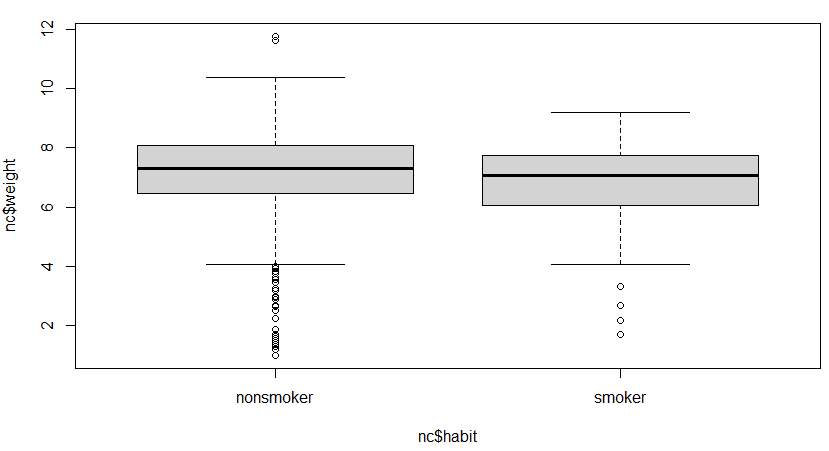
Lab 6

**1) What are the cases in this data set? How many cases are there in our sample?**

The cases are the babies born in North Carolina in 2004. There are 1000 cases in the sample.

**2) Make a side-by-side boxplot of habit and weight. What does the plot highlight about the relationship between these two variables?**

boxplot(nc$weight ~ nc$habit)

Nonsmokers tend to have a higher median baby’s weight than smokers, as well as a higher 1st and 3rd quartile, which is shown by the nonsmoker box being above the smoker box. The plot highlights that nonsmoking mothers tend to have babies that weigh more than smoking mothers’ babies.

**3) Check if the conditions necessary for inference are satisfied. Note that you will need to obtain sample sizes to check the conditions. You can compute the group size using the same by command above but replacing mean with length.**

by(nc$weight, nc$habit, length)

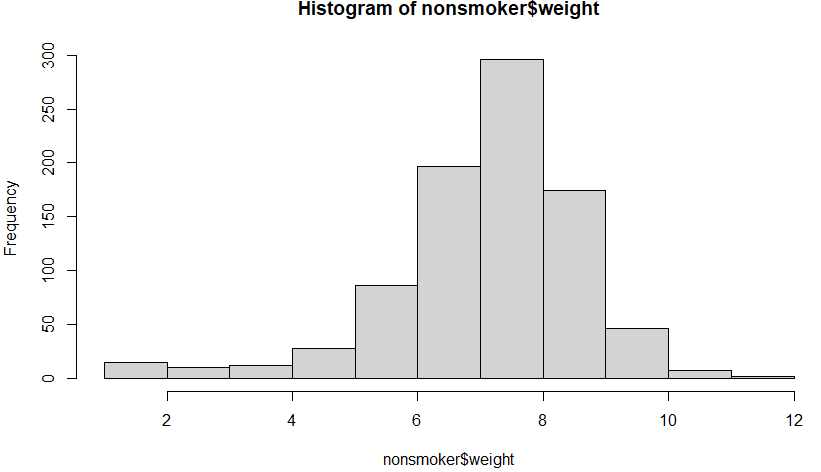
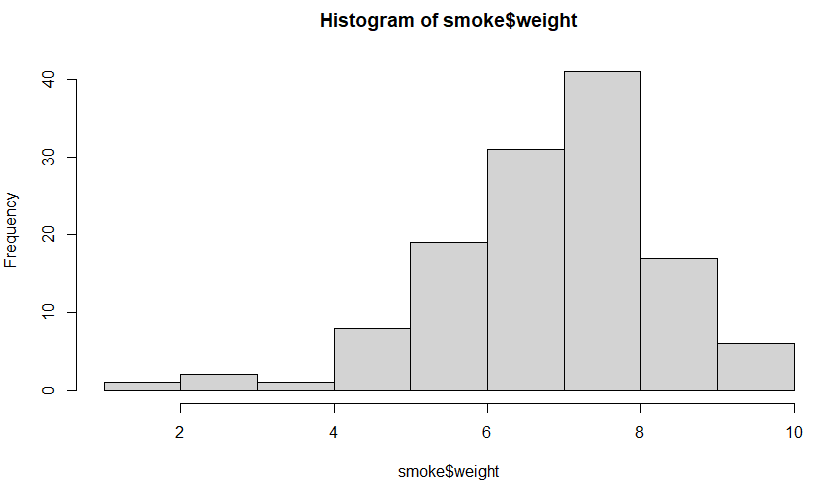
nc$habit: nonsmoker = 873

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

nc$habit: smoker = 126

hist(smoke$weight)

hist(nonsmoker$weight)

  
The ‘by’ command shows that both categories of habits have sample sizes that are independent and statistically significant, as they are both greater than 30. The sample size should be less than 10% of the population, and 1000 cases seems to follow that category. The next condition should be normality. Looking at the histograms of both, they both are slightly left skewed, but not enough to feel that they are not nearly normal. Both conditions are met, so we can move on.

**4) Write the hypotheses for testing if the average weights of babies born to smoking and non-smoking mothers are different.**

Ho: Avg. birthweight non-smoking = Avg. birthweight smoking

The average weights of babies born to smoking and non-smoking mothers is the same.

Ha: Avg. birthweight non-smoking =/= Avg. birthweight smoking

The average weights of babies born to smoking and non-smoking mothers are different.

**5) Change the type argument to "ci" to construct and record a confidence interval for the difference between the weights of babies born to smoking and non-smoking mothers.**

inference(y = nc$weight, x = nc$habit, est = "mean", type = "ci", null = 0,

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

Summary statistics:

n\_nonsmoker = 873, mean\_nonsmoker = 7.1443, sd\_nonsmoker = 1.5187

n\_smoker = 126, mean\_smoker = 6.8287, sd\_smoker = 1.3862

Observed difference between means (nonsmoker-smoker) = 0.3155

Standard error = 0.1338

95 % Confidence interval = ( 0.0534 , 0.5777 )

Based on the confidence intervals, we are 95% confident that on average, babies born to non-smoking mothers weigh between 0.05 to 0.58 lbs. more than babies born to smoking mothers at birth.

**ON YOUR OWN**

1. **Calculate a 95% confidence interval for the average length of pregnancies (weeks) and interpret it in context. Note that since you’re doing inference on a single population parameter, there is no explanatory variable, so you can omit the x variable from the function.**

inference(y = nc$weeks, est = "mean", type = "ci", null = 0, alternative = "twosided", method = "theoretical")

Single mean

Summary statistics: mean = 38.3347 ; sd = 2.9316 ; n = 998

Standard error = 0.0928

95 % Confidence interval = ( 38.1528 , 38.5165 )

We are 95% confident that the average length of pregnancy for mothers in the population can be estimated to be between 38.15 – 38.52 weeks in length, based on this study.

1. **Calculate a new confidence interval for the same parameter at the 90% confidence level. You can change the confidence level by adding a new argument to the function: conflevel = 0.90.**

inference(y = nc$weeks, est = "mean", type = "ci", null = 0,

alternative = "twosided", method = "theoretical", conflevel= 0.90)

Single mean

Summary statistics: mean = 38.3347 ; sd = 2.9316 ; n = 998

Standard error = 0.0928

90 % Confidence interval = ( 38.182 , 38.4873 )

We are 90% confident that the average length of pregnancy for mothers in the population can be estimated to be between 38.18 – 38.49 weeks in length, based on this study.

1. **Conduct a hypothesis test evaluating whether the average weight gained by younger mothers is different than the average weight gained by mature mothers.**

by(nc$gained, nc$mature, length)

nc$mature: mature mom = 133

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

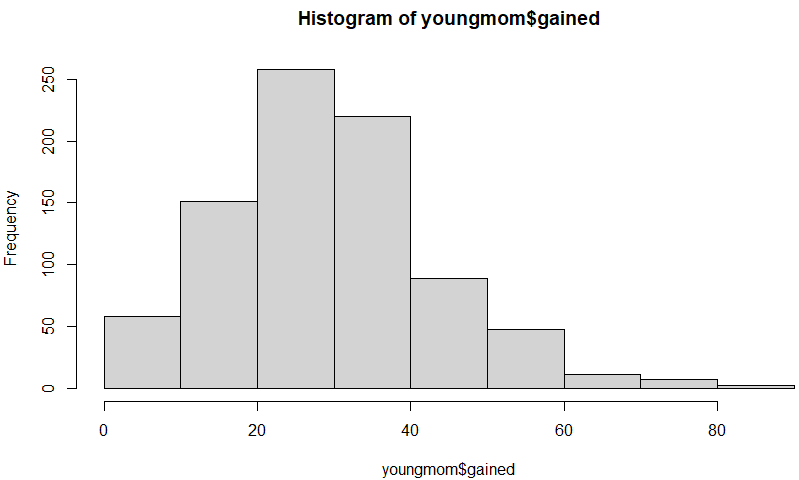
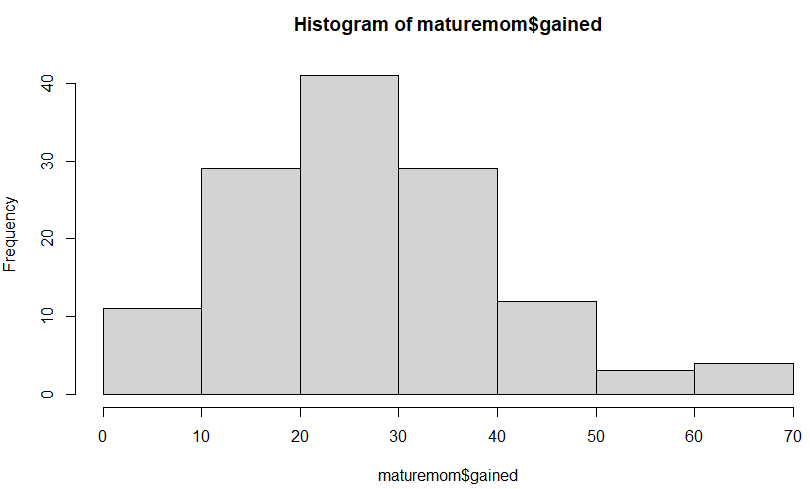
nc$mature: younger mom = 867

youngmom <- subset(nc, nc$mature == 'younger mom')

maturemom <- subset(nc, nc$mature == 'mature mom')

hist(youngmom$gained)

hist(maturemom$gained)

To make a hypothesis test, we must check for independence and near-normality. For independence, the samples sizes are greater than 30, and are less than 10% of the overall population, so we can say they are independent. For normality, both graphs appear slightly right skewed, but not enough to say they are not nearly normal, so we can say they are near-normal.

Ho: Avg. weight gain young mom = Avg. weight gain mature mom

Ha: Avg. weight gain young mom =/= Avg. weight gain mature mom

inference(y = nc$gained, x = nc$mature, est = "mean", type = "ht", null = 0, alternative = "twosided", method = "theoretical")

Summary statistics:

n\_mature mom = 129, mean\_mature mom = 28.7907, sd\_mature mom = 13.4824

n\_younger mom = 844, mean\_younger mom = 30.5604, sd\_younger mom = 14.3469

Observed difference between means (mature mom-younger mom) = -1.7697

H0: mu\_mature mom - mu\_younger mom = 0

HA: mu\_mature mom - mu\_younger mom != 0

Standard error = 1.286

Test statistic: Z = -1.376

p-value = 0.1686

Since the p-value is greater than 0.05, we fail to reject the null hypothesis, so the average weight gained for young and mature moms is not different overall.

1. **Now, a non-inference task: Determine the age cutoff for younger and mature mothers. Use a method of your choice and explain how your method works.**

To figure this out, we must know the ages of all the mothers and categorize them into the younger and mature categories. We can do this by using the ‘by’ function, nc$mages, and nc$mature, and a summary of mins and maxes to see if there is a cut-off.

by(nc$mage, nc$mature, summary)

nc$mature: mature mom

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

35.00 35.00 37.00 37.18 38.00 50.00

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

nc$mature: younger mom

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

13.00 21.00 25.00 25.44 30.00 34.00

Also I created a subset of mature and young moms and summarized them to get the same thing.

summary(maturemom$mage)

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

35.00 35.00 37.00 37.18 38.00 50.00

summary(youngmom$mage)

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

13.00 21.00 25.00 25.44 30.00 34.00

The methods both work because there is no overlap in ages between young moms and mature moms, which means we found the cutoff.

1. **Pick a pair of numerical and categorical variables and come up with a research question evaluating the relationship between these variables. Formulate the question in a way that it can be answered using a hypothesis test and/or a confidence interval. Answer your question using the inference function, report the statistical results, and also provide an explanation in plain language.**

Numerical: weight of baby at birth. Categorical: whitemom- whether the mom is white or not.

Research question: Conduct a hypothesis test and find the 99% confidence interval on the average birthweight difference of newborns between white and non-white moms.

Ho: Avg. birthweight white mom = Avg. birthweight non-white mom

Ha: Avg. birthweight white mom =/= Avg. birthweight non-white mom

by(nc$weight, nc$whitemom, length)

nc$whitemom: not white = 284

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

nc$whitemom: white = 714

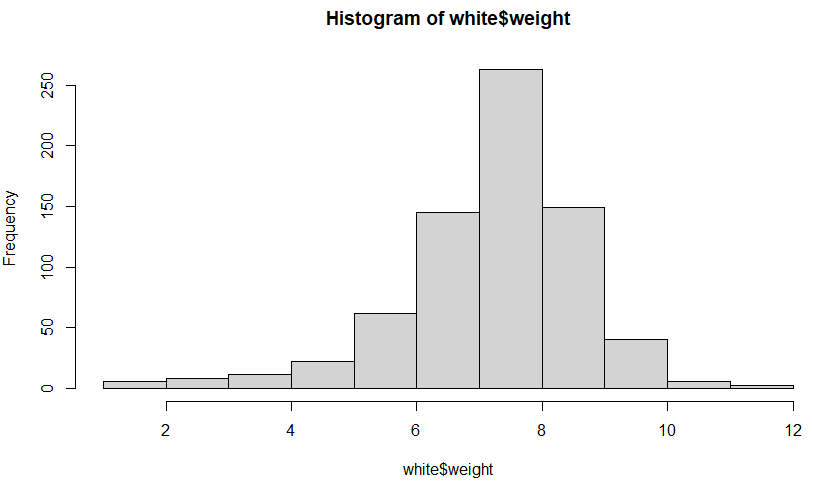
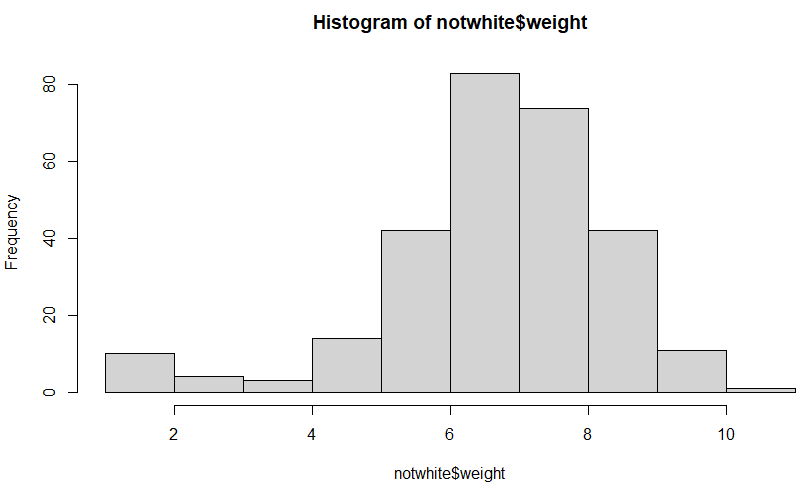
Independence check: The sample sizes are greater than 30 and they consist of less than 10% of the total population, so independence is checked.

white <- subset(nc, nc$whitemom == 'white')

notwhite <- subset(nc, nc$whitemom == 'not white')

hist(white$weight)

hist(notwhite$weight)

Normality check: Both histograms do not appear to be extremely skewed, as both are slightly left skewed, but nearly normal. This means that they are near-normal, and are checked.

inference(y = nc$weight, x = nc$whitemom, est = "mean", type = "ht", null = 0, alternative = "twosided", method = "theoretical")

Summary statistics:

n\_not white = 284, mean\_not white = 6.7195, sd\_not white = 1.6207

n\_white = 714, mean\_white = 7.2505, sd\_white = 1.4333

Observed difference between means (not white-white) = -0.5309

H0: mu\_not white - mu\_white = 0

HA: mu\_not white - mu\_white != 0

Standard error = 0.11

Test statistic: Z = -4.821

p-value = 0

The p-value is 0, which is < 0.05, so we reject the null hypothesis. This means that there is a difference in average birthweight between white and non-white mom’s newborns.

inference(y = nc$weight, x = nc$whitemom, est = "mean", type = "ci", null = 0, alternative = "twosided", method = "theoretical", conflevel = 0.99)

Standard error = 0.1101

99 % Confidence interval = ( -0.8146 , -0.2473 )

Based on this, we are 99% confident that non-white moms have newborns that weigh 0.25 – 0.81 lbs. less than white moms’ newborns at birth for the population.