Computers Skills Project

Sanjay Kadari

BIOS 600

Honors Pledge:

**Saved Code**

Code that was run first and saved so that variables such as gender and party could have proper labels; therefore all code presented is assumed to have been run with this code already preset:

use "\\tsclient\C\BioStatProject\example.dta", clear

encode p, generate(p2) (give numbers to the categorical values)

label define gen 1 Males 0 Females (label gender)

label values g gen

rename p2 Party

label define par 1 Democrats 2 Other 3 Republicans 4 Unaffiliated (label party)

label values Party par

**Note: Everything with \*\* before it indicates annotations to the code and is not actual code**

**Summary Tables and Statistics**

Statistical Concepts

* Summary statistics describe the shape, spread, symmetry, modality, and central tendency of the data.
* Summary tables include summary statistics such as mean, median, variance, min, max, and number of observations.
* A summary table should always include information about the unites of the variables being presented. Data is only meaningful if units are present, therefore this inclusion is vital.
* Data type is important when calculating summary statistics since continuous data will give a summary table with statistics such as mean, median, and variance. Categorical data will depict a frequency table instead.

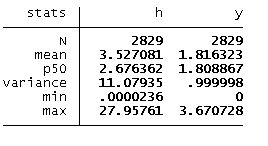
Code:

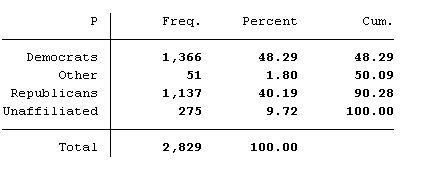
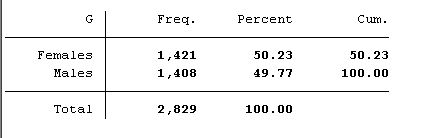
tabstat h y, stats(N mean med var min max) \*\* (create summary table)

tab g \*\* (create frequency table)

tab p

Output:





**Scatterplot of X and Y**

Statistical Concepts

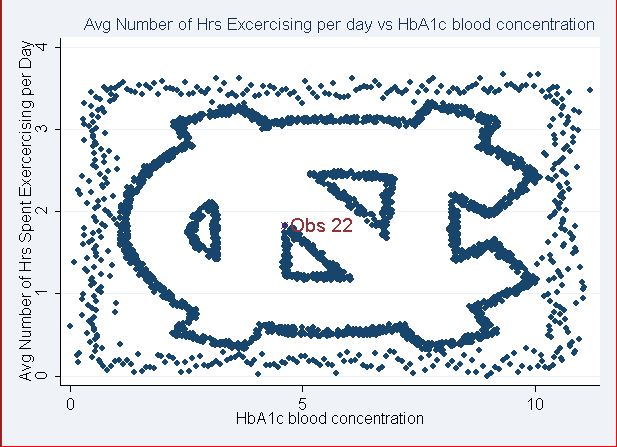
* X and Y typically have to be continuous data types so that trends as association factors can be seen properly. While categorical data points can be plotted, better plots than a scatter plot can be used to depict them.
* Goal of a scatter plot is to see the relationship between two variables.
* Scatter plots are helpful when identifying outliers and trends in data.

Code:

twoway (scatter y x if obs != 22) (scatter y x if obs == 22, mcolor(purple) msymbol(lgx) mlabel(obs) mlabsize(medlarge)), ytitle(Avg Number of Hrs Spent Exercercising per Day) xtitle(HbA1c blood concentration) title(Avg Number of Hrs Excercising per day vs HbA1c blood concentration) legend(off)

\*\*(plot scatterplot)

Output:



**Histogram of X**

Statistical Concepts:

* X must be a continuous variable to be able to be sorted into bins.
* Histograms communicate a visual impression of the distribution of the data.

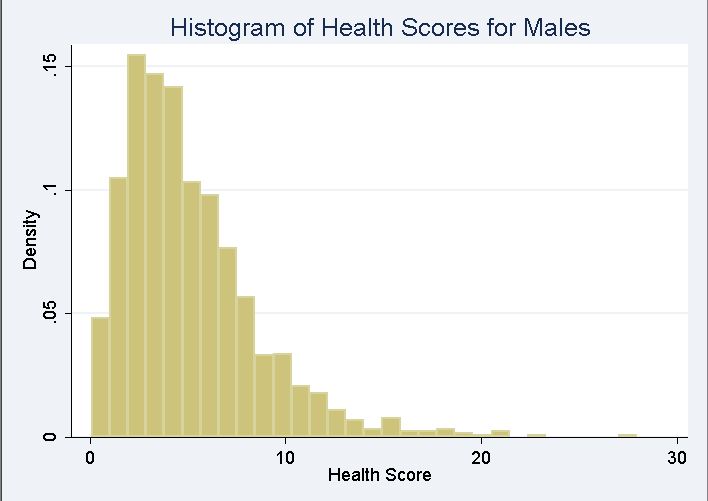
Code:

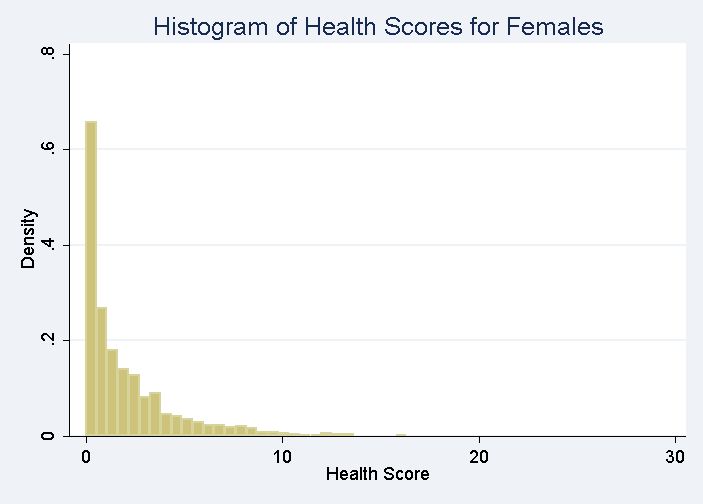
histogram h if g==1, title(Histogram of Health Scores for Males) xtitle(Health Score)

histogram h if g==0, bin(30) xscale(range(0 30)) xlabel(#4) title(Histogram of Health Scores for Females) xtitle(Health Score)

\*\*(plot histogram)

Output:





**Box plot of X**

Statistical Concepts:

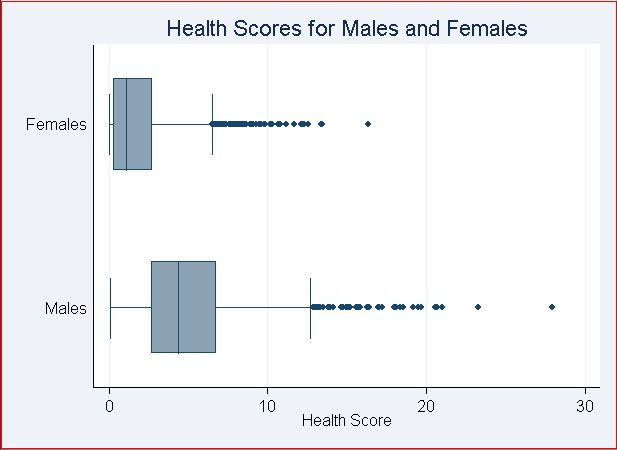
* Box plots must be continuous variables since numerical summaries are depicted (mean and quantiles)
* Box plots are a convenient way of graphically depicting groups of numerical data through their five number summaries: minimum, Q1, median, Q3, and maximum.
* Box plots are helpful in comparing shape of multiple variables.

Code:

graph hbox h, over(g, relabel(2"Males" 1"Females")) ytitle(Health Score) title(Health Scores for Males and Females)

\*\*(plot boxplot)

Output:



**Line Plot (Time Series)**

Statistical Concepts:

* Line plots communicate the changes over time for a variable(s).

Code:

gen Males = r if g ==1 \*\*(Separate genders)

gen Females = r if g==0

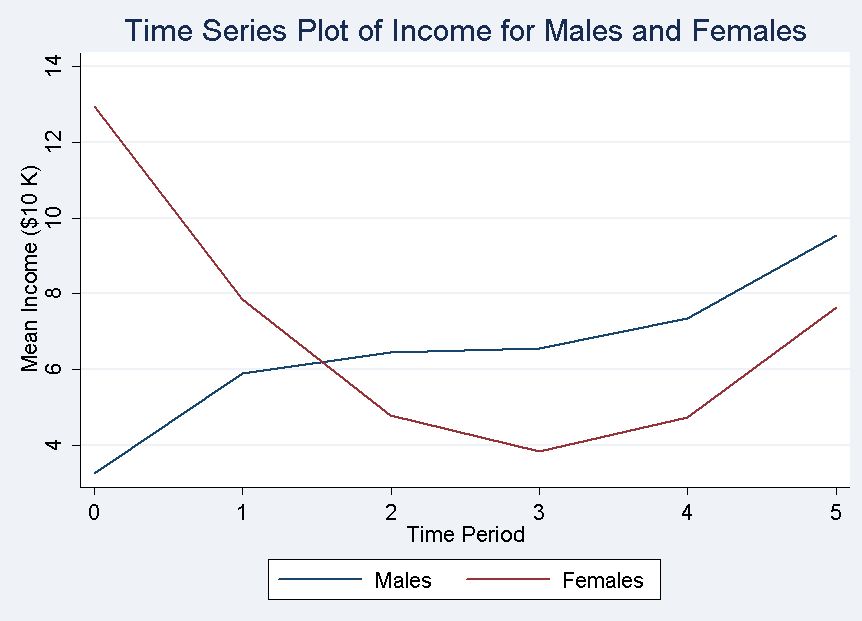
egen mean = mean(Males), by(t)

egenmean2 = mean(Females), by(t) \*\*(Calculate mean by time)

twoway (line mean t, sort) (line mean2 t, sort), ytitle(Mean Income ($10 K)) xtitle(Time Period) title(Time Series Plot of Income for Males and Females) legend(order(1 "Males" 2 "Females"))

\*\*(plot line plot)

Output:



**Probabilities from Standard Distributions**

Statistical Concepts:

* The probability density function of a continuous random variable is a function that describes the relative likelihood for this random variable to occur at a given point.
* The probability mass function is a function that gives the probability that a discrete random variable is exactly equal to some value.
* Distribution parameters are constants which stand for a property of a random variable.

Code:

1. gen z = (1.8-1)/sqrt(2)

display normal(z)

1. bitesti 44 23 0.4
2. display 1 - poisson(.75,3)
3. display 1-ttail(46,2.3)-(1-ttail(46,1.2))
4. display chi2tail(7,2)
5. display Ftail(4,38,3.2)

Output:

1. 0.71419619
2. 0.964104
3. 0.00729217
4. 0.10512396
5. 0.95984037
6. 0.02327884

**Generate Random Numbers**

Statistical Concepts:

* A random number generation is helpful whenever a study design involves separating a sample group into multiple sub groups. Suppose a study involves separating a group of males randomly into two groups of equal size; a random number generation is helpful in performing such a task.

Code:

gen random = runiform() \*\*(generate random numbers)

sort g random \*\*(sort random numbers and gender)

by g: gen group = group(2) \*\*( half in each group are 1s and 0s)

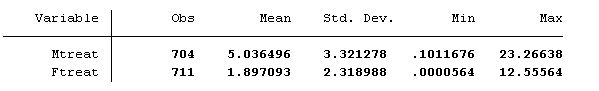
gen Mtreat = h if g==1 & group==1

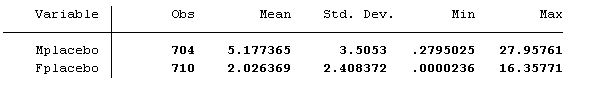
gen Mplacebo = h if g==1 & group==2

gen Ftreat = h if g==0 & group ==1

gen Fplacebo = h if g==0 & group ==1

Output:





**Perform a t-Test on X**

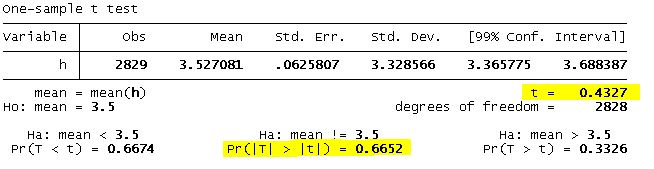
Statistical Concepts:

* X must be a continuous variable and also has to be either normal or have a large sample size.
* The hypothesis of interest is a statement about what we believe is actually true for the population we are studying. For a one sample t test, the hypothesis of interest usually involves testing the mean of a variable against a value. The null hypothesis is that the mean of the health score is equal to 3.5.
* Using the degrees of freedom and the t-value, one can find the p-value for the test. Comparing the p-value to the risk value, we can determine whether there is not enough information to reject the null hypothesis or enough information to reject it. For this test we fail to reject the null hypothesis since 0.665 > 0.01
* There is no sample size consideration for t tests, however, the smaller the sample size the less reliable the test becomes.

Code:

ttest h == 3.5, level(99) \*\*(conduct t test)

Output:



**Perform Two Sample t-test**

Statistical Concepts:

* The data has to be continuous and must be normal or have large sample sizes.
* The hypothesis of interest for two sample t tests usually involves the means of two populations being equal. The hypothesis in this case is that the mean difference of the health score between males and females is equal to 3.
* Using the t statistic and the degrees of freedom, we can determine the p value for the test. Comparing the p value to the risk level we can determine whether there is not enough evidence to reject the hypothesis of interest or that there is enough evidence to reject it. For this test we fail to reject the null hypothesis since 0.1881 > 0.1.

Code:

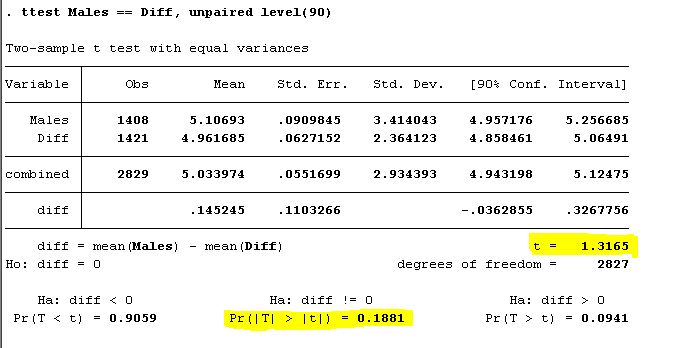
gen Males = h if g==1

gen Females = h if g==0 \*\*(Separate genders)

gen Diff = Females + 3 \*\*(Shift variable by 3)

ttest Males == Diff, unpaired level(90) \*\*(run t test)

Output:



**Binomial Proportion Test**

Statistical Concepts:

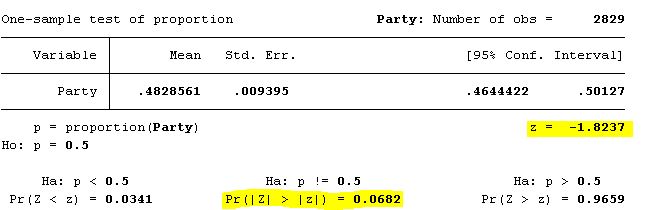
* Data for a binomical proportion test has to be numerical data however data has to be binomial; therefore values should be 0 or 1.
* The parameter of interest is the number of times a particular outcome occurs. In this case the parameter of interest is the number of Democrats in the population.
* The hypothesis of interest is whether the proportion of Democrats in the population is equal to 0.5.
* The results of a two sample t test are interpreted in the same way as a one sample t test. In this test we do reject the null hypothesis since 0.0682 < 0.1.

Code:

Replace Party =0 if Party > 1 \*\*(change data to binary)

prtest Party == 0.5 \*\*(run proportion test)

Output:



**Two Sample Binomial Proportion Test**

Statistical Concepts:

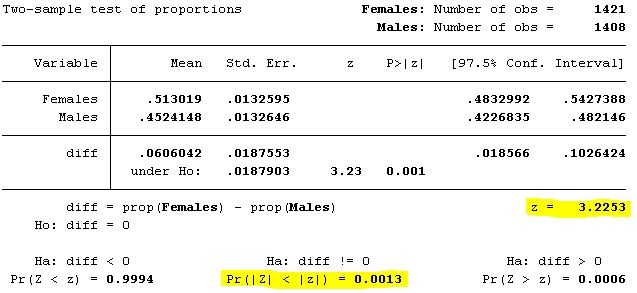
* Binomial proportion test contains numerical data however the data is binomial; which means data is either 0 or 1.
* The hypothesis of interest for this test is that the proportion of democrats are the same among males and females.

Code:

Replace Party =0 if Party > 1

prtest Party, by(g) level(97.5) \*\*(run proportion test)

Output:



**Confidence Interval**

Statistical Concepts:

* A confidence interval with a particular confidence level is intended to give the assurance that, if the statistical model is correct, then taken over all the data, the procedure for constructing the interval would deliver a confidence interval that included the true value of the parameter the proportion of the time set by the confidence level. The interval is used to indicate the reliability of an estimate.

Code:

Gen Males = h if g==1

Gen Females = h if g==0 \*\*(Separate gender)

Replace Party =0 if Party > 1 \*\*(Convert to binary)

Mean of H:

ci h, level(90)

Difference in means of H between males and females:

ttest Males == Females, unpaired level(90)

Proportion of Democrats:

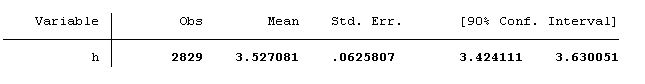
ci Party, level(90)

Difference in proportions of Democrats between males and females:

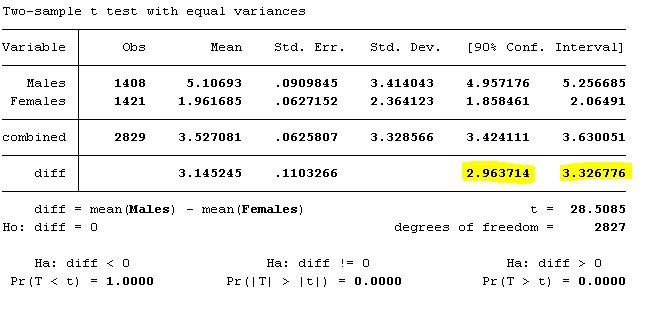
prtest Party, by(g) level(90)

Output:

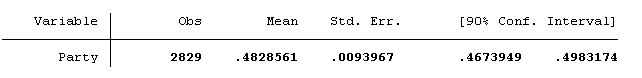
Mean of H:



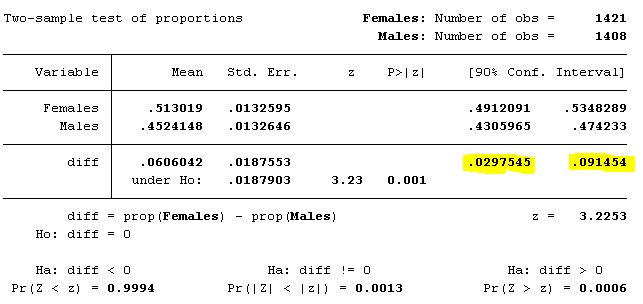
Difference of H between males and females:



Proportion of Democrats:



Difference in proportions of Democrats between males and females:



**Tabulate r x c Tables**

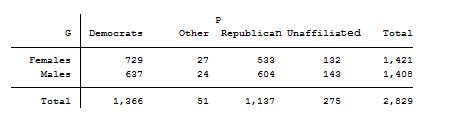
Statistical Concepts:

* Data for a contingency table has to be categorical as it tabulates the total number of occurrences for each group.

Code:

tabulate g Party \*\*(create contingency table)

Output:



**Test of no Association for r x c Tables**

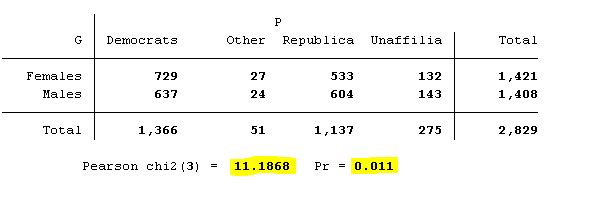
Statistical Concepts:

* The test of no association is used to determine if there is any relationship between two or more variables.
* We use the chi squared value to obtain a p value for the test. We use that to either reject or fail to reject the null hypothesis. The null hypothesis is that of no association; since the p value is less than the risk level we reject the null hypothesis. Therefore there is evidence to suggest that there is a relationship on the gender by party.

Code:

tabulate g Party, chi2 \*\*(test using chi squared)

Output:



**Goodness of Fit Test**

Statistical Concepts:

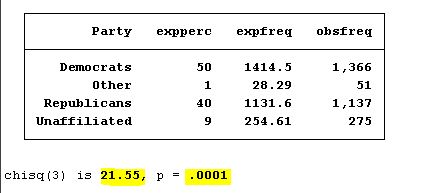
* Measure of goodness of fit typically summarizes the discrepancy between observed values and the values expected under the model in question.
* Typically categorical values are used in the goodness of fit test. However continuous variables can be used in regression analysis.
* The hypothesis of interest for a goodness of fit test states the expected proportions of each population under study. Therefore the hypothesis is that the proportion of Democrats is 0.5, Republicans is 0.4, Unaffiliated is 0.09, and other is 0.01.
* Since the p value is less than the risk level we reject the null hypothesis. Therefore there is large discrepancy between the observed and expected values.

Code:

findit csgof \*\*(download the package necessary)

csgof Party, expper(50 1 40 9) \*\*(calculate gof)

Output:



**Fisher's Exact Test**

Statistical Concepts:

* The Fisher’s exact test is used to determine if there are nonrandom associations between two categorical variables.
* The test applies to categorical data.
* The null hypothesis is that the school of the student does not affect the residency of the student.
* The Fisher's exact test is used when you want to conduct a chi-square test but one or more of your cells has an expected frequency of five or less.

Code:

generate int School = .

generate int Residency=.

set obs 1

replace School = 1 in 1

set obs 2

replace School = 1 in 2

set obs 3

replace School = 1 in 3

set obs 4

replace School = 1 in 4

set obs 5

replace School = 1 in 5

set obs 6

replace School = 1 in 6

set obs 7

replace School = 1 in 7

set obs 8

replace School = 1 in 8

set obs 9

replace School = 0 in 9

set obs 10

replace School = 0 in 10

set obs 11

replace School = 0 in 11

set obs 12

replace School = 0 in 12

set obs 13

replace School = 0 in 13

set obs 14

replace School = 0 in 14

set obs 15

replace School = 0 in 15

set obs 16

replace School = 0 in 16

replace Residency = 1 in 1

replace Residency = 1 in 2

replace Residency = 1 in 3

replace Residency = 1 in 4

replace Residency = 1 in 5

replace Residency = 0 in 6

replace Residency = 0 in 7

replace Residency = 0 in 8

replace Residency = 1 in 9

replace Residency = 1 in 10

replace Residency = 1 in 11

replace Residency = 0 in 12

replace Residency = 0 in 13

replace Residency = 0 in 14

replace Residency = 0 in 15

replace Residency = 0 in 16 \*\*(create data)

label define sch 1 UNC 0 Duke

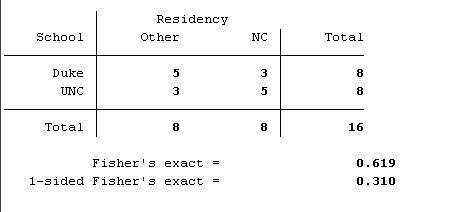
label values School sch

label define res 1 NC 0 Other

label values Residency res \*\*(define labels)

tabulate School Residency, exact \*\*(create table)

Output:



**Wilcoxon-Signed Rank Test**

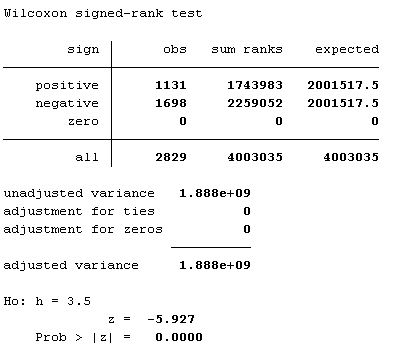
Statistical Concepts:

* The signed rank test is used when comparing two related samples or repeated measurements on a single sample to assess whether their population mean ranks differ. It is equivalent to the t-test of the parametric tests.
* The test is preferred when population cannot be assumed to be normally distributed or the data is on the ordinal scale.
* Since the p value is extremely low, we reject the null hypothesis. Therefore the mean rank is not equal to 3.5.

Code:

signrank h = 3.5 \*\*(create test)

Output:



**Wilcoxon-Rank-Sum Test**

Statistical Concepts:

* The rank sum test is a non-parametric statistical hypothesis test for assessing whether one of two samples of independent observations tends to have larger values than the other. It is equivalent to the two sample t-test.
* Test is preferred when the data is not normal or when n is not a large population.
* The p-value is just slightly greater than the significance level, therefore we fail to reject the null hypothesis which states that the mean difference between the genders is equal to 3.

Code:

gen Females = h if g==0

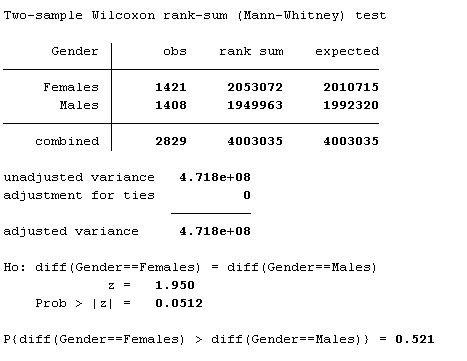
gen Males = h if g==1 \*\*(Separate genders)

gen diff = Females + 3 \*\*(Shift female variable)

replace diff = Males if g==1

ranksum diff, by(g) porder \*\*(create test)

Output:



**ANOVA**

Statistical Concepts:

* ANOVA uses both continuous variables and categorical variables. Summary statistics are performed in each group and compared between and among each categorical group.
* The hypothesis of interest is that means of the groups are equal.
* Using the F statistic and the critical value we can either reject or fail to reject the null hypothesis. If F is greater than the critical value, which is found through a table, then we reject the null hypothesis.
* Multiple comparison correction tests are tests that reduce the type 1 error that is associated with running too many tests. It is important in order to keep the type one error down. It is appropriate to use when many tests have been run between groups under study.

Code:

loneway h Party \*\*(create ANOVA table)

Output:

