SPSS II

Inferential Statistics

2026-03-11

After attending this training session, you will be able to run and interpret the following tests in SPSS:

1. Chi-square
2. Independent samples t-test
3. Levene's test for the equality of variances
4. One-way ANOVA
5. Bivariate Statistics: Correlation
6. Bivariate Statistics: Regression
7. Multiple regression

Introduction

Code and Dataset

For this training session, please download the following data:

SPSS Workshop.sav

Recap of SPSS I

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

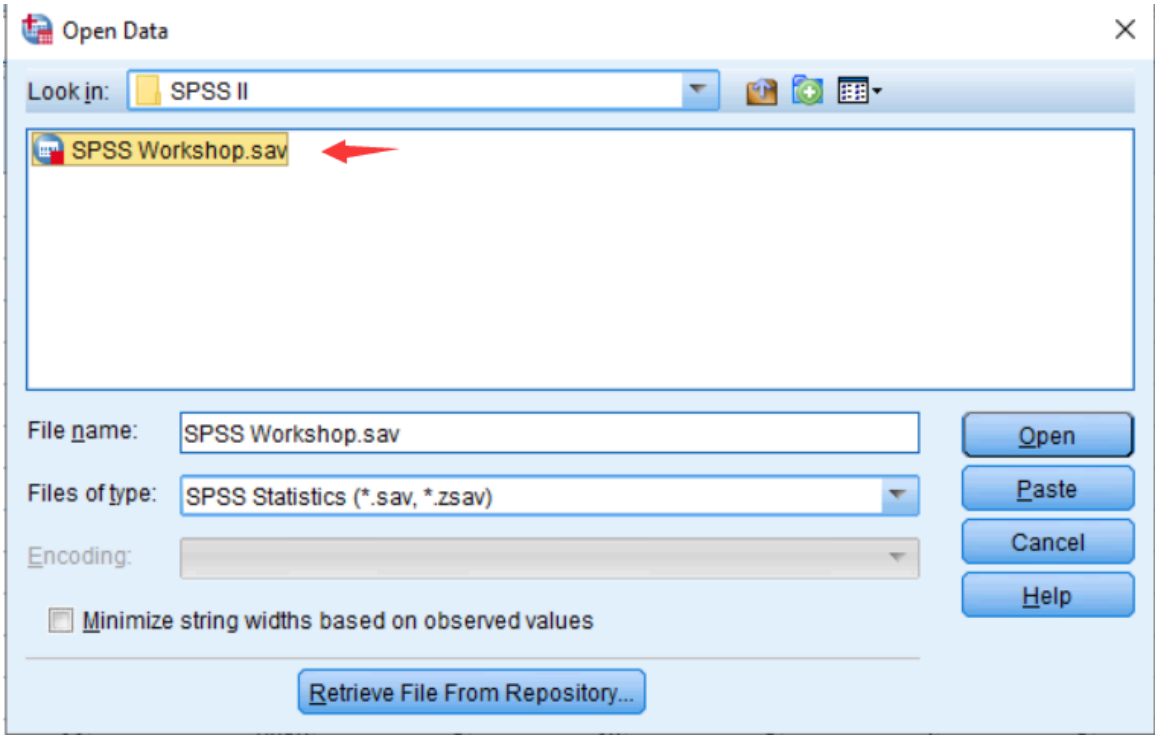
- Starting SPSS
- The SPSS Environment
- Descriptive Statistics
- Data Transformation
- Multivariate Tables
- Graphics
- Saving and Printing

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

Loading the data

1. In the opening SPSS window click **OK** to **Open an existing data source**; an **Open Data** window will appear.

2. Navigate to and open the data file (`SPSS Workshop.sav`) you downloaded from our website. Please, communicate any problem you may have with the data during this training to the consultant or afterward to citl-statstraining@illinois.edu.



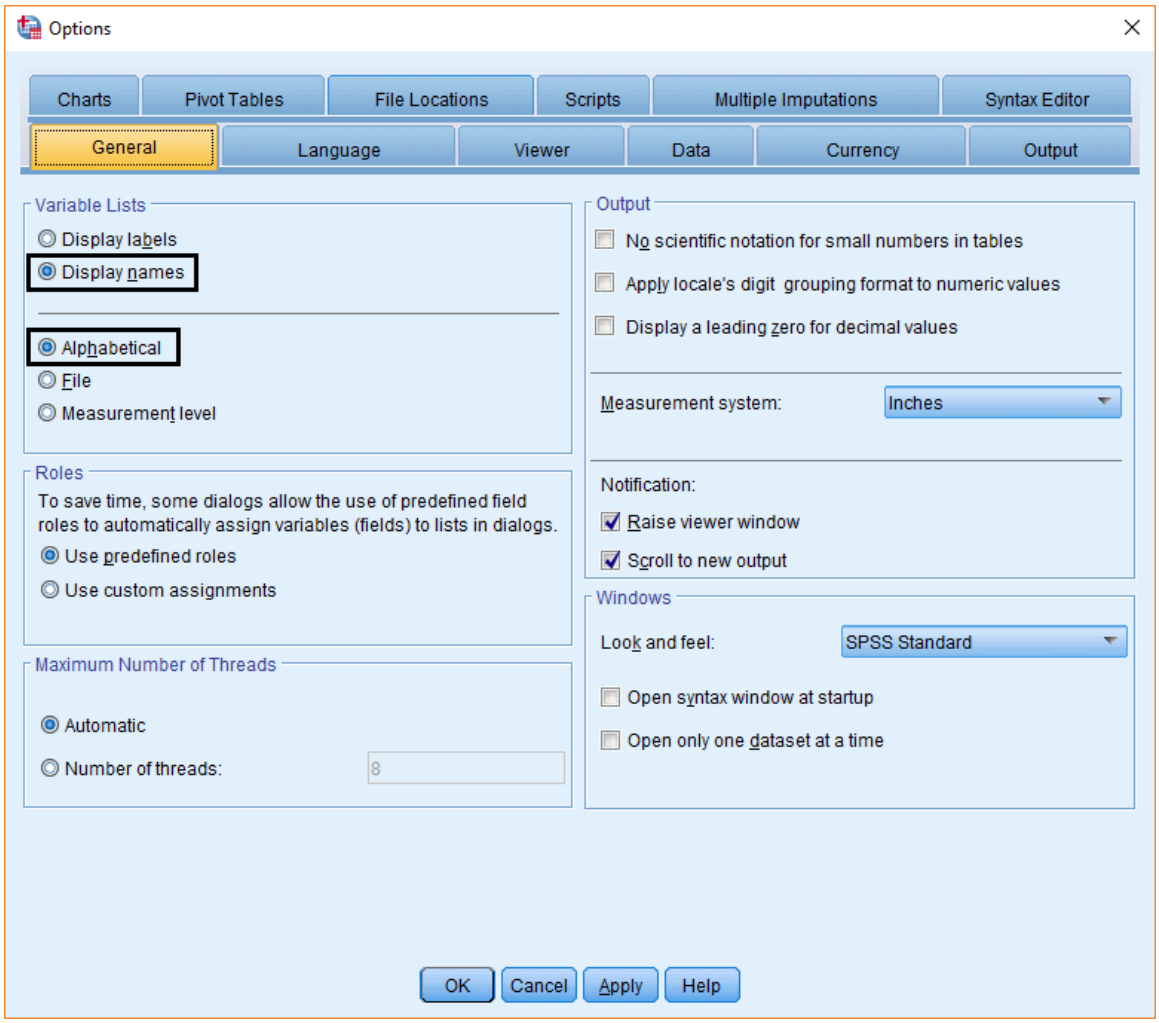
3. If you are unable to locate your data file, it is likely that the file is not in SPSS format. SPSS data files usually end in the suffix **.sav** so SPSS looks for these kinds of files first. If the data file is in a different format, change **Files of type:** to **All files (.)**, and the file should appear. Double-click on the file and SPSS will open it if it is compatible with SPSS.

> **i** Note
>
> Changing the extension of a file name to `.sav` does not convert the data file into an SPSS formatted data file. Rather, it must be imported through SPSS or another program for SPSS to be able to read it. Please contact CITL Data Analytics Group if you need assistance with this task.

Options

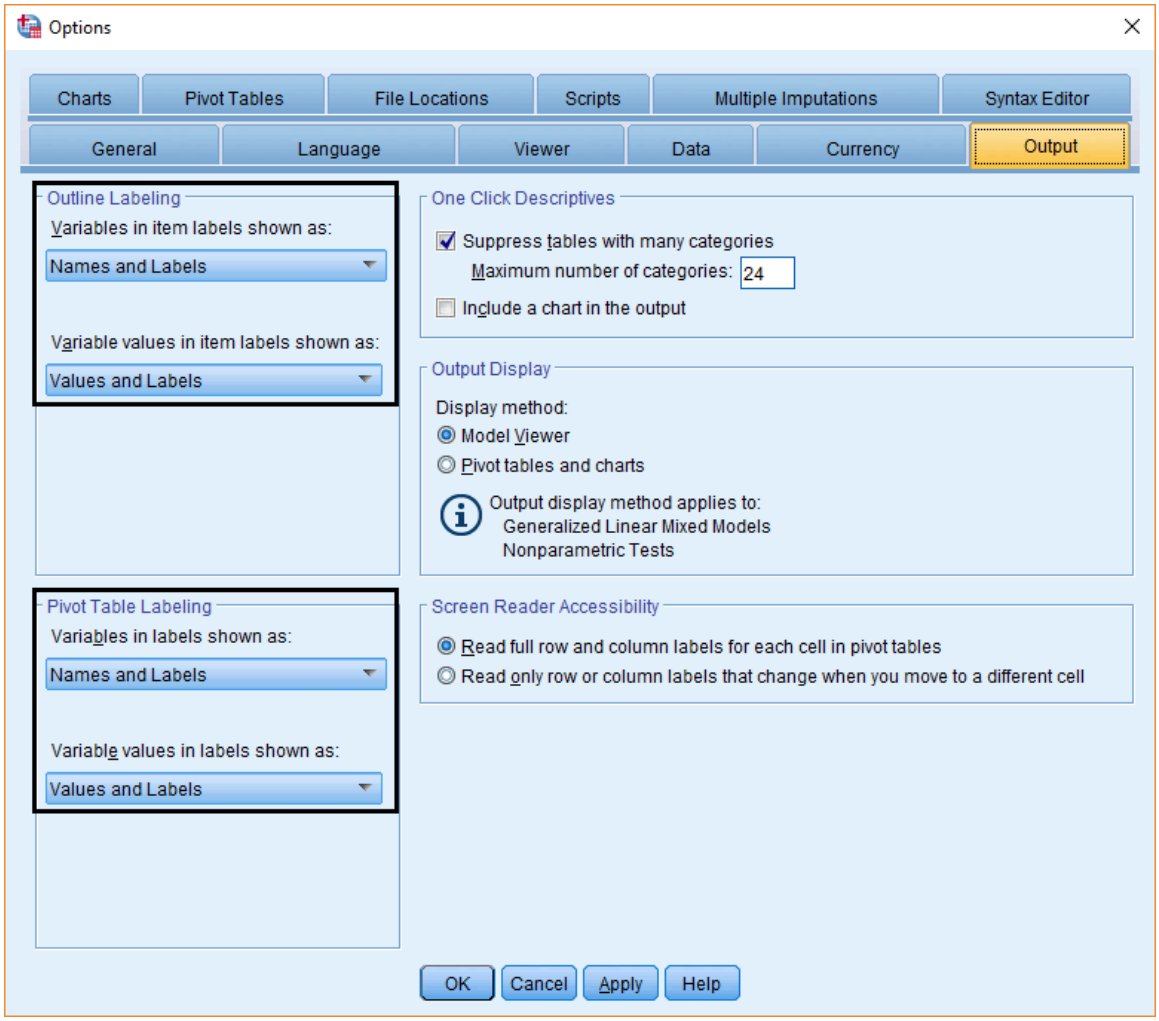
You may have noticed that after opening our data there are two windows popping up. The first window is the **Data Editor** interface, in which the data and variables are presented. The second window is the **Output** interface, in which all results of the computations will be presented. We are concentrating for now on the former as we have to specify some program settings. Let's start changing the following options in the **Data Editor** window:

1. In the **Menu bar**, select **Edit** > **Options**.
2. In the **General** tab, under **Variable Lists**, select **Display Names** and **Alphabetical**.



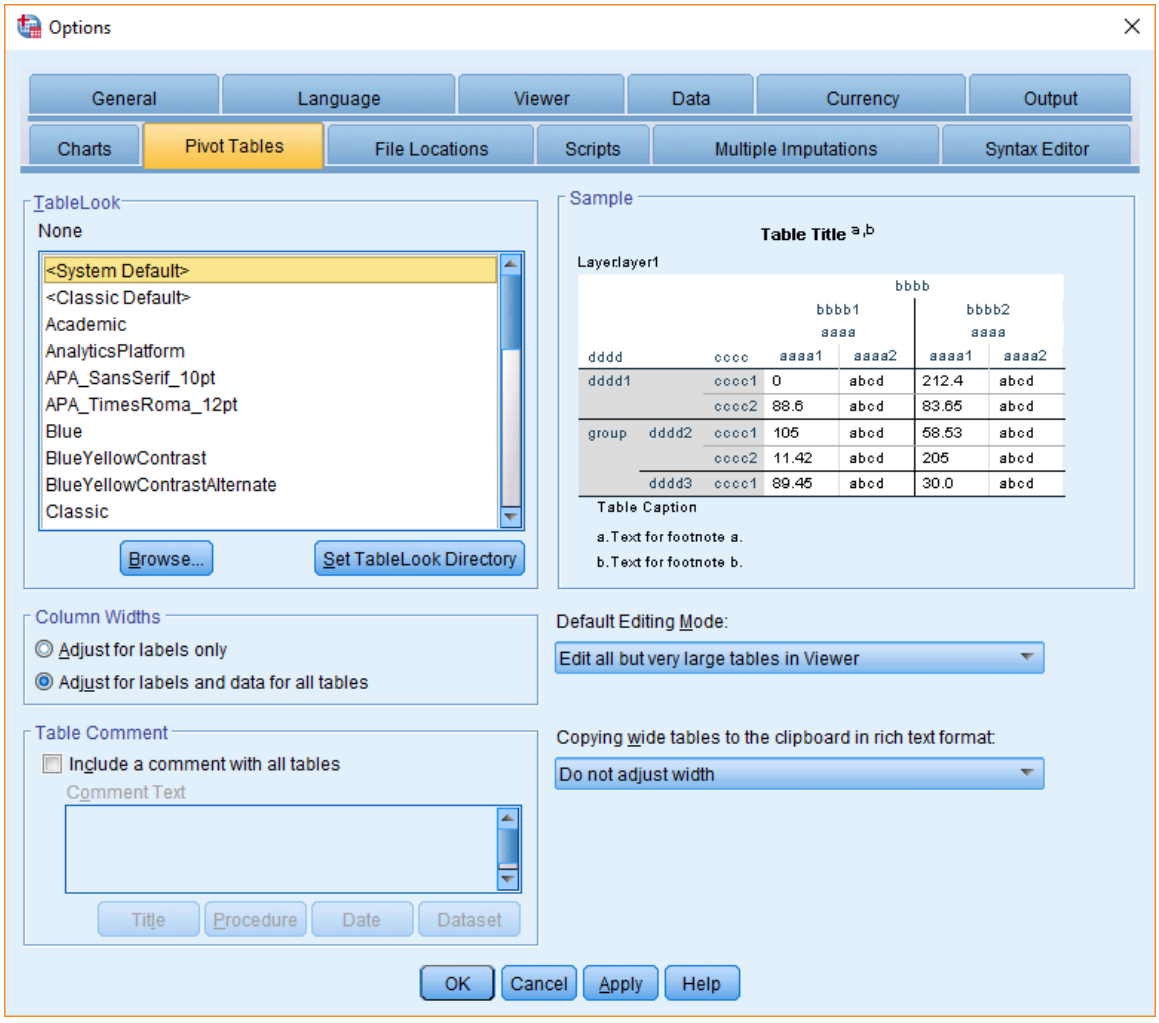
Note: These options change how the variables are displayed in the dialog boxes. They do not change the underlying data file in any way.

3. To change the labeling of the output so that both variable names and labels are displayed in output tables and graphs:

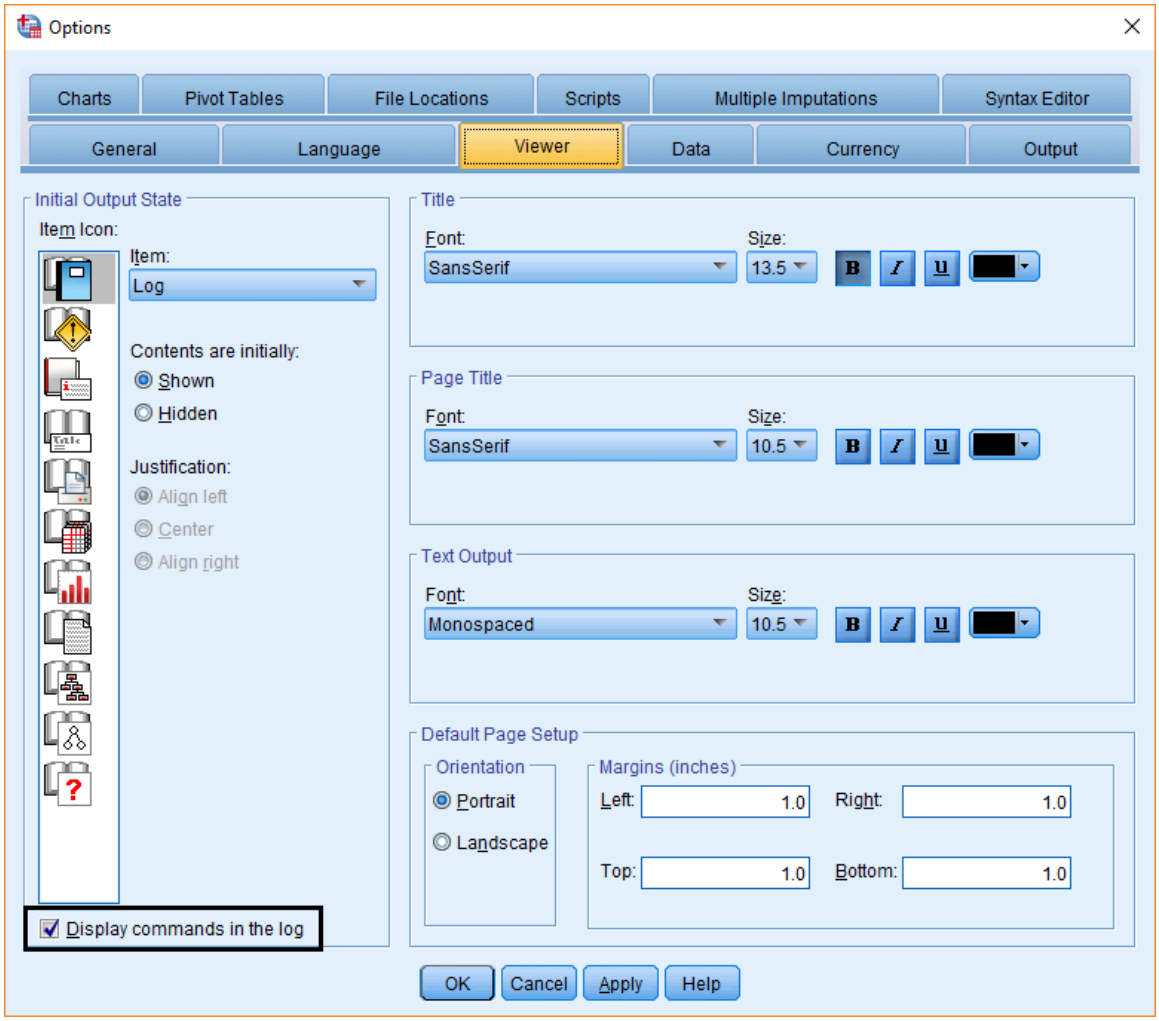


4. Click on the **Output** tab.
5. Under both **Outline Labeling** and **Pivot Table Labeling**, change the pull-down menus to specify **Names and Labels** or **Values and Labels**.

Now, click on the **Pivot Tables** tab and select the **TableLook** setting you prefer.



To display program commands in the output viewer: click on the **Viewer** tab, and select **Display commands in log**. This is useful when documenting your work.



When you are done making changes, click **Apply** and then **Ok** to finish.

Chi-square test

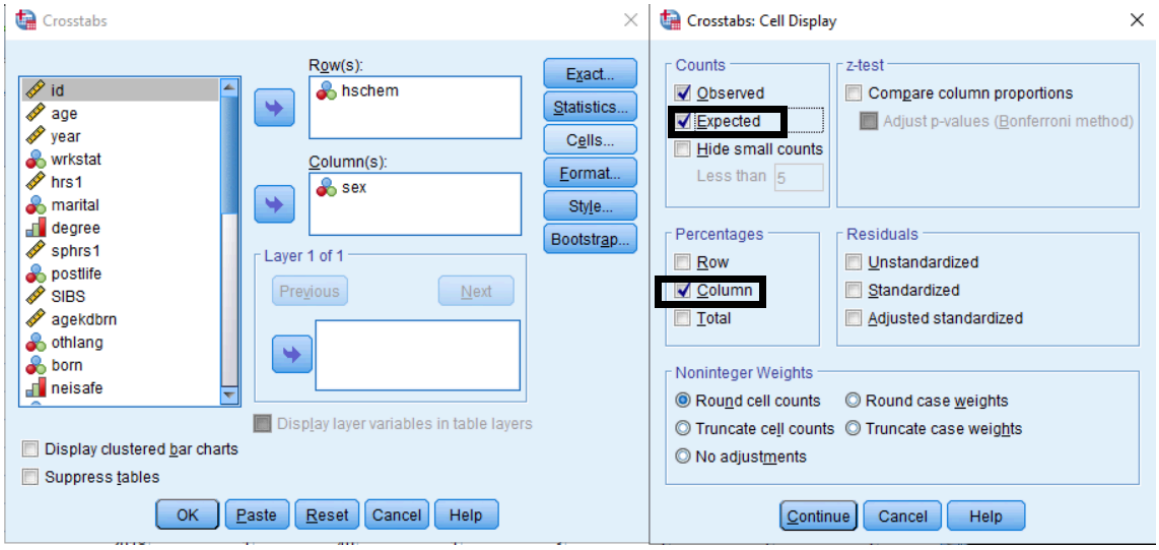
A **chi-square test** is used to determine whether there is a statistically significant association between two categorical variables.

Example 1: 2x2 Table

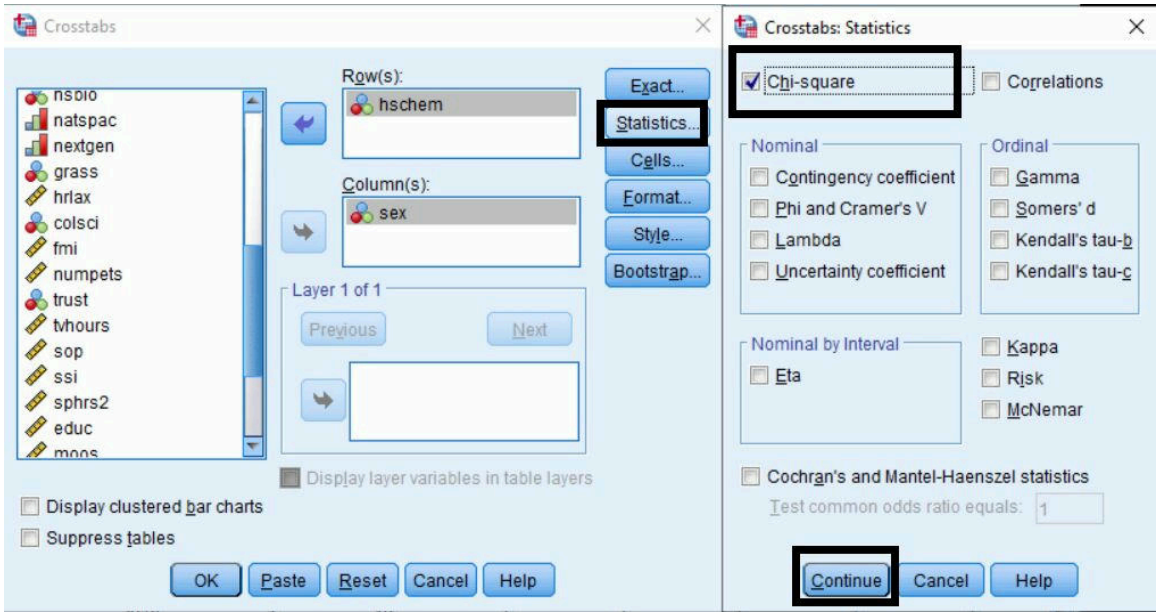
For example, you may want to know if there is a significant association between the variables *sex* (Respondent's gender) and *hschem* (Respondent ever took chemistry class in high school). The chi-square test statistic is requested in the **Crosstabs** procedure.

1. In the **Menu bar**, select **Analyze** > **Descriptive Statistics** > **Crosstabs**.
2. Move the dependent variable *hschem* into **Row(s)** and the independent variable *sex* into **Column(s)**.
3. Click **Cells**, select **Column** for **Percentages** and select **Expected** for **Counts**, and click **Continue**.

When selecting whether you want column or row percentages, select whichever contains the independent variable. In this case, our independent variable *sex* is in the column, so select column percentages.



4. Click **Statistics**, and select **Chi-square**.



5. Click **Continue** and **OK** to finish.

Interpretation of Output: The following three tables are created after requesting a crosstabulation table with the chi-square test statistic.

R ever took a high school chemistry course * Respondents sex Crosstabulation

			Respondents sex		
			MALE	FEMALE	Total
R ever took a high school chemistry course	Yes	Count	271	363	634
		Expected Count	258.8	375.2	634.0
		% within Respondents sex	59.7%	55.2%	57.0%
	No	Count	183	295	478
		Expected Count	195.2	282.8	478.0
		% within Respondents sex	40.3%	44.8%	43.0%
Total		Count	454	658	1112
		Expected Count	454.0	658.0	1112.0
		% within Respondents sex	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	2.244[a]	1	.134		
Continuity Correction[b]	2.063	1	.151		
Likelihood Ratio	2.248	1	.134		
Fisher's Exact Test				.139	.075
Linear-by-Linear Association	2.242	1	.134		
N of Valid Cases	1112				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 195.15.

b. Computed only for a 2x2 table

SPSS generates the following test statistics for chi-square:

> **i** Note
>
> The levels of measurement, the number of categories within the variables, and the sample size will determine which test statistic is appropriate.

- **Pearson 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$

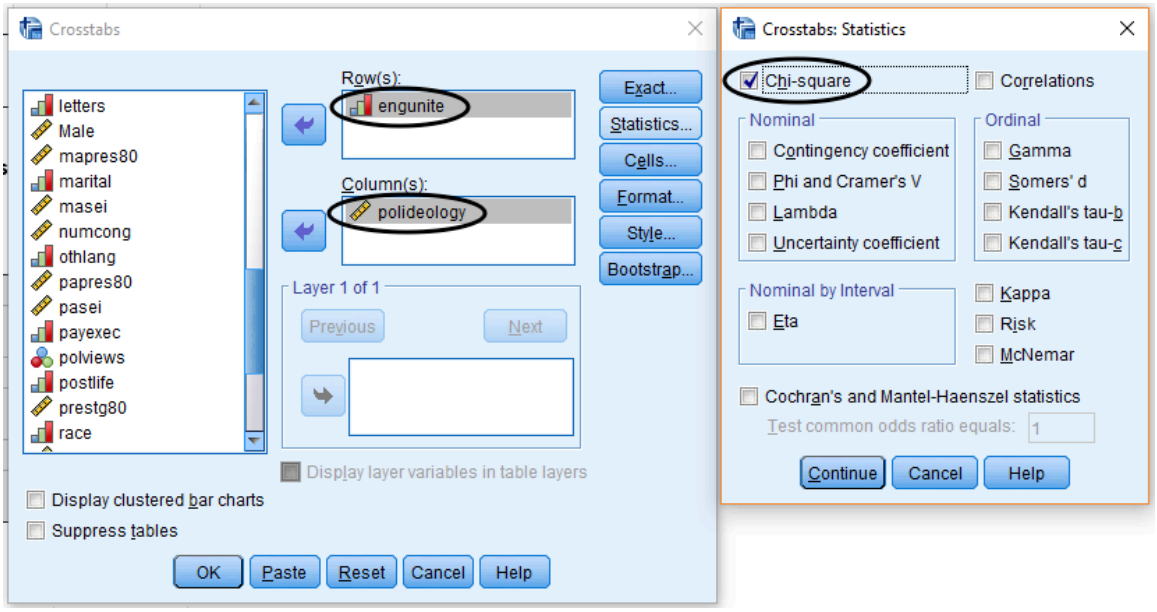
- **Continuity Correction**: This correction is applied in the computation of chi-square for 2x2 tables to improve the approximation. Corrected chi-square values are always smaller than uncorrected values.
- **Likelihood Ratio**: A goodness-of-fit statistic similar to Pearson's chi-square, this statistic is appropriate to report for smaller sample sizes. The likelihood ratio approaches the Pearson's chi-square as *n* increases, so for large sample sizes, the two statistics are the same and either can be reported.
- **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.
- **Linear-by-Linear Association**: A measure of the linear association between the row and column variables. It is also referred to as the Mantel-Haenszel chi-square and the Mantel-Haenszel test for linear association. *Note: This statistic should not be used for nominal data.*

Because this example uses a 2X2 table, we will use **Fisher's Exact** test. Here, the **Fisher's Exact** test has a probability of 1.139 that the null hypothesis is true, so we fail to reject the null hypothesis (using two-tail, as gender's value doesn't represent the actual quantity). Thus, we can conclude that there is NO statistically significant relationship between the respondents' sex and taking a chemistry class in high school.

Example 2: More than 2X2 Table

In this example, we want to know if there is a significant association between the dependent variable *natspac* (attitude towards the space exploration program) and the independent variable *colsci* (whether respondents took college-level science courses). It should produce a crosstab analysis for a 2X3 Table.

1. In the **Menu bar**, select **Analyze > Descriptive Statistics > Crosstabs**.
2. Move the dependent variable *nextspac* into **Row(s)** and the independent variable *colsci* into **Column(s)**.
3. Click **Cells**, select **Column** for **Percentages** and and select **Expected** for **Counts**, and click **Continue**.
4. Click **Statistics**, and select **Chi-square**.



5. Click **Continue** and **OK** to finish.

Interpretation of Output:

Space exploration program * R has taken any college-level sci course Crosstabulation

			R has taken any college-level sci course		
			Yes	No	Total
Space exploration program	TOO LITTLE	Count	96	53	149
		Expected Count	69.2	79.8	149.0
		% within R has taken any college-level sci course	30.3%	14.5%	21.8%
	ABOUT RIGHT	Count	160	188	348
		Expected Count	161.5	186.5	348.0
		% within R has taken any college-level sci course	50.5%	51.4%	51.0%
	TOO MUCH	Count	61	125	186
		Expected Count	86.3	99.7	186.0
		% within R has taken any college-level sci course	19.2%	34.2%	27.2%
Total		Count	317	366	683
		Expected Count	317.0	366.0	683.0
		% within R has taken any college-level sci course	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	33.340[a]	2	.000
Likelihood Ratio	33.803	2	.000
Linear-by-Linear Association	32.818	1	.000
N of Valid Cases	683		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 69.16.

Because 0 cells have expected counts less than 5, it is appropriate to use the **Pearson Chi-Square** test statistic in this example. The level of significance is 0.000. Therefore, it may be concluded that there is a statistically significant relationship between attitudes towards the space exploration program and those who took college level science classes.

> Tip
>
> Measures of the strength of the association can be requested in addition to the `Chi-Square statistic` in the `Crosstabs: Statistics` dialog box.

Exercise 1

Perform a chi-square test to determine if the variable *degree* (Respondent's highest degree) is significantly associated with the variable *grass* (belief that marijuana should be legal). 1. How many

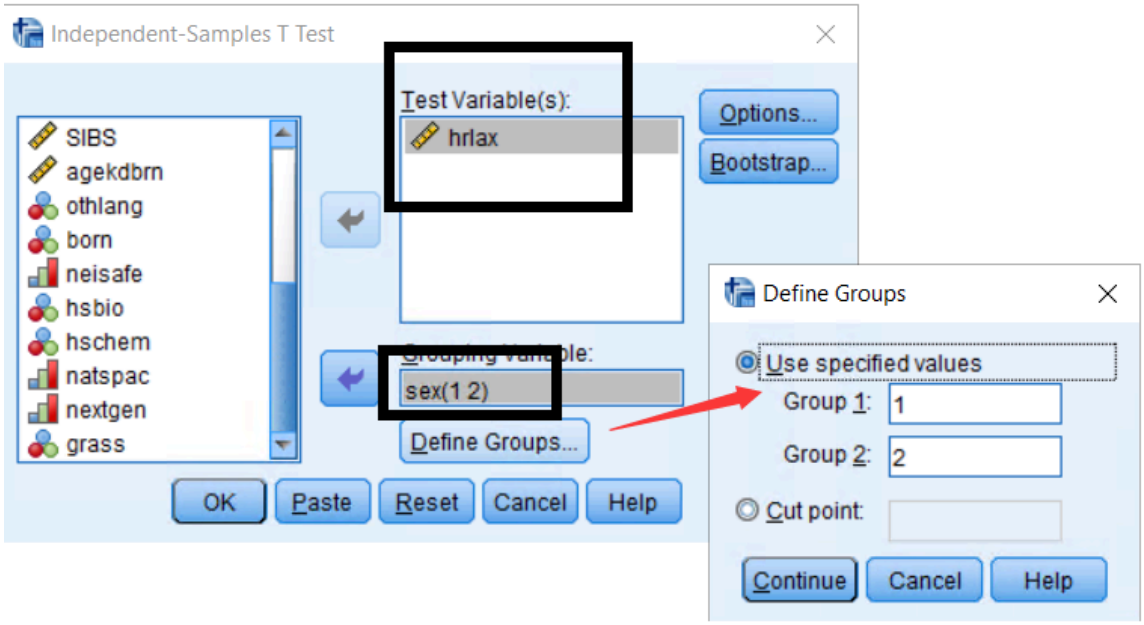
high school graduates said “Yes” to legalization? 2. Is there a statistically significant relationship between the two variables? Which chi-square test statistic is appropriate?

(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*). 1. In the **Menu bar**, select **Analyze > Compare Means > Independent Sample T Test**. 2. Place the dependent variable *hrlax* into the **Test Variable(s)** box. 3. Place the variable *sex* in the **Grouping Variable** box.



4. Click on **Define Groups** and set **Group 1 = 1** (Male) and **Group 2 = 2** (Female).
5. Click **Continue** and **OK**.

> **Note**
>
> To determine how the grouping variable is coded, generate a frequency table of the variable or in the `Independent-Samples T Test` dialog box, right-click the grouping variable, and select **Variable Information**.

The two independent samples t-test generates two output tables.

Group Statistics

	Respondents sex	N	Mean	Std. Deviation	Std. Error Mean
Hours per day R have to relax	MALE	676	3.97	2.836	.109
	FEMALE	729	3.50	2.714	.101

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Hours per day R have to relax	Equal variances assumed	.069	.792	3.200	1403	.001	.474	.148	.183	.764
	Equal variances not assumed			3.195	1383.327	.001	.474	.148	.183	.765

The **Group Statistics** table provides descriptive statistics of the independent variable for each group. According to the statistics in this box, males have to relax 3.97 hours per day while women have to relax 3.50 hours per day.

The **Independent Samples Test** table is used to decide whether there is a statistically significant difference by sex in beliefs about how much corporate heads make.

There are two sets of statistics generated by SPSS:

- **Equal variances assumed**: Assumes that the variances in the distribution of the dependent variable are equal for both groups.
- **Equal variances not assumed**: Does not assume that the variances in the distributions are equal.

To determine which statistics are appropriate, check the significance level for **Levene's Test for Equality of Variances**, located in the output table for the independent samples t-test. This tests the null hypothesis that the variances are equal.

Because the significance of Levene's test in this example is larger than 0.05, we fail to reject the null hypothesis, and the assumption of equal variance is met. In such a case, the statistics for **Equal variances assumed** should be used.

The significance level of the t-statistic is less than 0.05, therefore you may conclude that there is a statistically significant difference in the number of hours men and women have to relax per day.

Exercise 2

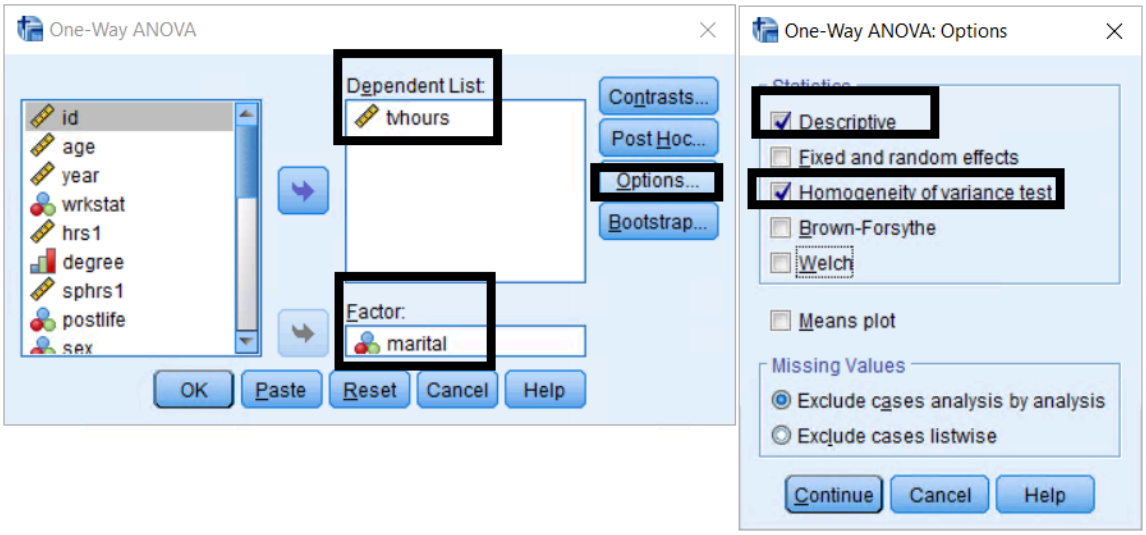
Find out if there is a statistically significant difference in terms of family income (*fmi*)between respondents who took college-level science courses and those who didn't (*colsci*). (See Appendix A for answers.)

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 more than two groups.

For example, this procedure may be used to find out if there is a statistically significant difference in the number of hours spent watching TV (*tvhours*) based on marital status (*marital*).

1. In the **Menu bar**, select **Analyze > Compare Means > One-Way ANOVA**.
2. Place the dependent variable *tvhours* into the **Dependent** List and the independent variable *marital* into the **Factor** box.



3. Click on **Options**, and select **Descriptive** and **Homogeneity of Variance Test**.
4. Click **Continue** and **OK** to finish.

First Step of Interpreting ANOVA Output: Levene Test

Descriptives

Hours per day watching TV

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
MARRIED	678	2.58	2.208	.085	2.41	2.75	0	20
WIDOWED	136	4.03	3.313	.284	3.47	4.59	0	24
DIVORCED	284	3.13	3.079	.183	2.77	3.49	0	20
SEPARATED	39	3.13	3.861	.618	1.88	4.38	0	16
NEVER MARRIED	417	3.01	3.183	.156	2.71	3.32	0	24
Total	1554	2.94	2.838	.072	2.80	3.08	0	24

Test of Homogeneity of Variances

Hours per day watching TV

Levene Statistic	df1	df2	Sig.
8.150	4	1549	.000

Test of Homogeneity of Variances

tvhours HOURS PER DAY WATCHING TV

Levene Statistic	df1	df2	Sig.
.806	3	1820	.491

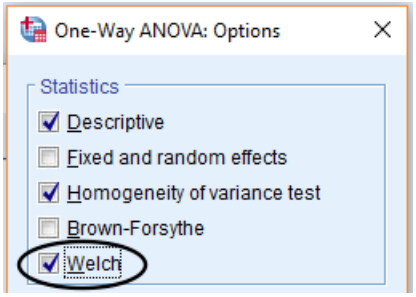
ANOVA

tvhours HOURS PER DAY WATCHING TV

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	213.064	3	71.021	10.793	.000
Within Groups	11975.776	1820	6.580		
Total	12188.840	1823			

As the significant threshold is **0.05** generally, the **Levene's Test of Homogeneity of Variances** is smaller than this threshold and thus significant ($p = 0.000$). As seen in the second output table, we reject the null hypothesis that the variances are equal, and not-equal variances are assumed. This is important in determining which F-test will be used and which post hoc tests will be best suited for any further analysis.

Second Step of Interpreting ANOVA Output: Choose Test



When Levene's test shows that the variances are not homogeneous, choose the Welch test of equality of means to report ANOVA results. To request this test, select **Options** in the **One-Way ANOVA** dialog box, select **Welch**, and click **Continue**. This will generate the **Robust Tests of Equality of Means** table, which gives a more accurate significance level than the default F-test. In this case, $p = 0.000$ for Welch Test, therefore we conclude that there are significant differences in the number of hours spent watching TV based on marital status. If the variances are found to be homogeneous following the Levene's test, we should otherwise choose the regular ANOVA table with homogeneous options.

ANOVA

Hours per day watching TV

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	263.412	4	65.853	8.331	.000
Within Groups	12243.534	1549	7.904		
Total	12506.945	1553			

Robust Tests of Equality of Means

Hours per day watching TV

	Statistic[a]	df1	df2	Sig.
Welch	7.642	4	217.388	.000

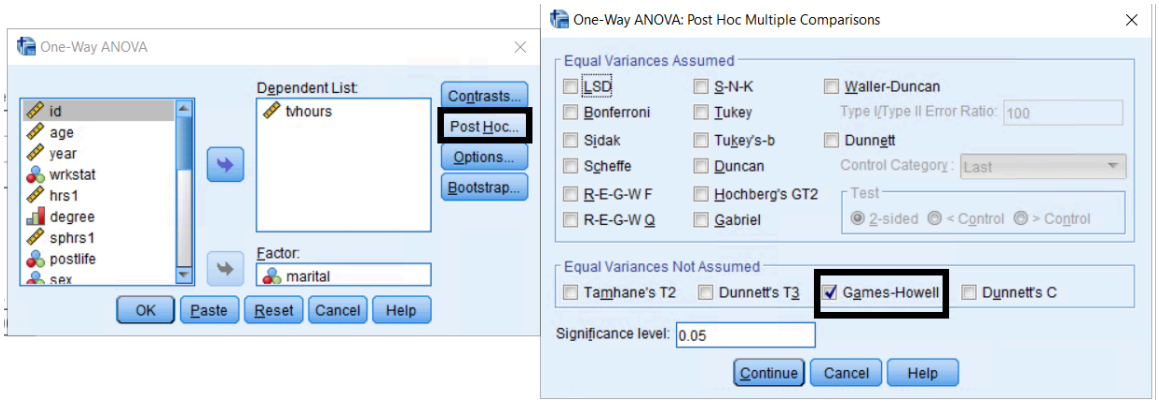
a. Asymptotically F distributed.

ANOVA Post Hoc Tests

While the ANOVA procedure itself determines whether differences exist among the group means, post hoc tests are needed to determine which means differ, and which post hoc test you use depends on the homogeneity of the variances.

As we found above, when looking at differences in number of TV hours watching per day (*tvhours*) by marital status in people (*marital*), equal variances cannot be assumed. Therefore, we must select a post hoc test that assumes not-equal variances.

1. In the **Menu bar**, select **Analyze > Compare Means > One-Way ANOVA**.
2. In the **One-Way ANOVA** dialog box, click **Post Hoc**, and select **Games-Howell**.



3. **Games-Howell** is a pairwise test – Pairwise tests compare the differences between each pair of means – with the null hypothesis that the paired means are equal.
4. Click **Continue** and **OK** to finish.

> **Note**
>
> If equal variances can be assumed, you should select the Tukey or Tukey's-b post hoc test.

Multiple Comparisons

Post Hoc Tests

Multiple Comparisons

Dependent Variable: Hours per day watching TV

Games-Howell

(I) Marital status	(J) Marital status	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
MARRIED	WIDOWED	-1.450*	.296	.000	-2.27	-.63
	DIVORCED	-.551	.201	.051	-1.10	.00
	SEPARATED	-.549	.624	.903	-2.33	1.23
	NEVER MARRIED	-.435	.177	.104	-.92	.05
WIDOWED	MARRIED	1.450*	.296	.000	.63	2.27
	DIVORCED	.899	.338	.063	-.03	1.83
	SEPARATED	.901	.680	.677	-1.02	2.82
	NEVER MARRIED	1.015*	.324	.017	.12	1.91
DIVORCED	MARRIED	.551	.201	.051	.00	1.10
	WIDOWED	-.899	.338	.063	-1.83	.03
	SEPARATED	.002	.645	1.000	-1.83	1.83
	NEVER MARRIED	.116	.240	.989	-.54	.77
SEPARATED	MARRIED	.549	.624	.903	-1.23	2.33
	WIDOWED	-.901	.680	.677	-2.82	1.02
	DIVORCED	-.002	.645	1.000	-1.83	1.83
	NEVER MARRIED	.114	.638	1.000	-1.70	1.93
NEVER MARRIED	MARRIED	.435	.177	.104	-.05	.92
	WIDOWED	-1.015*	.324	.017	-1.91	-.12
	DIVORCED	-.116	.240	.989	-.77	.54
	SEPARATED	-.114	.638	1.000	-1.93	1.70

*. The mean difference is significant at the 0.05 level.

The **Multiple Comparisons** table provides the results of the **Games-Howell** post hoc test.

From this table it may be concluded that:

1. Group Married mean is significantly different from the group Widowed mean.
2. Group Widowed mean is significantly different from group Never Married mean.
3. All the other groups based on marital status have no significant difference from each other.

After identifying statistically significant differences, look back at the table of descriptive statistics to determine which marital groups spend 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.

Descriptives

Hours per day watching TV

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
MARRIED	678	2.58	2.208	.085	2.41	2.75	0	20
WIDOWED	136	4.03	3.313	.284	3.47	4.59	0	24
DIVORCED	284	3.13	3.079	.183	2.77	3.49	0	20
SEPARATED	39	3.13	3.861	.618	1.88	4.38	0	16
NEVER MARRIED	417	3.01	3.183	.156	2.71	3.32	0	24
Total	1554	2.94	2.838	.072	2.80	3.08	0	24

Homogeneous Subsets

tvhours HOURS PER DAY WATCHING TV

	AgeGrouped Grouped age (Recoded)	N	Subset for alpha = 0.05	
			1	2
Tukey HSD[a,b]	1 teens and 20s	335	2.76	
	2 30s and 40s	824	2.78	
	3 50s and 60s	442	3.06	
	4 70 and above	223		3.83
	Sig.		.397	1.000
Tukey B[a,b]	1 teens and 20s	335	2.76	
	2 30s and 40s	824	2.78	
	3 50s and 60s	442	3.06	
	4 70 and above	223		3.83

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 385.450.
b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

Exercise 3

Using the **One-Way ANOVA procedure**, find out whether the mean level of family income (*fmi*) is different among different education groups (*degree*). Remember to check the homogeneity of variances. (See Appendix A for answers.)

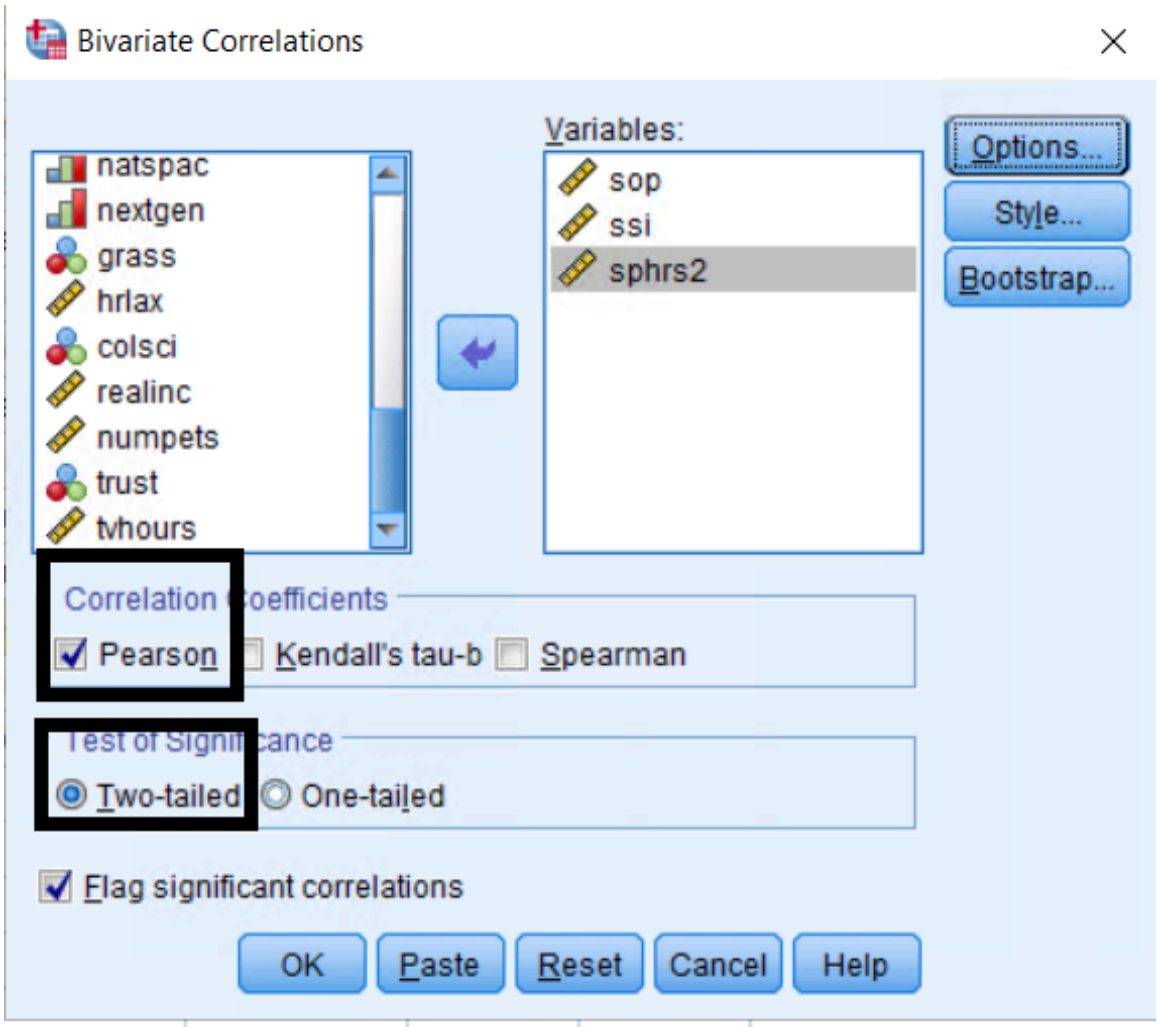
Bivariate Correlation

Correlation is a measure of 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. On the other hand, 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 hours the spouse usually works a week (*sphrs2*).

> **i** Note
>
> A bivariate correlation is used to determine the relationship between only two continuous variables.

1. In the **Menu bar**, select **Analyze > Correlate > Bivariate**.
2. In the **Bivariate Correlations** dialog box, move the variables *sop*, *ssi*, and *sphrs2* to the **Variables:** box.



3. Click **OK** to finish.

Correlations

		Spouse occupational prestige score	Spouse socioeconomic index score	No. of hrs spouse usually works a week
Spouse occupational prestige score	Pearson Correlation	1	.833**	-.152
	Sig. (2-tailed)		.000	.488
	N	1169	1169	23
Spouse socioeconomic index score	Pearson Correlation	.833**	1	-.094
	Sig. (2-tailed)	.000		.668
	N	1169	1169	23
No. of hrs spouse usually works a week	Pearson Correlation	-.152	-.094	1
	Sig. (2-tailed)	.488	.668	
	N	23	23	23

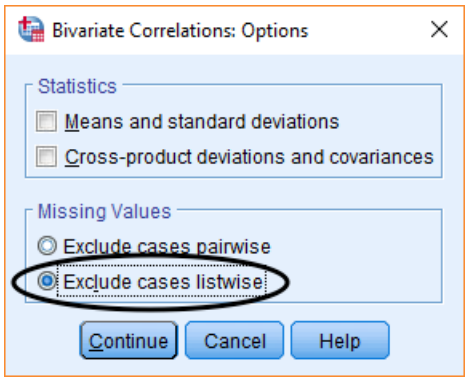
**. Correlation is significant at the 0.01 level (2-tailed).

Each cell contains:

- a correlation coefficient (**Pearson Correlation**),
- the p-value associated with the correlation coefficient (**Sig.**),
- the number of cases (**N**).

The footnote below the correlation table defines the level of significance associated with an asterisk - the more asterisks displayed, the higher the level of significance.

In this example, the correlation coefficient between *sop* and *ssi* is 0.833 and the significance of the coefficient is 0.000, which suggests that these two variables are significantly and positively correlated.



Note on the number of cases (**N**): According to the third row of the cell for the correlation of *sop* and *ssi*, 1,169 valid cases are used. The correlation of *sphrs2* and *ssi* or *sop* only used 23 valid cases.

The default treatment of missing cases in the **Correlation** procedure is to use the maximum number of non-missing values for each pair of variables. This is known as **pairwise deletion**.

An alternative treatment of missing cases is **listwise deletion**, which can be enabled by clicking **Options** in the **Bivariate Correlations** dialog box. When this option is selected, cases with a missing value in any of the variables will be excluded from the analysis. Thus, 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. *Note that the interpretation of the correlation coefficient for listwise deletion is the same as that for pairwise deletion.*

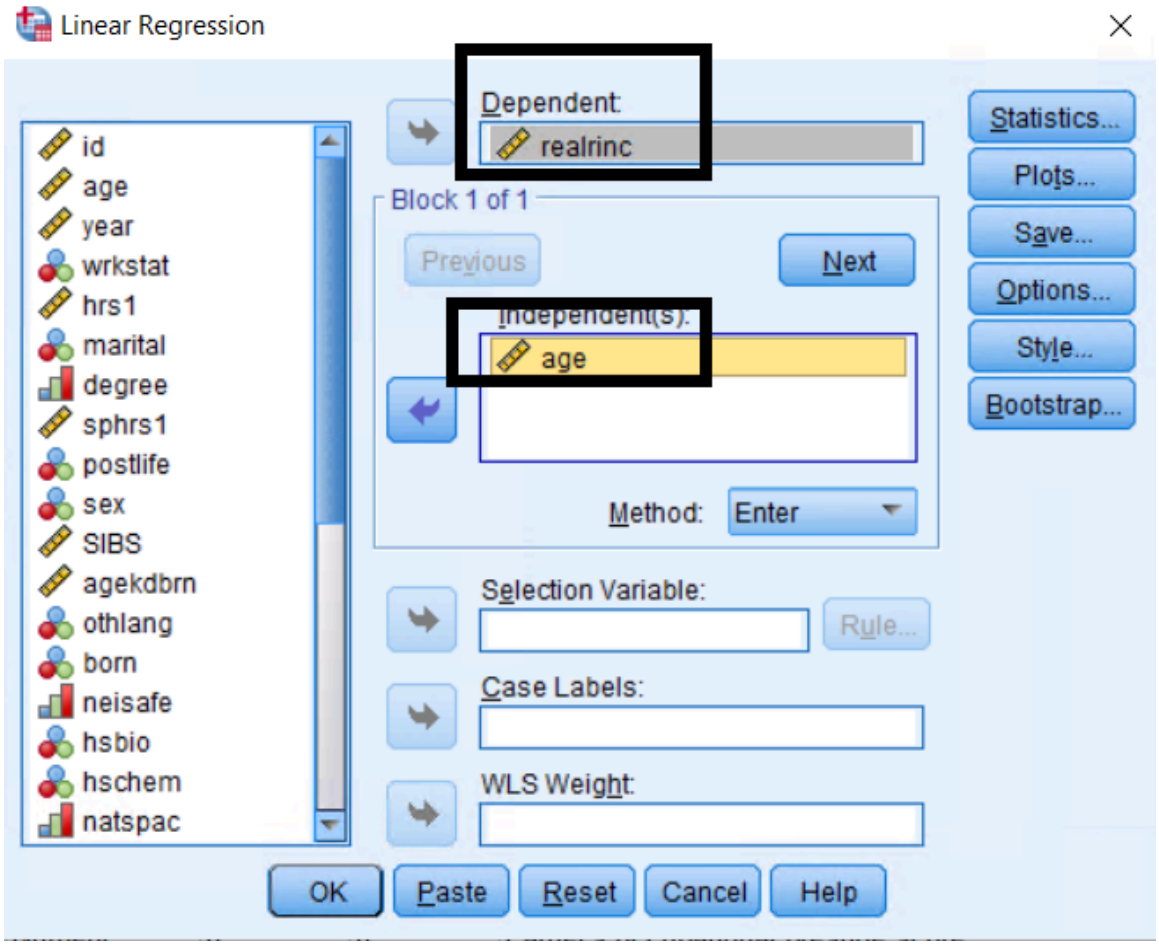
Exercise 4

Measure the association between the variables highest year of school completed (*educ*), respondent's father's occupational prestige index (*faos*) and respondent's mother's occupational prestige index (*moos*), once using pairwise deletion and again using listwise deletion, to compare how the sample size changes. (See Appendix A for answers.)

Bivariate Regression

Bivariate regression is used to assess the effect of a continuous or dichotomous independent variable on a continuous dependent variable. This example uses bivariate regression to measure the effect of respondent's age (*age*) on respondent's income (*realrinc*).

1. In the **Menu bar**, select **Analyze > Regression > Linear**.
2. Place *realrinc* into the **Dependent** box and *age* into the **Independent(s)** box.



3. Click **OK** to finish.

The regression procedure generates four tables: **Variables Entered/Removed**, **Model Summary**, **ANOVA**, and **Coefficients**.

- The **Variables Entered/Removed** table lists the independent variable included in the model. Only one independent variable, *age*, was included in this model.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Age of respondent[b]	.	Enter

a. Dependent Variable: R's income in constant $

b. All requested variables entered.

- The **Model Summary** table provides information about the overall fit of the model.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.182[a]	.033	.033	28447.874

a. Predictors: (Constant), Age of respondent

- **R** (aka Pearson's R) is the correlation between the dependent variable and independent variable. In this example, income and age are positively correlated.
- **R square** (aka the coefficient of determination) tells you the proportion of variation in the dependent variable explained by the independent variable. R square ranges from 0 (no effect of an independent variable on a dependent variable) to 1 (perfect prediction of dependent variable by given independent variable).
  - In this example, 3.3% of the variation in income (dependent variable) is explained by the age (independent variable).
- **Adjusted R square** is adjusted for the number of independent variables used in the model. However, since there is only one independent variable in this example, the R square and Adjusted R square are the same.

The **ANOVA** table measures whether the independent variable reliably predicts the dependent variable. In this example, the **F statistic** is statistically significant ($p < 0.05$). Thus, age is a significant predictor of income.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3.780E+10	1	3.780E+10	46.712	.000[b]
	Residual	1.097E+12	1356	809281516.1		
	Total	1.135E+12	1357			

a. Dependent Variable: R's income in constant $

b. Predictors: (Constant), Age of respondent

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

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	8620.324	2516.285		3.426	.001
	Age of respondent	368.137	53.864	.182	6.835	.000

a. Dependent Variable: R's income in constant $

The regression coefficient for the constant is 8620.324 and the unstandardized coefficients for the independent variable *educ* is 368.137. From this, it can be said that the predicted level of income will increase by $368.137 for each additional year of age. Because this effect is measured in the original scale of the dependent variable, it is referred to as an unstandardized coefficient. Thus, when you have multiple independent variables measured in the different units in the model, you cannot compare the effects of independent variables using unstandardized coefficients. This will be addressed further in the section on Multiple Regression.

The standard error (**Std. Error**) is used to calculate the t-value (**t**). This statistic is used to test the null hypothesis that the unstandardized coefficient is equal to 0. In this example, *age* is statistically significant at the 0.000 level. This means that you can reject the null hypothesis that the age variable's coefficient is equal to 0 and conclude that age is a significant predictor of income.

Exercise 5

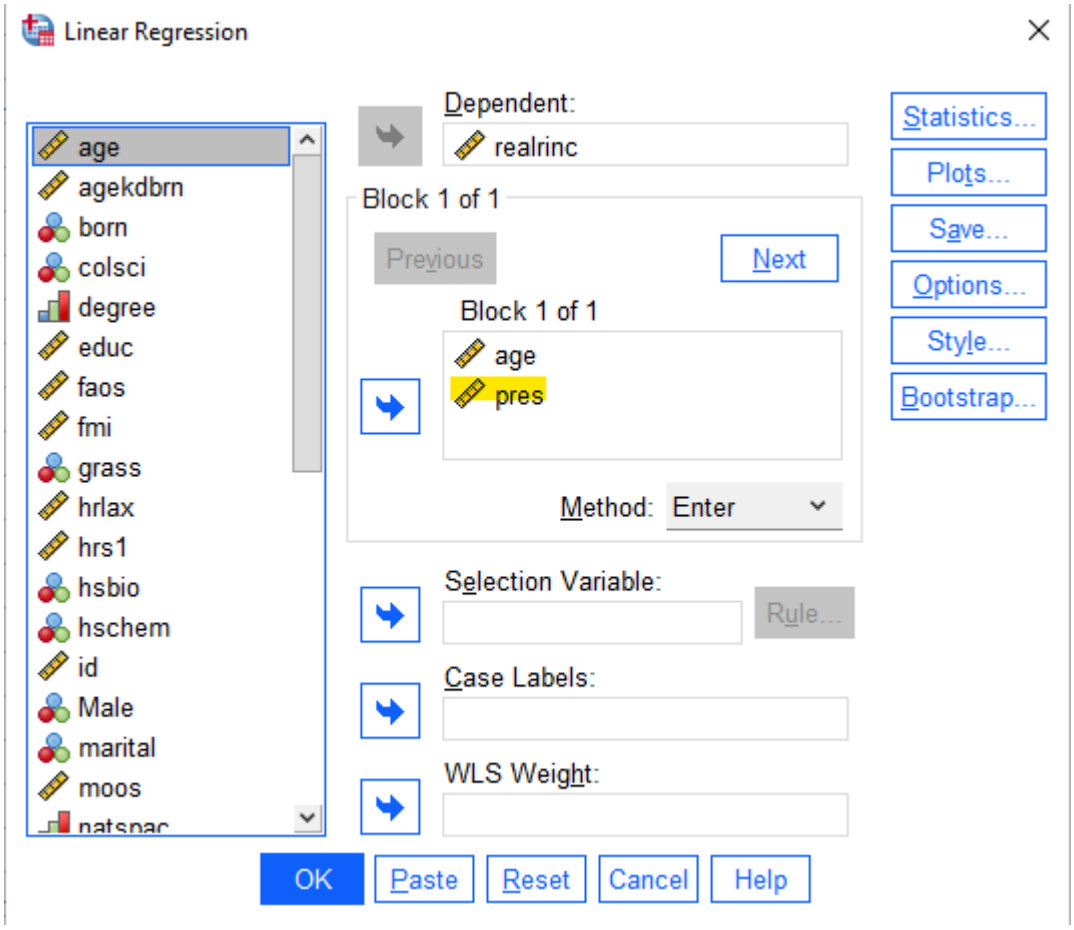
Does independent variable age (*age*) affect dependent variable of performance on a vocabulary test (*wordsum*)? (See Appendix A for answers.)

Multiple Regression

Multiple regression is used to assess the effect of more than one continuous or dichotomous independent variable on another continuous dependent variable. 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 respondent's occupational prestige score (*pres*) and age (*age*) on respondent's income (*realrinc*).

1. In the **Menu bar**, select **Analyze > Regression > Linear**.
2. Place *realrinc* into the **Dependent** box and *pres* and *age* into the **Independent(s)** box.



3. Click **OK** to finish.

The regression procedure generates four tables: **Variables Entered/Removed**, **Model Summary**, **ANOVA**, and **Coefficients**.

- The **Variables Entered/Removed** table lists the independent variables included in the model. Two independent variables, *pres* and *age*, were included in this model.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	pres R's occupational prestige score , age Age of respondent[b]	.	Enter

a. Dependent Variable: realrinc R's income in constant $

b. All requested variables entered.

- The **Model Summary** table provides information about the overall fit of the model. Since the denominator of **R-square** is fixed, each additional variable used in the equation can only increase the size of the numerator. As a result, the introduction of additional variables produces a higher R square value, regardless of the efficiency of the model.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.355[a]	.126	.125	26917.092

a. Predictors: (Constant), pres R's occupational prestige score , age Age of respondent

Since R Square will never decrease by adding more independent variables into the model, some researchers report an **Adjusted R Square**, which corrects this bias by adjusting both the numerator and the denominator by their respective degrees of freedom. Adjusted R Square is calculated using the equation below, where n is the number of observations and k is the number of independent variables.

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

Unlike R Square, the Adjusted R Square can decline in value. It is important to keep in mind, however, that the R Square value is a proportion, while the Adjusted R Square is not. Therefore, after adding one independent variable, the Adjusted R Square may decrease while the R Square may increase. If the Adjusted R Square is much lower than the R Square, this usually indicates that some relevant explanatory variables are not specified in the model.

In this example, compared to the bivariate model in the previous section, both R Square and Adjusted R Square have increased. Hence, this new model explains more variation of the dependent variable than the old model with one independent variable.

- The **ANOVA** table measures whether the independent variables significantly predicts the dependent variable. In this example, the **F** statistic is below 0.05. Thus, respondent's occupational prestige score and age are significant predictors of income.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.412E+11	2	7.060E+10	97.445	<.001[b]
	Residual	9.767E+11	1348	724529840.3		
	Total	1.118E+12	1350			

a. Dependent Variable: realrinc R's income in constant $

b. Predictors: (Constant), pres R's occupational prestige score , age Age of respondent

- The **Coefficients** table provides the values of the unstandardized regression coefficients (**B**), standardized coefficients (**Beta**), t-test statistics (**t**), and p-values (**Sig.**). It shows that a unit increase in *age* will increase predicted income by $285.290 while holding the occupational prestige score constant. The t-test statistic is 5.534, and the p-value associated with the t-test statistic is .001. Thus, *age* is statistically significant while controlling for *pres*. Likewise, a unit increase in *pres* will decrease the predicted income by $635.071 while holding *age* constant. The t-test statistic is 11.981, and the p-value associated with the t-test statistic is .001. Hence, *pres* is statistically significant while controlling for *age*.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-16646.225	3186.716		-5.224	<.001
	age Age of respondent	285.290	51.555	.142	5.534	<.001
	pres R's occupational prestige score	635.071	53.005	.308	11.981	<.001

a. Dependent Variable: realrinc R's income in constant $

To compare the magnitude of the effect of each independent variable, the coefficients must be standardized using the same scale. The standardized coefficient (**Beta**) measures change in the dependent variable (measured in standard deviations) per one standard deviation change in the independent variable. Predicted income will increase by 0.142 standard deviation units for each one standard deviation increase in *age* while it will increase 0.308 standard deviation units for

each one standard deviation increase in *pres*. Because the standardized coefficient for *pres* is larger than that for *age* we can conclude that occupational prestige has a larger impact on real income than age.

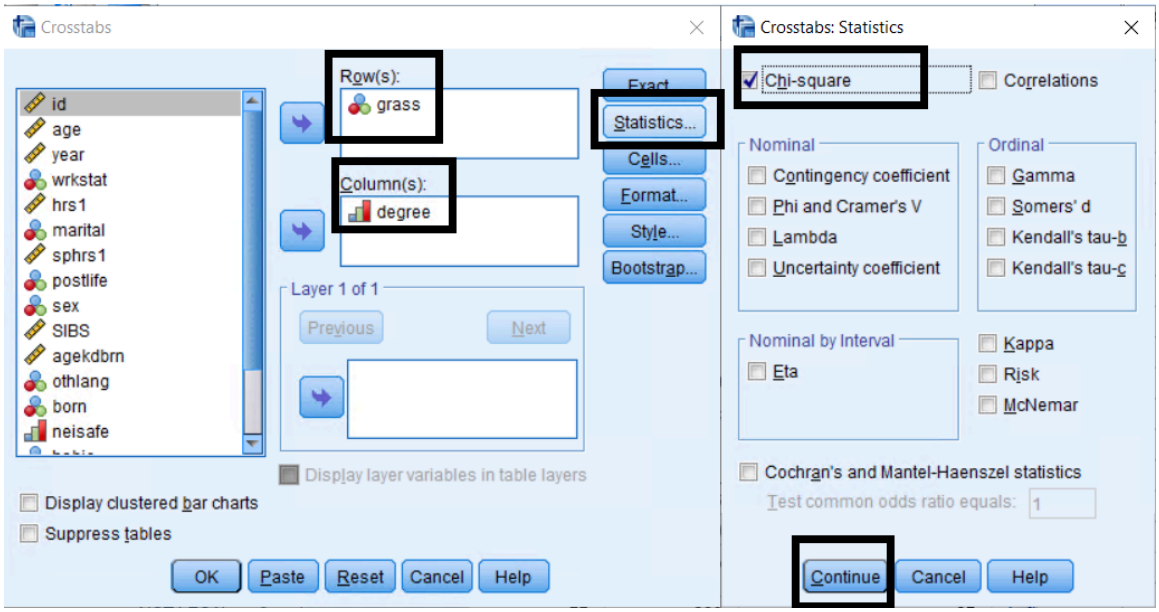
Exercise 6

Does the occupational prestige of a respondent's parents (*faos* and *moos*) and respondent's sex (*Male*) affect the respondent's occupational prestige (*pres*)? (See Appendix A for answers.)

Appendix A: Answers for Exercises

Exercise 1

1. In the **Menu bar**, select **Analyze** > **Descriptive Statistics** > **Crosstabs**.
2. Move the dependent variable *grass* into **Row(s)** and the independent variable *degree* into **Column(s)**.



3. Click **Cells**, select **Column** for **Percentages**, and click **Continue**.
4. Click **Statistics**, and select **Chi-square**.
5. Click **Continue** and **OK** to finish.

Results:

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Should marijuana be made legal * R's highest degree	1447	61.6%	901	38.4%	2348	100.0%

Should marijuana be made legal * R's highest degree Crosstabulation

			R's highest degree					
			LT HIGH SCHOOL	HIGH SCHOOL	JUNIOR COLLEGE	BACHELOR	GRADUATE	Total
Should marijuana be made legal	LEGAL	Count	87	466	85	201	99	938
		Expected Count	106.3	457.0	83.0	186.7	105.0	938.0
		% within R's highest degree	53.0%	66.1%	66.4%	69.8%	61.1%	64.8%
	NOT LEGAL	Count	77	239	43	87	63	509
		Expected Count	57.7	248.0	45.0	101.3	57.0	509.0
		% within R's highest degree	47.0%	33.9%	33.6%	30.2%	38.9%	35.2%
Total		Count	164	705	128	288	162	1447
		Expected Count	164.0	705.0	128.0	288.0	162.0	1447.0
		% within R's highest degree	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	14.712[a]	4	.005
Likelihood Ratio	14.391	4	.006
Linear-by-Linear Association	2.052	1	.152
N of Valid Cases	1447		

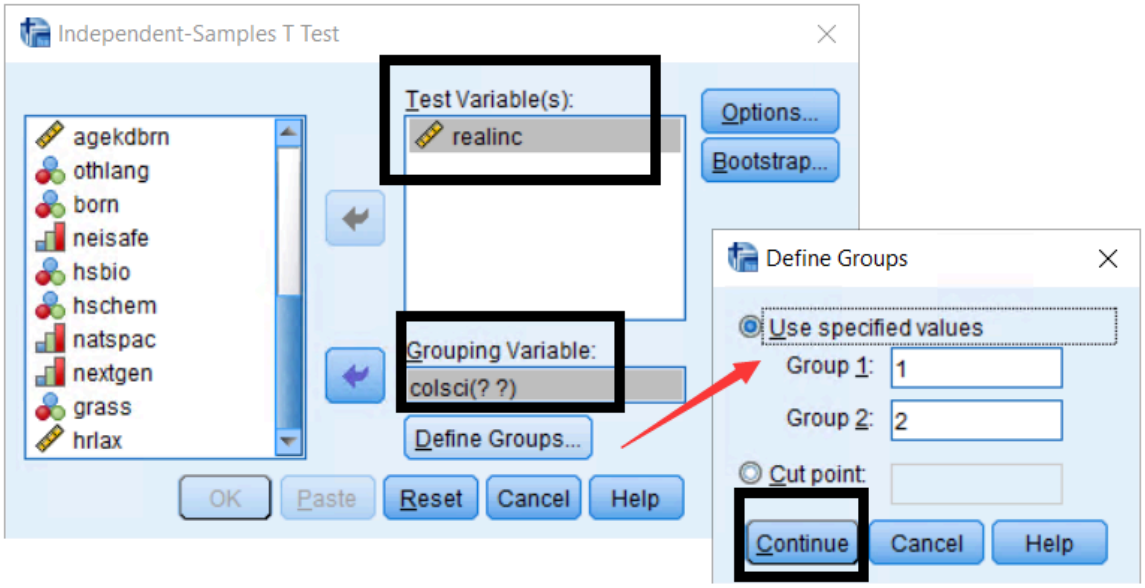
a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 45.03.

1. 466 high school graduates said **Yes** to legalization.

2. All expected cells are larger than 5, so we should use Pearson Chi-Square. The p-value associated with the Pearson chi-square statistic (0.005) is less than 0.05. Thus, there is a statistically significant relationship between the two variables.

Exercise 2

1. In the **Menu bar**, select **Analyze > Compare Means > Independent Sample T Test**.
2. Place the dependent variable *fmi* into the **Test variable(s)** box and the variable *colsci* in the **Grouping variable** box.



3. Click **Define groups** and set **Group 1 = 1** (Yes) and **Group 2 = 2** (No).
4. Click on **Continue**, and **OK**.

Results:

Group Statistics

	R has taken any college-level sci course	N	Mean	Std. Deviation	Std. Error Mean
Family income in constant $	Yes	486	43180.81	34536.592	1566.612
	No	583	24504.81	23563.453	975.899

Independent Samples Test

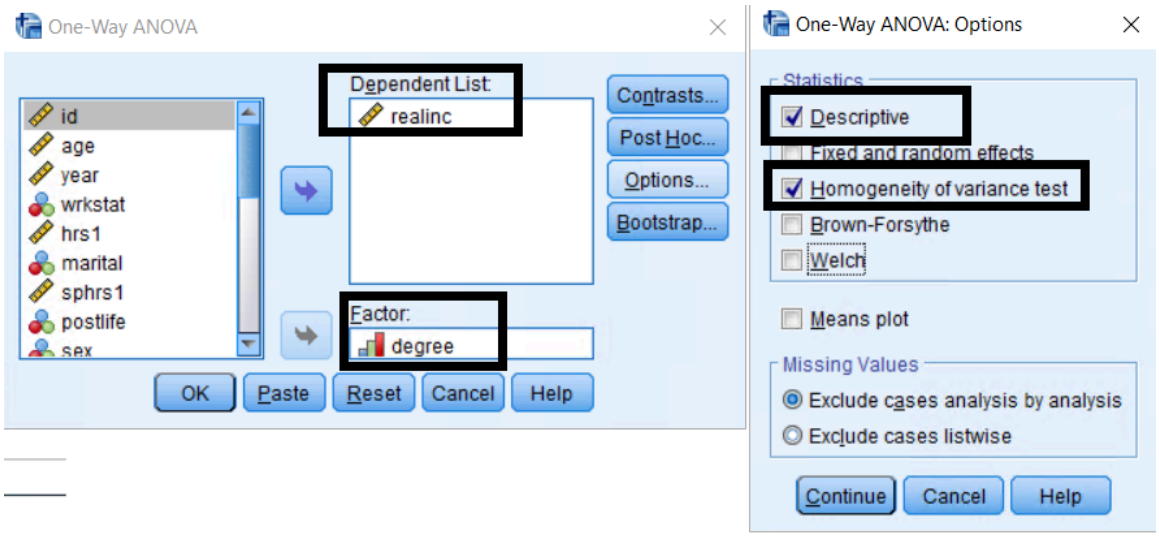
		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Family income in constant $	Equal variances assumed	76.401	.000	10.460	1067	.000	18676.008	1785.549	15172.422	22179.594
	Equal variances not assumed			10.119	830.255	.000	18676.008	1845.711	15053.199	22298.816

- Because the **Levene's Test for Equality of Variances** is significant ($p < 0.05$), the **Equal variances not assumed** test statistic is appropriate to use.

- The **t-test** statistic (10.119) is significant ($p < 0.05$), therefore the amount of family income significantly differs based on if they took or didn't take college-level science courses.

Exercise 3

1. In the **Menu bar**, select **Analyze > Compare Means > One-Way ANOVA**.
2. Place the dependent variable *fmi* into the **Dependent List** and the independent variable *degree* into the **Factor box**.



3. Click **Options** and select **Descriptive** and **Homogeneity of Variance Test**.
4. Click **Continue** and **OK** to finish.

Results:

Descriptives

Family income in constant $

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
LT HIGH SCHOOL	223	15636.13	16309.257	1092.149	13483.82	17788.43	227	119879
HIGH SCHOOL	1076	26569.32	25187.575	767.857	25062.66	28075.99	227	119879
JUNIOR COLLEGE	184	33192.75	28184.858	2077.815	29093.19	37292.30	227	119879
BACHELOR	436	49125.52	35515.750	1700.896	45782.52	52468.51	227	119879
GRADUATE	233	55912.97	36351.629	2381.474	51220.89	60605.05	908	119879
Total	2152	33749.70	31155.618	671.607	32432.64	35066.77	227	119879

Test of Homogeneity of Variances

Family income in constant $

Levene Statistic	df1	df2	Sig.
49.728	4	2147	.000

ANOVA

realinc FAMILY INCOME IN CONSTANT $

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	4.297E+11	4	1.074E+11	109.483	.000
Within Groups	2.396E+12	2442	981233248.1		
Total	2.826E+12	2446			

Because the **Test of Homogeneity of Variances** is significant the homogeneity of variances assumption is not met, and the ANOVA procedure will have to be run a second time, requesting the **Welch test** under **Options** in the **One-way ANOVA** dialog box. This will generate the **Robust Tests of Equality of Means** table.

ANOVA

Family income in constant $

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	3.462E+11	4	8.656E+10	106.700	.000
Within Groups	1.742E+12	2147	811219113.1		
Total	2.088E+12	2151			

Robust Tests of Equality of Means

Family income in constant $

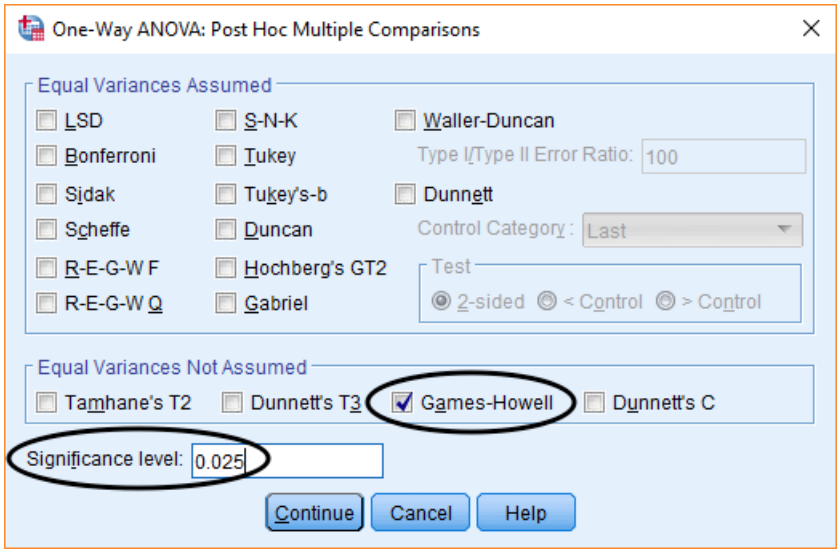
	Statistic[a]	df1	df2	Sig.
Welch	106.997	4	608.763	.000

a. Asymptotically F distributed.

From the output, you can conclude that differences in average income among educational groups are statistically significant.

To conduct a Post Hoc Test to determine which means differ:

1. In the **Menu bar**, select **Analyze > Compare Means > One-Way ANOVA**.
2. Place the dependent variable *fmi* into the **Dependent List** and the independent variable *degree* into the **Factor box**.
3. In the **One-Way ANOVA** dialog box, click **Post Hoc**.
4. Change the **Significance level** from 0.05 to 0.25 for the sake of this example.
5. Because the results of the **Test of Homogeneity of Variances** is significant ($p < 0.05$), you must select the **Games-Howell** post hoc test. This test does not assume equality of variances.



6. Click **Continue** and **OK** to finish.

Results:

The **Multiple Comparisons** table provides the results of the **Games-Howell** post hoc test. From this table it may be concluded that:

Multiple Comparisons

Dependent Variable: Family income in constant $

Games-Howell

(I) R's highest degree	(J) R's highest degree	Mean Difference (I-J)	Std. Error	Sig.	97.5% Confidence Interval	
					Lower Bound	Upper Bound
LT HIGH SCHOOL	HIGH SCHOOL	-10933.194*	1335.063	.000	-14913.97	-6952.42
	JUNIOR COLLEGE	-17556.618*	2347.361	.000	-24578.28	-10534.96
	BACHELOR	-33489.387*	2021.345	.000	-39508.67	-27470.10
	GRADUATE	-40276.843*	2619.964	.000	-48105.46	-32448.22
HIGH SCHOOL	LT HIGH SCHOOL	10933.194*	1335.063	.000	6952.42	14913.97
	JUNIOR COLLEGE	-6623.424	2215.157	.025	-13259.60	12.76
	BACHELOR	-22556.193*	1866.186	.000	-28114.38	-16998.00
	GRADUATE	-29343.649*	2502.204	.000	-36828.14	-21859.15
JUNIOR COLLEGE	LT HIGH SCHOOL	17556.618*	2347.361	.000	10534.96	24578.28
	HIGH SCHOOL	6623.424	2215.157	.025	-12.76	13259.60
	BACHELOR	-15932.769*	2685.212	.000	-23943.03	-7922.51
	GRADUATE	-22720.225*	3160.496	.000	-32150.00	-13290.45
BACHELOR	LT HIGH SCHOOL	33489.387*	2021.345	.000	27470.10	39508.67
	HIGH SCHOOL	22556.193*	1866.186	.000	16998.00	28114.38
	JUNIOR COLLEGE	15932.769*	2685.212	.000	7922.51	23943.03
	GRADUATE	-6787.456	2926.511	.141	-15514.11	1939.20
GRADUATE	LT HIGH SCHOOL	40276.843*	2619.964	.000	32448.22	48105.46
	HIGH SCHOOL	29343.649*	2502.204	.000	21859.15	36828.14
	JUNIOR COLLEGE	22720.225*	3160.496	.000	13290.45	32150.00
	BACHELOR	6787.456	2926.511	.141	-1939.20	15514.11

*. The mean difference is significant at the 0.025 level.

1. Group Less than High School is significantly different from all other groups at the 0.025 level.
2. Group High School is significantly different from all other groups except group Junior College at the 0.025 level.

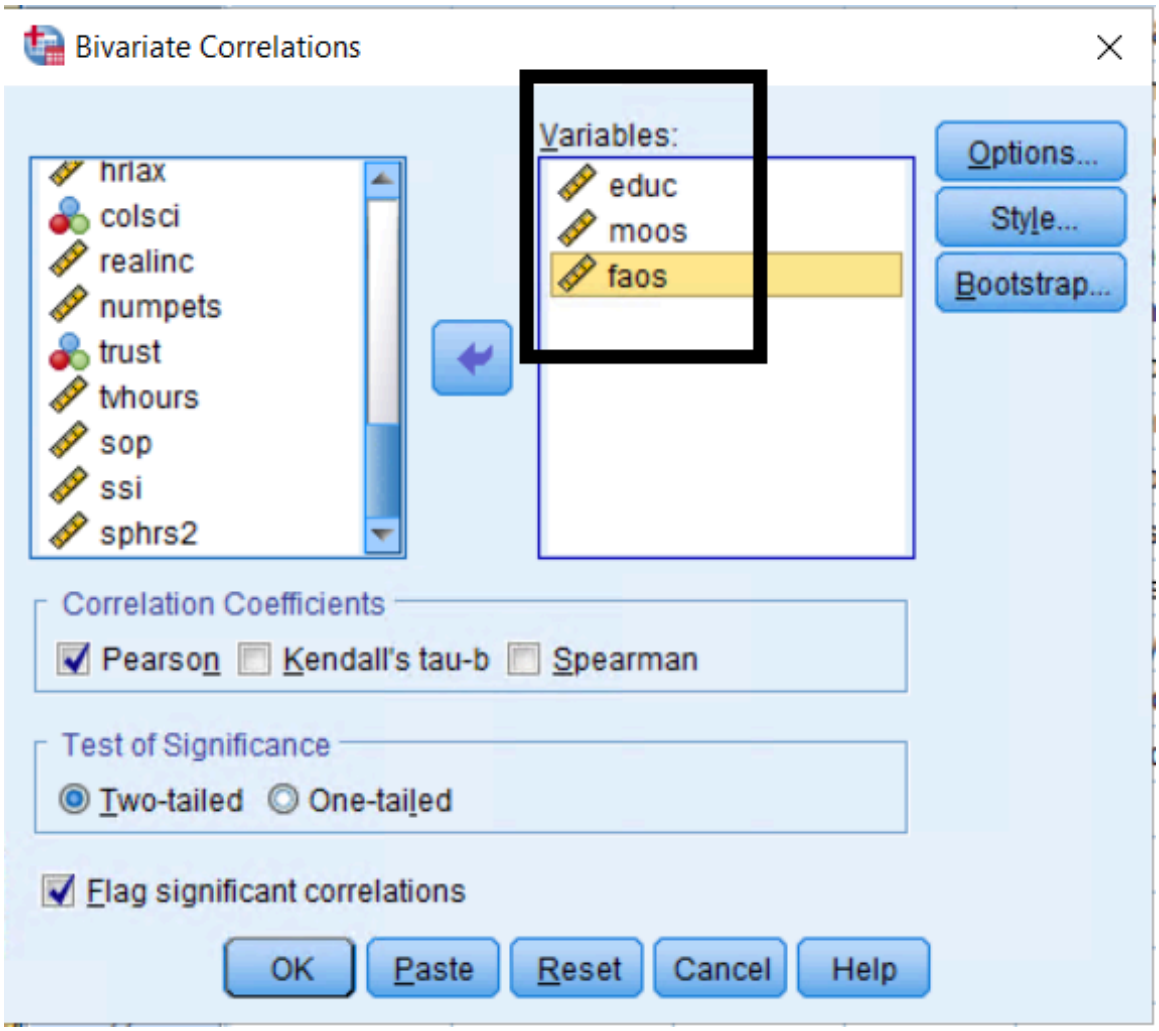
More specifically, those with a high school degree make $10,933.194 more in family income than those with less than a high school education, but the -$6,623.424 difference in family income between those with a high school degree and those with a junior college education is not significant. 3. Group Junior college is significantly different from all other groups except group High school at the 0.025 level. 4. Group Bachelor is significantly different from all other groups except group Graduate (and vice versa) at the 0.025 level.

Exercise 4

Correlation with Pairwise Deletion:

1. In the **Menu bar**, select **Analyze > Correlate > Bivariate**.

2. In the **Bivariate Correlations** dialog box, move the variables *educ*, *faos*, and *moos* to the **Variables box**.



3. Click **Continue** and **OK** to finish.

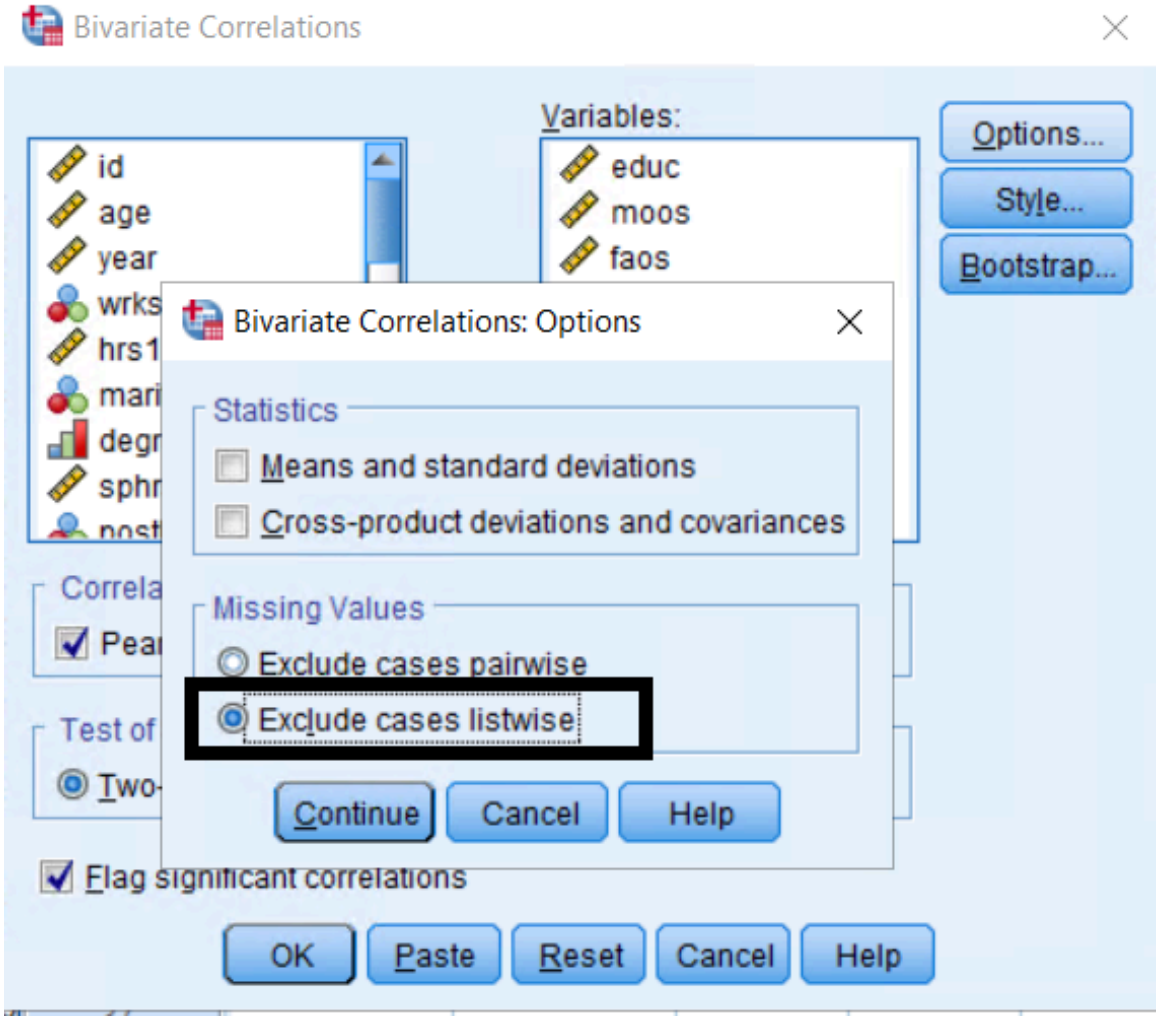
Correlations

		Highest year of school completed	Mothers occupational prestige score	Father's occupational prestige score
Highest year of school completed	Pearson Correlation	1	.269**	.261**
	Sig. (2-tailed)		.000	.000
	N	2345	1656	1840
Mothers occupational prestige score	Pearson Correlation	.269**	1	.236**
	Sig. (2-tailed)	.000		.000
	N	1656	1657	1226
Father's occupational prestige score	Pearson Correlation	.261**	.236**	1
	Sig. (2-tailed)	.000	.000	
	N	1840	1226	1842

**. Correlation is significant at the 0.01 level (2-tailed).

Correlation with Listwise Deletion:

1. In the **Menu bar**, select **Analyze > Correlate > Bivariate**.
2. In the **Bivariate Correlations** dialog box, move the variables *educ*, *faos*, and *moos* to the **Variables box**.



3. Click **Options**, and select **Exclude cases listwise**.
4. Click **Continue** and **OK** to finish.

Correlations[b]

		Highest year of school completed	Mothers occupational prestige score	Father's occupational prestige score
Highest year of school completed	Pearson Correlation	1	.242**	.247**
	Sig. (2-tailed)		.000	.000
Mothers occupational prestige score	Pearson Correlation	.242**	1	.236**
	Sig. (2-tailed)	.000		.000
Father's occupational prestige score	Pearson Correlation	.247**	.236**	1
	Sig. (2-tailed)	.000	.000	

**. Correlation is significant at the 0.01 level (2-tailed).

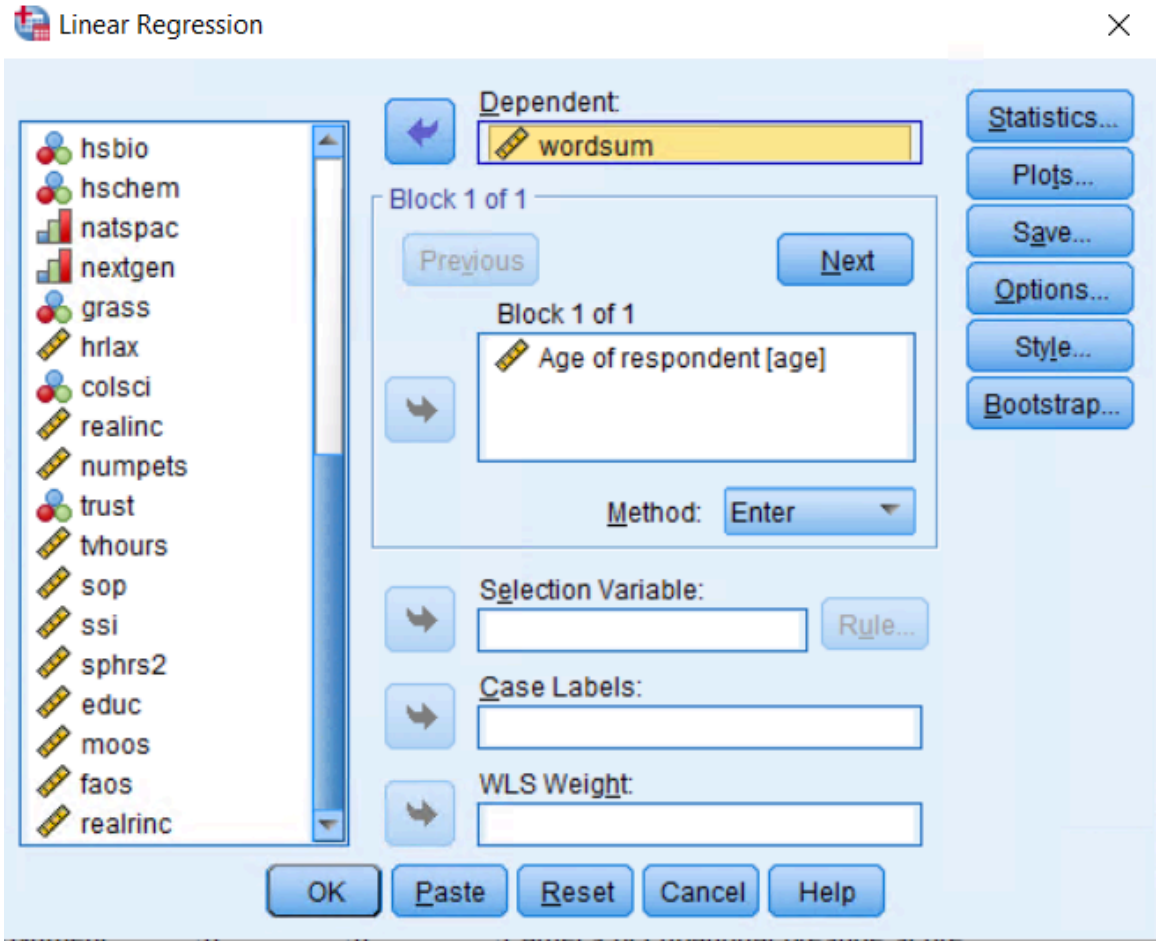
b. Listwise N=1225

Results:

- Educational attainment (*educ*) is significantly and positively correlated with mother's and father's occupational prestige score (*moos* and *faos*) at the 0.01 level.
- Mother's and father's occupational prestige score are also significantly and positively correlated with each other at the 0.01 level.
- While the direction and strength of the relationship do not change significantly if you use either pairwise or listwise deletion, the absolute value of the correlation coefficient does change depending on the treatment of missing cases.

Exercise 5

1. In the **Menu bar**, select **Analyze** > **Regression** > **Linear**.
2. Place *wordsum* into the **Dependent box** and *age* into the **Independent(s) box**.



3. Click **OK** to finish.

Results:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.086[a]	.007	.007	2.017

a. Predictors: (Constant), Age of respondent

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	46.979	1	46.979	11.553	.001[b]
	Residual	6254.047	1538	4.066		
	Total	6301.026	1539			

a. Dependent Variable: Number words correct in vocabulary test

b. Predictors: (Constant), Age of respondent

Coefficients[a]

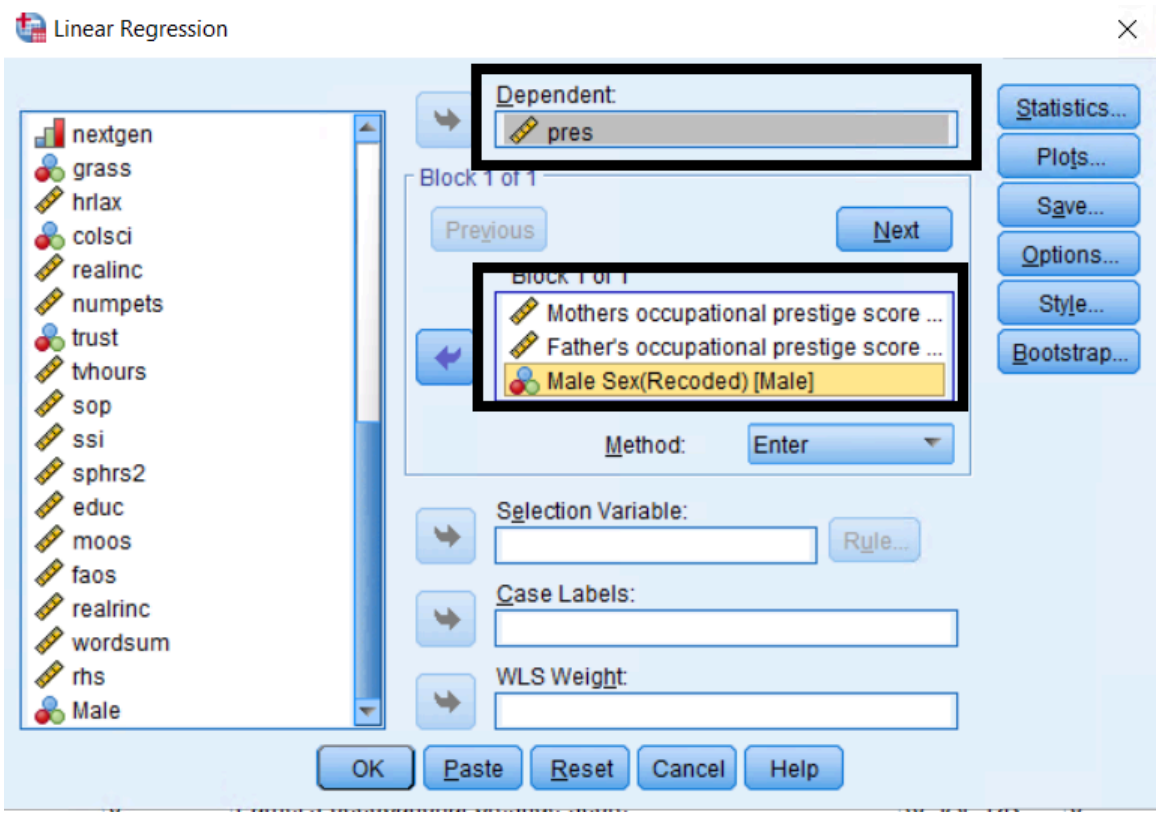
Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	5.426	.153		35.484	.000
	Age of respondent	.010	.003	.086	3.399	.001

a. Dependent Variable: Number words correct in vocabulary test

- The **R Square** is 0.007, meaning age can explain 0.7% of variation in the number of words correct in vocabulary test.
- Because the **F statistic** is statistically significant, age is a significant predictor of vocabulary test score.
- The regression coefficient for the constant is 5.426, and the one for the independent variable *age* is 0.010. The predicted vocabulary test score will increase by 0.010 for each additional year of age.
- The effect of *age* on *wordsum* is statistically significant at the 0.01 level.

Exercise 6

1. In the **Menu bar**, select **Analyze** > **Regression** > **Linear**.
2. Place *pres* into the **Dependent box** and *faos*, *moos*, and *Male* into the **Independent(s) box**.



3. Click **OK** to finish.

Results:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.236[a]	.056	.053	13.130

a. Predictors: (Constant), Male Sex(Recoded), Mothers occupational prestige score , Father's occupational prestige score

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	12090.246	3	4030.082	23.376	.000[b]
	Residual	204646.148	1187	172.406		
	Total	216736.395	1190			

a. Dependent Variable: R's occupational prestige score

b. Predictors: (Constant), Male Sex(Recoded), Mothers occupational prestige score , Father's occupational prestige score

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	32.881	1.692		19.434	.000
	Mothers occupational prestige score	.159	.030	.154	5.309	.000
	Father's occupational prestige score	.154	.030	.147	5.077	.000
	Male Sex(Recoded)	-.244	.763	-.009	-.319	.750

a. Dependent Variable: R's occupational prestige score

- Regression result yields an adjusted R-squared of 0.53 and an **R-squared** of 0.56.It means that 56% of variance in respondent's occupational prestige score can be explained by mother's occupational prestige score, father's occupational prestige score, and sex.
- Because the **F statistic** is statistically significant, mother's occupational prestige score, father's occupational prestige score, and sex are significant predictors of respondent's occupational prestige score.
- The predicted occupational prestige score will increase by 0.159 for each unit increase of mother's occupational prestige score, 0.154 for each unit increase of father's occupational prestige score, and −0.244 for being male.
- The effects of *moos* on *faos* are statistically significant at the 0.001 level. The effect of *sex* (male) is not statistically significant.