STA 2210 Final (Due on Mon 7/27 by 11:59pm)

<- read.csv(file.choose(), header = TRUE)

Directions: Investigate each research question below, including any R codes and plots/figures as they are relevant. To receive full credit, you must include all of the following in your presentations.

1. Formally check if the conditions for inference are satisfied.
2. Perform a hypothesis test using a 0.05 level of significance.
3. Interpret the decision of your test in context. Also describe the type of error that may have been made, and what that error means in the context.

Research Questions:

1. **The set Arthritis.csv contains data from the study “Arthritis Treatment Data.” Do the data provide significant evidence that the treatment used in the study has an effect on the symptoms of arthritis?**

Both variables (treatment and improved) are categorical. Explanatory: Treatment type (Placebo vs. Treated), Response: Improvement type (Marked, Some, None).

Based on this, the appropriate test is Chi Squared Test for Association/Independence

table(Arthritis$Treatment, Arthritis$Improved)

Marked None Some

Placebo 7 29 7

Treated 21 13 7

**a) Conditions:**

1) Independence: Sample size 84 is less than 10% of the population and the sample is randomly generated, so independence is satisfied.

2) Expected Count: The sample size of each cell’s expected value has to be at least 5.

chisq <- chisq.test(table(Arthritis$Treatment, Arthritis$Improved))

chisq$expected

Marked None Some

Placebo 14.33333 21.5 7.166667

Treated 13.66667 20.5 6.833333

All expected values are greater than 5, so the second condition is satisfied.

Since both conditions are satisfied, we can move forward with the chi squared test for association/independence.

**b) Hypotheses:**

Ho: There is no evidence that the treatment used in the study has an effect on the symptoms of arthritis.

Ha: There is evidence that the treatment used in the study has an effect on the symptoms of arthritis.

chisq

Pearson's Chi-squared test

data: table(Arthritis$Treatment, Arthritis$Improved)

X-squared = 13.055, df = 2, p-value = 0.001463

P-Value = 0.001463, which is less than the 0.05 alpha value. Based on this, we reject the null hypothesis.

**c)** **Conclusion:** Since we reject the null hypothesis, we conclude that there is evidence that the treatment used in the study does have an effect on the symptoms of arthritis. We may have made a Type 1 error, where we rejected the null hypothesis, when we should have failed to reject it. In context, this error would be where we conclude that the treatment in the study does have an effect on the symptoms of arthritis, when we should have said that the treatment does NOT have an effect on the symptoms of arthritis.

1. **The set cd4.csv contains data from the study “CD4 Counts for HIV-Positive Patients.” Do the data provide significant evidence that the anti-viral drug tested in the study increases the CD4 counts of HIV-positive patients?**

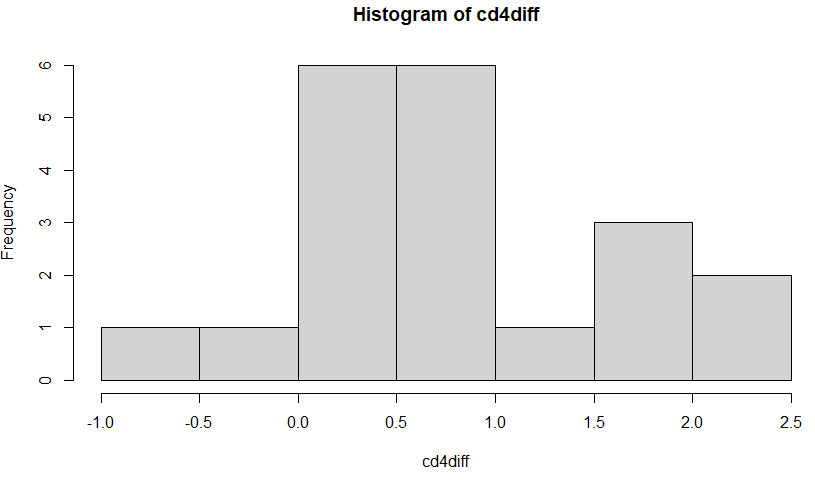
Both variables (baseline & oneyear) are numerical, paired data. We are looking at the difference between cd4 count before and after treatment, so I believe the test we are performing is a Paired 1-sided T-Test.

**a) Conditions:**

1) Independence: The sample size must be less than 10% of the population, and the samples must be randomly obtained. There are 84 observations which are based on a simple random sample, and consist of less than 10% of all HIV-positive patients, so independence is reasonable.

2) Normality: The distribution of the difference between post-treatment and pre-treatment needs to be nearly normally distributed. Based on the histogram shown below, it appears to be nearly normally distributed, but with a slight right skew. We say that this distribution is not strongly skewed, so normality’s condition is satisfied.

cd4diff <- (cd4$oneyear - cd4$baseline)

hist(cd4diff)

Both conditions are satisfied, so we can move forward with the 1-sided T-Test.

**b) Hypotheses:**

Ho: The anti-viral drug in the study decreases or does not change the cd4 cell count of HIV-positive patients (mu\_oneyear <= mu\_baseline).

Ha: The anti-viral drug in the study increases the cd4 cell count of HIV-positive patients (mu\_oneyear> mu\_baseline).

t.test(cd4$oneyear, cd4$baseline, paired = TRUE, alternative = "greater")

Paired t-test

data: cd4$oneyear and cd4$baseline

t = 4.4908, df = 19, p-value = 0.0001252

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

0.4950417 Inf

sample estimates:

mean of the differences

0.805

I used a t-test where x is the one-year variable (post-treatment), y is the baseline variable (pre-treatment), the data is paired, and the alternative hypothesis is greater than 0. The reason the y and x are swapped is because R-studio looks if x has a larger mean than y, which was the opposite of what I expected and took me a while to correct.

Based on this data, the p-value is 0.0001252, which is less than the alpha value of 0.05, so we reject the null hypothesis.

**c) Conclusion:** Since we reject the null hypothesis, we conclude that the anti-viral drug in the study increases the cd4 cell count of HIV-positive patients. We may have made a Type 1 error, where we rejected the null hypothesis when we should have failed to reject it. In the context of this study, the error would look like us saying that the anti-viral drug increased the cd4 cell count in HIV-positive patients, when in fact the drug DECREASED OR DID NOT CHANGE the cd4 cell count in these patients.

1. **The set GlowWorms.csv contains data from the study “Female glow-worms.” Do the data provide significant evidence that glow worms with longer lanterns tend to lay more eggs?**

The study is looking at two numeric variables (Lantern length vs. Eggs quantity). We would be reduced to choosing either 1 sample Paired T-Test, 2 sample T-Test, or Linear Regression. The paired data would need a difference between the x and y variables, so this is not feasible. This means we would use Linear Regression.

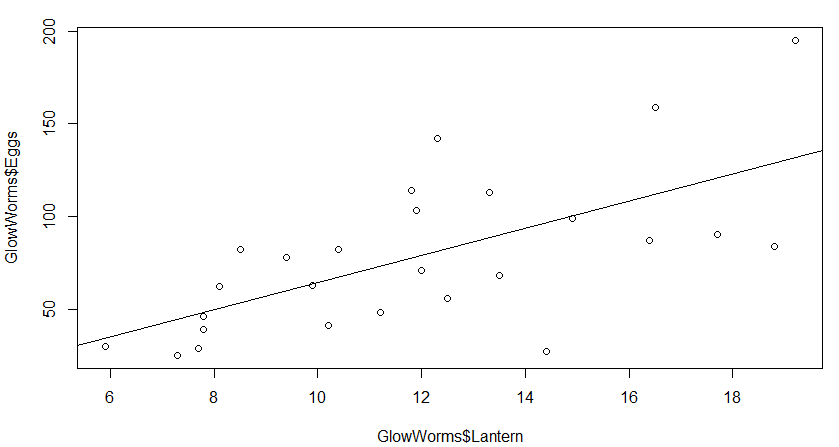
m1 <- lm(Eggs ~ Lantern, data = GlowWorms)

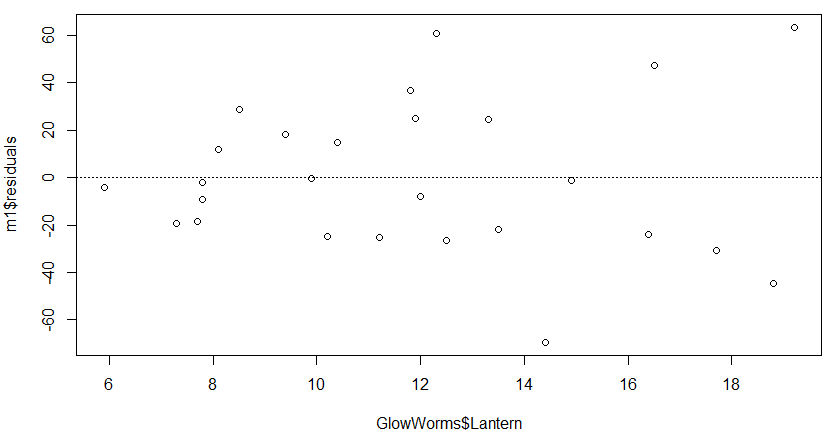
plot(GlowWorms$Lantern, GlowWorms$Eggs)

abline(m1)

plot(m1$residuals ~ GlowWorms$Lantern)

> abline(h = 0, lty = 3)

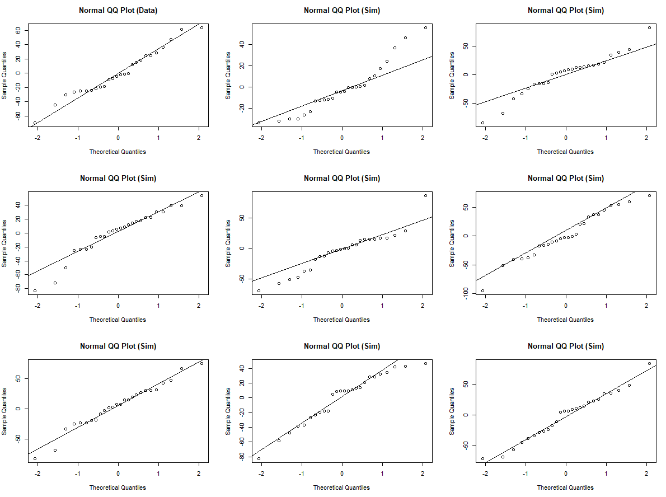
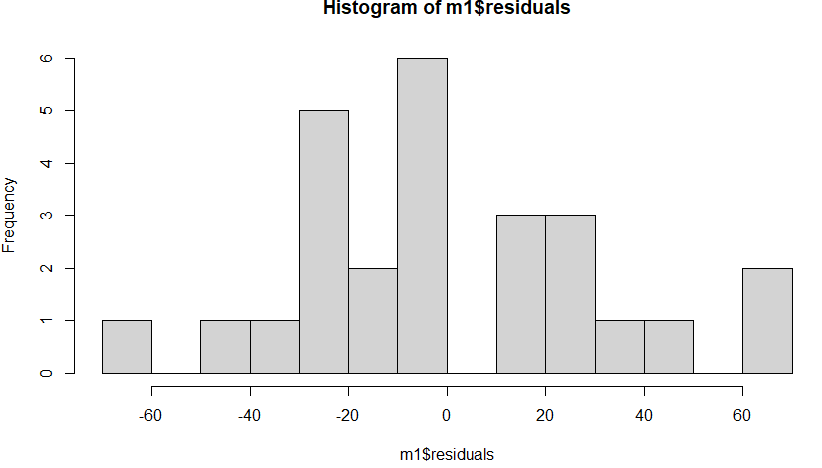
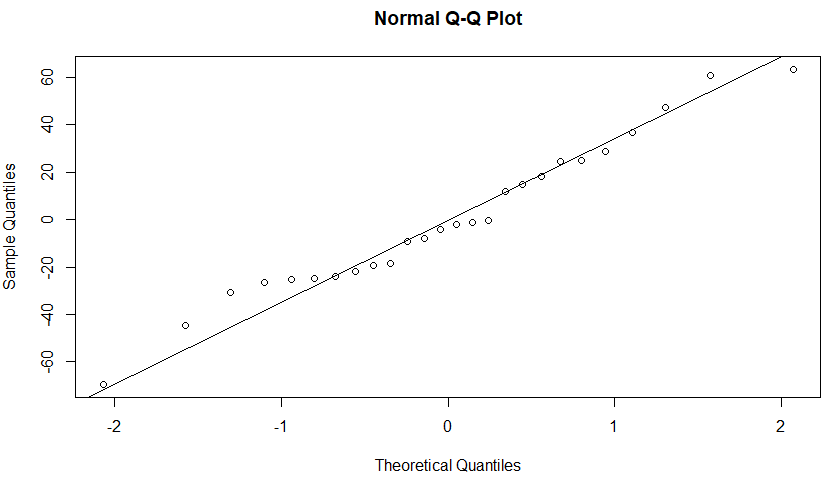
plot(m1$residuals ~ GlowWorms$Lantern)

abline(h = 0, lty = 3)

hist(m1$residuals, breaks = 15)

qqnorm(m1$residuals)

qqline(m1$residuals)

qqnormsim(m1$residuals)

**a) Conditions:**

1) Linearity: For linearity, the scatterplot should appear linear, and the residual plot should have no apparent patterns. This is true for both cases, as the scatterplot has a positive and weak-to-moderate linear relationship, and the residual plot has no pattern.

2) Nearly Normal Residuals: The residuals must be outputted on different graphs and show they are nearly normal. The histogram above appears to be nearly normal in distribution, although it appears to have some gaps. This can be explained by the small sample size. The qq-plot shows a slightly stepwise pattern, but there is no visible skew on the tails of the plot. Coupled with the simulated normal qq-plots, the qq-plot appears to be nearly normal. Both show normality, so this condition is met.

3) Constant Variability: For constant variance, we look at the residual plot and make sure that both sides of the y = 0 line have residuals that are consistent in standard error and distance from the line. The variance of the predicted error is very similar on both sides, so we say that constant variability is met.

Since all three conditions are met, a linear model would be reliable.

**b) Hypotheses:**

Ho: Glow worms with longer lanterns do not tend to lay more eggs than worms with shorter lanterns (Beta <= 0).

Ha: Glow worms with longer lanterns tend to lay more eggs than worms with shorter lanterns (Beta > 0).

summary(m1)

Call:

lm(formula = Eggs ~ Lantern, data = GlowWorms)

Residuals:

Min 1Q Median 3Q Max

-69.50 -23.59 -3.20 22.95 63.33

Coefficients:

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

(Intercept) -8.977 21.869 -0.410 0.685087

Lantern 7.325 1.757 4.169 0.000343 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32.71 on 24 degrees of freedom

Multiple R-squared: 0.4201, Adjusted R-squared: 0.3959

F-statistic: 17.38 on 1 and 24 DF, p-value: 0.0003431

Equation of Regression Line: Eggs = -8.977 + 7.325 (lantern length)

R-squared = 42.01%

P-value = 0.0003431

Since the p-value is less than 0.05, we reject the null hypothesis.

**c) Conclusion:** Since the null hypothesis was rejected, we conclude that there is significant evidence that the longer a glow worm’s lantern is, the more eggs they will lay. We may have made a Type 1 error, where the null hypothesis was rejected when it should not have been rejected. In this context, the error would be where we concluded that longer lanterns correlated to more eggs laid, when we should have concluded that a longer lantern does not correlate to laying more eggs.

1. **It was reported that, in the United States, 40% have blood type A, 11% have blood type B, 4% have blood type AB, and 45% have blood type O. The set transplant.csv contains data from the study “Liver Transplant Waiting List.” Do the data provide significant evidence that the distribution of blood types of those in need of liver transplants is different than the rest of the country?**

This study appears to want a Chi Squared Goodness of Fit test because it gives an expected value for all Americans and wants to compare this expected value to the outcome blood types of those in need of a liver.

Observed values:

table(transplant$abo)

A AB B O

325 41 103 346

table(transplant$abo)/815

A AB B O

0.39877301 0.05030675 0.12638037 0.42453988

Expected Values:

815\*0.40 = 326 A

815\*0.04 = 32.6 AB

815\*0.11 = 89.65 B

815\*0.45 = 366.75 O

**a) Conditions:**

1) Independence: The cases must come from a random sample, represent less than 10% of the total population, and have one case belong to only one cell. In this study, the samples are random, the sample size of 815 is less than 10% of the US’s population, and each person in the study only has one blood type. All conditions for independence are met, so independence is met.

2) Expected Count: The expected count of each cell must be at least 5, and looking at the above table, all four blood types have at least 5 cases. We can tell because when we multiply the total case count by the expected probabilities for each blood type given in the problem, all values are greater than 5. This condition is satisfied as well.

Since both cases are satisfied, we can move forward with the hypothesis test with Chi Squared goodness of fit.

**b) Hypotheses:**

Ho: Distribution of blood types for those in need of a liver transplant is not different from the rest of the country.

Ha: Distribution of blood types for those in need of a liver transplant is different from the rest of the country.

bloodType <- c(325,41,103,346)

chisq <- chisq.test(bloodType, p = c(0.40,0.04,0.11,0.45))

chisq

Chi-squared test for given probabilities

data: bloodType

X-squared = 5.3295, df = 3, p-value = 0.1492

Since the p-value of 0.1492 is greater than the alpha value of 0.05, we fail to reject the null hypothesis.

**c) Conclusion:** Since we fail to reject the null hypothesis, we conclude that the distribution of blood types for those in need of a liver transplant is not different from the rest of the country. We may have made a Type 2 error, where we failed to reject the null hypothesis when we should have rejected it. In the context of this study, the error would appear as if we said there is no difference in blood type distribution for those in need of a liver transplant vs. the rest of the country, when there IS a difference between the blood type distributions.

1. **The set urine.csv contains data from the study “Urine Analysis Data.” Do the data provide significant evidence of a difference in the average pH levels between subjects that do and do not have calcium oxalate crystals (kidney stones) present?**

I believe we should use a two-tailed T test to compare the categorical explanatory variable (subject having kidney stones present-Y/N) to the numeric response variable (average pH levels).

by(urine$ph, urine$r, length)

urine$r: 0 = 45

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

urine$r: 1 = 34

by(urine$ph, urine$r, summary)

urine$r: 0

Min. 1st Qu. Median Mean 3rd Qu. Max.

4.900 5.620 5.970 6.099 6.560 7.920

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

urine$r: 1

Min. 1st Qu. Median Mean 3rd Qu. Max.

4.760 5.433 5.815 5.936 6.255 7.940

**a) Conditions:**

1) Independence: For independence, we need to see that the sample sizes are less than 10% of the population and randomly sampled. We see that there are 79 cases, which is less than 10% of the population and they are randomly sampled. These cases all check off, and independence is met.

2) Not Paired: We see that the two groups are not paired, because those with kidney stones are not related to those without kidney stones, and the x and y variables are not paired,

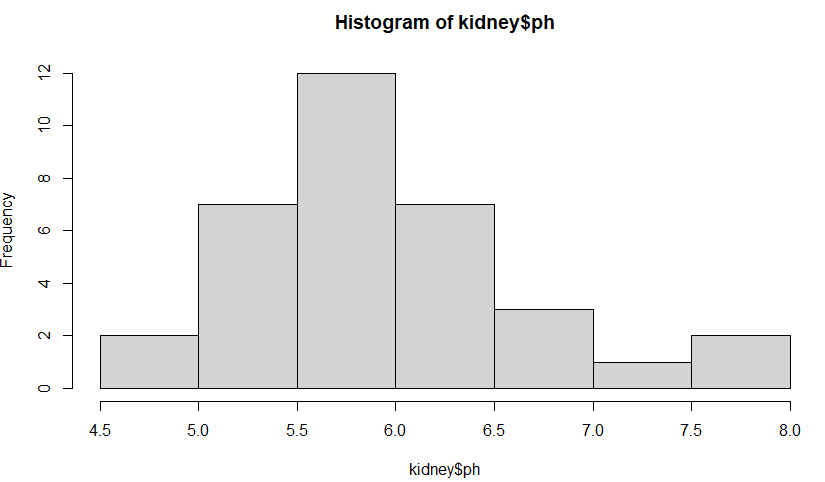
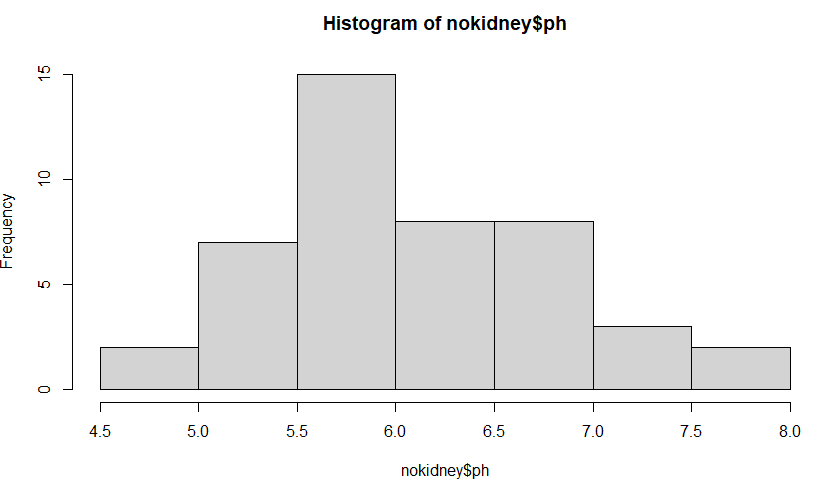
and this case is checked.

kidney <- subset(urine, urine$r == '1')

nokidney <- subset(urine, urine$r == '0')

hist(kidney$ph, breaks = 10)

hist(nokidney$ph, breaks = 10)



3) Normality: The histograms for the pH of the two categories have to be nearly normal. When we analyze both histograms, they both appear to be nearly normal with both having a slight left skew. The distribution is not extremely skewed to either side, so we can conclude that they are both nearly normal, and the condition of normality is checked off.

Since all three conditions appear to be checked off, we can continue with the two-tailed T Test.

**b) Hypotheses:**

Ho: There is no difference in average pH for those with and without kidney stones.

(mu\_kidneystones = mu\_nokidneystones)

Ha: There is a difference in average pH for those with and without kidney stones.

(mu\_kidneystones =/= mu\_nokidneystones)

inference(y = urine$ph, x = urine$r, est = "mean", type = "ht", null = 0,

alternative = "twosided", method = "theoretical")

Response variable: numerical, Explanatory variable: categorical

Difference between two means

Summary statistics:

n\_0 = 45, mean\_0 = 6.0987, sd\_0 = 0.702

n\_1 = 34, mean\_1 = 5.9356, sd\_1 = 0.7532

Observed difference between means (0-1) = 0.1631

H0: mu\_0 - mu\_1 = 0

HA: mu\_0 - mu\_1 != 0

Standard error = 0.166

Test statistic: Z = 0.981

p-value = 0.3266

Since the p-value of 0.3266 is greater than 0.05, we fail to reject the null hypothesis.

**c) Conclusion:** Since we fail to reject the null hypothesis, we conclude that there is no difference in average pH for those with and without kidney stones. We may have made a Type 2 error, where we failed to reject the null hypothesis when we should have rejected it. In the context of this study, the error would be where we conclude there is no difference in average pH for those with and without kidney stones, when in fact there IS a difference in the average pH between the two categories.

1. **The set Whickham2.csv contains data from the study “Whickham Health Study.” Do the data provide significant evidence that less than 50% of all women in Whickham who are age 18-64 are non-smokers?**

Since we are dealing with one population of women in Whickham 18-64 and looking at the categorical data proportion of them who smoke and do not smoke, we are dealing with a 1-prop Z test.

agegroup <- subset(Whickham2, Whickham2$AgeGroup == "18-64")

table(agegroup$Smoker)

No Yes

539 532

table(agegroup$Smoker)/1071

No Yes

0.503268 0.496732

Success = non-smoker

P^ = 0.503 (proportion of successes, since we are not given one in the problem).

**a) Conditions:**

1) Independence: The sample size must be random and consist of less than 10% of the population. In this case the sample size is random, and the size of 1071 is less than 10% of the total population, so independence is met.

2) The successes and failures for the 1071 women have to both be greater than 10.

N \* p^ = 1071 \* 0.503 = 539 successes

N \* (1-p^) = 1071 \* 0.497 = 532 failures

Both cases are above 10, so this condition is met.

Since both conditions are met, we can continue with the 1-prop Z test.

**b) Hypotheses:**

Ho: More than or 50% of 18-64 yr. old women in Whickham are non-smokers (p^ >= 0.5).

Ha: Less than 50% of 18-64 yr. old women in Whickham are non-smokers (p^ < 0.5).

prop.test(539, 1071, p = 0.5, alternative = "less", correct = TRUE)

1-sample proportions test with continuity correction

data: 539 out of 1071, null probability 0.5

X-squared = 0.033613, df = 1, p-value = 0.5727

alternative hypothesis: true p is less than 0.5

95 percent confidence interval:

0.0000000 0.5288236

sample estimates:

p

0.503268

Since our p-value of 0.5727 is greater than 0.05, we fail to reject the null hypothesis.

**c) Conclusion:** Since we fail to reject the null hypothesis, we conclude that 50% or more of 18-64 yr. old women in Whickham are non-smokers. We may have made a Type 2 error, where we failed to reject the null hypothesis, when we should have rejected it. In the context of this study, the error would be where we say 50% or more of 18-64 yr. old women in Whickham are non-smokers, when we should have said that less than 50% of 18-64 yr. old women in Whickham are non-smokers.