

A Gentle Introduction to Stata

5th Edition

ALAN C. ACOCK
Oregon State University



A Stata Press Publication
StataCorp LP
College Station, Texas



Copyright © 2006, 2008, 2010, 2012, 2014, 2016 by StataCorp LP
All rights reserved. First edition 2006
Second edition 2008
Third edition 2010
Revised third edition 2012
Fourth edition 2014
Fifth edition 2016

Published by Stata Press, 4905 Lakeway Drive, College Station, Texas 77845

Typeset in L^AT_EX 2_ε

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Print ISBN-10: 1-59718-185-4

Print ISBN-13: 978-1-59718-185-3

Library of Congress Control Number: 2016935690

No part of this book may be reproduced, stored in a retrieval system, or transcribed, in any form or by any means—electronic, mechanical, photocopy, recording, or otherwise—without the prior written permission of StataCorp LP.

Stata, **stata**, Stata Press, Mata, **mata**, and NetCourse are registered trademarks of StataCorp LP.

Stata and Stata Press are registered trademarks with the World Intellectual Property Organization of the United Nations.

L^AT_EX 2_ε is a trademark of the American Mathematical Society.

Contents

	List of figures	xv
	List of tables	xxiii
	List of boxed tips	xxv
	Preface	xxix
	Support materials for the book	xxxv
1	Getting started	1
	1.1 Conventions	1
	1.2 Introduction	4
	1.3 The Stata screen	7
	1.4 Using an existing dataset	9
	1.5 An example of a short Stata session	11
	1.6 Video aids to learning Stata	18
	1.7 Summary	19
	1.8 Exercises	19
2	Entering data	21
	2.1 Creating a dataset	21
	2.2 An example questionnaire	23
	2.3 Developing a coding system	24
	2.4 Entering data using the Data Editor	29
	2.4.1 Value labels	33
	2.5 The Variables Manager	34
	2.6 The Data Editor (Browse) view	40
	2.7 Saving your dataset	41
	2.8 Checking the data	43

2.9	Summary	49
2.10	Exercises	49
3	Preparing data for analysis	51
3.1	Introduction	51
3.2	Planning your work	51
3.3	Creating value labels	57
3.4	Reverse-code variables	60
3.5	Creating and modifying variables	65
3.6	Creating scales	70
3.7	Saving some of your data	73
3.8	Summary	74
3.9	Exercises	75
4	Working with commands, do-files, and results	77
4.1	Introduction	77
4.2	How Stata commands are constructed	78
4.3	Creating a do-file	82
4.4	Copying your results to a word processor	88
4.5	Logging your command file	89
4.6	Summary	91
4.7	Exercises	92
5	Descriptive statistics and graphs for one variable	93
5.1	Descriptive statistics and graphs	93
5.2	Where is the center of a distribution?	94
5.3	How dispersed is the distribution?	98
5.4	Statistics and graphs—unordered categories	100
5.5	Statistics and graphs—ordered categories and variables	110
5.6	Statistics and graphs—quantitative variables	112
5.7	Summary	119
5.8	Exercises	120

6	Statistics and graphs for two categorical variables	123
6.1	Relationship between categorical variables	123
6.2	Cross-tabulation	124
6.3	Chi-squared test	127
6.3.1	Degrees of freedom	129
6.3.2	Probability tables	129
6.4	Percentages and measures of association	132
6.5	Odds ratios when dependent variable has two categories	135
6.6	Ordered categorical variables	137
6.7	Interactive tables	140
6.8	Tables—linking categorical and quantitative variables	142
6.9	Power analysis when using a chi-squared test of significance	145
6.10	Summary	147
6.11	Exercises	148
7	Tests for one or two means	151
7.1	Introduction to tests for one or two means	151
7.2	Randomization	154
7.3	Random sampling	156
7.4	Hypotheses	156
7.5	One-sample test of a proportion	157
7.6	Two-sample test of a proportion	159
7.7	One-sample test of means	164
7.8	Two-sample test of group means	166
7.8.1	Testing for unequal variances	171
7.9	Repeated-measures t test	172
7.10	Power analysis	174
7.11	Nonparametric alternatives	182
7.11.1	Mann–Whitney two-sample rank-sum test	182
7.11.2	Nonparametric alternative: Median test	183
7.12	Video tutorial related to this chapter	184

7.13	Summary	184
7.14	Exercises	185
8	Bivariate correlation and regression	189
8.1	Introduction to bivariate correlation and regression	189
8.2	Scattergrams	190
8.3	Plotting the regression line	196
8.4	An alternative to producing a scattergram, binscatter	197
8.5	Correlation	201
8.6	Regression	207
8.7	Spearman's rho: Rank-order correlation for ordinal data	212
8.8	Power analysis with correlation	213
8.9	Summary	215
8.10	Exercises	215
9	Analysis of variance	219
9.1	The logic of one-way analysis of variance	219
9.2	ANOVA example	220
9.3	ANOVA example with nonexperimental data	229
9.4	Power analysis for one-way ANOVA	232
9.5	A nonparametric alternative to ANOVA	234
9.6	Analysis of covariance	237
9.7	Two-way ANOVA	249
9.8	Repeated-measures design	255
9.9	Intraclass correlation—measuring agreement	260
9.10	Power analysis with ANOVA	262
9.10.1	Power analysis for one-way ANOVA	263
9.10.2	Power analysis for two-way ANOVA	265
9.10.3	Power analysis for repeated-measures ANOVA	267
9.10.4	Summary of power analysis for ANOVA	269
9.11	Summary	270
9.12	Exercises	270

10	Multiple regression	273
10.1	Introduction to multiple regression	273
10.2	What is multiple regression?	274
10.3	The basic multiple regression command	275
10.4	Increment in R-squared: Semipartial correlations	279
10.5	Is the dependent variable normally distributed?	281
10.6	Are the residuals normally distributed?	284
10.7	Regression diagnostic statistics	288
10.7.1	Outliers and influential cases	289
10.7.2	Influential observations: DFbeta	291
10.7.3	Combinations of variables may cause problems	292
10.8	Weighted data	294
10.9	Categorical predictors and hierarchical regression	296
10.10	A shortcut for working with a categorical variable	305
10.11	Fundamentals of interaction	306
10.12	Nonlinear relations	313
10.12.1	Fitting a quadratic model	315
10.12.2	Centering when using a quadratic term	321
10.12.3	Do we need to add a quadratic component?	323
10.13	Power analysis in multiple regression	325
10.14	Summary	328
10.15	Exercises	329
11	Logistic regression	333
11.1	Introduction to logistic regression	333
11.2	An example	334
11.3	What is an odds ratio and a logit?	338
11.3.1	The odds ratio	340
11.3.2	The logit transformation	340
11.4	Data used in the rest of the chapter	341
11.5	Logistic regression	342

11.6	Hypothesis testing	353
11.6.1	Testing individual coefficients	353
11.6.2	Testing sets of coefficients	354
11.7	More on interpreting results from logistic regression	356
11.8	Nested logistic regressions	360
11.9	Power analysis when doing logistic regression	362
11.10	Next steps for using logistic regression and its extensions	365
11.11	Summary	365
11.12	Exercises	366
12	Measurement, reliability, and validity	369
12.1	Overview of reliability and validity	369
12.2	Constructing a scale	370
12.2.1	Generating a mean score for each person	371
12.3	Reliability	372
12.3.1	Stability and test–retest reliability	375
12.3.2	Equivalence	376
12.3.3	Split-half and alpha reliability—internal consistency	376
12.3.4	Kuder–Richardson reliability for dichotomous items	379
12.3.5	Rater agreement—kappa (κ)	380
12.4	Validity	383
12.4.1	Expert judgment	383
12.4.2	Criterion-related validity	384
12.4.3	Construct validity	385
12.5	Factor analysis	386
12.6	PCF analysis	391
12.6.1	Orthogonal rotation: Varimax	394
12.6.2	Oblique rotation: Promax	396
12.7	But we wanted one scale, not four scales	397
12.7.1	Scoring our variable	398
12.8	Summary	399

12.9	Exercises	400
13	Working with missing values—multiple imputation	401
13.1	The nature of the problem	401
13.2	Multiple imputation and its assumptions about the mechanism for missingness	403
13.3	What variables do we include when doing imputations?	405
13.4	Multiple imputation	406
13.5	A detailed example	407
13.5.1	Preliminary analysis	408
13.5.2	Setup and multiple-imputation stage	410
13.5.3	The analysis stage	413
13.5.4	For those who want an R^2 and standardized β s	414
13.5.5	When impossible values are imputed	416
13.6	Summary	418
13.7	Exercises	419
14	The sem and gsem commands	421
14.1	Linear regression using sem	421
14.1.1	Using the SEM Builder to fit a basic regression model	423
14.2	A quick way to draw a regression model and a fresh start	430
14.2.1	Using sem without the SEM Builder	433
14.3	The gsem command for logistic regression	433
14.3.1	Fitting the model using the logit command	434
14.3.2	Fitting the model using the gsem command	436
14.4	Path analysis and mediation	442
14.5	Conclusions and what is next for the sem command	446
14.6	Exercises	448
15	An introduction to multilevel analysis	451
15.1	Questions and data for groups of individuals	451
15.2	Questions and data for a longitudinal multilevel application	452
15.3	Fixed-effects regression models	453

15.4	Random-effects regression models	454
15.5	An applied example	456
15.5.1	Research questions	456
15.5.2	Reshaping data to do multilevel analysis	457
15.6	A quick visualization of our data	460
15.7	Random-intercept model	461
15.7.1	Random intercept—linear model	461
15.7.2	Random-intercept model—quadratic term	464
15.7.3	Treating time as a categorical variable	468
15.8	Random-coefficients model	471
15.9	Including a time-invariant covariate	474
15.10	Summary	479
15.11	Exercises	479
16	Item response theory (IRT)	481
16.1	How are IRT measures of variables different from summated scales? .	482
16.2	Overview of three IRT models for dichotomous items	484
16.2.1	The one-parameter logistic (1PL) model	484
16.2.2	The two-parameter logistic (2PL) model	486
16.2.3	The three-parameter logistic (3PL) model	487
16.3	Fitting the 1PL model using Stata	487
16.3.1	The estimation	490
16.3.2	How important is each of the items?	492
16.3.3	An overall evaluation of our scale	494
16.3.4	Estimating the latent score	495
16.4	Fitting a 2PL IRT model	496
16.4.1	Fitting the 2PL model	497
16.5	The graded response model—IRT for Likert-type items	502
16.5.1	The data	502
16.5.2	Fitting our graded response model	504
16.5.3	Estimating a person’s score	509

<i>Contents</i>	xiii
16.6 Reliability of the fitted IRT model	509
16.7 Using the Stata menu system	512
16.8 Extensions of IRT	515
16.9 Exercises	516
A What's next?	519
A.1 Introduction to the appendix	519
A.2 Resources	519
A.2.1 Web resources	520
A.2.2 Books about Stata	522
A.2.3 Short courses	525
A.2.4 Acquiring data	525
A.2.5 Learning from the postestimation methods	526
A.3 Summary	527
References	531
Author index	535
Subject index	537

(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. The 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 any 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 in many chapters, as well as the National Survey of Youth, 1997. 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.

The fifth edition includes two new chapters. 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.

I would like to point out the use of punctuation after quotes in this book. While the standard U.S. style of punctuation calls for periods and commas at the end of a quote to always be enclosed within the quotation marks, Stata Press follows a style typically used in mathematics books and British literature. In this style, any punctuation mark at the end of a quote is included within the quotation marks only if it is part of the quote. For instance, the pleased Stata user said she thought that Stata was a “very powerful program”. Another user simply said, “I love Stata.”

I assume that the reader is running Stata 14, 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. In preparing this book, I have used both a Windows-based PC and a Mac.

(Pages omitted)

6 Statistics and graphs for two categorical variables

- 6.1 Relationship between categorical variables
 - 6.2 Cross-tabulation
 - 6.3 Chi-squared test
 - 6.3.1 Degrees of freedom
 - 6.3.2 Probability tables
 - 6.4 Percentages and measures of association
 - 6.5 Odds ratios when dependent variable has two categories
 - 6.6 Ordered categorical variables
 - 6.7 Interactive tables
 - 6.8 Tables—linking categorical and quantitative variables
 - 6.9 Power analysis when using a chi-squared test of significance
 - 6.10 Summary
 - 6.11 Exercises
-

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, 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.

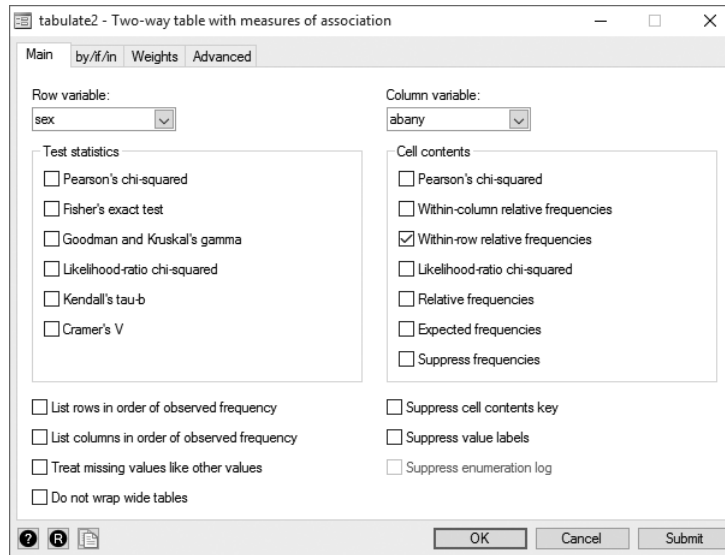


Figure 6.1. The Main tab for creating a cross-tabulation

If we had three variables and wanted to see all possible two-by-two tables (`varb` 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.

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
<i>frequency</i>
<i>row percentage</i>

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

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			
	<i>frequency</i>		
	<i>expected frequency</i>		
	<i>row percentage</i>		
Gender	ABORTION IF WOMAN WANTS FOR ANY REASON		Total
	YES	NO	
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.

(Pages omitted)

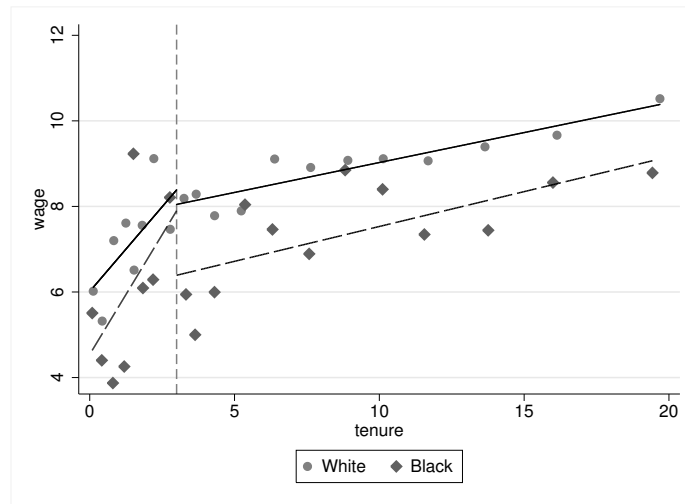


Figure 8.8. Relationship between wages and tenure with a discontinuity in the relationship at 3 years; whites shown with solid lines and blacks shown with dashed lines

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 zero). 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 zero 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 zero.

- 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 zero.

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 http://www.ats.ucla.edu/stat/stata/notes/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 zero 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 or 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.