**BIOS 600 HONORS PROJECT**

**NAME:**

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

(Sign and date the submitted paper copy.)

**Calculate Summary Tables and Statistics**

Statistical Concepts

* What are some of the properties that summary statistics describe?
* Which summary statistics are usually included in a summary table?
* A summary table should always include information about \_\_\_\_\_\_?
* Does data type matter when calculating summary statistics?

Code

Create a summary table of H, Y, G, and P. Calculate a measure of center and spread for the continuous variables. Calculate the frequency table for the categorical variables.

Output

**Create a Scatterplot of X and Y**

Statistical Concepts

* What type of data must X and Y be?
* What do scatterplots communicate to the reader?
* Scatterplots are helpful when identifying \_\_\_\_\_ ?

Code

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

Output

**Create a Histogram of X**

Statistical Concepts

* What type of data must X be?
* What do histograms communicate to the reader?

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.

Output

**Create a Boxplot of X**

Statistical Concepts

* What type of data must X be?
* What do boxplots communicate to the reader?
* Scatterplots are helpful when identifying \_\_\_\_\_ ?

Code

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

Output

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

Statistical Concepts

* What do line plots communicate to the reader?

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.

Output

**Calculate Probabilities from Standard Distributions**

Statistical Concepts

* What is a probability distribution function (pdf)?
* What is a probability mass function (pmf)?
* What is a distribution parameter?

Code

Calculate

* P( W < 1.8 ) when W is normally distributed with mean 1 and variance 2
* P( V < 23 ) when V is BIN( n = 44, p = 0.4 )
* P( U > 3 ) when U is Poisson with mean .75
* P( 1.2 < T < 2.3 ) when T is t( df = 46 )
* P( S > 2 ) when S is χ2 with df = 7
* P( R > 3.2 ) when R is F with df1 = 4 and df2 = 38

Output

**Generate Random Numbers**

Statistical Concepts

* In which situations is random number generation important?

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.

Output

(Only show the summary tables.)

**Perform a t-Test on X**

Statistical Concepts

* What type of data must X be?
* What is the hypothesis of interest?
* How does one interpret the results of a t-test?
* Are there sample size considerations?

Code

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

Output

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

**Perform Two Sample t-Test**

Statistical Concepts

* What type of data is needed for a two sample t-test?
* What is the hypothesis of interest?
* How does one interpret the results of a two-sample t-test?

Code

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

Perform the test at a 0.1 significance level.

Output

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

**Perform a Binomial Proportion Test**

Statistical Concepts

* What type of data is involved in a binomial proportion test?
* What is the parameter of interest?
* What is the hypothesis of interest?
* How does one interpret the results of the test?

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.

Output

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

**Two Sample Binomial Proportion Test**

Statistical Concepts

* What type of data is involved in a binomial proportion test?
* What is the hypothesis of interest?

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.

Output

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

**Confidence Intervals**

Statistical Concepts

* What do confidence intervals communicate?

Code

Calculate the 90% confidence interval for

* The mean of H
* The difference in means of H between males and females
* The proportion of democrats
* The difference in proportions of democrats between males and females

Output

**Tabulate r X c Tables**

Statistical Concepts

* What type of data is required for r X c tables?

Code

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

Output

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

Statistical Concepts

* What is the Test of No Association?
* How does one interpret the test results?

Code

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

Output

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

**Goodness-of-Fit Test**

Statistical Concepts

* What is the Goodness-of-Fit Test?
* What type of data does the test apply to?
* What is the hypothesis of interest?
* How does one interpret the test results?

Code

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

Perform the test at a 0.05 significance level.

Output

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

**Fisher’s Exact Test**

Statistical Concepts

* What is Fisher’s Exact Test?
* What type of data does the test apply to?
* What is the hypothesis of interest?
* When might we use Fisher’s Exact test?

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.

Output

**Wilcoxon-Signed-Rank Test**

Statistical Concepts

* When is the Signed Rank Test Appropriate? What is it equivalent to among the parametric tests?
* When is this test preferred the parametric equivalent?
* How does one interpret the test result?

Code

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

Output

**Wilcoxon-Rank-Sum Test**

Statistical Concepts

* When is the Rank Sum Test Appropriate? What is it equivalent to among the parametric tests?
* When is this test preferred the parametric equivalent?
* How does one interpret the test result?

Code

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

Perform the test at a 0.05 significance level.

Output

**ANOVA**

Statistical Concepts

* What type of data does ANOVA require?
* What is the hypothesis of interest?
* How does one interpret the results of ANOVA?
* What is multiple comparison correction, why is it important, and when is it appropriate?

Code

Compare the mean of H between political parties.

Output

(Report the ANOVA table).

**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**

