Stata II

Inferential Statistics

2026-04-01

After attending this training session, you will be able to:

1. Get to know CITL and Stata
2. Understand how to open data sets, set memory size
3. Install user-written programs
4. Learn about the General Social Survey (GSS)
5. Obtain inferential statistics including chi-square test, ANOVA, bivariate correlation, linear regression, multiple regression
6. Do practice exercises
7. Print output
8. Save and export output

Introduction

Code and Dataset

For this training session, please download the following data and code:

STATA workshop.dta

Recap of Stata I

In the **Stata I** training session, we covered the following topics. Please feel free to refer back to the Stata I workbook using the following links if you would like a refresher.

- What is Stata?
  - Starting Stata
  - Opening an Existing Data File
- Using the Data Editor
- Descriptive Statistics
- Data Transformation
- Graphics
- Saving and Printing

As in the Stata I training session, we will be using data from the General Social Survey.

Installing User-Written Programs in Stata

User-written programs can be installed by using the `findit` command. Your computer needs be connected to the Internet before submitting the **findit** statement. We will use this command to install the Levene's test (`levene`), as this test will be used later in this tutorial but is not preinstalled in Stata. To search for this program, type:

```
findit levene
```

When the `Viewer` window opens, click on the link "levene from http://fmwww.bc.edu..."

```
2 packages found (Stata Journal and STB listed first)
-----------------------------------------------------
robvar from http://www.stata.com/users/mcleves
    robvar.  Robust test for equality of variances. / Program by Mario A.
    Cleves, Stata Corporation <mcleves@stata.com>. / STB 26 distribution,
    June 1996. / robar performes robust tests for equality of variances. It
    calculates / Levene's test statistic along with two reformulations by

levene from http://fmwww.bc.edu/RePEc/bocode/l
    'LEVENE': module to test for equal population variances / Levene's T-test
    is used to test the null hypothesis that multiple / population variances
    (corresponding to multiple samples) are / equal. The test is a one-way
    ANOVA on absolute deviations for / each group. This is version 1.1 of the
```

To install `levene`, click on the link that says `"(click here to install)"`

```
package levene from http://fmwww.bc.edu/RePEc/bocode/l

TITLE
      'LEVENE': module to test for equal population variances

DESCRIPTION/AUTHOR(S)
      Levene's T-test is used to test the null hypothesis that multiple
      population variances (corresponding to multiple samples) are
      equal. The test is a one-way ANOVA on absolute deviations for
      each group. This is version 1.1  of the software.

      Author: Herve M. Caci
      Support: email hcaci@wanadoo.fr

      Distribution-Date: 20040503
INSTALLATION FILES                                  (click here to install)
      levene.ado
      levene.hlp
```

You will be directed to the screen below if installation was successful:

```
package installation

package name:  levene.pkg
        from:  http://fmwww.bc.edu/RePEc/bocode/l/

checking levene consistency and verifying not already installed...
installing into c:\ado\plus\...
installation complete.
```

Chi-square test

A **chi-square test** is used to determine whether there is a statistically significant association between two categorical variables.

Example 1: 2x2 Table

Example 1 tests whether there is a significant association between the variables *sex* (Respondent's sex) and the variable *hschem* (Respondent ever took chemistry class in high school). The syntax generates a crosstabulation table with chi-square test statistics. Enter:

```
tabulate sex hschem, all exact
```

Output:

```
. tabulate sex hschem, all exact

           |   R ever took a high
           |   school chemistry
Respondent |        course
     s sex |       Yes         No |     Total
-----------+----------------------+----------
      MALE |       271        183 |       454
    FEMALE |       363        295 |       658
-----------+----------------------+----------
     Total |       634        478 |     1,112

          Pearson chi2(1) =   2.2439   Pr = 0.134
 likelihood-ratio chi2(1) =   2.2484   Pr = 0.134
               Cramér's V =   0.0449
                    gamma =   0.0923  ASE = 0.061
          Kendall's tau-b =   0.0449  ASE = 0.030
           Fisher's exact =                 0.139
   1-sided Fisher's exact =                 0.075
```

Stata generates the following test statistics for chi-square:

For all categorical variables:

- **Pearson's Chi-Square**: Used to test the hypothesis that the row and column variables are independent. This should not be used if any cell has an expected value less than 1, or if more than 20% of the cells have expected values less than 5.
- **Likelihood Ratio**: A goodness-of-fit statistic similar to Pearson's chi-square. This statistic is appropriate to report for smaller sample sizes. The likelihood ratio approaches the chi-square distribution as n increases, so for large sample sizes, the two statistics are the same and either can be reported.

- **Cramér's V**: It shows the magnitude of negative or positive relationship between two variables which ranges from −1 to +1.
- **Fisher's Exact Test**: A test for independence in a 2x2 table. This statistic is most useful when the total sample size and expected values are small.
- **1-sided Fisher's Exact Test**: The interpretation is the same as Fischer's Exact Test except this result shows a p-value for 1-sided test.

For ordinal variables only: - **gamma**: This result is appropriate for ordinal variables because it uses the ordering within a variable to compute a test statistic. It is analogous to a correlation coefficient. The statistic ranges from −1 to +1. The gamma statistic will be close to zero when the relationship of two variables is independent. - **Kendall's tau-b**: Similar to gamma, but makes a correction for pairs tied in the dependent variable - **ASE**: Asymptotic Standard Error, a standard error specifically computed for categorical (i.e. nominal and ordinal) data.

In this example, **Fisher's Exact Test** indicates a p-value of .000. Thus, data like this are relatively inconsistent with the null hypothesis, and we can conclude that there is not a statistically significant relationship between the respondent's sex and whether they took a chemistry course in high school.

Example 2: 3x2 Table

Example 2 tests whether is a significant association between the variables *colsci* (Respondents who took college-level science courses or not) and *natspac* (Attitudes towards the space exploration program). The syntax generates a crosstabulation table with chi-square test statistics. Enter:

```
tabulate colsci natspac, all expected
```

Output:

Key
frequency
expected frequency

R has taken any college-level sci course	Space exploration program TOO LITTL	ABOUT RIG	TOO MUCH	Total
Yes	96 69.2	160 161.5	61 86.3	317 317.0
No	53 79.8	188 186.5	125 99.7	366 366.0
Total	149 149.0	348 348.0	186 186.0	683 683.0

```
         Pearson chi2(2) =  33.3400   Pr = 0.000
likelihood-ratio chi2(2) =  33.8026   Pr = 0.000
              Cramér's V =   0.2209
                   gamma =   0.3669  ASE = 0.059
         Kendall's tau-b =   0.2076  ASE = 0.035
```

It is appropriate to use the Chi-Square test statistic in this example. For example, using Pearson Chi-Square, the level of significance is `p < .05` level. Thus, there is a statistically significant relationship between attitudes towards the space program and taking college-level science courses.

Exercise 1

Perform a chi-square test to determine if educational attainment (*degree*) is significantly associated with the belief that marijuana should be legal (*grass*).

- How many high school graduates said "Yes" to legalization?
- Which chi-square test statistic is appropriate? Is there a statistically significant relationship between the two variables?

See Appendix A for answers.

Independent Samples t-test

An **independent samples t-test** determines whether there is a statistically significant difference in the mean of a continuous dependent variable between two groups.

For example, this procedure may be used to find out if there is a statistically significant difference between the responses of men and women (*sex*) for how much they have to relax per day (*hrlax*).

Before running an independent samples t-test, a test of the homogeneity of variance (**Levene's Test**) must be conducted to determine whether the variances of the dependent variable *hrlax* between groups are equal or unequal. Enter:

```
levene hrlax, by (sex)
```

Output:

```
                 Analysis of Variance
    Source              SS          df       MS
--------------------------------------------------
Between groups      .27761395     1      .27761395
 Within groups      5614.8024     1403   4.0019974
--------------------------------------------------
    Total           5615.08       1404

Levene's T-test for equal variances: T= 0.07  Prob > T = 0.7923


.
```

{fig-alt="An analysis of variance for the Levene's Test showing the results at the bottom where T=0.07 and Prob > T = 0.7923}

Because the p-value of Levene's test is `> 0.05`, we fail to reject the null hypothesis and conclude that the two groups have equal variances. Therefore, the **ttest** syntax should not use the unequal option. Enter:

```
ttest hrlax, by(sex)
```

Output:

```
Two-sample t test with equal variances
```

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
MALE	676	3.970414	.1090614	2.835596	3.756274	4.184555
FEMALE	729	3.496571	.100518	2.713985	3.299231	3.69391
combined	1,405	3.724555	.0742269	2.78227	3.578948	3.870163
diff		.4738436	.148073		.1833752	.7643119

```
    diff = mean(MALE) - mean(FEMALE)                          t =   3.2001
Ho: diff = 0                                 degrees of freedom =     1403

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9993         Pr(|T| > |t|) = 0.0014          Pr(T > t) = 0.0007
```

Because the t-test statistic significance level is 0.0014 (`Ha:Diff!=0` for two tail) which is less than 0.05, we reject the null hypothesis and conclude that there is a statistically significant difference in hours have to relax per day between men and women.

Exercise 2

Find out if there is a statistically significant difference in terms of respondents' family income (*fmi*) between respondents who took college-level science courses and those who did not?(*colsci*). See Appendix A for answers.

ANOVA

A **one-way ANOVA** is used to determine whether there is a statistically significant difference in the mean of a continuous dependent variable among two or more groups.

For example, this procedure may be used to test if there is a statistically significant difference in the average number of hours spent watching television per day (*tvhours*) by marital status (*marital*).

First, the `levene` command must be requested to determine if the equal variance assumption is met. (By default, the `oneway` command computes **Bartlett's Test for Equal Variances**. However, we recommend **Levene's Test for Equality of Variances** because it is relatively more robust to non-normality than other homogeneity tests.) Enter:

```
levene tvhours, by(marital)
```

Output:

```
                Analysis of Variance
    Source              SS         df      MS
--------------------------------------------------
Between groups      141.66498     4     35.416244
 Within groups      6731.3961     1549  4.3456398
--------------------------------------------------
    Total           6873.061      1553

Levene's T-test for equal variances: T= 8.15  Prob > T = 0.0000
```

Because `p < .05`, as seen in the output table, we reject the null hypothesis that the variances are equal, and not-equal variances are assumed. The result confirms that we should use the `Welch` test for non-equal variances.

Note: Because of the result from Levene's test above, we cannot perform a standard equal variance ANOVA test for these variables. If we want to use equal variance anova, we should use the code:

```
oneway tvhours marital, tabulate
```

Welch Test

Because the homogeneity of variances assumption is not met, the **Welch test of equality of means** should be requested when performing a one-way ANOVA. To request the Welch test, `wtest` must be installed in Stata. Submit the below syntax and install "wtest from http://www.ats.ucla.edu...":

```
findit wtest
```

```
4 packages found (Stata Journal and STB listed first)
-----------------------------------------------------

st0158 from http://www.stata-journal.com/software/sj9-1
    SJ9-1 st0158.  A general-purpose method for two-group... / A
    general-purpose method for two-group randomization tests / by Johannes
    Kaiser, Lab. for Experimental Economics, / Univ. of Bonn, Germany /
    Michael G. Lacy, Sociology Dept., Colo. State. Univ., / Fort Collins, CO,

fstar from http://www.ats.ucla.edu/stat/stata/ado/analysis
    F* test for test of equal means / Allows you to use perform tests similar
    to / standard ANOVA F tests, but often with better / control of Type I
    error rates when / heterogeneity of variance is present. / See also
    simanova and wtest. / Michael N. Mitchell / Statistical Computing and

simanova from http://www.ats.ucla.edu/stat/stata/ado/analysis
    simanova. Simulation for ANOVA / Allows you to use simulation to study /
    Type I error rates and power in standard ANOVA, / as well as using the F*
    and W test. / See also fstar and wtest. / Michael N. Mitchell /
    Statistical Computing and Consulting / UCLA Academic Technology Services /

wtest from http://www.ats.ucla.edu/stat/stata/ado/analysis
    W test for test of equal means / Allows you to use perform tests similar
    to / standard ANOVA F tests, but often with better / control of Type I
    error rates when / heterogeneity of variance is present. / See also
    simanova and fstar. / Michael N. Mitchell / Statistical Computing and
```

To perform the Welch test, enter:

```
wtest tvhours marital
```

Output:

```
--------------------------------------------------------------------
Dependent Variable is tvhours and Independent Variable is marital
WStat(   5, 317.02) =    6.118, p= 0.0000
--------------------------------------------------------------------
```

Because `p < .05`, we conclude *tvhours* significantly differs based on *marital* groups.

ANOVA Post Hoc Tests

Once you have determined that differences exist among the group means, post hoc tests can determine which means differ. Because the significance of the statistic found previously in `levene tvhours, by(marital)` was < .05, we reject the null hypothesis that the variances are equal. Therefore, we can select a post hoc test that assumes unequal variances.

Because the homogeneity of variances assumption is not met, the **Games and Howell comparison of group means** should be requested for post hoc. To request the Games and Howell test, `pwmc` must be installed in Stata. Submit the syntax below and install "pwmc from http://fmwww.bc.edu/...":

```
findit pwmc
```

```
6 packages found (Stata Journal and STB listed first)
-----------------------------------------------------

pwmc from http://fmwww.bc.edu/RePEc/bocode/p
    'PWMC': module to compute pairwise multiple comparisons (unequal
    variances) / pwmc performs pairwise comparisons of means. Confidence /
    intervals are derived using procedures allowing for unequal / variances
    across groups. Dunnett's C, GH and T2 procedures are / implemented.  / KW:
```

To perform the Games and Howell post hoc test, enter:

```
pwmc tvhours, over(marital)
```

Output:

tvhours	Diff.	Std.Err	Games and Howell [95% Conf.	Interval]
marital				
2 vs 1	1.449766	.2964798	.63173	2.267801
3 vs 1	.5506357	.2013946	-.0011762	1.102448
4 vs 1	.5485591	.6239762	-1.234741	2.331859
5 vs 1	.4347425	.1774241	-.0505657	.9200506
3 vs 2	-.8991301	.3377641	-1.827252	.028992
4 vs 2	-.9012066	.6803463	-2.819942	1.017528
5 vs 2	-1.015023	.3240434	-1.906251	-.123796
4 vs 3	-.0020766	.6446166	-1.833923	1.82977
5 vs 3	-.1158932	.2401326	-.77285	.5410636
5 vs 4	-.1138166	.6375344	-1.928862	1.701229

The **Multiple Comparisons** table provides the results of the Games and HoWell post hoc test. The values for the variable `Marital`: 1=**Married**, 2=**Widowed**, 3=**Divorced**, 4=**Separated**, 5=**Never married**. We use this confidence interval in the same way that we use a confidence interval for a regular independent samples t -test: if it contains 0.00, the groups are not significantly different, but if it does not contain 0.00 then the groups are significantly different.

From this table it can be concluded that:

- Group **Married**'s mean is significantly different from the group **Widowed**'s mean.
- Group **Widowed**'s mean is significantly different from group **Never married**'s mean.
- All the other groups based on marital status have no significant different with each other.

After identifying statistically significant differences, look back at the table of descriptive statistics to determine which marital groups spends more (or less) time watching TV per day. It is clear that the mean hours of time spent watching TV per day is significantly higher for group **Widowed** than othe groups.

For equal variance, The `oneway` command includes the capability of performing multiple-comparison tests, using either Bonferroni (`bonferroni`), Scheffé (`scheffe`), or Šidák (`sidak`) corrections. For example, if we use Bonferroni post hoc test for standard anonva with equal variance of homogeneity, we should enter :

```
oneway tvhours age, noanova bonferroni
```

Exercise 3

Using the one-way ANOVA procedure, test whether respondents' family income (*fmi*) is different based on respondents' education level (*degree*). See Appendix A for answers.

Bivariate Correlation

A **(bivariate) correlation coefficient** measures the linear relationship between two continuous variables. A correlation coefficient ranges from −1 to 1. A positive correlation coefficient indicates that there is a positive linear relationship, whereas a negative correlation coefficient indicates that there is a negative linear relationship. A correlation coefficient of 0 indicates that there is no linear relationship between two variables.

For example, this procedure may be used to measure the relationship between the continuous variables spouse occupation prestige score (*sop*), spouse socioeconomic index score (*ssi*), and number of hrs spouse usually works a week (*sphrs2*). Enter:

```
pwcorr sop ssi sphrs2, obs sig star(0.05)
```

Output:

	sop	ssi	sphrs2
sop	1.0000		
	1169		
ssi	0.8328*	1.0000	
	0.0000		
	1169	1169	
sphrs2	-0.1523	-0.0944	1.0000
	0.4879	0.6683	
	23	23	23

The output is a 3x3 table. Each cell contains (1) the correlation coefficient (Pearson Correlation), (2) the p-value associated with the correlation coefficient (Sig.), and (3) the number of cases (N). In this example, the correlation coefficient between *sop* and *ssi* is 0.8328. and the significance of the coefficient is reported as 0.0000, which suggests that these two variables are positively correlated at 0.05 level.

Note on sample size (N): The third row of the cells for the correlation of *sphrs2* and *sop*, and *sphrs2* and *ssi* show that only 23 cases are used for computation for both pairs. Meanwhile, the correlation of *ssi* and *sop* used 1,169 cases.

The default treatment of missing cases is to use the maximum number of non-missing values for each pair of variables, which is known as **pairwise deletion**. The `pwcorr` command uses pairwise deletion.

An alternative treatment of missing cases is **listwise deletion**. The `correlate` command calculates a correlation coefficient using listwise deletion, which omits all cases with at least one missing value in all given variables. When only two variables are included, the sample size is the same whether pairwise or listwise deletion is used. However, these options make a difference in the sample size when three or more variables are included. Enter:

```
correlate sop ssi sphrs2
```

Output:

```
(obs=23)
```

	sop	ssi	sphrs2
sop	1.0000		
ssi	0.9672	1.0000	
sphrs2	-0.1523	-0.0944	1.0000

To compare enter:

```
pwcorr sop ssi sphrs2
```

Output:

	sop	ssi	sphrs2
sop	1.0000		
ssi	0.8328	1.0000	
sphrs2	-0.1523	-0.0944	1.0000

Exercise 4

Compute the correlation between: the highest year of school completed (*educ*), the respondent's father's occupational prestige index (*faos*), and the respondent's mother's occupational prestige index (*moos*) using (1) pairwise deletion and (2) listwise deletion and compare how the sample size changes. See Appendix A for answers.

Bivariate Regression

Bivariate linear regression is used to model the effect of a numeric independent variable on a numeric dependent variable.

This example uses linear regression to measure the effect of educational attainment (*educ*) on respondent's income (*realrinc*). In Stata, the `regress` command estimates a linear regression model with the dependent variable listed first and the independent variable(s) listed second. Enter:

```
regress realrinc educ
```

Output:

Source	SS	df	MS			
Model	1.0780e+11	1	1.0780e+11	Number of obs	=	1,362
Residual	1.0281e+12	1,360	755933237	F(1, 1360)	=	142.60
				Prob > F	=	0.0000
Total	1.1359e+12	1,361	834581390	R-squared	=	0.0949
				Adj R-squared	=	0.0942
				Root MSE	=	27494

realrinc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	3121.341	261.3855	11.94	0.000	2608.579	3634.104
_cons	-19105.83	3768.318	-5.07	0.000	-26498.18	-11713.48

The ANOVA table (upper left, box 1) tests whether the independent variable (*educ*) has a linear relationship with the dependent variable (*realrinc*). The ANOVA F-score (box 2, upper right) is statistically significant, F(1, 1362) = 142.60, Sig = 0.0000, thus the education variable has a significant linear relationship with the income variable.

We can also use the information in box 2 to assess the overall fit of the model. **R-squared**, also called the **coefficient of determination**, tells us the proportion of all variation in the dependent variable that is explained by the independent variable. The R-squared ranges ranges from 0 (no effect of an independent variable on a dependent variable) to 1 (perfect prediction of the dependent variable by the independent variable). From box 2, we see that 9.49% of the variation in income (dependent variable) is explained by the years of education (independent variable).

> **Note**
>
> The **adjusted R-squared** is used for model selection and is adjusted for the number of independent variables in the model.

Box 3 (lower) presents the unstandardized regression coefficients, their standard errors, t-statistics, significance levels, and the 95% confidence intervals for the coefficients. The regression coefficient for the constant is −19105.83 and the gregression coefficient for *educ* is 3121.341. We can use these values to express the linear regression equation: $Y = -19105.83 + 3121.341(educ)$. The slope, 3121.341 is called an **unstandardized coefficient** because the effect is measured in the original scale of the variable. In this example with a one unit increase in the years of education, predicted income will increase by $3,121.341.

The **standard error** (Std.Err.) is used to calculate the **t-value** (coefficient/Std. Error). This is the test statistic for the significance test where the null hypothesis states that the unstandardized coefficient is equal to 0. In this example, the coefficient for *educ* is statistically significant. This means that you can reject the null hypothesis that the education variable's coefficient is equal to 0 and conclude that education is a significant predictor of income.

Exercise 5

Run a bivariate linear regression to assess the effect of respondent's age (*age*) on their performance of the vocabulary test score (*wordsum*). See Appendix A for answers.

Multiple Regression

Multiple regression is the same as bivariate regression, except it uses two or more independent variables to explain the variation of the dependent variable. This example uses Multiple regression to measure the effect of respondents' health score (*rhs*) and age (*age*) on respondent's income (*realrinc*). The `beta` option requests standardized regression coefficients. Enter:

```
regress realrinc rhs age, beta
```

Output:

Source	SS	df	MS
Model	1.9125e+10	2	9.5624e+09
Residual	3.1996e+11	585	546941330
Total	3.3909e+11	587	577658345

Number of obs	=	588
F(2, 585)	=	17.48
Prob > F	=	0.0000
R-squared	=	0.0564
Adj R-squared	=	0.0532
Root MSE	=	23387

realrinc	Coef.	Std. Err.	t	P>|t|	Beta
rhs	-3643.138	1446.845	-2.52	0.012	-.1026611
age	380.3832	66.68653	5.70	0.000	.2325603
_cons	11211.44	3632.251	3.09	0.002	.

The top left hand side shows the results of the ANOVA tests. The ANOVA tests examines whether the independent variables (*rhs* and *age*) have a linear relationship with the dependent variable (*realrinc*). In this example, the ANOVA F-score is statistically significant, F = 17.48, Sig = 0.0000; thus a respondent's health score and age have a significant linear relationship with income.

The introduction of additional variables to a model produces a higher R-squared value, regardless of the efficiency of the model. Thus, R-squared as a model selection tool always favors models with additional predictors.

The Adjusted R-Squared value corrects this bias by adjusting both the numerator and the denominator of the R-Squared by their respective degrees of freedom. The adjusted R-squared is calculated using the equation below, where *n* is the number of observations and *k* is the number of independent variables.

$$\text{Adjusted R-Squared } = 1 - (1 - R)^{2} * \frac{n-1}{n-k-1}$$

Unlike R-squared, the Adjusted R-squared can decline in value as more predictors are added. It is important to keep in mind, however, that the R-Squared value is a proportion of total variance, while the Adjusted R-Squared is not.

The **Coefficients table** provides the values of unstandardized regression coefficients (**B**), standardized coefficients (**Beta**), t-test statistics (**t**), and p-values (**Sig.**).

- The value of unstandardized regression coefficient shows that a unit increase in *rhs* (worse health condition) will decrease predicted income by $3643.138 while holding *age* as constant.
- The t-test statistic is -2.52, and the p-value associated with the t-test statistic is 0.012. This results show that *rhs* is statistically significant at the 0.05 level while controlling for *age*.

The standardized coefficient (**Beta**) measures change in the dependent variable in standard deviations per one standard deviation change in the independent variable. The purpose of using standardized coefficients is to compare the magnitude of the effect of each independent variable. Because the absolute value of standardized coefficient for *age* is larger than that for *rhs*, we can conclude that age has a larger impact on real income than health rating.

Exercise 6

Run multiple regression to assess the effect of occupational prestige of a respondent's parents (*faos* and *moos*) and respondent's gender (*Male*) on the respondent's occupational prestige (*pres*). See Appendix A for answers.

Appendix A: Answers for Exercises

Exercise 1

Enter:

```
tabulate grass degree, all expected
```

```
+-------------------+
| Key               |
|-------------------|
|     frequency     |
| expected frequency|
+-------------------+

    Should |
 marijuana |
   be made |                   R's highest degree
     legal |  LT HIGH S  HIGH SCHO  JUNIOR CO   BACHELOR   GRADUATE |     Total
-----------+-------------------------------------------------------+----------
     LEGAL |         87        465         85        201         99 |       937
           |      106.3      456.2       82.9      186.6      105.0 |     937.0
-----------+-------------------------------------------------------+----------
 NOT LEGAL |         77        239         43         87         63 |       509
           |       57.7      247.8       45.1      101.4       57.0 |     509.0
-----------+-------------------------------------------------------+----------
     Total |        164        704        128        288        162 |     1,446
           |      164.0      704.0      128.0      288.0      162.0 |   1,446.0

          Pearson chi2(4) =  14.6690   Pr = 0.005
 likelihood-ratio chi2(4) =  14.3519   Pr = 0.006
               Cramér's V =   0.1007
                    gamma =  -0.0791  ASE = 0.044
          Kendall's tau-b =  -0.0447  ASE = 0.025
```

- 465 high school graduates said ``Yes'' to legalization.
- Because the expected cell counts are greater than five, Pearson's chi-square test is appropriate. Because the p-value associated with the Pearson chi-square statistic (0.005) is less than 0.05, there is a statistically significant relationship between the two variables.

Exercise 2

Enter:

```
levene fmi, by(colsci)
```

```
                 Analysis of Variance
    Source              SS         df       MS
--------------------------------------------------
Between groups      2.822e+10      1     2.822e+10
 Within groups      3.942e+11    1067   3.694e+08
--------------------------------------------------
    Total           4.224e+11    1068

Levene's T-test for equal variances: T= 76.40  Prob > T = 0.0000
```

Because the Levene's Test for Equality of Variances is significant ($p < 0.05$), the "unequal" option is necessary.

Enter:

```
ttest fmi, by(colsci) unequal\
```

```
Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     Yes |     486    43180.81    1566.612    34536.59    40102.63       46259
      No |     583    24504.81    975.8985    23563.45     22588.1    26421.52
---------+--------------------------------------------------------------------
combined |   1,069    32995.49    933.1226    30508.97    31164.53    34826.45
---------+--------------------------------------------------------------------
    diff |            18676.01    1845.711                 15053.2    22298.82
------------------------------------------------------------------------------
    diff = mean(Yes) - mean(No)                                   t =  10.1186
Ho: diff = 0                     Satterthwaite's degrees of freedom =  830.255

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

The P statistic (.0000) is significant ($p < 0.05$), therefore the amount of respondent's family income respondents made significantly differ based on if they took or didn't take college-level science courses.

Exercise 3

Enter:

```
levene fmi, by(degree)
```

```
                 Analysis of Variance
    Source              SS         df       MS
--------------------------------------------------
Between groups      7.046e+10      4     1.761e+10
 Within groups      7.607e+11   2142     3.551e+08
--------------------------------------------------
    Total           8.312e+11   2146

Levene's T-test for equal variances: T= 49.60   Prob > T = 0.0000
```

We use the similar levene test step as we did in Exercise 2. Because the Levene's Test is significant (`p < .05`), we reject the homogeneity of variance assumption. Thus, we can conduct a unequal variance for ANOVA using Welch test. Enter:

```
wtest fmi degree
```

```
---------------------------------------------------------------------
Dependent Variable is fmi and Independent Variable is degree
WStat(   5,   53.69) = 112.493, p= 0.0000
---------------------------------------------------------------------
```

Because `p < .05`, we conclude that differences in average income among educational groups are statistically significant.

We now explore the nature of those differences using Games and Howell post hoc tests. Enter:

```
pwmc fmi, over(degree)
```

fmi	Diff.	Std.Err	Games and Howell [95% Conf.	Interval]
degree				
1 vs 0	11030.53	1336.397	7371.043	14690.01
2 vs 0	17556.62	2347.361	11111.68	24001.55
3 vs 0	33489.39	2021.345	27960.12	39018.66
4 vs 0	40276.84	2619.964	33089.86	47463.83
2 vs 1	6526.09	2215.962	434.4419	12617.74
3 vs 1	22458.86	1867.141	17350.72	27567
4 vs 1	29246.31	2502.916	22374.6	36118.03
3 vs 2	15932.77	2685.212	8576.881	23288.66
4 vs 2	22720.22	3160.496	14061.06	31379.39
4 vs 3	6787.456	2926.511	-1226.858	14801.77

The **Multiple Comparisons** table provides the results of the Games and Howell post hoc test. The values for the variable `degree`: 0=**Less than High School**, 1=**High School**, 2=**Junior College**, 3=**Bachelor**, 4=**Graduate**.

To decide significance:if confidence interval contains 0.00, the groups are not significantly different, but if it does not contain 0.00 then the groups are significantly different.

- Group Less than High School is significantly different from all other groups at 0.05 level.
- Group High School is significantly different from all other groups at 0.05 level.
- Group Junior college is significantly different from all other groups at 0.05 level.
- Group Bachelor is significantly different from all other groups except group Graduate(and vice versa).

Exercise 4

Correlation with Pairwise Deletion. Enter:

```
pwcorr educ faos moos, obs
```

	educ	faos	moos
educ	1.0000 2345		
faos	0.2614 1840	1.0000 1842	
moos	0.2691 1656	0.2358 1226	1.0000 1657

Correlation with Listwise Deletion. Enter:

```
correlate educ faos moos
```

	educ	faos	moos
educ	1.0000		
faos	0.2470	1.0000	
moos	0.2424	0.2357	1.0000

Educational attainment (*educ*) is positively correlated with mother's (*moos*) and father's (*faos*) occupational prestige score index (PSI). Mother's and father's PSI are also positively correlated. The direction of the relationship does not change much if you use either pairwise or listwise deletion. The strength of the relationship does change slightly depending on the treatment of missing cases.

Exercise 5

Enter:

```
regress wordsum age
```

Source	SS	df	MS
Model	46.9785059	1	46.9785059
Residual	6254.04747	1,538	4.06635076
Total	6301.02597	1,539	4.0942339

Number of obs	=	1,540
F(1, 1538)	=	11.55
Prob > F	=	0.0007
R-squared	=	0.0075
Adj R-squared	=	0.0068
Root MSE	=	2.0165

wordsum	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0097889	.0028799	3.40	0.001	.0041398	.0154379
_cons	5.426053	.152916	35.48	0.000	5.126107	5.725998

- 0.75% of the variation in vocabulary test score is explained by age.
- Because the F-statistic is statistically significant, age has a significant positive prediction with vocabulary test score.
- The regression coefficient for the constant is 5.426 and the coefficient for the independent variable *age* is 0.01. The predicted vocabulary test score will increase by 0.01 for each additional year of age.
- The effect of *age* on *wordsum* is statistically significant at the 0.01 level.

Exercise 6

Enter:

```
regress pres moos faos Male
```

Source	SS	df	MS			
Model	12090.2463	3	4030.08208	Number of obs	=	1,191
Residual	204646.148	1,187	172.406191	F(3, 1187)	=	23.38
				Prob > F	=	0.0000
Total	216736.395	1,190	182.131424	R-squared	=	0.0558
				Adj R-squared	=	0.0534
				Root MSE	=	13.13

pres	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
moos	.1587009	.0298944	5.31	0.000	.1000491	.2173527
faos	.1536201	.0302608	5.08	0.000	.0942495	.2129907
Male	-.2437124	.7634073	-0.32	0.750	-1.74149	1.254066
_cons	32.88052	1.691936	19.43	0.000	29.561	36.20004

- 5.58% of the variation in the respondent's occupational prestige is explained by the other variables.
- Because the F-statistic is statistically significant, at least one independent variable of mother's occupational prestige score, father's occupational prestige score, and gender is a significant predictor of the respondent's occupational prestige score.
- The predicted occupational prestige score will increase by 0.1587 for each unit increase of mother's occupational prestige score, 0.1536 for each unit increase of father's occupational prestige score, and −0.244 for being males.
- The effects of *faos* and *moos* are statistically significant at the 0.001 level. The effect of gender (*male*) is not statistically significant.