STA 2210 Homework 6 (Due on Mon. 7/6 by 11:59pm)

Directions: Write your R codes, in addition to your answer, to the following problems. You may find the “t.test()” function to be especially helpful on this assignment. Refer to the R reference card for more helpful commands.

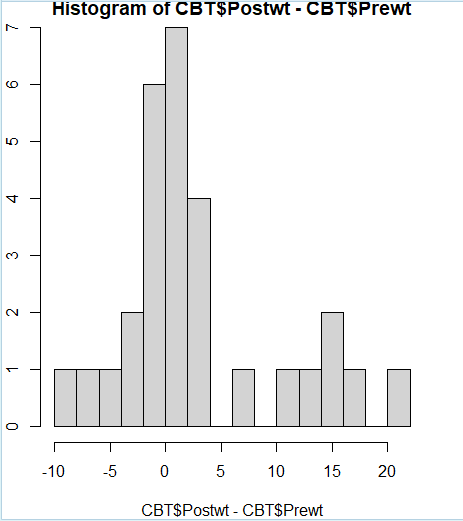
1. **The data set anorexia, a .csv file, contains data on the effect of different treatments for anorexia, for a sample of 72 young female anorexia patients. Do the data provide significant evidence that CBT (Cognitive Behavioral Treatment) is an effective treatment for anorexia (i.e. is there significant evidence of weight gain for subjects undergoing that treatment)? Test the relevant hypotheses at a significance level of . Be sure to formally check if the conditions for inference are satisfied.**

CBT <- subset(anorexia, anorexia$Treat == 'CBT') = 29 observations

Ho: Weight gain <= 0 for CBT patients. (CBT patients will lose weight or not change in weight at all on average).

Ha: Weight gain > 0 for CBT patients. (CBT patients will gain more than 0 lbs. on average).

We need to check the conditions for inference, which are independence and normality. The data is paired because it looks at the same subjects’ pre-treatment weights and post-treatment weights, so it is not independent. Independence is checked by seeing if the sample size is less than 10% of the population and if the samples are not paired. These are both true, because every sample is an independent patient who does not have any effect on another patient, and the sample size is not too big.

hist(CBT$Postwt-CBT$Prewt, breaks = 20)

The graph appears slightly right skewed, but it is not extremely skewed, so we can say that it is nearly normal, and both conditions are checked off.

To calculate if CBT is effective, we have to see that the sample mean is greater than 0. This means we are using a one-sided T test, looking at every sample’s weight difference, and testing it against a mean of 0.

t.test((CBT$Postwt - CBT$Prewt), alternative = "greater", mu = 0)

One Sample t-test

data: (CBT$Postwt - CBT$Prewt)

t = 2.2156, df = 28, p-value = 0.01751

alternative hypothesis: true mean is greater than 0

95 percent confidence interval:

0.6981979 Inf

sample estimates:

mean of x

3.006897

Based on these calculations, the p-value is less than 0.05, so we reject the null hypothesis. This means that we accept the alternative hypothesis and conclude that CBT is an effective treatment for anorexia, as patients gain more than 0 lbs. on average from the treatment, which is better than no treatment.

1. **The data set uis, a .csv file, contains data on the treatment of drug addiction for a sample of 628 Americans. Do the data provide significant evidence of a relationship between a subject’s race and their time to relapse (i.e. is there a statistically significant difference in relapse time between the two groups)? Test the relevant hypotheses at a significance level of . Be sure to formally check if the conditions for inference are satisfied.**

by(uis$TIME, uis$RACE, length)

uis$RACE: 0 = 430 (white)

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

uis$RACE: 1 = 145 (non-white)

by(uis$TIME, uis$RACE, summary)

uis$RACE: 0

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

4.0 76.5 163.0 228.4 346.8 1172.0

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

uis$RACE: 1

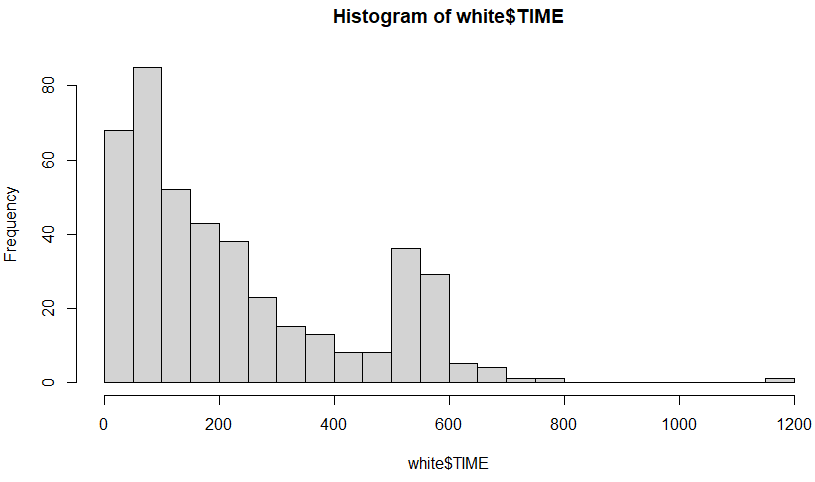
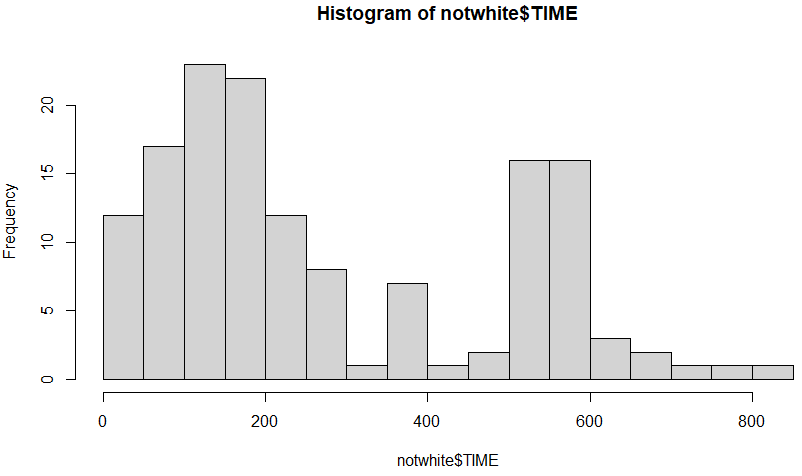
Min. 1st Qu. Median Mean 3rd Qu. Max.

6.0 115.0 198.0 280.7 537.0 805.0

Ho: White relapse time = Non-white relapse time

Ha: White relapse time =/= Non-white relapse time

The formal conditions we need to check are independence and normality. For independence, we see that the sample sizes are less than 10% the overall population, greater than 30, and not paired, so this is fine. For normality, we need to see if the distributions of the two race groups and their relapse times are near normally distributed.

Looking at these histograms, it is clear that they are not remotely close to being normally distributed. For the white race box, it is strongly right skewed, with no data to the left side of the center. For the non-white histogram, there is a strong right skew, and the graph is almost bimodal. Both graphs have strong outliers, so we can conclude that the condition of normality is not checked. The data we are looking at makes it difficult for a normal distribution to be depicted, so a Two-sided T Test may not be the best way to go forth with the hypothesis testing. However, it is known from the Central Limit Theorem that when the sample size increases, the data moves to become more normal to represent the population mean as there is less standard error. Since the sample size is very large, we can continue to move forward with the hypothesis testing.

inference(y = uis$TIME, x = uis$RACE, est = "mean", type = "ht", null = 0,

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

Summary statistics:

n\_0 = 430, mean\_0 = 228.3791, sd\_0 = 196.0301

n\_1 = 145, mean\_1 = 280.669, sd\_1 = 207.7786

Observed difference between means (0-1) = -52.2899

H0: mu\_0 - mu\_1 = 0

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

Standard error = 19.675

Test statistic: Z = -2.658

p-value = 0.0078

Based on the p-value which is below 0.05, we reject the null hypothesis and state that there is a difference between a subject’s race and their relapse time.