Stata for the Behavioral Sciences

Michael N. Mitchell

A Stata Press Publication
STATA CORPORATION
College Station, Texas

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

ISBN-10: 1-59718-173-0
ISBN-13: 978-1-59718-173-0

Library of Congress Control Number: 2015947163

(Pages omitted)

Contents

(Pages omitted)

Preface

I worked as a statistical consultant at the University of California, Los Angeles, ATS statistical consulting group for over 12 years. Before the start of every walk-in consulting session, I would wonder about the questions that would walk into our office. Every session was different, with people bringing questions from all parts of the campus. When a new client walked in the door, one of my first questions was, What department are you from? It might have seemed like a polite question as we got to know each other before diving into the heart of his or her problem. But this was an essential question for me: the answer would guide the way I handled the entire visit.

In working as a consultant who served so many schools and departments, I discovered that there are many regional dialects of statistics. Depending on your school or department, there are certain types of statistical models you emphasize, customs you embrace, and types of terminology you favor. This is why there are statistics books that are written specifically for certain disciplines—to address the statistical customs and traditions within that discipline.

My home discipline is psychology, in which I received my bachelor's, master's, and doctorate. As I was taught statistics, my professors focused on forming specific hypothesis tests for the exact predicted pattern of results. They emphasized taking a laserlike focus on specific contrasts that would directly test the hypothesis of interest. This was especially the case when I was taught about factorial analysis of variance (ANOVA), where a finding of an overall interaction effect was nothing to get excited about because the significant interaction could be consistent with a variety of patterns of results, some of which could be contrary to our hypotheses. We were taught to graph the interactions and probe and dissect the interaction using planned contrasts to test for the exact pattern of results that we hypothesized.

I brought this training to my statistical consulting and my use of statistical packages. However, I frequently found that my training regarding ways of dissecting interactions was not so easily supported by statistical packages—that is, until the most recent versions of Stata. Unlike any other statistical package that I have ever used, Stata provides a suite of tools that allows us to probe, interpret, understand, and graph the results of ANOVA models. These tools are incredibly powerful; they are also very simple and intuitive. In this book, I show how you can use this suite of ANOVA tools to easily form contrasts among groups and dissect interactions with surgical precision. This allows you to present tests of hypotheses regarding the specific pattern of your results, establishing not only that your results are significant but also that they are in the pattern predicted. The suite of tools also integrates graphing tools so that you can use graphics as a means of interpreting your results and for presenting results to others.

The heart and spirit of this book is about showcasing this suite of ANOVA tools that Stata offers, but that does not mean this book is limited to just the presentation of ANOVA. This is because this suite of ANOVA tools can be applied to a wide variety of designs, including analysis of covariance, analysis of covariance with interactions, repeated measures designs, longitudinal designs, and the analysis of survey data to name just a few. This suite of tools can also be used in the context of a wide variety of regression modeling methods, including ordinary regression, robust regression, multilevel models, logistic regression, and Poisson regression.

As I see it, one of the strengths of learning statistics from a behavioral science perspective is seeing how factorial designs can help us understand how the effect of one variable is moderated by another variable through testing of interactions. With most statistical packages, you are handcuffed to using these tools only in the context of a traditional ANOVA. Once you extend your reach outside that realm, these tools are taken away from you. In Stata, you carry this suite of tools with you as you run a multilevel model, a robust regression, a logistic regression, or even a regression based on complex survey sampling. In this book, my aim is to show how you can use these familiar tools and to enable you to apply them across a wide variety of designs and modeling methods.

While this book draws upon my statistical training from the perspective of psychology, it is written for anyone in the behavioral sciences and anyone who would like to learn how to apply ANOVA (and ANOVA-like tools) to a variety of designs and modeling techniques using Stata. Regardless of your home discipline, I hope this book shows how you can use Stata to understand your results so that you can interpret and present them with clarity and confidence.

Valencia, California Michael N. Mitchell
July 2015

(Pages omitted)

5 Contrasts for a one-way ANOVA

5.1 Chapter overview

This chapter is devoted to showing you how to perform specific comparisons among groups (contrasts) after performing a one-way analysis of variance (ANOVA). Such contrasts allow you to test for the exact pattern of results that you hypothesized before the start of your study. This chapter illustrates many different kinds of contrasts that you can perform using Stata's built-in contrast operators. Further, it shows how you can perform custom contrasts (if you want to perform a contrast that is not among the built-in contrasts).

5.2 Introducing contrasts

It turns out that Professor Cheer is interested in studying happiness as well as studying optimism. She is interested in exploring the effect of marital status on happiness. She decides to explore this question using the General Social Survey (GSS) dataset from the year 2012. The `use` command reads this dataset into memory. (Section 1.1 shows how to download the example datasets.)

```
. use gss2012_sbs, clear
```

Let's start by looking at a frequency distribution of the variable `marital`, which contains the marital status of the respondent. The distribution of this variable is shown with the user-written `fre` command (see section 1.1 to see how to download this command).

```
. fre marital
marital — marital status
```

		Freq.	Percent	Valid	Cum.
Valid	1 married	900	45.59	45.59	45.59
	2 widowed	163	8.26	8.26	53.85
	3 divorced	317	16.06	16.06	69.91
	4 separated	68	3.44	3.44	73.35
	5 never married	526	26.65	26.65	100.00
	Total	1974	100.00	100.00	

This shows that 1,974 people responded, and their responses were coded as married, widowed, divorced, separated, or never married. The `fre` command shows the numeric codes associated with each of these responses.

Let's now use the `fre` command to show the frequency distribution for the variable `happy7`, which is the happiness rating of the respondent on a 1 to 7 scale. We can see that a response of `1` represents `Completely unhappy` and that a response of `7` represents `Completely happy`. For this question, there are 1,284 valid responses and 690 missing values. This large number of missing values reflects the fact that this question was only asked of some of the respondents.

```
. fre happy7
```

happy7 — how happy R is (recoded)

		Freq.	Percent	Valid	Cum.
Valid	1 Completely unhappy	5	0.25	0.39	0.39
	2 Very unhappy	16	0.81	1.25	1.64
	3 Fairly unhappy	35	1.77	2.73	4.36
	4 Neither happy nor unhappy	77	3.90	6.00	10.36
	5 Fairly happy	440	22.29	34.27	44.63
	6 Very happy	563	28.52	43.85	88.47
	7 Completely happy	148	7.50	11.53	100.00
	Total	1284	65.05	100.00	
Missing	.c Cannot choose	11	0.56		
	.i Inapplicable	672	34.04		
	.n No Answer	7	0.35		
	Total	690	34.95		
Total		1974	100.00		

Let's now perform a one-way ANOVA in which `happy7` is the DV[1] and marital status (`marital`) is the IV.

```
. anova happy7 marital
                        Number of obs =      1,284    R-squared     =  0.0408
                        Root MSE      =    .977102    Adj R-squared =  0.0378
```

Source	Partial SS	df	MS	F	Prob>F
Model	51.89968	4	12.97492	13.59	0.0000
marital	51.89968	4	12.97492	13.59	0.0000
Residual	1221.0972	1,279	.95472807		
Total	1272.9969	1,283	.99220334		

The F statistic above tests the null hypothesis that the mean of happiness is equal for all five levels of marital status. We can express this null hypothesis as shown below.

$$H_0\colon \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$$

The overall test of this null hypothesis is significant: $F(4, 1279) = 13.59$, $p < 0.001$. We can reject the null hypothesis that the average happiness is equal among the five marital status groups.

1. Some might be bothered by analyzing a Likert scale like `happy7` as though it were measured on an interval scale. For the sake of these examples, let's assume that `happy7` is measured on an interval scale.

5.2.1 Computing and graphing means

Let's use the `margins` command to compute the mean of happiness by marital status. This output shows, for example, that the mean of happiness for those who are married (group 1) is 5.71 with a 95% confidence interval (CI) of 5.63 to 5.79.

```
. margins marital, nopvalues
Adjusted predictions                            Number of obs     =      1,284
Expression   : Linear prediction, predict()

------------------------------------------------------------------
              |            Delta-method
              |     Margin   Std. Err.     [95% Conf. Interval]
--------------+---------------------------------------------------
      marital |
      married |   5.714038   .0406773      5.634237     5.79384
      widowed |   5.429907     .09446      5.244593     5.61522
     divorced |   5.252475   .0687486      5.117603    5.387348
    separated |   5.119048   .1507701      4.823264    5.414831
never married |   5.365169   .0517863      5.263573    5.466764
------------------------------------------------------------------
```

The `marginsplot` command can be used to make a graph of the means and CIs reported by the `margins` command. This produces the graph shown in figure 5.1.

```
. marginsplot
  Variables that uniquely identify margins: marital
```

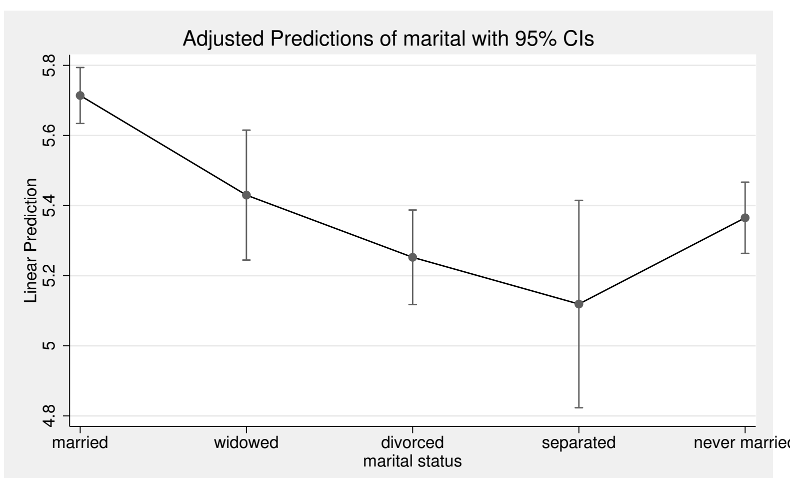

Figure 5.1: Mean happiness by marital status

5.2.2 Making contrasts among means

Let's probe this finding in more detail. Suppose that Professor Cheer predicted (before even seeing the data) that those who are married will be happier than each of the four other marital status groups. We can frame this as four separate null hypotheses, shown below.

$$H_0\#1\colon \mu_2 = \mu_1$$
$$H_0\#2\colon \mu_3 = \mu_1$$
$$H_0\#3\colon \mu_4 = \mu_1$$
$$H_0\#4\colon \mu_5 = \mu_1$$

The first null hypothesis states that the mean happiness is the same for group 2 and group 1 (widowed versus married). The second states that the mean happiness is the same for groups 3 and 1 (divorced versus married). The third states that the mean happiness is the same for groups 4 and 1 (separated versus married). Finally, the fourth states that the mean happiness is the same for groups 5 and 1 (never married versus married).

We can test each of these four null hypotheses using the `contrast` command shown below. (I will discuss the syntax of the `contrast` command shortly.)

```
. contrast r.marital
Contrasts of marginal linear predictions
Margins      : asbalanced
```

	df	F	P>F
marital			
(widowed vs married)	1	7.63	0.0058
(divorced vs married)	1	33.39	0.0000
(separated vs married)	1	14.52	0.0001
(never married vs married)	1	28.07	0.0000
Joint	4	13.59	0.0000
Denominator	1279		

	Contrast	Std. Err.	[95% Conf.	Interval]
marital				
(widowed vs married)	-.2841316	.1028462	-.4858973	-.0823659
(divorced vs married)	-.4615629	.0798813	-.6182756	-.3048502
(separated vs married)	-.5949905	.156161	-.9013504	-.2886306
(never married vs married)	-.3488696	.0658518	-.478059	-.2196801

This `contrast` command compares each marital status group to group 1. The first test compares those who are widowed with those who are married. The upper portion of the output shows the F statistic for the test of this hypothesis. This test is statistically significant: $F(1, 1279) = 7.63$, $p = 0.0058$. The lower portion of the output shows that the average difference in the happiness for those who are widowed versus married is -0.28, (95% CI $= [-0.49, -0.08]$). Those who are widowed are significantly less happy than those who are married. Put another way, those who are married are significantly happier than those who are widowed. We can reject $H_0\#1$ and say the results are consistent with our prediction that those who are married are significantly happier.

Let's now consider the output for the second, third, and fourth contrasts. These test the second, third, and fourth null hypotheses contrasting those who are divorced versus married, separated versus married, and never married versus married. The upper portion of the `contrast` output shows that each of these contrasts is statistically significant (ps ≤ 0.001). Further, the differences in the means (as shown in the lower portion of the output) are always negative, indicating that those who are married are happier than the group they are compared with. Thus we can reject the second, third, and fourth null hypotheses.

To summarize, we can reject all four null hypotheses. Further, each difference was in the predicted direction. Those who are married are significantly happier than those who are widowed, significantly happier than those who are divorced, significantly happier than those who are separated, and significantly happier than those who have never been married.

5.2.3 Graphing contrasts

Let's create a graph that visually depicts these contrasts. We first use the `margins` command to replicate the results we found above via the `contrast` command. (This will allow us to then use the `marginsplot` command to graph the results from the `margins` command.)

(Pages omitted)

23 Stata equivalents of common IBM SPSS Commands

23.1 Chapter overview

When taking a language class, you often would have a book that would quickly translate English words into another language (say, Spanish). In some cases, there would be one exact word that was the equivalent of the English word, and in other cases, the translation would be a bit more murky. There might be two or more equivalent words, or the words would have similar but not exact meanings. Well, think of this chapter as an IBM® SPSS® to Stata translation, where sometimes, the SPSS command has an exact equivalent, and sometimes, there is a little bit more in the translation.

This chapter lists commonly used SPSS commands (in alphabetical order) and shows the equivalent (or near equivalent) Stata command along with one or more examples of the Stata command. The aim of this chapter is to help you learn, for example, that the Stata equivalent of the SPSS `AGGREGATE` command is the `collapse` command and to show a simple example of the `collapse` command. To that end, the examples are intentionally simple (and sometimes might be overly simplistic). But once you know that the command you want is, for example, the `collapse` command, you can then type `help collapse` and access the Stata documentation for all the details about the use of the `collapse` command.

23.2 ADD FILES

Stata equivalent: `append`
Example: Consider the file `moms.dta`, which contains a family ID variable, the mother's age, race, and whether she graduated from high school.