the letter α. Note that a type I error is not the result of an inappropriate application of a statistical procedure. Rather, it occurs as a result of our selection of a test which will lead to a false positive a certain percentage of the time. We can adjust this probability of false positives to our liking, but doing so involves a tradeoff against the probability of failing to accept the alternative hypothesis when it is in fact true. See also alpha, type I error.

**type II error**: The error of incorrectly of failing to reject the null
hypothesis ($H_0$) when the alternative hypothesis ($H_a$) is in fact true. For example, our statistical test might lead us to determine that there is not enough evidence to conclude from our sample that athletes drink more soft drinks than non-athletes; our sample may show differences, but those differences may not be large enough to be statistically significant. Note that a type II error is not the result of an improperly applied statistical procedure. Rather, it is the result of the finite sensitivity of the test.

**value labels**: An optional SPSS/PC+ command which allows us to assign labels to the values (i.e., in a questionnaire, the answers to a question) associated with a variable. For example, for the question "Did you consume one or more soft drinks today?" we might assign the label of "Yes" to the value of "1" and "2" to the value of "No." Some variables, such as age in years, should not have value labels assigned since the answers are self-explanatory. However, value labels might be used if the variable is collapsed, as in the following example:

```
value labels
age 1 '0-15' 2 '16-25' 3 '26-40' 4 '40-55'
5 'Over 55'.
```

**variable labels**: A command which allows us to assign a "label," or brief description, to a variable. The variable label is intended to help us identify the meaning of the variable and has no effect on the computer's computations. If one question on a survey asks how many soft drinks a person consumes on an average day, a good variable label might be "Soft drinks consumed on an average day" but, if by some accident we typed in "Expenditures on Christmas presents," the computer would not know that anything was wrong.

**variable name**: A name used by the computer to identify a variable. For example, we might assign the name AGE to the variable describing the subject's age. A variable name has to comply with approximately the same rules as the ones that apply for naming DOS files; i.e. a filename must be no longer than eight
characters long, must begin with a letter of the alphabet, and the remaining letters must be either letters, numbers, or certain other characters such as underscore (_).

Computer Terms

**ASCII file**: A plain text file that can be edited. SPSS/PC+ data files are ASCII files while system files, word processing files, and executable files are normally not ASCII files.

**cursor**: The small blinking character that tells you where on the screen you are editing.

**file**: An electronic means of storing data, normally on a disk. A file has a file name and the data is accessed by referring to the filename.

**system file**: An SPSS/PC+ file in which the data, value labels, and variable labels are compressed so that they can be more quickly accessed by SPSS/PC+. System files are normally only worth the effort if the data set is very large. For more information, see Appendix D.
Appendix

Additional Sources of Information

There are both official and unofficial SPSS/PC+ manuals on the market. The official manual put out by SPSS, Inc., comes in five volumes, all of which are written by Marija Norusis:

- **SPSS PC/ +Base Manual**: Covers basic topics such as frequencies, crosstabs, correlation, condescriptive analysis, and data recoding and variable computing.

- **SPSS/ PC+ Advanced Statistics Manual**: Covers more advanced topics such as factor analysis, discriminant analysis, and cluster analysis.

- **SPSS/ PC+Version 3.0 Update Manual**: Covers additions to SPSS/ PC+ made in Version 3.0 such as the inclusion of Spearman Rank-Order Correlation, non-linear regression, and a new data checking program.

- **SPSS PC/ +Trend**: Covers forecasting and time series programs available within SPSS/ PC+.

- **SPSS PC/ +Tables**: Covers special table and presentation tools available within SPSS/ PC+.

While the official manuals are good for reference purposes, some more readable (but less detailed) unofficial books are also available:


- **Neil Frude, SPSS/ PC+Introductory Guide** (New York:
Springer Verlag, 1987).