3.9 Exercises

1. Open `relate.dta`. The `R3828700` variable represents the gender of the adolescent and is coded as a 1 for males and 2 for females. There are 775 people who are missing because they dropped out of the study after a previous wave of data collection. These people have a missing value on `R3828700` of `-5`. Run a tabulation on `R3828700`. Then modify the variable so that the code of `-5` will be recognized by Stata as a missing value. Label the variable so that a 1 is male and a 2 is female. Finally, run another tabulation and compare this with your first tabulation.

2. Open `relate.dta`. Using the `R3483600` variable, repeat the process you did for the first exercise. Go to the webpage for the book and examine the dictionary file `relate.dct` to see how this variable is coded. Modify the missing values so that `-5` is `.a`, `-4` is `.b`, `-3` is `.c`, `-2` is `.d`, and `-1` is `.e`. Then label the values for the variable and run a tabulation.

3. Using the result of the second exercise, run the command `numlabel, add`. Repeat the tabulation of `R3483600`, and compare it with the tabulation for the second exercise. Next run the command `numlabel, remove`. Finally, repeat the tabulation, but add the `missing` option; that is, insert a comma and the word "missing" at the end of the command. Why is it good to include this option when you are screening your data?

4. `relate.dta` is data from the third year of a long-term study. Because of this, some of the youth who were adolescents the first year (1997) are more than 18 years old. Assign a missing value on `R3828100` (age) for those with a code of `-5`. Drop observations that are 18 or older. Keep only `R0000100`, `R3483600`, `R3483800`, `R3485200`, `R3485400`, and `R3828700`. Save this as a dataset called `positive.dta`.

5. Using `positive.dta`, assign missing values to four of the variables: `R3483600`, `R3483800`, `R3485200`, and `R3485400`. Copy these four variables to four new variables called `mompraise`, `momhelp`, `dadpraise`, and `dadhelp`, respectively.

6. Create a scale called `parents` of how youth relate to their parents using the four items (`R3483600`, `R3483800`, `R3485200`, and `R3485400`). Do this separately for boys (`if R3828700 == 1`) and girls (`if R3828700 == 2`). Use the `rowmean()` function to create your scale. Do a tabulation of `parents`.

4.7 Exercises

1. Open `firstsurvey_chapter4.dta` by selecting File ▷ Open.... Open a Do-file Editor, and copy into the Editor the command that opened the dataset. Open the dialog box to summarize the dataset, run the `summarize` command for all the variables, and copy this command from the Results window into the Do-file Editor. Save this do-file as `4-1.do` in a place where you can find it.

2. Open the do-file you created in the first exercise, and add appropriate comments. Save the new do-file under the name `4-2.do`.

3. Open the file `4-2.do`. Put your cursor in the Do-file Editor just below the command that opened the dataset and above the command that summarized the variables (you will need to insert a new line to do this). Type the `describe` command in the Do-file Editor. Add a command at the bottom of the file that gives you the median score on education. Save the new do-file under the name `4-3.do`.

4. Open `4-3.do`, run the `describe` command and the command that gave you the median score on education, highlight the results, paste the results into your word processor, and change the font so that it looks nice.

5. Open `4-3.do`. Open a log file with the log file type. Call the file `4results.log`. Run the entire `4-3.do` file, and exit Stata. Open a new session in your word processor, and open your log file into this session. Format it appropriately.

- How to compute measures of central tendency (averages), including the mean, median, and mode
- When to use the different measures of central tendency based on level of measurement, distribution characteristics, and your purposes
- How to describe the dispersion of a distribution by using
  - Statistics (standard deviations)
  - Tables (frequency distributions)
  - Graphs (pie charts, bar charts, histograms, and box plots)
- How to use Stata to give you these results for nominal-, ordinal-, and interval-level variables
- How to use Stata's Graph Editor

The graphs we have introduced in this chapter show just a few of the graph capabilities offered by Stata. We will cover a few more types of graphs in later chapters, but if you are interested in producing high-quality graphs, see *A Visual Guide to Stata Graphics, Third Edition* by Michael Mitchell (2012), which is available from Stata Press.

This is just the start of the useful output you can produce using Stata. Statistics books tend to get harder and harder as you move toward more complicated procedures. I cannot help that, but Stata is just the opposite. Managing data and doing graphs are the two hardest tasks for statistical programs because Stata is designed primarily to do statistical analysis. In the next chapter, we will examine how to use graphs and statistics when we are examining the relationship between two or more variables.

5.8 Exercises

1. Open `descriptive_gss.dta` and do a detailed summary of the variable `hrs1` (hours worked last week). Also create a histogram of the variable. Interpret the mean and median. Looking at the histogram, explain why the skewness value is close to zero. What does the value of kurtosis tell us? Looking at the histogram, explain why the kurtosis is a positive value.

2. Open `descriptive_gss.dta` and do a detailed summary of the variable `satjob7` (job satisfaction). Type the command `numlabel satjob7, add`, and then do a tabulation of `satjob7`. Interpret the mean and median values. Why would some researchers report the median? Why would other researchers report the mean?

3. Open `descriptive_gss.dta` and do a tabulation of `deckids` (who makes decisions about how to bring up children). Do this using the by/if/in tab to select by `sex`. Create and interpret a bar chart by using the `histogram` dialog box. Why would it make no sense to report the mean, median, or standard deviation for `deckids`? Use Stata's Graph Editor to make the bar chart look as nice as you can.

4. Open `descriptive_gss.dta` and do a tabulation of `strsswrk` (job is rarely stressful) and a detailed summary. Do this using the by/if/in tab to select by `sex`. Create and interpret a histogram, using the By tab to do this for males and females. In the Main tab, be sure to select the option *Data are discrete*. Carefully label your histogram to show value labels and the percentage in each response category. Each histogram should be similar to figure 5.9. Interpret the median and mean for men and women.

5. Open `descriptive_gss.dta` and do a tabulation of `trustpeo`, `wantbest`, `advantge`, and `goodlife`. Use the `tabstat` command to produce a table that summarizes descriptive statistics for this set of variables by gender. Include the median, mean, standard deviation, and count for each variable. Interpret the means by using the variable labels you get with the tabulation command.

6. Open `descriptive_gss.dta` and do a tabulation of `polviews`. Create a bar chart for this variable showing the percentage of people who are in each of the seven categories. Next create a chart that has labels at a 45-degree angle for each of the bars. Finally, change the chart by using the X axis tab's *Minor tick/label properties* and *Major tick/label properties* (using the *Custom* option) so that the histogram does not have a null category and all the categories are labeled. Compare this final figure with figure 5.9.

7. Open `descriptive_gss.dta`. In figure 5.13, we created a box plot for the hours women and men spend surfing the World Wide Web. First, create a similar box plot for `hrs1` (hours worked last week). Now add a second grouping variable, `marital`. Using this graph, give a detailed interpretation of how marital status and gender are related to hours a person works for pay.

8. Open `descriptive_gss.dta`. Execute the following commands for `educ` and compare the two sets of results. What problem are you illustrating here?

```
. summarize educ, detail
. sktest educ
. histogram educ
. preserve
. sample 10, count
. summarize educ, detail
. sktest educ
. histogram educ
. restore
```

9. Open `descriptive_gss.dta`. Construct a graph showing two histograms, one for women and one for men for the `educ` variable. Use the Graph Editor to improve the appearance of the graph by adding an overall title and making other changes you can think of. Is this graph helpful for comparing the educational achievement of women and men? How so?

10. Repeat all parts of exercise 9, but construct a box plot instead of histograms.

11. Repeat exercise 10 using the `tabstat` command, and ask for the mean, median, standard deviation, skewness, kurtosis, and interquartile range. Interpret each of these statistics to compare the women and men, and explain how the differences are reflected (or not) in the graphs done in exercises 9 and 10.

power to show that an important result is statistically significant. We only touched the surface of what Stata can do with power analysis, and I encourage you to check the Stata reference manuals for more ways to estimate power and sample-size requirements.

When we have statistical significance, we are confident that what we observed in our data represents a real effect that should not be attributed to chance. It is still essential to evaluate how substantively significant the result is. With a large sample, we may find a statistically significant difference of means when the actual difference is small and substantively insignificant. Finding a significant result begs the question of how important the result is. A statistically significant result may or may not be large enough to be important.

Finally, we learned briefly about nonparametric alternatives to z tests and t tests. These alternatives usually have less power but may be more easily justified in the assumptions they make.

Sometimes we have more than two groups, and the t test is no longer adequate. Chapter 9 discusses analysis of variance (ANOVA), which is an extension of the t test that allows us to work with more than two groups. Before we do analysis of variance, however, we will cover bivariate correlation and regression in chapter 8.

7.13 Exercises

When doing these exercises, you should create a do-file for each assignment; you might name the do-file for the first exercise `c7_1.do`. Put these do-files in the directory where you are keeping the Stata programs related to this book. Having these do-files will be useful in the future if you need to redo one of these examples or want to do a similar task. For example, you could use the do-file for the second exercise anytime you need to do randomization of participants in a study.

1. According to the registrar's office at your university, 52% of the students are women. You do a web-based survey of attitudes toward student fees supporting the sports program. You have 20 respondents, 14 of whom are men and 6 of whom are women. Is there a gender bias in your sample? To answer this, create a new dataset that has one variable, namely, `sex`. Enter a value of `1` for your first 14 observations (for your 14 males) and a value of `0` for your last 6 observations (for your 6 females). Your data will have one column and 20 rows. Then do a one-sample z test against the null hypothesis that $p = 0.52$. Explain your null hypothesis. Interpret your results.

2. You have 30 volunteers to participate in an exercise program. You want to randomly assign 15 of them to the control group and 15 to the treatment group. You list them by name—the order being arbitrary—and assign numbers from 1 to 30. What are the Stata commands you would use to do the random assignment (randomization without replacement)? Show how you would do this.

3. You learned ho
sample with rep
use the comma
were selected m

4. Using the appr
from a large cla
has a number f
953. List the nu

5. Open `nlsy97_c`
dren than do ot
under age 18)
test whether th
significant, how

6. Use the same
to check the di
non-Hispanics.
difference signifi
(Think about ti

7. You want to con
rights. From ea
50 and the stan
difference that
is pretty subjec
deviation differ
many cases do
have 99% powe

8. A friend believe
any circumstan
sion about whe
dataset (`gss20`
ence between w
(`abany`).

9. You are planning
the body mass i
of the interventi
past research, y
standard deviati
effect size is med

10. You are planning
weight loss prog

3. You learned how to do a random sample without replacement. To do a random sample with replacement, you use the `bsample` command. Repeat exercise 2, but use the command `bsample 15`. Then do a tabulation to see if any observations were selected more than once.

4. Using the approach you used in exercise 1, draw a random sample of 10 students from a large class of 200 students. You use the class roster in which each student has a number from 1 to 200. Show the commands you use, and set the seed at 953. List the numbers for the 10 students you select.

5. Open `nlsy97_chapter7.dta`. A friend says that Hispanic families have more children than do other ethnic groups. Use the variables `hh18_97` (number of children under age 18) and `ethnic97` (0 being non-Hispanic and 1 being Hispanic) to test whether this is true. Are the means different? If the result is statistically significant, how substantively significant is the difference?

6. Use the same variables as exercise 5. Use `summarize`, `detail`, and `tabulate` to check the distribution of `hh18_97`. Do this separately for Hispanics and for non-Hispanics. What are the medians of each group? Run a median test. Is the difference significant? How can you reconcile this with the medians you computed? (Think about ties and the distributions.)

7. You want to compare Democrats and Republicans on their mean scores on abortion rights. From earlier uses of the scale, you know that the mean is somewhere around 50 and the standard deviation is about 15. Select an alpha level and the minimum difference that you would find important. Justify your minimum difference (this is pretty subjective, but you might think in terms of a proportion of a standard deviation difference). How many cases do you need to have 80% power? How many cases do you need to have 90% power? How many cases do you need to have 99% power?

8. A friend believes that women are more likely to feel that abortion is okay under any circumstances than are men because women have more at stake in a decision about whether to have an abortion. Use the General Social Survey 2002 dataset (`gss2002_chapter7.dta`), and test whether there is a significant difference between women and men (`sex`) on whether abortion is acceptable in any case (`abany`).

9. You are planning to evaluate the effectiveness of Overeaters Anonymous to reduce the body mass index (BMI) of participants. You will weigh each person at the start of the intervention and then again after they have participated for 5 weeks. From past research, you expect the average BMI to be 30 prior to the intervention. The standard deviation is about 4. You want to be able to detect a difference if the effect size is medium. How many people do you need to have in your intervention?

10. You are planning to compare how satisfied women and men are with a particular weight loss program. You have 20 items measuring satisfaction and each item is

scored from 1 for very dissatisfied to 5 for very satisfied. You are interested in being able to detect a small effect size with alpha = 0.05 and power = 0.90. How big of a sample will you need?

11. A researcher is planning a study of the effectiveness of a new method of teaching a course on statistics. Students in one class get a traditional course, where all the work is done without computers. Students in a different course (new method) do half the work by hand and half using Stata. If you want a moderate effect size, alpha of 0.05, and a power of 0.80, how many students are needed in each class? The final examination is used to evaluate performance, and the mean score has been around 80 with a standard deviation of 10.

12. You are planning an intervention to reduce test-taking anxiety. You have a 100-point scale to measure anxiety and a pretest you did without any intervention had a mean of 70 and a standard deviation of 15. You want to be able to show that a true difference of means between your intervention group and your control group should lower the anxiety by one third of a standard deviation or 5 points. You want a power of 0.80.

    a. How many students do you need in each group (control group and intervention group) using a two-tail test?

    b. How many using a one-tail test?

13. Suppose you are doing question 12, it is much more expensive to get measures for the intervention group. Because of the cost of running the intervention, let's say it is two times as expensive to obtain students' scores for the intervention group. How many students do you need in the intervention group? The control group?

    a. For a two-tail test?

    b. For a one-tail test?

8 Bivari

8.1 Introduc

Bivariate correl
variables. They
sured on the int
and then regress
of these, but wh
first. In this cha

1. Construct
between t
2. Superimp
form of th
3. Estimate
4. Estimate
of the rela
5. Estimate
data.

Spearman's rho is a correlation of ranked data. To save the time of converting the variables to ranks and then doing a Pearson's correlation, Stata has a special command: `spearman`. Here we could run the command `spearman age liberal` to yield $\rho_s = -0.82$.

8.8 Summary

Scattergrams, correlations, and regressions are great ways to evaluate the relationship between two variables.

- The scattergram helps us visualize the relationship between two variables and is usually most helpful when there are relatively few observations.
- The correlation is a measure of the strength of the relationship between two variables. Here it is important to recognize that it measures the strength of a particular form of the relationship. The other examples have used a linear regression line as the form of the relationship. Although it is not covered here, it is possible for regression to have other forms of the relationship.
- The regression analysis tells us the form of the relationship. Using this line, we can estimate how much the dependent variable changes for each unit change in the independent variable.
- The standardized regression coefficient, β, measures the strength of a relationship and is identical to the correlation for bivariate regression.
- You have learned how to compute Spearman's rho for rank-order data, and you now understand its relationship to Pearson's r.

8.9 Exercises

1. Use `gss2006_chapter8.dta`. Imagine that you heard somebody say that there was no reason to provide more educational opportunities for women because so many of them just stay at home anyway. You have a variable measuring education, `educ`, and a variable measuring hours worked in the last week, `hrs1`. Do a correlation and regression of hours worked in the last week on years of education. Then do this separately for women and for men. Interpret the correlation and the slope for the overall sample and then for women and for men separately. Is there an element of truth to what you heard?
2. Use `gss2006_chapter8.dta`. What is the relationship between the hours a person works and the hours his or her spouse works? Do this for women and for men separately. Compute the correlation, the regression results, and the scattergrams. Interpret each of these. Next test if the correlation is statistically significant and interpret the results.

3. Use gss2006_cl
observations. T
figures. Set you
4. Use gss2006_c
and health by
results slightly
get the signific
repeat the pwc
casewise/listwi
5. Use gss2002_c
Run the codeb
with being unh
mean more of
that reverse th
and a score of
very dissatisfie
happiness on fa
variables? Wri
6. Use spearman
and liberal,
the variables r
the Spearman'
should involve
7. Use depressi
. use http
From this hyp
pression (varia
aid in reading
to remember t
need to remem
would be to l
more or less d
a. Type sca
b. Interpret
8. You suspect th
those over abo
a. Type bir
b. Interpret
cise 7.

3. Use gss2006_chapter8.dta. Repeat figure 8.2 using your own subsample of 250 observations. Then repeat the figure using a jitter(3) option. Compare the two figures. Set your seed at 111.

4. Use gss2006_chapter8.dta. Compute the correlations between happy, hapmar, and health by using correlate and then again by using pwcorr. Why are the results slightly different? Then estimate the correlations by using pwcorr, and get the significance level and the number of observations for each case. Finally, repeat the pwcorr command so that all the Ns are the same (that is, there is casewise/listwise deletion).

5. Use gss2002_chapter8.dta. There are two variables called happy7 and satfam7. Run the codebook command on these variables. Notice how the higher score goes with being unhappy or being dissatisfied. You always want the higher score to mean more of a variable, so generate new variables (happynew and satfamnew) that reverse these codes so that a score of 1 on happynew means very unhappy and a score of 7 means very happy. Similarly, a score of 1 on satfamnew means very dissatisfied and a score of 7 means very satisfied. Now do a regression of happiness on family satisfaction with the new variables. How correlated are these variables? Write the regression equation. Interpret the constant and the slope.

6. Use spearman.dta. Plot a scattergram, including the regression line for age and liberal, treating liberal as the dependent variable. Repeat this using the variables rankage and ranklib. Interpret this scattergram to explain why the Spearman's rho is smaller than the Pearson's correlation. Your explanation should involve the idea of one observation being an outlier.

7. Use depression.dta from the Stata Press website; that is, type

```
. use http://www.stata-press.com/data/r13/depression.dta
```

From this hypothetical data, you are interested in the relationship between depression (variable TotalScore) and age. (This dataset uses capitalization as an aid in reading the total score variable. This is rarely a good idea because it is hard to remember these conventions, and if you always use all lowercase, you do not need to remember when and how you used capitalization. Perhaps better options would be to label the variable totalscore or total_score.) Are older people more or less depressed?

   a. Type scatter and binscatter to describe the relationship.
   b. Interpret these results. Why is the binscatter graph easier to interpret?

8. You suspect that the relationship may be nonlinear with a gradual increase among those over about 50 years of age.

   a. Type binscatter to fit a curve.
   b. Interpret these results and compare them with the graphs created in exercise 7.