A Gentle Introduction to Stata

Revised Sixth Edition

ALAN C. ACOCK
Oregon State University

A Stata Press Publication
StataCorp LLC
College Station, Texas

First edition 2006
Second edition 2008
Third edition 2010
Revised third edition 2012
Fourth edition 2014
Fifth edition 2016
Sixth edition 2018
Revised sixth edition 2023

Published by Stata Press, 4905 Lakeway Drive, College Station, Texas 77845
Typeset in LaTeX 2ε
Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

Print ISBN-10: 1-59718-367-9
Print ISBN-13: 978-1-59718-367-3
ePub ISBN-10: 1-59718-368-7
ePub ISBN-13: 978-1-59718-368-0

Library of Congress Control Number: 2022949742

Contents

(Pages omitted)

Preface

This book was written with a particular reader in mind. This reader is learning social statistics and needs to learn Stata but has no prior experience with other statistical software packages. When I learned Stata, I found there were no books written explicitly for this type of reader. There are certainly excellent books on Stata, but they assume extensive prior experience with other packages, such as SAS or IBM SPSS Statistics; they also assume a fairly advanced working knowledge of statistics. These books moved quickly to advanced topics and left my intended reader in the dust. Readers who have more background in statistical software and statistics will be able to read chapters quickly and even skip sections. My goal is to move the true beginner to a level of competence using Stata.

With this target reader in mind, I make far more use of the menus and dialog boxes in Stata's interface than do other books about Stata. Advanced users may not see the value in using the interface, and the more people learn about Stata, the less they will rely on the interface. Also, even when you are using the interface, it is still important to save a record of the sequence of commands you run. Although I rely on the commands much more than the dialog boxes in the interface in my own work, I still find value in the interface. The dialog boxes in the interface include many options that I might not have known or might have forgotten.

To illustrate the interface as well as graphics, I have included more than 100 figures, many of which show dialog boxes. I present many tables and extensive Stata "results" as they appear on the screen. I interpret these results substantively in the belief that beginning Stata users need to learn more than just how to produce the results—users also need to be able to interpret them.

I have tried to use real data. There are a few examples where it is much easier to illustrate a point with hypothetical data, but for the most part, I use data that are in the public domain. For example, I use the General Social Surveys for 2002 and 2006 (Smith et al. 2015) in many chapters, as well as the National Survey of Youth, 1997 (Bureau of Labor Statistics, U.S. Department of Labor 2015). I have simplified the files by dropping many of the variables in the original datasets, but I have kept all the observations. I have tried to use examples from several social-science fields, and I have included a few extra variables in several datasets so that instructors, as well as readers, can make additional examples and exercises that are tailored to their disciplines. People who are used to working with statistics books that have contrived data with just a few observations, presumably so work can be done by hand, may be surprised to see more than 1,000 observations in this book's datasets. Working with these files provides better experience for other real-world data analysis. If you have your own data and the dataset has a variety of variables, you may want to use your data instead of the data provided with this book.

The exercises use the same datasets as the rest of the book. Several of the exercises require some data management prior to fitting a model because I believe that learning data management requires practice and cannot be isolated in a single chapter or single set of exercises.

This book takes the student through much of what is done in introductory and intermediate statistics courses. It covers descriptive statistics, charts, graphs, tests of significance for simple tables, tests for one and two variables, correlation and regression, analysis of variance, multiple regression, logistic regression, reliability, factor analysis, and path analysis. There are chapters on constructing scales to measure variables and on using multiple imputation for working with missing values.

By combining this coverage with an introduction to creating and managing a dataset, the book will prepare students to go even further on their own or with additional resources. More advanced statistical analysis using Stata is often even simpler from a programming point of view than what we will cover here. If an intermediate course goes beyond what we do with logistic regression to multinomial logistic regression, for example, the programming is simple enough. The `logit` command can simply be replaced with the `mlogit` command. The added complexity of these advanced statistics is the statistics themselves and not the Stata commands that implement them. Therefore, although more advanced statistics are not included in this book, the reader who learns these statistics will be more than able to learn the corresponding Stata commands from the Stata documentation and help system.

Two new chapters were added in the fifth edition. Chapter 15 introduces multilevel analysis for longitudinal and panel models. This chapter touches only on the capabilities of multilevel analysis using Stata, but it provides a starting point. Chapter 16 covers item response theory (IRT), which was added in Stata 14. IRT offers an advanced approach to measurement that has many advantages over the use of summated scales, which are discussed in chapters 3 and 12.

In the sixth edition, an extension to the discussion of using the `sem` command (structural equation modeling) to fit regression models to account for missing values was added. This also includes using auxiliary variables to help justify the missing-at-random assumption.

The revised sixth edition has been updated to Stata 17, including Stata command syntax, updated output, and all screenshots of Stata's interface. Discussion and examples include new features added to Stata since Stata 15, including the new `table` command, the `collect` suite of commands for creating, customizing, and exporting tables, and the use of transparencies in graphs.

I assume that the reader is running Stata 17, or a later version, on a Windows-based PC. Stata works equally as well on Mac and on Unix systems. Readers who are running Stata on one of those systems will have to make a few minor adjustments to some of the examples in this book. I will note some Mac-specific differences when they are important.

Corvallis, OR
November 2022

Alan C. Acock

(Pages omitted)

6 Statistics and graphs for two categorical variables

6.1 Relationship between categorical variables

Chapter 5 focused on describing single variables. Even there, it was impossible to resist some comparisons, and we ended by examining the relationship between gender and hours per week spent using the web. Some research simply requires a description of the variables, one at a time. You do a survey for your agency and make up a table with the means and standard deviations for all the quantitative variables. You might include frequency distributions and bar charts for each key categorical variable. This information is sometimes the extent of statistical research your reader will want. However, the more you work on your survey, the more you will start wondering about possible relationships.

- Do women who are drug dependent use different drugs from those used by drug-dependent men?

- Are women more likely to be liberal than men?
- Is there a relationship between religiosity and support for increased spending on public health?

You know you are "getting it" as a researcher when it is hard for you to look at a set of questions without wondering about possible relationships. Understanding these relationships is often crucial to making policy decisions. If 70% of the nonmanagement employees at a retail chain are women, but only 20% of the management employees are women, there is a relationship between gender and management status that disadvantages women.

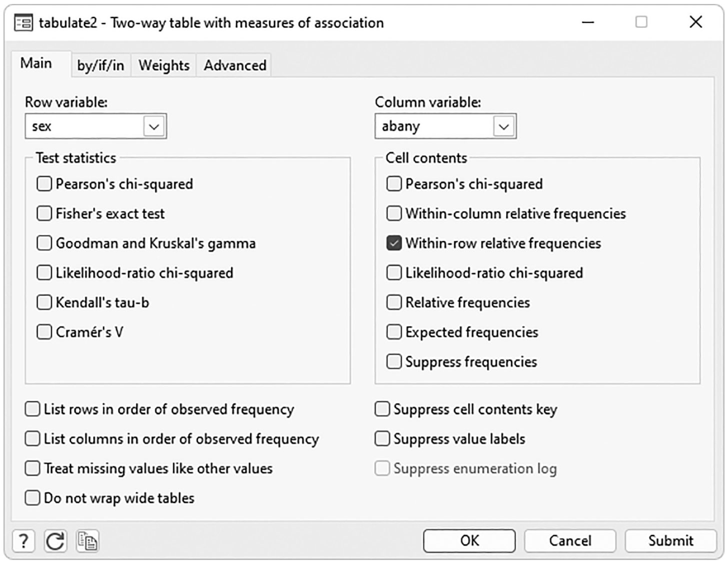
In this chapter, you will learn how to describe relationships between categorical variables. How do you define these relationships? What are some pitfalls that lead to misinterpretations? In this chapter, the statistical sophistication you will need increases, but there is one guiding principle to remember: the best statistics are the simplest statistics you can use—as long as they are not too simple to reflect the inherent complexity of what you are describing.

6.2 Cross-tabulation

Cross-tabulation is a technical term for a table that has rows representing one categorical variable and columns representing another. These tables are sometimes called contingency tables because the category a person is in on one of the variables is contingent on the category the person is in on the other variable. For example, the category people are in on whether they support a particular public health care reform may be contingent on their gender. If you have one variable that depends on the other, you usually put the dependent variable as the column variable and the independent variable as the row variable. This layout is certainly not necessary, and several statistics books do just the opposite. That is, they put the dependent variable as the row variable and the independent variable as the column variable.

Let's start with a basic cross-tabulation of whether a person says abortion is okay for any reason and their gender. Say you decide that whether a person accepts abortion for any reason is more likely if the person is a woman because a woman has more at stake when she is pregnant than does her partner. Therefore, whether a person accepts abortion will be the dependent variable, and gender will be the independent variable.

We will use `gss2006_chapter6.dta`, which contains selected variables from the 2006 General Social Survey (Smith et al. 2015), and we will use the cross-tabulation command, `tabulate`, with two categorical variables, `sex` and `abany`. To open the dialog box for `tabulate`, select Statistics ▷ Summaries, tables, and tests ▷ Frequency tables ▷ Two-way table with measures of association. This dialog box is shown in figure 6.1.



Row variable: sex

Column variable: abany

Test statistics
- □ Pearson's chi-squared
- □ Fisher's exact test
- □ Goodman and Kruskal's gamma
- □ Likelihood-ratio chi-squared
- □ Kendall's tau-b
- □ Cramér's V

Cell contents
- □ Pearson's chi-squared
- □ Within-column relative frequencies
- ☑ Within-row relative frequencies
- □ Likelihood-ratio chi-squared
- □ Relative frequencies
- □ Expected frequencies
- □ Suppress frequencies

- □ List rows in order of observed frequency
- □ List columns in order of observed frequency
- □ Treat missing values like other values
- □ Do not wrap wide tables
- □ Suppress cell contents key
- □ Suppress value labels
- □ Suppress enumeration log

Figure 6.1. The Main tab for creating a cross-tabulation

If we had three variables and wanted to see all possible two-by-two tables (vara with varb, vara with varc, and varb with varc), we could have selected instead Statistics ▷ Summaries, tables, and tests ▷ Frequency tables ▷ All possible two-way tables. With our current data, we continue with the dialog box in figure 6.1 and select sex, the independent variable, as the *Row variable* and abany, the dependent variable, as the *Column variable*. We are assuming that abany is the variable that depends on sex. Also check the box on the right side under *Cell contents* for the *Within-row relative frequencies* option. This option tells Stata to compute the percentages so that each row adds up to 100%. Here are the resulting command and results:

```
. tabulate sex abany, row

  +----------------+
  | Key            |
  |----------------|
  |   frequency    |
  | row percentage |
  +----------------+

           |  ABORTION IF WOMAN
           | WANTS FOR ANY REASON
    GENDER |       YES         NO |     Total
-----------+----------------------+----------
      MALE |       350        478 |       828 
           |     42.27      57.73 |    100.00 
-----------+----------------------+----------
    FEMALE |       434        677 |     1,111 
           |     39.06      60.94 |    100.00 
-----------+----------------------+----------
     Total |       784      1,155 |     1,939 
           |     40.43      59.57 |    100.00 
```

Independent and dependent variables

Many beginning researchers get these terms confused. The easiest way to remember which is which is that the dependent variable "depends" on the independent variable. In this example, whether a person accepts abortion for any reason depends on whether the person is a man or a woman. By contrast, it would make no sense to say that whether a person is a man or a woman depends on whether they accept abortion for any reason.

Many researchers call the dependent variable an "outcome" and the independent variable the "predictor". In this example, `sex` is the predictor because it predicts the outcome, `abany`.

It may be hard or impossible to always know that one variable is independent and the other variable is dependent. This is because the variables may influence each other. Imagine that one variable is your belief that eating meat is a health risk (there could be four categories: strongly agree, agree, disagree, or strongly disagree). Then imagine that the other variable is whether you eat meat or not. It might seem easy to say that your belief is the independent variable and your behavior is the dependent variable. That is, whether you eat meat or not depends on whether you believe doing so is a health risk. But try to think of the influence going the other way. A person may stop eating meat because of a commitment to animal rights issues. Not eating meat over several years leads to look for other justifications, and they develop a belief that eating meat is a health risk. For them, the belief depends on their prior behavior.

There is no easy solution when we have trouble deciding which is the independent variable. Sometimes, we can depend on time ordering. Whichever came first is the independent variable. Other times, we are simply forced to say that the variables are associated without identifying which one is the independent variable and which is the dependent variable.

The independent variable `sex` forms the rows with labels of `MALE` and `FEMALE`. The dependent variable `abany`, accepting abortion under any circumstance, appears as the columns labeled `YES` and `NO`. The column on the far right gives us the total for each row. Notice that there are 828 males, 350 of whom find abortion for any reason to be acceptable, compared with 1,111 females, 434 of whom say abortion is acceptable for any reason. These frequencies are the top number in each cell of the table.

The frequencies at the top of each cell are hard to interpret because each row and each column has a different number of observations. One way to interpret a table is to use the percentage, which takes into account the number of observations within each category of the independent variable (predictor). The percentages appear just below the frequencies in each cell. Notice that the percentages add up to 100% for each row. Overall, 40.43%

of the people said “yes”, abortion is acceptable for any reason, and 59.57% said “no”. However, men were relatively more likely (42.27%) than women (39.06%) to report that abortion is okay, regardless of the reason. We get these percentages because we told Stata to give us the *Within-row relative frequencies*, which in the command is the `row` option.

Thus men are more likely to report accepting abortion under any circumstance. We compute percentages on the rows of the independent variable and make comparisons up and down the columns of the dependent variable. Thus we say that 42.27% of the men compared with 39.06% of the women accept abortion under any circumstance. This is a small difference, but interestingly, it is in the opposite direction from what we expected.

6.3 Chi-squared test

The difference between women and men seems small, but could we have obtained this difference by chance? Or is the difference statistically significant? Remember, when you have a large sample like this one, a difference may be statistically significant even if it is small.

If we had just a handful of women and men in our sample, there would be a good chance of observing this much difference just by chance. With such a large sample, even a small difference like this might be statistically significant. We use a chi-squared (χ^2) statistic to test the likelihood that our results occurred by chance. If it is extremely unlikely to get this much difference between men and women in a sample of this size by chance, you can be confident that there was a real difference between women and men, but you still need to look at the percentages to decide whether the statistically significant difference is substantial enough to be important.

The chi-squared test compares the frequency in each cell with what you would expect the frequency to be if there were no relationship. The expected frequency for a cell depends on how many people are in the row and how many are in the column. For example, if we asked a small high school group if they accept abortion for any reason, we might have only 10 males and 10 females. Then we would expect far fewer people in each cell than in this example, where we have 828 men and 1,111 women.

In the cross-tabulation, there were many options on the dialog box (see figure 6.1). To obtain the chi-squared statistic, check the box on the left side for *Pearson’s chi-squared*. Also check the box for *Expected frequencies* that appears in the right column on the dialog box. The resulting table has three numbers in each cell. The first number in each cell is the frequency, the second number is the expected frequency if there were no relationship, and the third number is the percentage of the row total. We would not usually ask for the expected frequency, but now you know it is one of Stata’s capabilities. The resulting command now has three options: `chi2 expected row`. Here is the command and the output it produces:

```
. tabulate sex abany, chi2 expected row

+--------------------+
| Key                |
|--------------------|
|     frequency      |
| expected frequency |
|   row percentage   |
+--------------------+

           |    ABORTION IF WOMAN
           |  WANTS FOR ANY REASON
    GENDER |       YES         NO |     Total
-----------+----------------------+----------
      MALE |       350        478 |       828
           |     334.8      493.2 |     828.0
           |     42.27      57.73 |    100.00
-----------+----------------------+----------
    FEMALE |       434        677 |     1,111
           |     449.2      661.8 |   1,111.0
           |     39.06      60.94 |    100.00
-----------+----------------------+----------
     Total |       784      1,155 |     1,939
           |     784.0    1,155.0 |   1,939.0
           |     40.43      59.57 |    100.00

          Pearson chi2(1) =   2.0254   Pr = 0.155
```

In the top left cell of the table, we can see that we have 350 men who accept abortion for any reason, but we would expect to have only 334.8 men here by chance. By contrast, we have 434 women who accept abortion for any reason, but we would expect to have 449.2. Thus we have $350 - 334.8 = 15.2$ more men accepting abortion than we would expect by chance and $434 - 449.2 = -15.2$ fewer women than we would expect. Stata uses a function of this information to compute chi-squared.

At the bottom of the table, Stata reports `Pearson chi2(1) = 2.0254` and `Pr = 0.155`, which would be written as $\chi^2(1, N = 1939) = 2.0254$; p not significant. Here we have one degree of freedom. The sample size of $N = 1939$ appears in the lower right part of the table. We usually round the chi-squared value to two decimal places, so 2.0254 becomes 2.03. Stata reports an estimate of the probability to three decimal places. We can report this, or we can use a convention found in most statistics books of reporting the probability as less than 0.05, less than 0.01, or less than 0.001. Because $p = 0.155$ is greater than 0.05, we say p not significant. What would happen if the probability were $p = 0.0004$? Stata would round this to $p = 0.000$. We would not report $p = 0.000$ but instead would report $p < 0.001$.

To summarize what we have done in this section, we can say that men are more likely to report accepting abortion for any reason than are women. In the sample of 1,939 people, 42.3% of the men say that they accept abortion for any reason compared with just 39.1% of the women. This relationship between gender and acceptance of abortion is not statistically significant.

6.3.1 Degrees of freedom

Because I assume that you have a statistics book explaining the necessary formulas, I have not gone into detail. Stata will compute the chi-squared, the number of degrees of freedom, and the probability of getting your observed result by chance.

You can determine the number of degrees of freedom yourself. The degrees of freedom refers to how many pieces of independent information you have. In a 2×2 table, like the one we have been analyzing, the value of any given cell can be any number between 0 and the smaller of the number of observations in the row and the number of observations in the column. For example, the upper left cell (350) could be anything between 0 and 784. Let's use the observed value of 350 for the upper left cell. Now how many other cells are free to vary? By subtraction, you can determine that 434 people must be in the female/yes cell because $784 - 350 = 434$. Similarly, 478 observations must be in the male/no cell ($828 - 350 = 478$), and 677 observations must be in the female/no cell ($1111 - 434 = 677$). Thus with four cells, only one of these is free, and we can say that the table has 1 degree of freedom. We can generalize this to larger tables where degrees of freedom $= (R - 1)(C - 1)$, where R is the number of rows and C is the number of columns. If we had a 3×3 table instead of a 2×2 table, we would have $(3-1)(3-1) = 4$ degrees of freedom.

6.3.2 Probability tables

Many experienced Stata users have made their own commands that might be helpful to you. Philip Ender made a series of commands that display probability tables for various tests. The `search` command finds user-contributed ado-files and lets you install them on your machine. Typing the command `search chitable`[1] produces the results shown in figure 6.2.

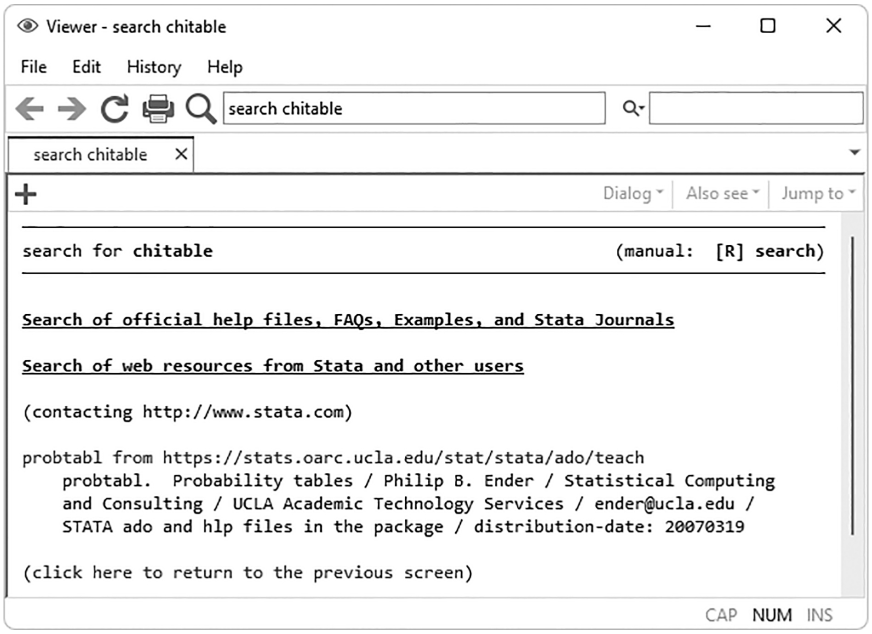
[1] You cannot install this community-contributed command by typing `ssc install chitable`, because the `ssc` command is limited to installing and uninstalling community-contributed commands that are located within the Statistical Software Components (SSC) Archive or often called the Boston College Archive at http://www.repec.org. Although the SSC Archive is by far the dominant depository of community-contributed Stata commands, it does not include all community-contributed commands.

Updating community-contributed commands

Stata will prompt you when you need to update Stata files, and you should do this when prompted. Community-contributed commands such as `fre` or `chitable` may be revised by their authors. You will not be prompted to update these, so you need to remember to do this, say, once a month. The command you type in the Command window is

```
. ado update, update
```

This command will list any community-contributed commands you have installed and then update any of them that have changed.



```
search for chitable                                   (manual:  [R] search)

Search of official help files, FAQs, Examples, and Stata Journals

Search of web resources from Stata and other users

(contacting http://www.stata.com)

probtabl from https://stats.oarc.ucla.edu/stat/stata/ado/teach
    probtabl.  Probability tables / Philip B. Ender / Statistical Computing
    and Consulting / UCLA Academic Technology Services / ender@ucla.edu /
    STATA ado and hlp files in the package / distribution-date: 20070319

(click here to return to the previous screen)
```

Figure 6.2. Results of `search chitable`

From here, click on the blue web link

```
probtabl from https://stats.oarc.ucla.edu/stat/stata/ado/teach
```

This link takes you to another screen where you can click on the blue link labeled `click here to install`. Once you have done this, anytime you want to see a chi-squared table, you merely type the command `chitable`. This installation also gives you other probability tables that we will use elsewhere in this book, including *t*-test tables (`ttable`) and *F*-test tables (`ftable`). Simply entering `chitable` is a lot more convenient than having to look up a probability in a textbook. Try it now.

```
. chitable

         Critical Values of Chi-square
 df     .50     .25     .10     .05    .025      .01     .001
  1    0.45    1.32    2.71    3.84    5.02     6.63    10.83
  2    1.39    2.77    4.61    5.99    7.38     9.21    13.82
  3    2.37    4.11    6.25    7.81    9.35    11.34    16.27
  4    3.36    5.39    7.78    9.49   11.14    13.28    18.47
  5    4.35    6.63    9.24   11.07   12.83    15.09    20.52
  6    5.35    7.84   10.64   12.59   14.45    16.81    22.46
  7    6.35    9.04   12.02   14.07   16.01    18.48    24.32
  8    7.34   10.22   13.36   15.51   17.53    20.09    26.12
  9    8.34   11.39   14.68   16.92   19.02    21.67    27.88
 10    9.34   12.55   15.99   18.31   20.48    23.21    29.59
 11   10.34   13.70   17.28   19.68   21.92    24.72    31.26
 12   11.34   14.85   18.55   21.03   23.34    26.22    32.91
 13   12.34   15.98   19.81   22.36   24.74    27.69    34.53
 14   13.34   17.12   21.06   23.68   26.12    29.14    36.12
 15   14.34   18.25   22.31   25.00   27.49    30.58    37.70
 16   15.34   19.37   23.54   26.30   28.85    32.00    39.25
 17   16.34   20.49   24.77   27.59   30.19    33.41    40.79
 18   17.34   21.60   25.99   28.87   31.53    34.81    42.31
 19   18.34   22.72   27.20   30.14   32.85    36.19    43.82
 20   19.34   23.83   28.41   31.41   34.17    37.57    45.31
 21   20.34   24.93   29.62   32.67   35.48    38.93    46.80
 22   21.34   26.04   30.81   33.92   36.78    40.29    48.27
 23   22.34   27.14   32.01   35.17   38.08    41.64    49.73
 24   23.34   28.24   33.20   36.42   39.36    42.98    51.18
 25   24.34   29.34   34.38   37.65   40.65    44.31    52.62
 26   25.34   30.43   35.56   38.89   41.92    45.64    54.05
 27   26.34   31.53   36.74   40.11   43.19    46.96    55.48
 28   27.34   32.62   37.92   41.34   44.46    48.28    56.89
 29   28.34   33.71   39.09   42.56   45.72    49.59    58.30
 30   29.34   34.80   40.26   43.77   46.98    50.89    59.70
 35   34.34   40.22   46.06   49.80   53.20    57.34    66.62
 40   39.34   45.62   51.81   55.76   59.34    63.69    73.40
 45   44.34   50.98   57.51   61.66   65.41    69.96    80.08
 50   49.33   56.33   63.17   67.50   71.42    76.15    86.66
 55   54.33   61.66   68.80   73.31   77.38    82.29    93.17
 60   59.33   66.98   74.40   79.08   83.30    88.38    99.61
 65   64.33   72.28   79.97   84.82   89.18    94.42   105.99
 70   69.33   77.58   85.53   90.53   95.02   100.43   112.32
 75   74.33   82.86   91.06   96.22  100.84   106.39   118.60
  (output omitted)
100   99.33  109.14  118.50  124.34  129.56   135.81   149.45
```

In a `chitable`, the first row shows the significance levels. The first column shows the degrees of freedom. You can see that with 1 degree of freedom, you need a chi-squared value of 3.84 to be significant at the 0.05 level. If you had 4 degrees of freedom, you would need a chi-squared of 9.49 to be significant at the 0.05 level. You might try the other commands for the other tables. If you have ever had to search for one of these tables in a textbook, you will appreciate the convenience of these commands.

Reporting chi-squared results

How do we report the significance level of chi-squared? How do we report that chi-squared varies somewhat from one field to another? A safe way to report the significance is as $p < 0.05$, $p < 0.01$, or $p < 0.001$. Suppose that we had a chi-squared of 10.05 with 1 degree of freedom and $p = 0.002$. This is less than 0.01 but not less than 0.001. We would say that $\chi^2(1) = 10.05$, $p < 0.01$. Notice that we put the degrees of freedom in parentheses. Some disciplines would also like you to list the sample size, for example, $\chi^2(1, N = 435) = 10.05$, $p < 0.01$. Still other disciplines would rather you report the probability value, for example, $p = 0.002$.

Reporting the probability value has the advantage of providing more-detailed information. For example, one result might have a $p = 0.052$ and another might have a $p = 0.451$. The first of these is almost statistically significant and is an unlikely result if the null hypothesis of no relationship is correct. The second of these is just about what you would expect to get by chance flipping a coin. Saying that both of these have the same classification of p not significant conceals the clear difference between these two results.

6.4 Percentages and measures of association

We have already discussed the use of percentages. These are often the easiest and best way to describe a relationship between two variables. In our last example, the percentage of men who said abortion was okay for any reason was slightly greater than, although not statistically significantly greater than, the percentage of women who said abortion was okay for any reason. Percentages often tell us what we want to know. There are other ways of describing an association called measures of association; they try to summarize the strength of a relationship with a number.

The value of chi-squared depends on two things. First, the stronger the association between the variables, the bigger chi-squared will be. Second, because we have more confidence in our results when we have larger samples, then the more cases we have, the bigger chi-squared will be. In fact, for a given relationship expressed in percentages, chi-squared is a function of sample size. If you had the same relationship as in our example but, instead of having 1,939 observations, you had 19,390 (10 times as many), then chi-squared would be 20.254, also 10 times as big. There would still be 1 degree of freedom, but here the results would be statistically significant, $p < 0.001$. With large samples, researchers sometimes misinterpret a statistically significant chi-squared value as indicating a strong relationship. With a large sample, even a weak relationship can be statistically significant. This makes sense because with a large sample we have the power to detect even small effects.

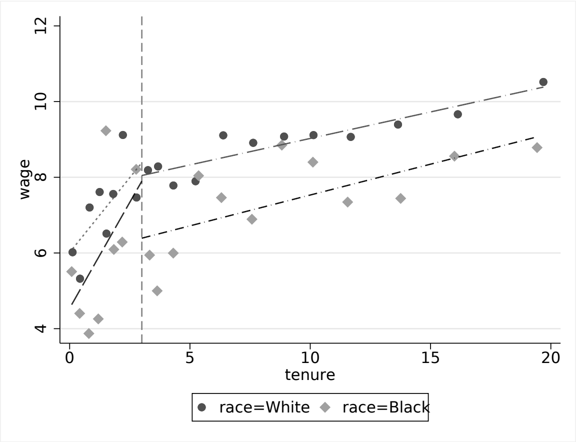


Figure 8.8. Relationship between wages and tenure with a discontinuity in the relationship at 3 years; whites shown with solid lines and blacks shown with dashed lines

8.5 Correlation

Your statistics textbook gives you the formulas for computing correlation, and if you have done a few of these by hand, you will love the ease of using Stata. We will not worry about the formulas. Correlation measures how close the observations are to the regression line. We need to be cautious in interpreting a correlation coefficient. Correlation does not tell us how steep the relationship is; that measure comes from the regression. You may have a steep relationship or an almost flat relationship, and both relationships could have the same correlation. Suppose that the correlation between education and income is $r = 0.3$ for women and $r = 0.5$ for men. Does this mean that education has a bigger payoff for men than it does for women? We really cannot know the answer from the correlation. The correlation tells us almost nothing about the form of the relationship. The fact that the r is larger for men than it is for women is not evidence that men get a bigger payoff from an additional year of education. Only the regression line will tell us that. The $r = 0.5$ for men means that the observations for men are closer to the regression line than are the observations for women ($r = 0.3$) and that the income of men is more predictable than that of women. Correlation also tells us whether the regression line goes up (r will be positive) or down (r will be negative). Strictly speaking, correlation measures the strength of the relationship only for how close the dots are to the regression line.

Bivariate correlation is used by social scientists in many ways. Sometimes we will be interested simply in the correlation between two variables. We might be interested in the relationship between calorie consumption per day and weight loss. If you discover that $r = -0.5$, this would indicate a fairly strong relationship. Generally, a correlation of $|r| = 0.1$ is a weak relationship, $|r| = 0.3$ is a moderate relationship, and $|r| = 0.5$ is a strong relationship. An r of -0.3 and an r of 0.3 are equally strong. The negative correlation means that as X goes up, Y goes down; the positive correlation means that as X goes up, Y goes up.

We might be interested in the relationship between several variables. We could compare three relationships between 1) weight loss and calorie consumption, 2) weight loss and the time spent in daily exercise, and 3) weight loss and the number of days per week a person exercises. We could use the three correlations to see which predictor is more correlated with weight loss.

Suppose that we wanted to create a scale to measure a variable, such as political conservatism. We would use several specific questions and combine them to get a score on the scale. We can compute a correlation matrix of the individual items. All of them should be at least moderately correlated with each other because they were selected to measure the same concept.

When we estimate a correlation, we also need to report its statistical significance level. The test of statistical significance of a correlation depends on the size or substantive significance of a correlation in the sample and depends on the size of the sample. An $r = 0.5$ might be observed in a very small sample, just by chance, even though there were no correlations in the population. On the other hand, an $r = 0.1$, although a weak substantive relationship, might be statistically significant if we had a huge sample.

Statistical and substantive significance

It is easy to confuse statistical significance and substantive significance. Usually, we want to find a correlation that is substantively significant (r is moderate or strong in our sample) and statistically significant (the population correlation is almost certainly not 0). With a very large sample, we can find statistical significance even when $r = 0.1$ or less. What is important about this is that we are confident that the population correlation is not 0 and that it is very small. Some researchers mistakenly assume that a statistically significant correlation automatically means that it is important when it may mean just the opposite—we are confident it is not very important.

With a very small sample, we can find a substantively significant $r = 0.5$ or more that is not statistically significant. Even though we observe a strong relationship in our small sample, we are not justified in generalizing this finding to the population. In fact, we must acknowledge that the correlation in the population might even be 0.

- Substantive significance is based on the size of the correlation.
- Statistical significance is based on the probability that you could get the observed correlation by chance if the population correlation is 0.

Now let's look at an example that we downloaded from the UCLA Stata Portal. As mentioned at the beginning of the book, this portal is an exceptional source of tutorials, including movies on how to use Stata. The data used are from a study called High School and Beyond. Here you will download a part of this dataset used for illustrating how to use Stata to estimate correlations. Go to your command line, and enter the command

```
. use https://stats.oarc.ucla.edu/stat/data/hsb2, clear
```

You will get a message back that you have downloaded 200 cases, and your listing of variables will show the subset of the High School and Beyond dataset. If your computer is not connected to the Internet, you should use one that is connected, download this file, save it to a flash disk, and then transfer it to your computer. This dataset is also available from this book's webpage.

Say that we are interested in the bivariate correlations between `read`, `write`, `math`, and `science` skills for these 200 students. We are also interested in the bivariate relationships between each of these skills and each students' socioeconomic status and between each of these skills and each students' gender. We believe that socioeconomic status is more related to these skills than gender is.

It is reasonable to treat the skills as continuous variables measured at close to the interval level, and some statistics books say that interval-level measurement is a critical assumption. However, it is problematic to treat socioeconomic status and gender as continuous variables. If we run a tabulation of socioeconomic status and gender, `tab1 ses female`, we will see the problem. Socioeconomic status has just three levels (low, middle, and high) and gender has just two levels (male and female). This dataset has all values labeled, so the default tabulation does not show the numbers assigned to these codes. We can run `codebook ses female` and see that `female` is coded `1` for girls and `0` for boys. Similarly, `ses` is coded `1` for low, `2` for middle, and `3` for high. If you have installed the `fre` command using `ssc install fre`, you can use the command `fre female ses` to show both the value labels and the codes, as we discussed in section 5.4. We will compute the correlations anyway and see if they make sense.

Stata has two commands for doing a correlation: `correlate` and `pwcorr`. The `correlate` command runs the correlation using a casewise deletion (some books call this listwise deletion) option. Casewise deletion means that if any observation is missing for any of the variables, even just one variable, the observation will be dropped from the analysis. Many datasets, especially those based on surveys, have many missing values. For example, it is common for about 30% of people to refuse to report their income. Some survey participants will skip a page of questions by mistake. Casewise deletion can introduce serious bias and greatly reduce the working sample size. Casewise deletion is a problem for external validity or the ability to generalize when there are a lot of missing data. Many studies using casewise deletion will end up dropping 30% or more of the observations, and this makes generalizing a problem even though the total sample may have been representative.

The `pwcorr` command uses a pairwise deletion to estimate each correlation based on all the people who answered each pair of items. For example, if Julia has a score on `write` and `read` but nothing else, she will be included in estimating the correlation between `write` and `read`. Pairwise deletion introduces its own problems. Each correlation may be based on a different subsample of observations, namely, those observations who answered both variables in the pair. We might have 500 people who answered both `var1` and `var2`, 400 people who answered both `var1` and `var3`, and 450 people who answered both `var2` and `var3`. Because each correlation is based on a different subsample, under extreme circumstances it is possible to get a set of correlations that would be impossible for a population.

To open the `correlate` dialog box, select Statistics ▷ Summaries, tables, and tests ▷ Summary and descriptive statistics ▷ Correlations and covariances. To open the `pwcorr` dialog box, select Statistics ▷ Summaries, tables, and tests ▷ Summary and descriptive statistics ▷ Pairwise correlations. Because the command is so simple, we can just enter the command directly.

```
. correlate read write math science ses female
(obs=200)

             |     read    write     math  science      ses   female
-------------+------------------------------------------------------
        read |   1.0000
       write |   0.5968   1.0000
        math |   0.6623   0.6174   1.0000
     science |   0.6302   0.5704   0.6307   1.0000
         ses |   0.2933   0.2075   0.2725   0.2829   1.0000
      female |  -0.0531   0.2565  -0.0293  -0.1277  -0.1250   1.0000
```

We can read the correlation table going either across the rows or down the columns. The $r = 0.63$ between `science` and `read` indicates that these two skills are strongly related. Having good reading skills is probably helpful to having good science skills. All the skills are weakly to moderately related to socioeconomic status, `ses` ($r = 0.21$ to $r = 0.29$). Having a higher socioeconomic status does result in higher expected scores on all the skills for the 200 adolescents in the sample.

A dichotomous variable, such as gender, that is coded with a `0` for one category (man) and `1` for the other category (woman) is called a dummy variable or indicator variable. Thus `female` is a dummy variable (a useful standard is to name the variable to match the category coded as `1`). When you are using a dummy variable, the stronger the correlation is, the greater impact the dummy variable has on the outcome variable. The last row of the correlation matrix shows the correlation between `female` and each skill. The $r = 0.26$ between being a girl and writing skills means that girls (they were coded `1` on `female`) have higher writing skills than boys (they were coded `0` on `female`), and this is almost a moderate relationship. You have probably read that girls are not as skilled in math as are boys. The $r = -0.03$ between `female` and `math` means that in this sample, the girls had just slightly lower scores than boys (remember an $|r| = 0.1$ is weak, so anything close to 0 is very weak). If, instead of having 200 observations, we had 20,000, this small of a correlation would be statistically significant. Still, it is best described as very weak, whether it is statistically significant or not. The math advantage that is widely attributed to boys is very small compared with the writing advantage attributed to girls.

Stata's correlate command does not give us the significance of the correlations when using casewise deletion. The pwcorr command is a much more general command to estimate correlations because it has several important options that are not available using the correlate command. Indeed, the pwcorr command can do casewise/listwise deletion as well as pairwise deletion. When you are generating a set of correlations, you usually want to know the significance level, and it would be nice to have an asterisk attached to each correlation that is significant at the 0.05 level. You can use the dialog box Statistics ▷ Summaries, tables, and tests ▷ Summary and descriptive statistics ▷Correlations and covariances for the correlate command or Statistics ▷ Summaries, tables, and tests ▷ Summary and descriptive statistics ▷ Pairwise correlations for the pwcorr command. You can see additional options on the dialog boxes, but we can also simply enter the command directly. We use the same command as we did for correlate, substituting pwcorr for correlate and adding listwise, sig, and star(5) as options:

```
. pwcorr read write math science socst ses female, listwise sig star(5)
             |     read    write     math  science    socst      ses   female
-------------+---------------------------------------------------------------
        read |   1.0000
             |
             |
       write |   0.5968*  1.0000
             |   0.0000
             |
        math |   0.6623*  0.6174*  1.0000
             |   0.0000   0.0000
             |
     science |   0.6302*  0.5704*  0.6307*  1.0000
             |   0.0000   0.0000   0.0000
             |
       socst |   0.6215*  0.6048*  0.5445*  0.4651*  1.0000
             |   0.0000   0.0000   0.0000   0.0000
             |
         ses |   0.2933*  0.2075*  0.2725*  0.2829*  0.3319*  1.0000
             |   0.0000   0.0032   0.0001   0.0000   0.0000
             |
      female |  -0.0531   0.2565* -0.0293  -0.1277   0.0524  -0.1250   1.0000
             |   0.4553   0.0002   0.6801   0.0714   0.4614   0.0778
```

In this table, the listwise correlation between science and reading is $r = 0.63$. The asterisk indicates this is significant at the 0.05 level. Below the correlation is the probability and we can say that the correlation is significant at the $p < 0.001$ level. The reported probability is for a two-tailed test. If you had a one-tailed hypothesis, you could divide the probability in half.

If you want the correlations using pairwise deletion, you would also want to know how many observations were used for estimating each correlation. The command for pairwise deletion that gives you the number of observations, the significance, and an asterisk for correlations significant at the 0.05 level is

```
pwcorr read write math science ses female, obs sig star(5)
```

Notice that the only change was to replace the listwise option with the obs option.

Multiple-comparison procedures with correlations

When you are estimating several correlations, the reported significance level given by the `sig` option can be misleading. If you made 100 independent estimates of a correlation that was 0 in the population, you would expect to get five significant results by chance (using the 5% level). In this example, we had 21 correlations, and because we are considering all of them, we might want to adjust the probability estimate. One of the ways to adjust the probability estimate in the `pwcorr` command is with the option `bon`, short for the Bonferroni multiple-comparison procedure. You can get this procedure simply by adding the `bon` option at any point after the comma in the `pwcorr` command. The complete command would be (add the `listwise` option if you want to use casewise/listwise deletion)

```
pwcorr read write math science socst ses female, bon obs sig star(5)
```

Without this correction, the correlation between `write` and `ses`, $r = 0.21$, had a $p = 0.0032$ and was significant at the 0.01 level. With the Bonferroni adjustment, the $r = 0.21$ does not change, but the correlation now has a $p = 0.067$ and is no longer statistically significant. (An alternative multiple-comparison procedure uses the `sidak` option, which produces the Šidák-adjusted significance level.) It is difficult to give simple advice on when you should or should not use a multiple-comparison adjustment. If your hypothesis is that a certain pattern of correlations will be significant and this involves the set of all the correlations (here, 21), the multiple-comparison adjustment is appropriate. If your focus is on individual correlations, as it probably is here, then the adjustment is not necessary.

8.6 Regression

Earlier you learned how to plot a regression line on a scattergram. Now we will focus on how to estimate the regression line itself. Suppose that you are interested in the relationship between how many hours per week a person works and how much occupational prestige he or she has. You expect that careers with high occupational prestige require more work rather than less. Therefore, you expect that the more hours respondents work, the more occupational prestige they will have. This is certainly not a perfect relationship, and we have all known people who work many hours, even doing two jobs, who do not have high occupational prestige. We will use the General Social Survey 2006 dataset (`gss2006_chapter8_selected.dta`) for this section. It has variables called `prestg80`, which is a scale of occupational prestige, and `hrs1`, which is the number of hours respondents worked last week in their primary jobs.