**BIOS 600 HONORS PROJECT**

**NAME: Ann Somers Wilton**

**HONOR PLEDGE:** I have neither given nor received unauthorized aid on this assignment.

(Sign and date the submitted paper copy.)

Note: For the Statistical Concepts portion of task, our course textbook, class notes, and Wikipedia were used.

**Calculate Summary Tables and Statistics**

Statistical Concepts

Summary statistics quantitatively describe a data set, such as the center and spread of data sets. Some example summary statistics are mean, median, mode, range, standard deviation, minimum, and maximum. This list is not exhaustive.

Summary tables typically include the number of observations/participants in the study, the percentage of observations missing, a measure of center (typically mean), the standard deviation, and the min and max. Even if it does not contain all of these components, a summary table should always include information about the center and spread of the data set. To calculated summary statistics, the data set must be comprised of continuous data.

Frequency tables show the results of observations at each level of a variable for discrete data, or for continuous data. When these tables are used with continuous data, it is usually grouped into categories.

Code

1. Create a summary table of H, Y, G, and P.

summarize H Y G /\*Note: Can’t make summary table of P because it is a categorical variable\*/

1. Calculate a measure of center and spread for the continuous variables.

summarize X Y H R

1. Calculate the frequency table for the categorical variables.

tabulate G

tabulate P

tabulate T

Output

1. Variable | Obs Mean Std. Dev. Min Max

----------+-------------------------------------------------

H | 2829 3.527081 3.328566 .0000236 27.95761

Y | 2829 1.816323 .999999 0 3.670728

G | 2829 .4977024 .5000831 0 1

2. Variable | Obs Mean Std. Dev. Min Max

-------------+--------------------------------------------------------

X | 2829 5.681694 3 0 11.20417

Y | 2829 1.816323 .999999 0 3.670728

H | 2829 3.527081 3.328566 .0000236 27.95761

R | 2829 6.684393 2.692916 0 15.67666

3. G | Freq. Percent Cum.

-------+-----------------------------------

0 | 1,421 50.23 50.23

1 | 1,408 49.77 100.00

------+-----------------------------------

Total | 2,829 100.00

P | Freq. Percent Cum.

-------+-----------------------------------

D | 1,366 48.29 48.29

O | 51 1.80 50.09

R | 1,137 40.19 90.28

U | 275 9.72 100.00

-------+-----------------------------------

Total | 2,829 100.00

T | Freq. Percent Cum.

-------+-----------------------------------

0 | 463 16.37 16.37

1 | 489 17.29 33.65

2 | 445 15.73 49.38

3 | 456 16.12 65.50

4 | 481 17.00 82.50

5 | 495 17.50 100.00

--------+-----------------------------------

Total | 2,829 100.00**Create a Scatterplot of X and Y**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

Statistical Concepts

To create an effective scatter plot of two variables, both variables should be made up of continuous data. Scatter plots inform the reader of the strength of the relationship, or correlation, between two variables. When points are tightly clustered together, the plot suggests a stronger correlation between the two variables. However, when points are loosely clustered, or distributed widely across the graph, this suggests a lesser correlation between the two variables. Correlations can be either positive or negative.

Scatter plots are most effective for small to moderate sample sizes, and can also be helpful when identifying outliers.

Code

Create a scatterplot of X & Y. Label observation 22 with a special character & the text “obs 22”.

label var X "HbA1c Level"

label var Y "Avg Hrs of Exercise per Day"

twoway scatter Y X if OBS>22 | OBS<22, msymbol(S) || scatter Y X if OBS==22, msymbol(T) text(1.834846722 4.598326639 "obs22") \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Output



**Create a Histogram of X**

Statistical Concepts

To create a histogram of “X”, the variable must be comprised of interval or ratio data, thus it must be continuous. Histograms inform the reader of the frequency distribution of the data of a given variable. If all intervals are the same width, then the height of the interval represents the frequency of that interval. A histogram can inform the reader if the data follow a normal distribution, or if the distribution is right skewed, left skewed, bimodal, etc.

Code

Create two histograms of H: one for women and one for males. Ensure the axes for both plots are identical. Be sure to clearly label each histogram. Do not label with 0 and 1.

label define genderlabel 0 "Female" 1 "Male"

label values G genderlabel

label var H "Health Score"

twoway histogram H, by(G)

Output



**Create a Boxplot of X**

Statistical Concepts

To create a box plot, data must be continuous. These plots inform the reader of five key summary statistics: the median, minimum, maximum, lower quartile, and upper quartile. They are also very helpful in identifying outliers.

Code

In a single plot, create two boxplots of H: one for women and one for males. Be sure to clearly label each boxplot. Do not label with 0 or 1.

graph box H, over(G)

Output



**Create a Line Plot (Time Series)**

Statistical Concepts

Line plots are effective at showing the value of a variable over time. The variable of interest is on the Y axis, while time is the X axis. Line graphs are useful for determining trends over time – to see how the value of the variable has changed at different points in time, and over time as a whole.

Code: In a single plot, create two time series of the mean of R: one for males and one for females. Time is indexed by T. Be sure to label the plots.

bysort G T: egen MeanR=mean(R)

generate MeanRF=MeanR if G==0

generate MeanRM=MeanR if G==1

label var MeanRF "Mean Female Income"

label var MeanRM "Mean Male Income"

label var T "Time"

label var R "Income"

twoway line MeanRF MeanRM T, ytitle("Income")

Output



**Calculate Probabilities from Standard Distributions**

Statistical Concepts

A probability distribution is a listing of the probabilities of all possible outcomes of a discrete variable. The probability density function for a continuous random variable, such as X, is a formula that is used to determine the probability of X being in a given interval.

A probability mass function is a function whose input is an observed value of X (X=x), which is a discrete RV, and whose output is P(X=x).

A distribution parameter is a value reflecting a given property of the population (Forthofer et al). More simply stated, it is a constant which stands for a property of a random variable.

Code

Calculate

1. P( W < 1.8 ) when W is normally distributed with mean 1 and variance 2

zcalc 1.8 1 2

display normal(.4)

1. P( V < 23 ) when V is BIN( n = 44, p = 0.4 ) :

display binomial(44,23,.4)

1. P( U > 3 ) when U is Poisson with mean .75:

pprob, mean(.75)

display 1-0.95949459

1. P( 1.2 < T < 2.3 ) when T is t( df = 46 ):

display ttail(47, 1.2) - ttail(47, 2.3)

1. P( S > 2 ) when S is χ2 with df = 7

display chi2(8,2)

1. P( R > 3.2 ) when R is F with df1 = 4 and df2 = 38

display 1 - F(4,38,3.2)

Output

1. z-score for sample observations

(X - m) (1.8 - 1)

z = --------- = ------------------ = 0.40

s 2

display normal(.4)

.65542174

1. .9641045
2. .04050541 Note: See chart below for detail on answer derivation

Poisson Probabilities for lambda = .75

+------------------------------+

| k pprob pcum |

|------------------------------|

1. | 0 0.47236654 0.47236654 |

2. | 1 0.35427493 0.82664144 |

3. | 2 0.13285309 0.95949459 |

4. | 3 0.03321327 0.99270785 |

5. | 4 0.00622749 0.99893534 |

|------------------------------|

6. | 5 0.00093412 0.99986947 |

7. | 6 0.00011677 0.99998623 |

8. | 7 0.00001251 0.99999875 |

9. | 8 0.00000117 0.99999988 |

10. | 9 0.00000010 1.00e+00 |

|------------------------------|

11. | 10 0.00000001 1.00e+00 |

+------------------------------+

display 1-0.95949459: .04050541

1. .10510908
2. .01898816
3. .02327884

**Generate Random Numbers**

Statistical Concepts

Random number generation is important for participant selection for surveys, assignment to experimental groups, etc. When using a simple random, stratified random, randomized block, or clustered sampling design process, random numbers are helpful and necessary for proper selection or group assignment. Using random numbers helps to reduce the possibility of selection bias for surveys and can also reduce possible confounding variables in an experimental design.

Code

Assign the individuals in the dataset to treatment and placebo groups. Assign the individuals using a Randomized Block Design, where gender is the blocking variable.

Create a summary table for each treatment group. Only include the variables G and H.

In this code section, show the code to create the two groups. You do not need to show the code to create the two summary tables.

sort G

generate RannumF=uniform() in 1/1421

generate RannumM=uniform() in 1422/2829

egen MeanRannumF=mean(RannumF)

egen MeanRannumM=mean(RannumM)

gen Trt=1

replace Trt=0 if RannumM<MeanRannumM & G==1

replace Trt=0 if RannumF<MeanRannumF & G==0

/\*Note: following is to name Trt groups and create summary table\*/

label define trtlabel 0 "Placebo" 1 "Treatment"

label values Trt trtlabel

tabulate Trt G, summarize(H)

Output

Means, Standard Deviations and Frequencies of Health Score

| Gender

Trt | Female Male | Total

-----------+----------------------+----------

Placebo | 2.0136702 5.1502016 | 3.5785968

| 2.3268986 3.3736726 | 3.2934683

| 706 703 | 1409

-----------+----------------------+----------

Treatment | 1.910355 5.0637821 | 3.4759649

| 2.4008355 3.4556841 | 3.3634104

| 715 705 | 1420

-----------+----------------------+----------

Total | 1.9616854 5.1069304 | 3.5270813

| 2.3641227 3.4140427 | 3.3285662

| 1421 1408 | 2829

**Perform a t-Test on X**

Statistical Concepts

To use the t-test for a mean, data must be continuous. It must also be either distributed normally, with an unknown variance, or the distribution can be unknown, if the sample size is sufficiently large (n>30). The t-test can be used with non-normal data, but in this situation, the sample size must be sufficiently large (n>30).

The hypothesis of interest is the alternative hypothesis, also known as the research hypothesis. In a one sample t-test of a mean, possible hypotheses of interest are:

* + H1: µ ≠ µ 0 (two-sided)
  + H1: µ > µ 0 (one-sided)
  + H1: µ < µ0 (one-sided)

To interpret the results of a t-test, one compares the resulting t-value to the t-distribution. t-distribution values of interest are tn-1,α/2 or tn-1,1-α/2, depending on the alternative hypothesis. For a two-sided test, you would reject the null hypothesis if t ≤ tn-1,α/2  or if t ≥ tn-1,1-α/2. Additionally, one can use the p-value to determine whether or not to reject the null hypothesis. If p > α, fail to reject the null. If p<α, reject the null in favor of the alternative hypothesis.

Code

Perform a two-sided t-Test on H. Let = 3.5. Perform the test at a 0.01 significance level.

ttest H=3.5, level(99)

Output

* Highlight the p-value
* Highlight the value of the test statistic.

One-sample t test

----------------------------------------------------------------------------- Variable | Obs Mean Std. Err. Std. Dev. [99% Conf. Interval]

---------+--------------------------------------------------------------------

H | 2829 3.527081 .0625807 3.328566 3.365775 3.688387

------------------------------------------------------------------------------

mean = mean(H) t = 0.4327

Ho: mean = 3.5 degrees of freedom = 2828

Ha: mean < 3.5 Ha: mean != 3.5 Ha: mean > 3.5

Pr(T < t) = 0.6674 Pr(|T| > |t|) = 0.6652 Pr(T > t) = 0.3326

**Perform Two Sample t-Test**

Statistical Concepts

For a two sample t-test, one variable must be continuous (such as health scores), and the other must be a binomial discrete variable (such as gender).

The hypothesis of interest looks at the difference between means and compares the difference to a stated value. It states that the difference between the mean for one population and the mean for the other population is either not equal to, greater than, or less than a stated amount (Δ0 ). It is stated as one of the following:

* H1: µ1 - µ 2 ≠ Δ0
* H1: µ1 - µ 2 > Δ0
* H1: µ1 - µ 2 < Δ0

Similar to a one sample t-test, if the resulting test statistic (td : t-stat associated with the difference between means) falls within the rejection region for the hypothesis of interest, the null hypothesis is rejected in favor of the alternative. For example if H1: µ1 - µ 2 < Δ0, then the rejection region includes values of td less than or equal to tn-1,α.

Code

Compare the values of H for males and females. Test the hypothesis that

Perform the test at a 0.1 significance level.

tabulate G, summarize(H)/\*Note: Look at resulting summary table to determine if F and M have equal or unequal variance. Determine unequal\*/

ttest H, by(G) unequal level(90)

Output

* Highlight the value of the test statistic; Highlight the p-value

Two-sample t test with unequal variances

----------------------------------------------------------------------------

Group | Obs Mean Std. Err. Std. Dev. [90% Conf. Interval]

---------+-------------------------------------------------------------------

Female | 1421 1.961685 .0627152 2.364123 1.858461 2.06491

Male | 1408 5.10693 .0909845 3.414043 4.957176 5.256685

---------+-------------------------------------------------------------------

combined | 2829 3.527081 .0625807 3.328566 3.424111 3.630051

---------+-------------------------------------------------------------------

diff | -3.145245 .1105051 -3.327077 -2.963413

-----------------------------------------------------------------------------

diff = mean(Female) - mean(Male) t = -28.4624

Ho: diff = 0 Satterthwaite's degrees of freedom = 2501.99

Ha: diff < 0 Ha: diff != 0 Ha: diff > 0

Pr(T < t) = 0.0000 Pr(|T| > |t|) = 0.0000 Pr(T > t) = 1.0000

Note: Since the hypothesis of interest was , but STATA assumes , the p-value for Ha < 0 is chosen, as it is closest to representing -3, which would be the hypothesis of interest if the original null hypothesis were reordered: . We would REJECT the null for this test.

**Perform a Binomial Proportion Test**

Statistical Concepts

A binomial proportion test involves discrete data where there are only two possible outcomes. The proportion (p) is the parameter of interest. For example, in the example below we are evaluating what proportion of participants are democrats, versus the proportion of participants who are every other party: republicans, other, or unaffiliated.

The hypothesis of interest is also called the alternative hypothesis and compares the proportion of interest with a stated proportion value (0.5, for example). In this situation it is one of the following:

* + H1: p ≠ p0
  + H1: p > p0
  + H1: p < p0

In a binomial proportion test using the Z-test method, if the calculated z-value falls within the rejection region, the alternative hypothesis is chosen and the null is rejected. Also, if the resulting p-value is greater than the stated alpha level, you fail to reject the null hypothesis. In that case, you would conclude that the proportion of interest (p, or proportion of democrats in this example) was either equal to, greater than or equal to, or less than or equal to the stated p0, depending on your original null hypothesis because there was not sufficient significant evidence to suggest otherwise.

Code: Test the two-sided hypothesis that the proportion of democrats is 0.5. Party affiliation is coded in the variable P. Perform the test at a 0.05 significance level.

gen PDummy=1 if P=="D"

replace PDummy=0 if PDummy==.

prtest PDummy=0.5

Output: Highlight the value of the test statistic; Highlight the p-value

One-sample test of proportion PDummy: Number of obs = 2829

----------------------------------------------------------------------------

Variable | Mean Std. Err. [95% Conf. Interval]

-------------+---------------------------------------------------------------

PDummy | .4828561 .009395 .4644422 .50127

-----------------------------------------------------------------------------

p = proportion(PDummy) z = -1.8237

Ho: p = 0.5

Ha: p < 0.5 Ha: p != 0.5 Ha: p > 0.5

Pr(Z < z) = 0.0341 Pr(|Z| > |z|) = 0.0682 Pr(Z > z) = 0.9659

Note: We would fail to reject the Null hypothesis because p > .05.

**Two Sample Binomial Proportion Test**

Statistical Concepts

In a two sample binomial proportion test, both variables must be discrete, binomial variables. For example, someone may be interested in whether the proportion of people who affiliate with a given political party (ex: democratic or republican) differs by gender (male or female).

For a two sample binomial proportion test, the hypothesis of interest looks at the ***difference*** (δ or p1 – p2) between the proportions of interest (political party in this example) by group (gender in this example), and compares this observed difference(p1 – p2) to an expected difference (δ0). The alternative hypotheses for two sample binomial proportion tests are one of the following:

* + H1: p1 – p2 ≠ δ0
  + H1: p1 – p2 > δ0
  + H1: p1 – p2 < δ0

Code

Compare the proportion of democrats among males and females. Test the hypothesis that the proportions are equal. Perform the test at a 0.025 significance level.

prtest PDummy, by(G) level(97.5)

Output

* Highlight the value of the test statistic
* Highlight the p-value

Two-sample test of proportions Female: Number of obs = 1421

Male: Number of obs = 1408

-----------------------------------------------------------------------------

Variable | Mean Std. Err. z P>|z| [97.5% Conf. Interval]

-------------+---------------------------------------------------------------

Female | .513019 .0132595 .4832992 .5427388

Male | .4524148 .0132646 .4226835 .482146

-------------+---------------------------------------------------------------

diff | .0606042 .0187553 .018566 .1026424

| under Ho: .0187903 3.23 0.001

-----------------------------------------------------------------------------

diff = prop(Female) - prop(Male) z = 3.2253

Ho: diff = 0

Ha: diff < 0 Ha: diff != 0 Ha: diff > 0

Pr(Z < z) = 0.9994 Pr(|Z| < |z|) = 0.0013 Pr(Z > z) = 0.0006

Note: We would Reject the Null hypothesis because p < .025.

**Confidence Intervals**

**Statistical Concepts**

Confidence intervals tell us that if we take a large number of samples, then about 100 \* (1-α) of the intervals would cover, or contain, the true parameter of interest (e.g. µ, σ2, p). For example, if we were looking for the mean Health Score at α=.05, and we took a large number of samples of Health Scores, we would expect the true mean of Health Scores to be covered by, or contained in, 95% of the calculated confidence intervals.

Code

Calculate the 90% confidence interval for

* The mean of H: ci H, level(90)
* The difference in means of H between males and females

ttest H, by(G) unequal level(90)

* The proportion of democrats: ci PDummy, binomial level(90)
* The difference in proportions of democrats between males and females

prtest PDummy, by(G) level(90)

Output

**Mean of H:**

Variable | Obs Mean Std. Err. [90% Conf. Interval]

-------------+---------------------------------------------------------------

H | 2829 3.527081 .0625807 3.424111 3.630051

**Difference in Means of H between males and females:**

Two-sample t test with unequal variances

----------------------------------------------------------------------------

Group | Obs Mean Std. Err. Std. Dev. [90% Conf. Interval]

---------+------------------------------------------------------------------

Female | 1421 1.961685 .0627152 2.364123 1.858461 2.06491

Male | 1408 5.10693 .0909845 3.414043 4.957176 5.256685

---------+------------------------------------------------------------------

combined | 2829 3.527081 .0625807 3.328566 3.424111 3.630051

---------+------------------------------------------------------------------

diff | -3.145245 .1105051 -3.327077 -2.963413

**The proportion of democrats:**

-- Binomial Exact --

Variable | Obs Mean Std. Err. [90% Conf. Interval]

-------------+---------------------------------------------------------------

PDummy | 2829 .4828561 .009395 .4672443 .4984943

**The difference of proportions of democrats between males and females:**

Two-sample test of proportions Female: Number of obs = 1421

Male: Number of obs = 1408

----------------------------------------------------------------------------

Variable | Mean Std. Err. z P>|z| [90% Conf. Interval]

-------------+--------------------------------------------------------------

Female | .513019 .0132595 .4912091 .5348289

Male | .4524148 .0132646 .4305965 .474233

-------------+---------------------------------------------------------------

diff | .0606042 .0187553 .0297545 .091454

| under Ho: .0187903 3.23 0.001

**Tabulate r X c Tables**

Statistical Concepts

r x c tables must be comprised of discrete data. Depending on the values of r and c, the discrete data can be either binomial or multinomial discrete data. The “r” describes the number of rows and “c” describes the number of columns in the table. The table below looking at gender and political party would be a 2 x 4 table.

Code

Tabulate the gender by party table. Be sure to correctly label each gender and party.

encode P, generate(P2) /\*Note: must encode the string variable in order to assign labels to it\*/

label define partylabel 1 "Democrat" 2 "Other" 3 "Republican" 4 "Unaffiliated"

label values P2 partylabel

tabulate G P2

Output

| Political Party

Gender | Democrat Other Republica Unaffilia | Total

-----------+--------------------------------------------+----------

Female | 729 27 533 132 | 1,421

Male | 637 24 604 143 | 1,408

-----------+--------------------------------------------+----------

Total | 1,366 51 1,137 275 | 2,829

**Test of No Association for r X c Tables**

Statistical Concepts

The Test of No Association is a test of independence between two discrete random variables, which can be either binomial or multinomial discrete random variables. Two variables are independent if P(Xi ᴖ Yj) = P(Xi) \* P(Yj) for all i = 1,2 and j = 1,2,3. The test is also known as Pearson’s Chi-Squared Test.

To interpret the results of this test, one compares the resulting test statistic to the Chi-squared (X2) distribution at the corresponding (r-1) and (c-1) degrees of freedom at 1 - α. With the Chi-squared distribution, the extreme values are always to the right. This means if the resulting test statistic is larger than the X2 (r-1)(c-1)(1-α) value, the null hypothesis of no association is rejected. Additionally, one can compare the resulting p-values. If the stated alpha is greater than the resulting p-value, one would reject the null hypothesis of no association.

To use the chi-squared test, the cell counts must be greater than 5 in every cell.

Code

Perform a test of no association on the gender by party table. Perform the test at a 0.025 significance level.

set level 97.5

tabulate G P2, chi2

Output

* Highlight the value of the test statistic
* Highlight the p-value

| Political Party

Gender | Democrat Other Republica Unaffilia | Total

-----------+--------------------------------------------+----------

Female | 729 27 533 132 | 1,421

Male | 637 24 604 143 | 1,408

-----------+--------------------------------------------+----------

Total | 1,366 51 1,137 275 | 2,829

Pearson chi2(3) = 11.1868 Pr = 0.011

Note: At alpha=.025, reject the null hypothesis of no association

**Goodness-of-Fit Test**

Statistical Concepts

The goodness of fit test determines how well a set of observed counts fits its expected outcomes, given the null hypothesis is correct. The null hypothesis is this case is that there is good agreement between the observed and expected counts. The alternative hypothesis (hypothesis of interest) is that the observed counts do not provide a good fit to the counts that are expected.

The Goodness of Fit Test applies to the data set associated with a single discrete, multinomial random variable, collected for a given number of individuals (“n”).

The goodness of fit test is a one-sided test, and it uses the Chi-squared distribution. One compares the resulting test statistic to the X2 distribution at (k-1) degrees of freedom. The null hypothesis is rejected if the test statistic is larger than X2(k-1). If the null is rejected, one would conclude that observed counts do not provide a good fit to expected counts.

Code

Perform a goodness-of-fit test on party. The null hypothesis is

Perform the test at a 0.05 significance level.

set level 95

csgof P2, expperc(50 1 40 9)

Output

* Highlight the value of the test statistic
* Highlight the p-value

+--------------------------------------------+

| P2 expperc expfreq obsfreq |

|--------------------------------------------|

| Democrat 50 1414.5 1,366 |

| Other 1 28.29 51 |

| Republican 40 1131.6 1,137 |

| Unaffiliated 9 254.61 275 |

+--------------------------------------------+

chisq(3) is 21.55, p = .0001

**Fisher’s Exact Test**

Statistical Concepts

Fisher’s Exact Test is another way to evaluate the hypothesis of no association. It considers the range of all possible table configurations of cell counts, given that the row and column totals are fixed. It computes the probabilities of the observed configuration occurring and more extreme configurations occurring by chance.

The data used must be measurements of two discrete, binomial variables. In the example below, students either go to UNC or Duke (Variable 1), and they are either from NC or Not North Carolina (Variable 2).

The hypothesis of interest is whether or not the two variables are independent. It essentially asks: Will knowing something about variable 1 give me any more information about the values of variable 2? The null hypothesis assumes the variables are independent, or that there is no association between the two, but the hypothesis of interest (alternative hypothesis) assumes there *is* an association between the variables, or they are NOT independent.

We would use Fisher’s Exact test when evaluating the hypothesis of no association between two discrete, binomial variables when cell counts are less than 5. I.e. you would use this test if the sample size was not large enough to use the X2 distribution, as discussed in the Test of No Association concept.

Code: Perform Fisher’s Exact Test on the following data: Of 8 UNC students, 5 are from NC. Of 8 Duke students, 3 are from NC.

tabi 3 5 \ 5 3, exact

Output

| col

row | 1 2 | Total

-----------+----------------------+----------

1 | 3 5 | 8

2 | 5 3 | 8

-----------+----------------------+----------

Total | 8 8 | 16

Fisher's exact = 0.619

1-sided Fisher's exact = 0.310

Note: Row 1=Duke, 2 = UNC. Column 1=NC, 2 = Not NC

**Wilcoxon-Signed-Rank Test**

Statistical Concepts

The signed rank test is a non-parametric, one sample median test that assesses if an observed sample median is significantly different from a hypothesized value. It is appropriate to use if the distribution of data is non-normal, or if the “n” is small. The test is used to compare the center of the population to a given value. Its parametric equivalent is a one sample t-test.

When the data is not normal, the variance of the data is very large or ∞, or there are outliers in the sample, the Wilcoxon-Signed-Rank Test is preferred to the t-test. The WSR is robust to outliers, while the t-test is not.

If the resulting test statistic falls within the rejection region, one rejects the null hypothesis in favor of the alternative hypothesis. In the example below, the alternative hypothesis is m0≠3.5. In a two sided test, the “rejection region” is at the two extreme ends of the data. Based on the results below, we would reject the null hypothesis. As always, it is also possible to assess the result based on the p-value. IF p<α, reject the null.

Code

Perform a Wilcoxon-Signed-Rank Test on H. Let = 3.5. Perform the test at a 0.01 significance level.

set level 99

signrank H=3.5

Output

Wilcoxon signed-rank test

sign | obs sum ranks expected

-------------+---------------------------------

positive | 1131 1743983 2001517.5

negative | 1698 2259052 2001517.5

zero | 0 0 0

-------------+---------------------------------

all | 2829 4003035 4003035

unadjusted variance 1.888e+09

adjustment for ties 0

adjustment for zeros 0

----------

adjusted variance 1.888e+09

Ho: H = 3.5

z = -5.927

Prob > |z| = 0.0000

**Wilcoxon-Rank-Sum Test**

Statistical Concepts

The Rank Sum Test is appropriate when comparing the middle of two populations. Its equivalent among the parametric tests is a two sample t-test.

When the data is not normal, the variance of the data is very large or ∞, or there are outliers in the sample, the Wilcoxon-Rank-Sum Test is preferred to the two sample t-test. The WRS is robust to outliers, while the t-test is not.

For a two-sided test (such as the example below), one interprets the test statistic (RWRS) as follows: if the calculated test statistic, which is the sum of the ranks for the smaller sample, deviates greatly from the expected RWRS , then the null hypothesis is rejected in favor of the alternative hypothesis.

Code

Compare the values of H for males and females with a rank sum test.

Perform the test at a 0.05 significance level.

ranksum H, by(G)

Output

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

G | obs rank sum expected

-------------+---------------------------------

Female | 1421 1365980 2010715

Male | 1408 2637055 1992320

-------------+---------------------------------

combined | 2829 4003035 4003035

unadjusted variance 4.718e+08

adjustment for ties 0

----------

adjusted variance 4.718e+08

Ho: H(G==Female) = H(G==Male)

z = -29.681

Prob > |z| = 0.0000

**ANOVA**

Statistical Concepts

To perform an ANOVA, one looks at two variables of interest. One of the variables must be continuous (health score in the example below) and the other should be a multinomial discrete variable (political party in the example below).

The null hypothesis assumes that all group means are equal. In the example below, it assumes that the mean health score is the same for all political parties. The alternative, or hypothesis of interest, says that there is ***at least*** one difference between means. I.e. ***at least*** one political party does not have an equivalent mean health score to the others.

To interpret the results of ANOVA, the F-table is used. The null hypothesis is rejected if the calculated test statistic, or F value, is greater than the expected F value (Fr-1, n-r, 1-α).

Multiple comparison correction is used to control Type I error. It is important to control for the possibility of identifying a significant difference when in fact, there is not one. It is appropriate to use when multiple tests are performed and a significant result is found (i.e. when the null hypothesis is rejected.) There are many types of multiple comparison procedures, including Fisher’s LSD method, Dunnett’s method, The Tukey-Kramer method, and the Bonferroni method. Bonferroni controls for Type I error by multiplying the resulting p-value by the number of comparisons being made. Then this multiplied p-value is used to compare to alpha.

Code

Compare the mean of H between political parties.

oneway H P

Output: (Report the ANOVA table).

Analysis of Variance

Source SS df MS F Prob > F

------------------------------------------------------------------------

Between groups 188.844531 3 62.948177 5.71 0.0007

Within groups 31143.5661 2825 11.0242712

------------------------------------------------------------------------

Total 31332.4106 2828 11.0793531

Bartlett's test for equal variances: chi2(3) = 29.8043 Prob>chi2 = 0.000

**Linear Regression**

**Statistical Concepts**

In its basic form, we can use linear regression to model the (linear) relationship between a continuous response variable (Y) and continuous predictor variables (X1, X2, … ). The basic form of the model is

If the Xs do predict Y in this way, the parameters estimate how Y changes for a unit change in X (holding all else constant). For example, the mean of Y goes up on average by for every unit increase of while holding all other Xs constant.

If there is linear association between the response and a predictor, we would expect the corresponding to be different than 0. It follows that the hypothesis of no association between Y and (while controlling for other predictors) is

The test of overall association, , is also of interest. Model fit may be assessed with graphical diagnostics.

**Code**

Regress X and Y onto H. Assess model fit with graphical diagnostics.

summary(lm1 <- lm(H~X+Y))

plot(lm1)

**Output**

Coefficients:

Estimate Std. Error t value Pr(>|t|)

Estimates

of betas. Tests of single beta

(Intercept) 3.5587546 0.1758252 20.240 <2e-16 \*\*\*

X -0.0002635 0.0208711 -0.013 0.990

Y -0.0166140 0.0626132 -0.265 0.791

---

Residual standard error: 3.33 on 2826 degrees of freedom

Multiple R-squared: 2.497e-05, Adjusted R-squared: -0.0006827

Test of

Overall Association

F-statistic: 0.03528 on 2 and 2826 DF, p-value: 0.9653

**Linear Regression (continued)**

**Output**

