Example 2: Bivariate Association and Significance Tests in SAS and STATA

The data for this example were selected from the 2012 General Social Survey dataset featured in Mitchell (2015). This example will demonstrate linear and nonlinear transformations of quantitative variables, Pearson's and Spearman correlations for quantitative and ordinal variables, and cross-tabulations and measures of bivariate association for binary variables.

SAS Syntax for Importing and Preparing Data for Analysis:

Comments (your notes the program will not interpret) are in green and start with * in SAS or // in STATA.

```
* Define placeholder for folder location to be used below;
* \\Client\ precedes path in Virtual Desktop outside H drive;
%LET filesave=C:\Dropbox\21SP_PSQF6242\PSQF6242_Example2;

* IMPORT GSS_Example.xlsx data using filesave reference and exact file name;
* from the Excel workbook in DATAFILE= location from SHEET= ;
* New SAS file is in "work" library place with name "Example2";
* "GETNAMES" reads in the first row as variable names;
* DBMS=XLSX (can also use EXCEL or XLS for .xls files);
PROC IMPORT DATAFILE="&filesave.\GSS_Example.xlsx"
            OUT=work.Example2 DBMS=XLSX REPLACE;
     SHEET="GSS_Example";
     GETNAMES=YES;
RUN;
```

Anything in PINKY-PURPLE is case- and space-sensitive in SAS (but not otherwise).

```
* Create formats: set of value labels for categorical variables;
PROC FORMAT;
       VALUE Fmarry    1="1.Unmarried" 2="2.Married";
       VALUE Fgender   1="1.Man" 2="2.Woman";
       VALUE Fhappy    1="1.Unhappy" 2="2. Neither" 3="3.Fairly Happy"
                           4="4.Very Happy" 5="5.Completely Happy";
        VALUE Ffriends  1="1.Never" 2="2.Once/Year" 3="3. Several/Year"
                           4="4.Once/Month" 5="5.Several/Month"
                           6="6.Several/Week" 7="7.Almost Daily";
RUN;

* DATA = create new dataset, SET = from OLD dataset;
* So DATA + SET means "save as itself" after these actions;
* All data transformations must happen inside a DATA+SET+RUN combo;
DATA work.Example2; SET work.Example2;
* Label variables and apply value formats for variables used below;
* LABEL name=    "name: Descriptive Variable Label";
  LABEL marry=   "marry: 2-Category Marital Status"
        gender=  "gender: 2-Category Gender Identity"
        happy=   "happy: 5-Category Happy Rating"
        friends= "friends: 7-Category Time with Friends"
        educ=    "educ: Years of Education"
        income=  "income: Annual Income in 1000s";
* Apply value labels created above: name Format.;
  FORMAT marry Fmarry. gender Fgender. happy Fhappy. friends Ffriends.;
* Make a copy of income to z-score next;
  incomeZ=income;
  LABEL incomeZ="incomeZ: Z-Scored Income";
* Example nonlinear transformation: natural-log transform;
  incomeLog=LOG(income);
  LABEL incomeLog="incomeLog: Log Annual Income";
RUN;

* Example linear transformation: z-scoring;
* This will over-write the original incomeZ with z-scored version;
PROC STANDARD DATA=work.Example2 OUT=work.Example2 MEAN=0 STD=1;
       VAR incomeZ;
RUN;
* Now dataset work.Example2 is ready to use;
```

All SAS commands and comments end in a semi-colon.

<u>STATA</u> Syntax for Importing and Preparing Data for Analysis:

```
// Paste in the folder address where "GSS_Example.xlsx" is saved between " "
   global filesave "C:\Dropbox\21SP_PSQF6242\PSQF6242_Example2"

// We can then refer to the syntax variable "filesave" by putting $ in front

// IMPORT GSS_Example.xlsx data using filesave reference and exact file name
// To change all variable names to lowercase, remove "case(preserve")
   clear // Clear before means close any open data
   import excel "$filesave\GSS_Example.xlsx", case(preserve) firstrow clear
// Clear after means re-import if it already exists (if need to start over)

// Create formats: set of value labels for categorical variables;
   label define Fmarry    1 "1.Unmarried" 2 "2.Married"
   label define Fgender   1 "1.Man" 2 "2.Woman"
   label define Fhappy    1 "1.Unhappy" 2 "2.Neither" 3 "3.Fairly Happy" ///
                            4 "4.Very Happy" 5 "5.Completely Happy"
   label define Ffriends  1 "1.Never" 2 "2.Once/Year" 3 "3. Several/Year" ///
                            4 "4.Once/Month" 5 "5.Several/Month"  ///
                            6 "6.Several/Week" 7 "7.Almost Daily"

// Label variables and apply value formats for variables used below
// label variable name    "name: Descriptive Variable Label"
   label variable marry   "marry: 2-Category Marital Status"
   label variable gender  "gender: 2-Category Gender Identity"
   label variable happy   "happy: 5-Category Happy Rating"
   label variable friends "friends: 7-Category Time with Friends"
   label variable educ    "educ: Years of Education"
   label variable income  "income: Annual Income in 1000s"

// Apply value labels created above: name Format
   label values marry Fmarry
   label values gender Fgender
   label values happy Fhappy
   label values friends Ffriends

// Example linear transformation: z-scoring
// EGEN does more complicated transformations
   egen incomeZ = std(income)
   label variable income "income: Z-Scored Annual Income"

// Example nonlinear transformation: natural-log transform
   gen incomeLog = log(income)
   label variable incomeLog "incomeLog: Log Annual Income"

// Now dataset is ready to use
```

STATA commands do not have a line terminator (like semi-colon in SAS).

However, if you need to continue a command across multiple lines, you need `///` at the end of each line to be continued (see example in `label define` above).

SAS Syntax for Descriptive Statistics for Quantitative Variables:
(I include old-school "listing" output below because it's easier to paste and annotate than HTML)

```
* Request descriptive statistics for quantitative variables;
* NDEC=3 print 3s digits after decimal, NOLABELS suppresses labels;
TITLE "Descriptive statistics for quantitative variables";
PROC MEANS NDEC=3 NOLABELS N MEAN STDDEV MIN MAX DATA=work.Example2;
     VAR income incomeZ incomeLog educ;
RUN;
```

Variable	N	Mean	Std Dev	Minimum	Maximum
income	734	17.303	13.792	0.245	68.600
incomeZ	734	-0.000	1.000	-1.237	3.719
incomeLog	734	2.422	1.116	-1.406	4.228
educ	734	13.812	2.909	2.000	20.000

```
* STDERR, ALPHA=.05, and CLM request SE and 95% CIs for mean;
PROC MEANS NDEC=3 NOLABELS N MEAN STDERR ALPHA=.05 CLM DATA=work.Example2;
     VAR income incomeZ incomeLog educ;
RUN; TITLE;
```

Variable	N	Mean	Std Error	Lower 95% CL for Mean	Upper 95% CL for Mean
income	734	17.303	0.509	16.303	18.302
incomeZ	734	-0.000	0.037	-0.072	0.072
incomeLog	734	2.422	0.041	2.342	2.503
educ	734	13.812	0.107	13.601	14.023

STATA Syntax for Descriptive Statistics for Quantitative Variables:
(STATA default output is unformatted text as shown below)

```
// Request descriptive statistics for quantitative variables
   display "Descriptive statistics for quantitative variables"
   summarize income incomeZ incomeLog educ
```

Variable	Obs	Mean	Std. Dev.	Min	Max
income	734	17.30287	13.79163	.245	68.6
incomeZ	734	6.29e-10	1	-1.236828	3.719439
incomeLog	734	2.422476	1.116169	-1.406497	4.228292
educ	734	13.81199	2.909282	2	20

```
// MEAN gives SE of mean and level% CIs
   mean income incomeZ incomeLog educ, level(95)
```

Mean estimation   Number of obs   =   734

	Mean	Std. Err.	[95% Conf.	Interval]
income	17.30287	.5090583	16.30349	18.30226
incomeZ	6.29e-10	.0369107	-.0724632	.0724632
incomeLog	2.422476	.0411985	2.341595	2.503357
educ	13.81199	.1073836	13.60117	14.02281

*What's the difference between the sample standard deviation (**SD**, labeled "Std Dev") and standard error of the mean (**SE**, labeled "Std Err")?*

SAS Syntax and Output for Pearson Correlations for Quantitative Variables:

```
* Request Pearson correlations and p-values;
* FISHER gives CI (%=1-alpha) using r-to-z transformation;
TITLE "Pearson correlations for quantitative variables";
PROC CORR DATA=work.Example2 PEARSON FISHER(BIASADJ=YES ALPHA=.05);
     VAR income incomeZ incomeLog educ;
RUN; TITLE;
```

Descriptive statistics are also given for each variable by default (not shown here).

Pearson Correlation Coefficients, N = 734
Prob > |r| under H0: Rho=0

	income	incomeZ	incomeLog	educ
income income: Annual Income in 1000s	1.00000	1.00000 <.0001	0.82497 <.0001	0.38471 <.0001
incomeZ incomeZ: Z-Scored Income	**1.00000** <.0001	1.00000	0.82497 <.0001	0.38471 <.0001
incomeLog incomeLog: Log Annual Income	**0.82497** <.0001	0.82497 <.0001	1.00000	0.30497 <.0001
educ educ: Years of Education	**0.38471** <.0001	**0.38471** <.0001	**0.30497** <.0001	1.00000

This is a **Pearson correlation matrix**, which standardizes the variances of each variable on the diagonal to 1. The correlations (r) for each pair of variables are given on the off-diagonals. **Below each r is the exact two-tailed p-value against a null hypothesis of H_0: $\rho = 0$.** (The t test-statistics that provided the exact p-values are not given.)

Pearson Correlation Statistics (Fisher's z Transformation)

Variable	With Variable	N	Sample Correlation	Fisher's z	Bias Adjustment	Correlation Estimate	95% Confidence Limits	
income	incomeZ	734	1.00000	16.44867	0.0006821	1.00000	1.000000	1.000000
income	incomeLog	734	0.82497	1.17220	0.0005627	0.82479	0.800190	0.846628
income	educ	734	0.38471	0.40558	0.0002624	0.38449	0.321055	0.444485
incomeZ	incomeLog	734	0.82497	1.17220	0.0005627	0.82479	0.800190	0.846628
incomeZ	educ	734	0.38471	0.40558	0.0002624	0.38449	0.321055	0.444485
incomeLog	educ	734	0.30497	0.31500	0.0002080	0.30479	0.237662	0.369012

"Sample correlation" is just the Pearson r.

"Fisher's Z" is given by:

$$z_r = 0.5\left[Log_e\left(\frac{1+r}{1-r}\right)\right],$$

"**Bias adjustment**" is the difference between original Pearson r and sample-size-adjusted r (labeled "**correlation estimate**") given by:

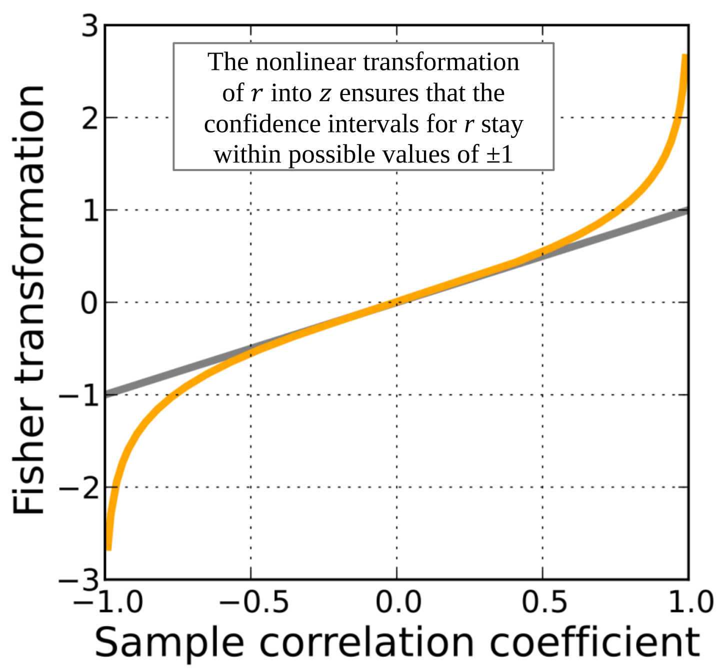
$$r_{adj} = \sqrt{1 - \frac{(1-r^2)(N-1)}{N-2}}$$

Steps to compute 95% CI for r_{adj}:
(a) convert r to z_r, (b) compute lower and upper bounds in z-scale using:

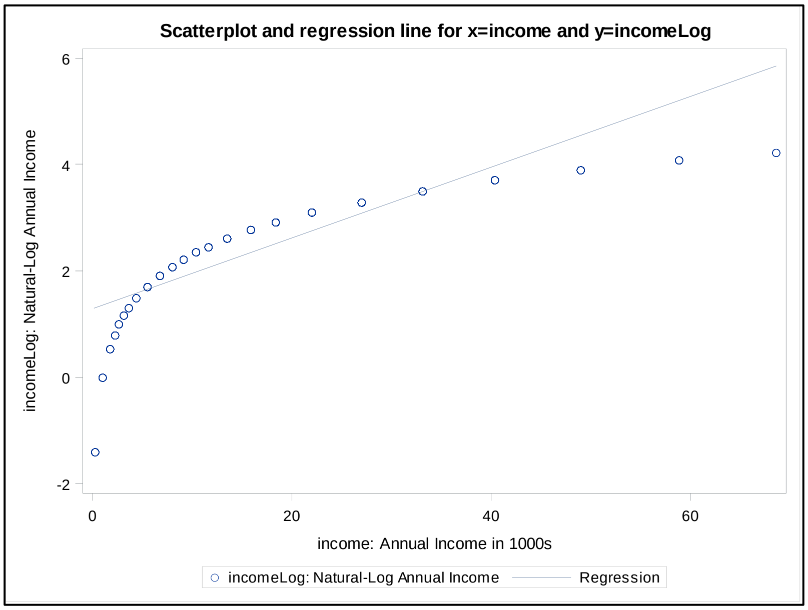
$$SE\ z_r = \frac{1}{\sqrt{N-3}},\ CI = z_r \pm z_{crit} * SE,$$

and (c) back-transform bounds to r-scale:

$$r = \frac{\exp(2z)-1}{\exp(2z)+1}$$



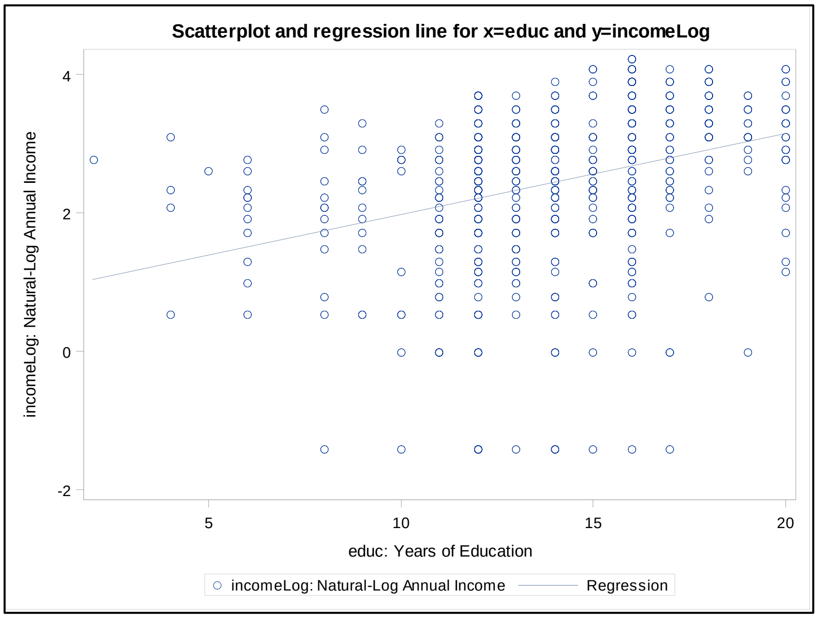
```
* Make scatterplot with regression lines to show predicted linear relations;
TITLE "Scatterplot and regression line for x=income and y=incomeLog";
PROC SGPLOT DATA=work.Example2;
     SCATTER x=income y=incomeLog;
     REG     x=income y=incomeLog;
RUN; TITLE;
```



This plot illustrates the effect of the nonlinear natural-log transformation of income. Relative to original income on the x-axis, the upper values of log-transformed income on the y-axis do not increase nearly as quickly, where the lower values of log-transformed income on the y-axis become more distinct (i.e., they spread out).

The Pearson correlation between these variables is only r = .82, because Pearson r only captures linear relationships.

```
TITLE "Scatterplot and regression line for x=educ and y=incomeLog";
PROC SGPLOT DATA=work.Example2;
     SCATTER x=educ y=incomeLog;
     REG     x=educ y=incomeLog;
RUN; TITLE;
```



The Pearson correlation between education and log-income is r = .30, as given by the "regression" line that best fits through the scatterplot points. This means that as education goes up, log-income is expected to go up, too.

Soon we will learn how to use general linear models to fit different kinds of relationships besides linear… stay tuned!

<u>STATA</u> Syntax and Output for Pearson Correlations for Quantitative Variables:

```
// Request Pearson correlations and p-values
   display "Pearson correlations for quantitative variables"
   pwcorr income incomeZ incomeLog educ, sig obs
```

```
             |   income  incomeZ incomeLog     educ
-------------+------------------------------------
      income |   1.0000 
             |
             |      734
             |
     incomeZ |   1.0000   1.0000 
             |   0.0000
             |      734      734
             |
   incomeLog |   0.8250   0.8250   1.0000 
             |   0.0000   0.0000
             |      734      734      734
             |
        educ |   0.3847   0.3847   0.3050   1.0000 
             |   0.0000   0.0000   0.0000
             |      734      734      734      734
```

This is a **Pearson correlation matrix**, which standardizes the variances on the diagonal to 1. The correlations (r) for each pair of variables are given on the off-diagonals. **Below each r is the exact two-tailed p-value against a null hypothesis of $H_0: \rho = 0$**, followed by the sample size for that correlation. (The t test-statistics that provided the exact p-values are not given.)

This STATA module ci2 uses the original Pearson's r rather than the adjusted r used by SAS, and thus this CI does not exactly match the CI given by SAS.

STATA's requested scatterplots are not show.

```
// To get CI using r-to-z, need to download and run a special module, ci2
   ssc install ci2

// CI2 to use Fisher r-to-z transform to get CI for correlation
   display "Fisher r-to-z transform to get CI for Pearson correlation"
   ci2 incomeLog educ, corr  // Need to list each pair of variables separately
```

```
Confidence interval for Pearson's product-moment correlation
of incomeLog and educ, based on Fisher's transformation.
Correlation = 0.305 on 734 observations (95% CI: 0.238 to 0.369)
```

```
// Make scatterplots with regression lines to show predicted linear relations
   display "Scatterplot and regression line for x=income and y=incomeLog"
   graph twoway (lfit incomeLog income) (scatter incomeLog income)

   display "Scatterplot and regression line for x=educ and y=incomeLog"
   graph twoway (lfit incomeLog educ) (scatter incomeLog educ)
```

<u>Example Results Section Using SAS Output:</u>

We estimated Pearson's correlations (r) to examine the extent of linear relationship for years of education (M = 13.81, SD = 2.91, range = 2 to 20) with two variants of personal annual income: original in thousands of dollars (M = 17.30, SD = 13.79, range = 0.25 to 68.60), or annual income after a natural-log transformation to reduce the influence of extreme values (M = 2.42, SD = 1.12, range = −1.41 to 4.23). We selected a two-tailed alpha = .05, and we used a Fischer r-to-z transformation to compute 95% confidence limits around the r estimates after adjusting for sample size (r_{Adj}). Given our sample size of N = 734, degrees of freedom = N – 2 = 732 for each pair of variables. Years of education was significantly positively related to annual income, r = .385, p < .0001, r_{Adj} = .384, 95% CI = .321 to .444, indicating that greater education was associated with greater income (i.e., a sample r more extreme than r = .385 would be expected < .01% of the time if the population correlation were 0). Similar results were found when using log-transformed annual income instead, r = .305, p < .0001, r_{Adj} = .305, 95% CI = .238 to .369.

[NOTE: I decided to use two digits for descriptive statistics, but three digits for correlations.]

SAS Syntax and Output for Spearman Correlations for Ordinal Variables:

```
* Request Spearman correlations and p-values;
* FISHER gives CI (%=1-alpha) using r-to-z transformation;
TITLE "Spearman correlations for rank-order quantitative and ordinal variables";
PROC CORR DATA=work.Example2 SPEARMAN FISHER(BIASADJ=YES ALPHA=.05);
     VAR income friends happy;
RUN; TITLE;
```

Simple Statistics

Variable	N	Mean	Std Dev	Median	Minimum	Maximum
income	734	17.30287	13.79163	13.47500	0.24500	68.60000
friends	734	4.23842	1.55011	4.00000	1.00000	7.00000
happy	734	3.55586	0.89504	4.00000	1.00000	5.00000

When requesting **Spearman** correlations, the **median** is also given as a descriptive statistic.

What do these univariate statistics imply about the likely distributions of *friends* (how often socialize with friends) and *happy* self-rating?

Spearman Correlation Coefficients, N = 734
Prob > |r| under H0: Rho=0

	income	friends	happy
income income: Annual Income in 1000s	1.00000	-0.04587 0.2146	0.04931 0.1821
friends friends: 7-Category Time with Friends	**-0.04587** 0.2146	1.00000	0.08272 0.0250
happy happy: 5-Category Happy Rating	**0.04931** 0.1821	**0.08272** 0.0250	1.00000

This is a **Spearman correlation matrix**, which standardizes the variances of each variable on the diagonal to 1. The correlations (r) for each pair of variables are given on the off-diagonals. **Below each r is the exact two-tailed p-value against a null hypothesis of H_0: $\rho = 0$.** (The t test-statistics that provided the exact p-values are not given.)

Spearman Correlation Statistics (Fisher's z Transformation)

Variable	With Variable	N	Sample Correlation	Fisher's z	Bias Adjustment	Correlation Estimate	95% Confidence Limits	
income	friends	734	-0.04587	-0.04590	-0.0000313	-0.04583	-0.117808	0.026620
income	happy	734	0.04931	0.04935	0.0000336	0.04928	-0.023172	0.121208
friends	happy	734	0.08272	0.08291	0.0000564	0.08266	0.010358	0.154104

Example Results Section Using SAS Output:

We estimated Spearman's rank-order correlations (*rho*) to examine the associations among annual income in thousands of dollars (M = 17.30, SD = 13.79, range = 0.25 to 68.60), ordinal amount of time with friends (1 to 7 scale; M = 4.24, SD = 1.55), and ordinal happiness (1 to 5 scale; M = 3.56, SD = 0.90). We selected a two-tailed alpha = .05, and we used a Fischer r-to-z transformation to compute 95% confidence limits around the Spearman correlation estimates after adjusting for sample size (rho_{Adj}). Given our sample size of N = 734, degrees of freedom = N – 2 = 732 for each pair of variables. Annual income was not significantly related to amount of time with friends, *rho* = −.046, p = .2146, r_{Adj} = −.046, 95% CI = −.118 to .027, indicating that greater income was nonsignificantly associated with lesser amount of time with friends (i.e., a sample *rho* more extreme than *rho* = −.046 would be expected 21.46% of the time if the population correlation were 0). Likewise, annual income was not significantly related to self-rated happiness, *rho* = .049, p = .1821, r_{Adj} = .049, 95% CI = −.023 to .121, indicating that greater income was nonsignificantly associated with greater happiness (i.e., a sample *rho* more extreme than *rho* = .049 would be expected 18.21% of the time if the population correlation were 0). However, amount of time spent with friends was significantly positively related to self-rated happiness, *rho* = .083, p = .0250, r_{Adj} = .083, 95% CI = .010 to .154, indicating that greater time spent with friends was associated with greater happiness (i.e., a sample *rho* more extreme than *rho* = .083 would be expected only 2.50% of the time if the population correlation were 0).

STATA Syntax and Output for Spearman Correlations for Ordinal Variables:

```
// To get median, download and run a special module, univar
   ssc install Univar

// Request median for quantitative or ordinal variables
   display "Median for quantitative and ordinal variables"
   univar income friends happy

                                        -------------- Quantiles --------------
Variable       n     Mean     S.D.      Min      .25      Mdn      .75      Max
-------------------------------------------------------------------------------
  income     734    17.30    13.79     0.25     6.74    13.48    22.05    68.60
 friends     734     4.24     1.55     1.00     3.00     4.00     5.00     7.00
   happy     734     3.56     0.90     1.00     3.00     4.00     4.00     5.00
-------------------------------------------------------------------------------

// Request Spearman correlations and p-values
   display "Spearman correlations for rank-order quantitative and ordinal variables"
   spearman income friends happy, stats(rho obs p)

             |   income  friends    happy
-------------+---------------------------
      income |   1.0000
             |      734
             |
             |
     friends |  -0.0459   1.0000
             |      734      734
             |   0.2146
             |
       happy |   0.0493   0.0827   1.0000
             |      734      734      734
             |   0.1821   0.0250
```

This is a **Spearman correlation matrix**, which standardizes the variances of each variable on the diagonal to 1. The correlations (r) for each pair of variables are given on the off-diagonals. **Below each r is the sample size for the pair of variables, followed by the exact two-tailed p-value against a null hypothesis of H_0: $\rho = 0$.** (The t test-statistics that provided the exact p-values are not given.)

```
// CI2 to use Fisher r-to-z transform to get CI for correlation
   display "Fisher r-to-z transform to get CI for Spearman correlation"
   ci2 income friends, corr spearman // Need to list each pair of variables separately

Confidence interval for Spearman's rank correlation
of income and friends, based on Fisher's transformation.
Correlation = -0.046 on 734 observations (95% CI: -0.118 to 0.027)
```

SAS Syntax and Output for Cross-Tabulations and Associations for Binary Variables:

```
* Request cross-tabulation of categorical variables with percentages;
* NOROW NOCOL suppresses row- and column-specific frequencies;
* Options request chi-square test and expected frequencies;
TITLE "Cross-tabulations for binary variables gender and marry";
PROC FREQ DATA=work.Example2;
     TABLE gender*marry / NOROW NOCOL ALPHA=.05 CHISQ EXPECTED;
RUN; TITLE;
```

gender(gender: 2-Category Gender Identity) by marry(marry: 2-Category Marital Status)

Frequency Expected Percent	1.Unmarried	2.Married	Total
1.Man	197 196.34 26.84	166 166.66 22.62	363 49.46
2.Woman	200 200.66 27.25	171 170.34 23.30	371 50.54
Total	397 54.09	337 45.91	734 100.00

In this cross-tabulation, the first row of each cell is the **frequency**. The second row is the **expected frequency**: the count that would have been observed just based on the marginal frequencies if there were no association. The third row is the **percentage** (out of the total).

By comparing the observed and expected frequencies, we see there are slightly more married women than expected.

Statistic	DF	Value	Prob
Chi-Square	**1**	**0.0097**	**0.9217**
Likelihood Ratio Chi-Square	1	0.0097	0.9217
Continuity Adj. Chi-Square	1	0.0006	0.9807
Mantel-Haenszel Chi-Square	1	0.0096	0.9218
Phi Coefficient		**0.0036**	
Contingency Coefficient		0.0036	
Cramer's V		0.0036	

There are many measures of association that can be used for categorical variables. The most common is just chi-square (χ^2). The *p*-value > .05 would be declared nonsignificant, meaning that we do not have evidence for an association that is different than 0 here (we retain the null hypothesis of 0 association).

```
TITLE "Pearson correlation and CI for binary variables gender and marry";
PROC CORR DATA=work.Example2 PEARSON FISHER(BIASADJ=NO ALPHA=.05);
     VAR gender marry;
RUN; TITLE;
```

Pearson Correlation Coefficients, N = 734
Prob > |r| under H0: Rho=0

	gender	marry
gender gender: 2-Category Gender Identity	1.00000	0.00363 0.9218
marry marry: 2-Category Marital Status	**0.00363** 0.9218	1.00000

The "phi" coefficient given above is Pearson's r. The *p*-value for its significance against $H_0: r \neq 0$ (found using PROC CORR) will be slightly bigger than the *p*-value for the χ^2 test-statistic. This is because Pearson's r is tested using a t test-statistic that uses denominator degrees of freedom (DF based on N), whereas the χ^2 test-statistic does not use denominator DF, just like z (for $DF_{num} = 1,\ \chi^2 = z^2$).

Pearson Correlation Statistics (Fisher's z Transformation)

Variable	With Variable	N	Sample Correlation	Fisher's z	95% Confidence Limits	
gender	marry	734	0.00363	0.00363	-0.068755	0.075973

For consistency in reporting I requested 95% confidence intervals around the original Pearson r (= phi correlation), not the adjusted correlation.

So what does a positive correlation of r = .004 mean between these variables???

Would the Spearman correlation be different than the Pearson correlation in this case?

Example Results Section Using SAS Output:

We estimated the association between respondent gender (49.46% men = 1; 50.54% women = 2) and marital status (54.09% unmarried = 1; 45.91% married = 2) using a two-tailed alpha = .05. The association between these variables was nonsignificant, Pearson's $\chi^2(1)$ = 0.010, *p* = .9217, Pearson *r* = .004, 95% CI = −.069 to .076, indicating that women were nonsignificantly more likely to be married than unmarried (i.e., a sample *r* more extreme than *r* = .004 would be expected 92.18% of the time if the population correlation were 0).

STATA Syntax and Output for Cross-Tabulations and Association for Binary Variables:

```
// Request cross-tabulation of categorical variables with percentages
   display "Cross-tabulations for binary variables gender and marry"
   tabulate gender marry, cell
```

In this cross-tabulation, the first row of each cell is the **frequency**, and the second row is the **percentage** (out of the total).

```
   gender: |
2-Category |   marry: 2-Category
    Gender |    Marital Status
  Identity | 1.Unmarri  2.Married |     Total
-----------+----------------------+----------
     1.Man |       197        166 |       363
           |     26.84      22.62 |     49.46
-----------+----------------------+----------
   2.Woman |       200        171 |       371
           |     27.25      23.30 |     50.54
-----------+----------------------+----------
     Total |       397        337 |       734
           |     54.09      45.91 |    100.00
```

```
// Request x2 test of association with expected frequencies
   display "Chi-square tests for binary variables gender and marry"
   tab2 gender marry, chi2 expected
```

```
   gender: |
2-Category |   marry: 2-Category
    Gender |    Marital Status
  Identity | 1.Unmarri  2.Married |     Total
-----------+----------------------+----------
     1.Man |       197        166 |       363
           |     196.3      166.7 |     363.0
-----------+----------------------+----------
   2.Woman |       200        171 |       371
           |     200.7      170.3 |     371.0
-----------+----------------------+----------
     Total |       397        337 |       734
           |     397.0      337.0 |     734.0

     Pearson chi2(1) =   0.0097   Pr = 0.922
```

In this table, the first row of each cell is the **frequency**, and the second row is the **expected frequency**: the count that would have been observed just based on the marginal frequencies if there were no association.

Note that for your homework, you will need to report expected frequencies to two digits after the decimal. So you will need to ask STATA to compute them as shown below.

Pearson $\chi^2(DF_{num}) = 0.0097$, *p* = .922, so the association is nonsignificant (so retain H_0)

```
// Make STATA compute expected frequencies manually: Nrow*Ncol/Ntotal
   display "Expected Frequency for Man, Unmarried"
   display 363*397/734

196.33651
```

```
// Request Pearson correlation, p-value, and CI
   display "Pearson correlation and CI for binary variables gender and marry"
   pwcorr gender marry, sig obs
   ci2 gender marry, corr
```

```
             |   gender    marry
-------------+------------------                       |
       marry |   0.0036   1.0000
             |   0.9218
             |      734      734
             |

Confidence interval for Pearson's product-moment correlation
of gender and marry, based on Fisher's transformation.
Correlation = 0.004 on 734 observations (95% CI: -0.069 to 0.076)
```