BIOS 600 HONORS PROJECT

NAME: [redacted]

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

Summary statistics such as mean, median, and mode are used to describe the center of a sample or population, while statistics like min and max, percentile, standard deviation, and variance are used to describe spread. Summary tables often include mean, min, max, variance or standard deviation, and the number of observations. Data type influences summary statistic calculations and guides what statistics are useful. For example, for continuous variables, measures like mean, variance (and standard deviation), min, max, median, and percentile are meaningful. However, these values are not meaningful for categorical variables. For example, there is no mean, min, or max of genders in a population. Conversely, values such as mode are useful for describing a population or sample. If the values of categorical variables can be ordered, such as in the case of ordinal variables, median can be useful to describe a the central tendency of the data as well.

Code

```
label define Gender 0 "Female" 1 "Male"
label values G Gender
Sum H Y G P
tab G
tab P
```

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

```
  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
         P |         0
```

Gender	Freq.	Percent	Cum.
Female	1,421	50.23	50.23
Male	1,408	49.77	100.00
Total	2,829	100.00	

PoliticalPa rty	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	

Create a Scatterplot of X and Y

Statistical Concepts

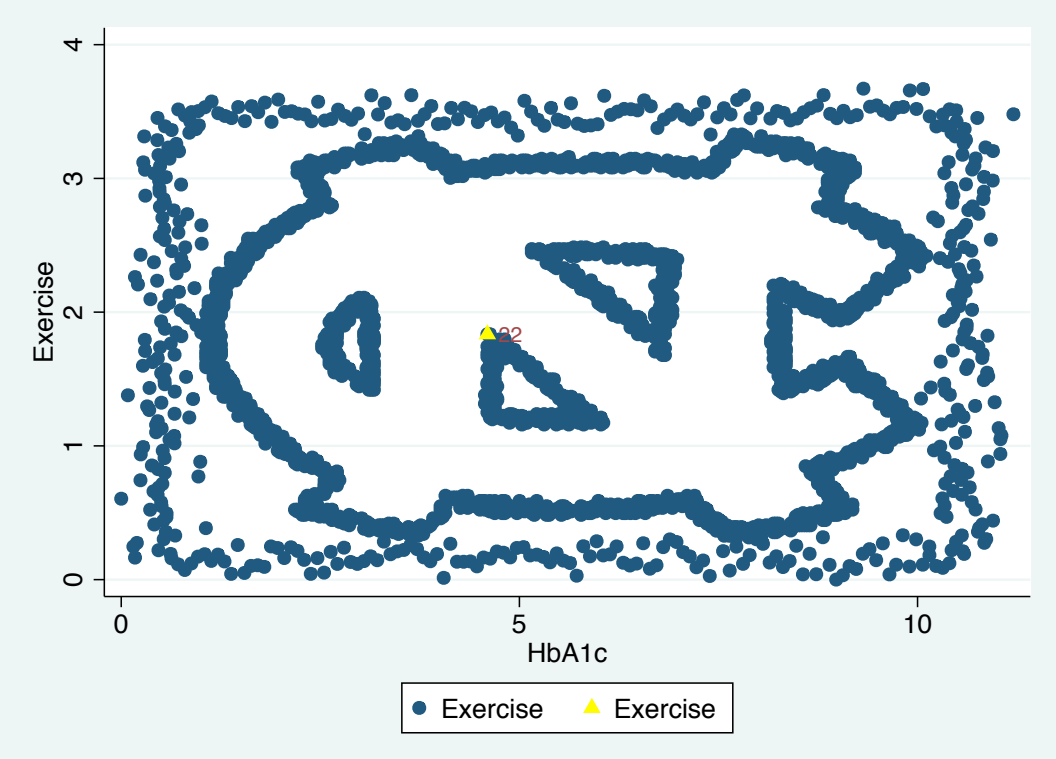
Scatterplots communicate potential linear associations between two continuous variables. While categorical variables can be used, the visual representation between two continuous variables is likely to provide more useful information about trends. If x increases but y does not change, this may signal to the user that a linear association between the two variables does not exist. Furthermore, the shape of the points can demonstrate a positive or negative association.

Code

```
scatter Y X || scatter Y X if obs==22, mlabel(obs) mcolor(yellow)
msymbol(triangle)
```

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

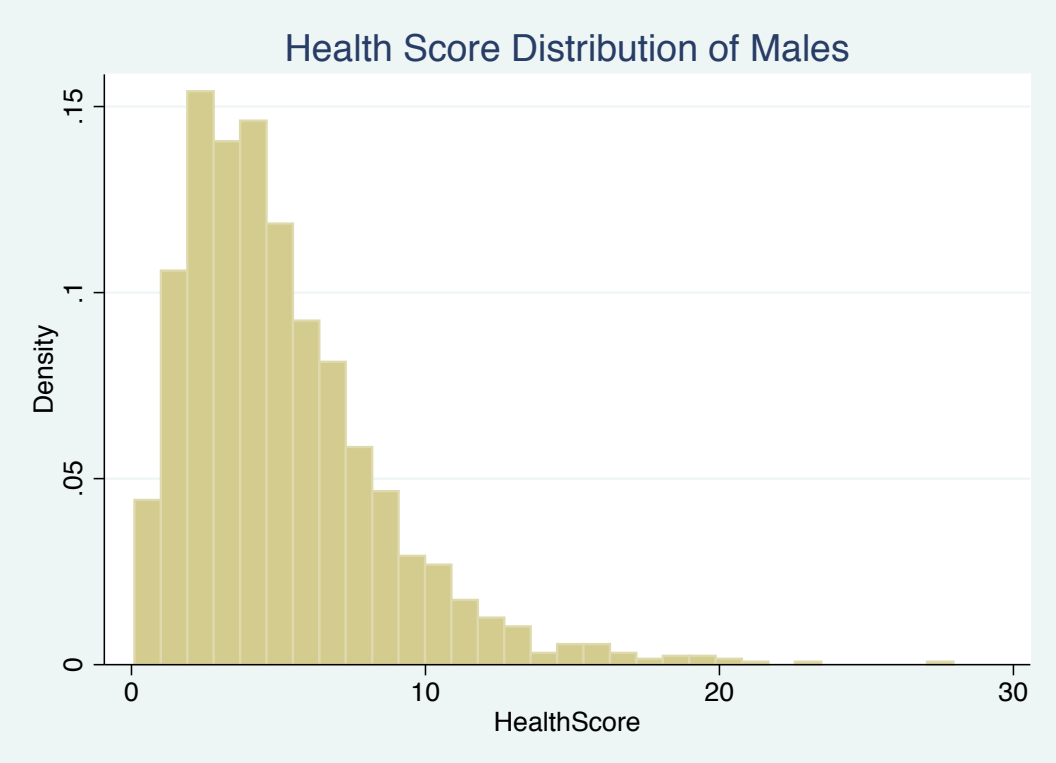
Histograms can depict continuous or categorical data (although bar charts are better suited for categorical data). Histograms communicate frequency of the values in a given bin. With these bins and corresponding frequencies displayed next to each other, histograms allow for easy visual comparisons of frequencies as well as preliminary conclusions regarding the center and spread of data.

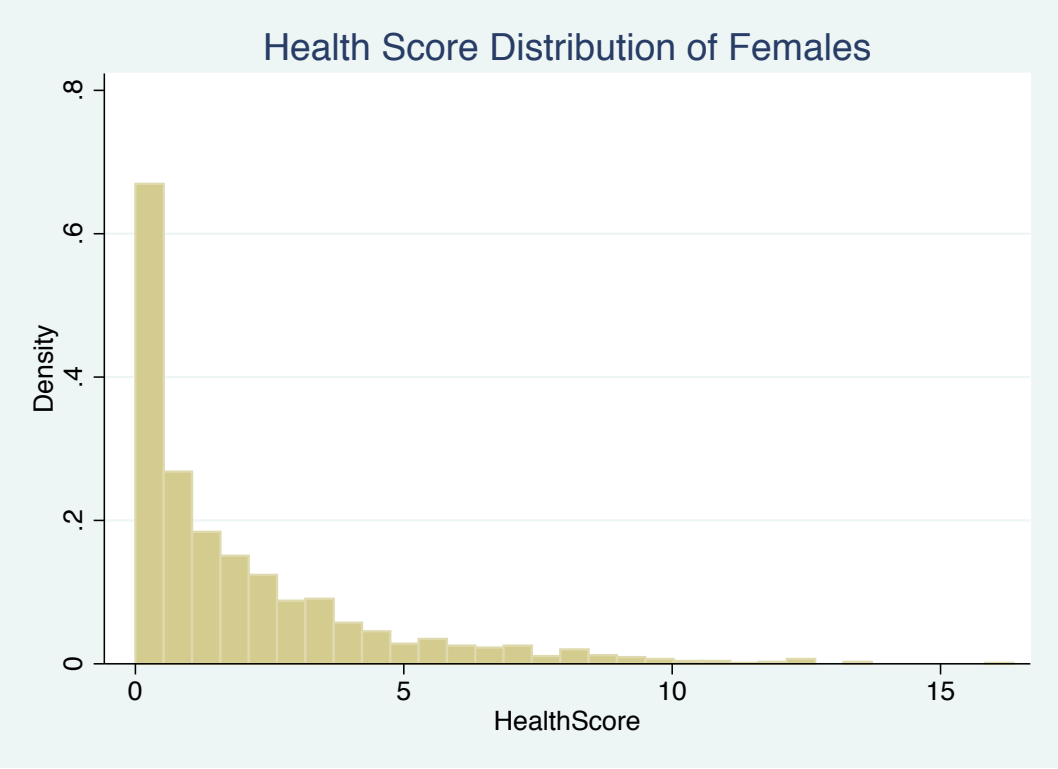
Code

```
histogram H if G, title("Health Score Distribution of Males")
histogram H if !G, title("Health Score Distribution of Females")
```

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




Health Score Distribution of Females
Density
0
.2
.4
.6
.8
0
5
10
15
HealthScore

Create a Boxplot of X

Statistical Concepts

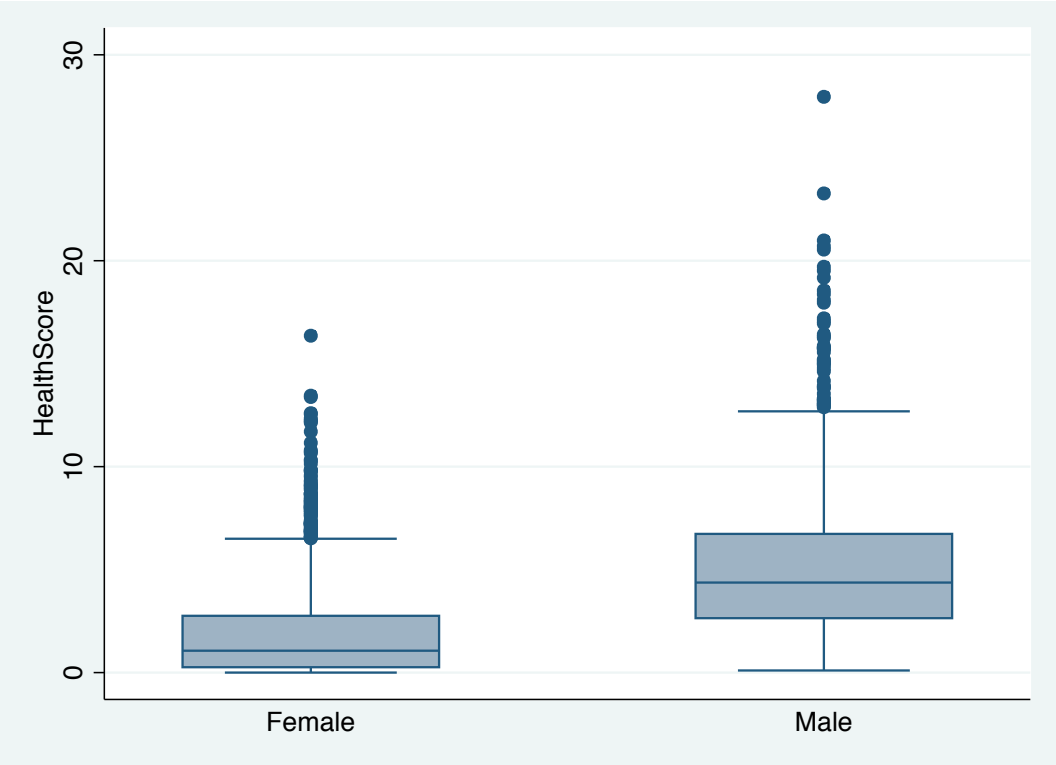
Boxplots can be used for continuous data and communicate the variance, min, max, median, and quartiles of a variable. The whiskers represent the range of data with the hinges of the box showing the 25th and 75th percentiles (or first and third quartiles). The line through the center of the box displays the median, or 50th percentile. Similar to the easy visual use of histograms, boxplots provide a means of describing both center and skewedness.

Code

graph box H, over (G)

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

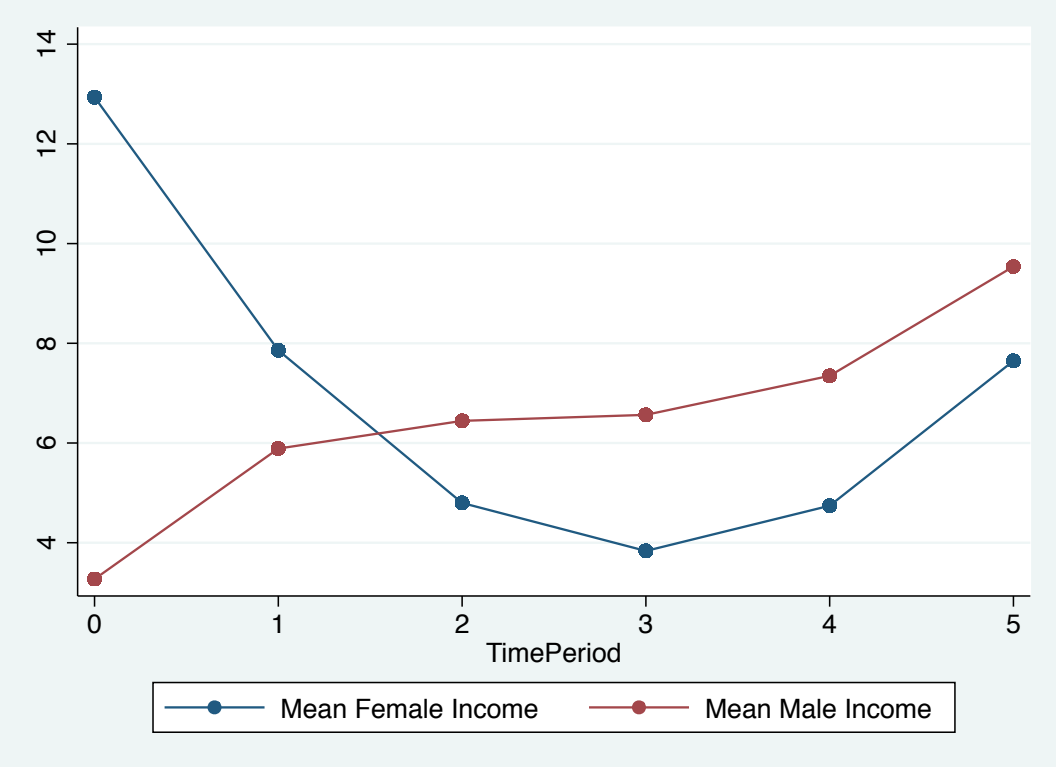
Line plots show how the value of a given variable changes across different values of another variable, usually time.

Code

```
egen Rmeanfem = mean(R) if G==0, by(T)
egen Rmeanmale = mean(R) if G==1, by(T)
label var Rmeanmale "Mean Male Income"
label var Rmeanfem "Mean Female Income"
twoway (connected Rmeanfem T, sort) (connected Rmeanmale T, sort), legend(on)
```

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

Probability distribution functions and probability mass functions shows the different values that a random variable can take and the probability of that variable being a certain value. PDFs are used for continuous variables, while PMFs are used to described discrete variables.

A distribution parameter describes a certain characteristic of a distribution such as the mean, standard variation, or degrees of freedom.

Code

Calculate

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

```
display normalden(1.8,1,2^0.5)
```

- P(V < 23) when V is BIN(n = 44, p = 0.4)

```
display binomial(44,23,0.4)
```

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

```
display poisson(0.75,3)
```

- P(1.2 < T < 2.3) when T is t(df = 46)

```
display (tden(47,1.2)-tden(47,2.3))
```

- P(S > 2) when S is χ2 with df = 7

```
display chi2(7,2)
```

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

```
display Fden(4,38,3.2)
```

Output

.24038532

.9641045

.99270783

.16164786

.04015963

.03032522

Generate Random Numbers

Statistical Concepts

Random number generation is important to control selection biases. For example, in an experimental study, subjects for control and treatment groups need to be selected randomly so that only the presence of the treatment is the true difference between the two groups. For example, at UNC Hospital, an audit of central line insertions is performed each month. As the Patient Safety Department's intern, I receive a list of patients that needed in a central line in a given month. I must randomly select 30 patients to audit for staff compliance with established protocols. If I were to just pick the first 30 patients in the list, I would probably be auditing 30 patients in the same unit or 30 patients who needed a central line early in the month, depending on how the sheet is sorted. Unit location or date of the procedure may somehow effect compliance and would not allow me to perform an accurate audit of current compliance practices. As a result, each patient must be tied to a random number and sorted to avoid selection bias.

Code

```
generate rannum = uniform()
sort rannum
egen trtmale = cut(rannum), group(2), if G==1
egen trtfemale = cut(rannum), group(2), if G==0
gen treatment= 0
replace treatment = 1 if trtmale==1 | trtfemale==1
sum H G if treatment==1
sum H G if treatment==0
```

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.

Output

Variable	Obs	Mean	Std. Dev.	Min	Max	**Treatment group
H	1415	3.583769	3.422468	.0000236	23.26638	
G	1415	.4975265	.5001707	0	1	

Variable	Obs	Mean	Std. Dev.	Min	Max	**Control group
H	1414	3.470354	3.232085	.0000564	27.95761	
G	1414	.4978784	.5001724	0	1	

Perform a t-Test on X

Statistical Concepts

Because t-Tests requires a mean, continuous data is needed for X. The hypothesis of interest is whether or not the mean of a population equals a certain number, for example the mean of the population. In the case below,

H_o (null hypothesis): xbar=3.5

H_1 (hypothesis of interest): xbar ≠ 3.5

In performing a t-Test, the resulting p-value is compared to the desired level of significance (e.g. 0.05) for the test. If the p-value is smaller than the significance level, H_o is rejected. If the p-value is larger than the significance level, one fails to reject H_o. In the case below, the p-value of the two-sided test is about 0.665, which is much greater than 0.01, the stated significance level. Therefore, we fail to reject H_o.

If the sample size is small, the data must be distributed normally. If the data is not distributed normally, a large sample size is needed to calculate the appropriate t-value.

Code

```
ttest H==3.5, level (99)
```

Perform a two-sided t-Test on H. Let μ_0 = 3.5. Perform the test at a 0.01 significance level.

Output

```
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

Again, continuous data that is normally distributed or is not normally distributed but covers a large sample size is needed. Here the two samples are assumed to be independent, meaning the health score responses for males are assumed not to affect the health scores of women (and vice versa). Two sample t-tests aim to demonstrate whether or two samples have the same mean or how the means differ. In this sample,

H_o(null hypothesis): $\mu_m - \mu_f = 3$

H_1 (hypothesis of interest): $\mu_m - \mu_f \neq 3$

To interpret the results of a two-sample t-test, the resulting p-values are again compared to the chosen level of significance. Because the hypotheses here simply use "does not equal" rather than "greater than" or "less than," the p-value for a two-sided test is used. If the p-value is less than the level of significance (0.1 below), then the null hypothesis is rejected. If the p-value is greater than the level of significance (as seen in the example below), then we fail to reject the null hypothesis .

Code

```
gen Hfemale = H if G==0
gen Hmale = H if G==1
gen Hfemale3=Hfemale+3
ttest Hmale==Hfemale3, level(90) unpaired
```

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

$$H_0 : \mu_m - \mu_f = 3$$

Perform the test at a 0.1 significance level.

Output

Two-sample t test with equal variances

```
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [90% Conf. Interval]
---------+--------------------------------------------------------------------
```

```
  Hmale |    1408    5.10693    .0909845    3.414043    4.957176    5.256685
Hfemale3 |    1421    4.961685    .0627152    2.364123    4.858461     5.06491
---------+--------------------------------------------------------------------
combined |    2829    5.033974    .0551699    2.934393    4.943198     5.12475
---------+--------------------------------------------------------------------
    diff |           .145245    .1103266           -.0362855    .3267756
------------------------------------------------------------------------------
    diff = mean(Hmale) - mean(Hfemale3)                           t =   1.3165
Ho: diff = 0                                     degrees of freedom =     2827

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9059         Pr(|T| > |t|) = 0.1881          Pr(T > t) = 0.0941
```

Perform a Binomial Proportion Test

Statistical Concepts

A binomial proportion test requires the total number of observations and the number of "successes" that occurred during those observations. The parameter of interest is the proportion of the observations that resulted in a success. The hypothesis of interest reflects a proportion that is not equal to some given value. For example, to test whether or not a success (the subject is a democrat) is as likely to occur as a failure (the subject is not a democrat), the hypothesis of interest would state that the success rate does not equal 0.5. The test is interpreted by examining the resulting p-value and comparing it to the chosen significance level. If the p-value (here, 0.0682) is less than the significance level (0.05), then one rejects the null hypothesis that half the subjects are democrats. If the p-value is greater than the significance level, then one fails to reject the null hypothesis, as in the case below because 0.068 > 0.05.

Code

```
encode P, gen(demornot)
recode demornot (1=1) (2=0) (3=0) (4=0)
prtest demornot=0.5 level(95)
```

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

```
One-sample test of proportion          demornot: Number of obs =     2829
------------------------------------------------------------------------------
    Variable |      Mean   Std. Err.                 [95% Conf. Interval]
-------------+----------------------------------------------------------------
    demornot |   .4828561    .009395                  .4644422     .50127
------------------------------------------------------------------------------
    p = proportion(demornot)                                  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

Two Sample Binomial Proportion Test

Statistical Concepts

Again, a binomial proportion test requires the total number of observations and the number of "successes" that occurred during those observations. However, in the came of a two-sample test, the different rates of "successes" are compared between two independent groups.

The hypothesis of interest is that the rates of "successes" are not equal to a given value or are the same. Depending on the status quo, the hypothesis of interest could also be that the rates of "successes" are equal to a given value or each other. In the example below, the hypothesis of interest is that the probability of selecting a democrat among female subjects is not equal to the probability of selecting a democrat among male subjects, implying that political party affiliation is influenced by gender.

Code

Prtest demornot, by (G) level(97.5) ← Note: Recall "demornot" was generated earlier to code political parties as Democrat or not Democrat

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

```
Two-sample test of proportions                0: Number of obs =     1421
                                              1: Number of obs =     1408
------------------------------------------------------------------------------
   Variable |      Mean   Std. Err.      z    P>|z|   [97.5% Conf. Interval]
------------+-----------------------------------------------------------------
          0 |    .513019   .0132595                      .4832992    .5427388
          1 |   .4524148   .0132646                      .4226835     .482146
------------+-----------------------------------------------------------------
```

```
        diff |   .0606042   .0187553                .018566    .1026424
             |  under Ho:   .0187903     3.23   0.001
------------------------------------------------------------------------------
        diff = prop(0) - prop(1)                      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
```

*reject null hypothesis (0.0013>0.025)

Confidence Intervals

Statistical Concepts

Confidence intervals communicate an established amount of uncertainty.

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 Hmale==Hfemale, level(90) unpaired
```

- The proportion of democrats

```
ci demornot, binomial level(90)
```

- The difference in proportions of democrats between males and females

```
prtest demornot, by (G) level(90)
```

Output

```
    Variable |       Obs       Mean    Std. Err.      [90% Conf. Interval]
-------------+---------------------------------------------------------------
           H |      2829    3.527081    .0625807       3.424111    3.630051
```

```
------------------------------------------------------------------------------
Variable |     Obs       Mean    Std. Err.   Std. Dev.   [90% Conf. Interval]
---------+--------------------------------------------------------------------
   Hmale |    1408    5.10693    .0909845    3.414043    4.957176    5.256685
 Hfemale |    1421    1.961685    .0627152    2.364123    1.858461     2.06491
---------+--------------------------------------------------------------------
combined |    2829    3.527081    .0625807    3.328566    3.424111    3.630051
---------+--------------------------------------------------------------------
    diff |           3.145245    .1103266                2.963714    3.326776

  diff = mean(Hmale) - mean(Hfemale)                      t =  28.5085
```

```
Ho: diff = 0                     degrees of freedom =     2827

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

```
                                  -- Binomial Exact --
    Variable |       Obs       Mean    Std. Err.      [90% Conf. Interval]
-------------+---------------------------------------------------------------
    demornot |      2829    .4828561     .009395       .4672443    .4984943
```

```
Two-sample test of proportions                0: Number of obs =     1421
                                              1: Number of obs =     1408
------------------------------------------------------------------------------
    Variable |       Mean   Std. Err.      z    P>|z|     [90% Conf. Interval]
-------------+----------------------------------------------------------------
           0 |    .513019   .0132595                      .4912091    .5348289
           1 |    .4524148  .0132646                      .4305965     .474233
-------------+----------------------------------------------------------------
        diff |    .0606042  .0187553                      .0297545     .091454
             |  under Ho:   .0187903     3.23   0.001
------------------------------------------------------------------------------
        diff = prop(0) - prop(1)                          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
```

Tabulate r X c Tables

Statistical Concepts

r x c tables require categorical data and discrete variables.

Code

```
tab G P
```

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

Output

```
    |          PoliticalParty
 Gender |         D         O         R         U |     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 examines or not one variable on the r X c table influences the variable. There is no association between two variables (i.e. the two variables are independent) if the probability of both variable outcomes occurring at the same time equals the probability of one of those outcomes multiplied by the probability of the other outcome.

$p(x_i \text{ and } y_j) = p(x_i) * p(y_j)$

If the above equation does not hold true, it is possible that the two variables are not independent. The test statistic used to assess this relationship is χ^2. If the p-value related to a specific χ^2 value and degrees of freedom is less than the significance level, a null hypothesis(of no association) will be rejected. If the p-value is higher than the significance level, one will fail to reject the null hypothesis of no association.

Code

```
tabulate G P, cchi2 chi2
```

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

Output

```
+------------------+
| Key              |
|------------------|
|    frequency     |
| chi2 contribution |
+------------------+
```

Gender	PoliticalParty D	O	R	U	Total
Female	729	27	533	132	1,421
	2.7	0.1	2.5	0.3	5.6
Male	637	24	604	143	1,408
	2.7	0.1	2.6	0.3	5.6
Total	1,366	51	1,137	275	2,829
	5.4	0.1	5.1	0.5	11.2

Pearson chi2(3) = 11.1868 ← test statistic Pr = 0.011 ← p-value

Conclusion: The p-value is less than level of significance, therefore the null hypothesis of no association should be rejected.

Goodness-of-Fit Test

Statistical Concepts

The Goodness-of-Fit Test describes differences between expected and observed values. The test utilizes discrete variables and categorical data. The hypothesis of interest is that actual probabilities do not equal the expected probability values. If the p-value is less than the significance level, than the null hypothesis of observed values equaling expected values is rejected. If the p-value is greater than the significance level, then one fails to reject the null hypothesis.

Code

csgof P, expperc(50 9 40 1)

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

$$H_o: p_{democrat} = 0.5 \; p_{republican} = 0.4 \; p_{unaffiliated} = 0.09$$

Perform the test at a 0.05 significance level.

Output

csgof P, expperc(50 9 40 1)

```
+--------------------------------+
| P   expperc   expfreq   obsfreq |
|--------------------------------|
| D       50    1414.5    1,366 |
| O        9    254.61       51 |
| R       40    1131.6    1,137 |
| U        1    28.29        275 |
+--------------------------------+
```

chisq(3) is 2316.01 ← test statistic, p = 0 ← p-value

Fisher's Exact Test

Statistical Concepts

Fisher's Exact Test utilizes categorical data to test a null hypothesis of no association. The hypothesis of interest expresses the existence of association or non-independence between two variables. Fisher's Exact Test is used for testing the hypothesis of no association in small sample sizes that do not necessarily approximate a chi-squared distribution.

Code

```
tabi 5 3 \ 3 5, chi2 exact
```

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

```
           |        col
       row |         1         2 |     Total
-----------+----------------------+----------
         1 |         5         3 |         8
         2 |         3         5 |         8
-----------+----------------------+----------
     Total |         8         8 |        16

          Pearson chi2(1) =   1.0000   Pr = 0.317
           Fisher's exact =                 0.619
  1-sided Fisher's exact =                 0.310
```

Wilcoxon-Signed-Rank Test

Statistical Concepts

The Signed Rank Test is used to compare measures of center (e.g. median) to a specific value, similar to t-tests. The Signed Rank Test is preferred when a small sample or a sample that is not distributed normally is used. Additionally, by analyzing median values, the test is also robust to outliers. If the test's resulting p-value is less than the significance level, than the null hypothesis below of m_0 = 3.5 is rejected. If the p-value is greater than the significance level, then one fails to reject the null hypothesis.

Code

signrank H = 3.5

Perform a Wilcoxon-Signed-Rank Test on H. Let m_0 = 3.5. Perform the test at a 0.01 significance level.

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 ← Reject null at 0.01 significance level

Wilcoxon-Rank-Sum Test

Statistical Concepts

The Rank Sum Test Appropriate is appropriate to use to compare the medians of two populations and preferred (relative to its parametric counterpart) when a sample size is small or is not distributed normally. By analyzing median values, the test is also robust to outliers. Among the parametric tests, the Rank Sum Test is very similar to the two-sample t-test. If the test's resulting p-value is less than the significance level, than the null hypothesis below of $m_m - m_f = 3$ is rejected. If the p-value is greater than the significance level, then one fails to reject the null hypothesis.

Code

```
gen H3=H
replace H3=H+3 if G==0
ranksum H3, by(G)
```

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

$$H_0 : m_m - m_f = 3$$

Perform the test at a 0.05 significance level.

Output

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

G	obs	rank sum	expected
Female	1421	2053072	2010715
Male	1408	1949963	1992320
combined	2829	4003035	4003035

unadjusted variance 4.718e+08
adjustment for ties 0
---
adjusted variance 4.718e+08

Ho: H3(G==Female) = H3(G==Male)
z = 1.950
Prob > |z| = 0.0512 ← P-value is greater than 0.05 significance level, thus fail to reject null

ANOVA

Statistical Concepts

ANOVA requires one variable of categorical data and one variable of continuous data. The hypothesis of interest is at lease one part of the null hypothesis is false. In the case below, the hypothesis of interest is that at least one party's mean health score is different from the other parties' health scores.

To interpret ANOVA results, the p-value is again compared to the significance level (or the F statistic is compared to its critical value at the given significance level. If the p-value is less than the significance level, then one can reject the null hypothesis that the mean health score of each political party is the same.

Multiple comparison corrections work to control for type I error. That is, if you keep taking samples from data, you are eventually going to find a significant result even if its not truly significant (i.e. in 95% of the samples). These correctional methods are appropriate when a large number of groups are being compared.

Code

```
replace Democrat=1 if P=="D"
anova H Democrat
gen Republican=0
replace Republican=1 if P=="R"
gen Unaffiliated=0
replace Unaffiliated=1 if P=="U"
anova H Democrat Republican Unaffiliated
```

Compare the mean of H between political parties.

Output

```
      Number of obs =    2829     R-squared     =  0.0060
           Root MSE     = 3.32028     Adj R-squared =  0.0050
```

Source	Partial SS	df	MS	F	Prob > F
Model	188.84453	3	62.9481768	5.71	0.0007
Democrat	7.61784437	1	7.61784437	0.69	0.4059
Republican	.054793697	1	.054793697	0.00	0.9438
Unaffilia~d	7.84850933	1	7.84850933	0.71	0.3989
Residual	31143.5661	2825	11.0242712		
Total	31332.4106	2828	11.0793531		

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

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + e.$$

If the Xs do predict Y in this way, the β 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 β_2 for every unit increase of X_2 while holding all other Xs constant.

If there is linear association between the response and a predictor, we would expect the corresponding β to be different than 0. It follows that the hypothesis of no association between Y and X_i (while controlling for other predictors) is $H_0 : \beta_i = 0$.

The test of overall association, $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$, 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|)
(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
F-statistic: 0.03528 on 2 and 2826 DF,  p-value: 0.9653
```

Estimates of betas. Tests of single beta

Test of Overall Association

Linear Regression (continued)

Output