MSDS5043 - Assignment 5

Olivia Samples, Lipscomb University

10/15/2019

Question 1: Problem 5.7

(a) The null hypothesis is that New Yorkers get 8 hours a sleep on average. The alternative hypothesis is that New Yorkers get less than 8 hours a sleep on average. We can write this as follows:

$H_0 : x = 8$
$H_A : x < 8.$

(b) We have a simple random sample of 25 New Yorkers which is less than 10% of the New York population, such that the observations are independent. Also, we have that the observations come from a nearly normal distribution because the summary statistics do not suggest any strong skews.

```
#given
n = 25
u = 8
xbar = 7.73
s = .77

#calculate standard error
se = s/sqrt(n)

#calculate degrees of freedom
df = n-1
df
```

```
## [1] 24
```

```
#calculate t distribution
t = (xbar - u)/se
t
```

```
## [1] -1.753247
```

(c) Now we will use t = -1.75 and degrees of freedom = 24 to find the p-value.

```
pt(-abs(t), df)
```

```
## [1] 0.04616261
```

```
#we use the absolute value times the negative to ensure that we are getting the proportion (p) associat
```

And so we have that the probability of getting an average of 7.73 hours of sleep for a random sample of 25 New Yorkers would be 4.6% if the true average is 8 hours of sleep for a random 25 New Yorkers.

(d) Since we have that our p-value < our significance level, we have sufficient evidence to reject the null hypothesis. Simarly, we have sufficient evidence that New Yorkers sleep less than an average of 8 hours a night.

(e) No, because we rejected the null hypothesis. We will prove this below:

```
t24 <- qt(.90, df)
lower <- xbar - t24*se
upper <- xbar + t24*se
c(lower, upper)
```

```
## [1] 7.527053 7.932947
```

And we can see that 8 is not within the 90% confidence interval.

Question 2: Problem 5.22

(a) We will calculate the confidence interval at 95% using a z-score of 1.96, because n is large.

```
#given
n = 200
xbar = -.545
sd = 8.887
z = 1.96

#standard error of difference
se = sd/sqrt(n)

lower <- xbar - z * se
upper <- xbar + z * se
c(lower, upper)
```

```
## [1] -1.7766754  0.6866754
```

(b) We have a 95% confidence interval of (-1.78, 0.69) for the average difference between the reading and writing scores of all students. In other words, 95% of all students will have a difference of reading and writing scores between this confidence interval.

(c) Since the confidence interval contains 0, there is not sufficient evidence to reject the null Hypothesis.

Question 3: Problem 5.29

(a) We will conduct the following hypothesis test: $H_0 : \mu_{diff} = 0$
$H_A : \mu_{diff} \neq 0$ with a 95% confidence interval such that the significance level is $\alpha = 0.05$. We calculate the t-score below.

```
#given differences
n = 6
xbar = -3.33
sd = 3.01

#standard error of difference
se = sd/sqrt(n)

#degrees of freedom
df = n - 1
df
```

```
## [1] 5
```

```
#t-score
t = (xbar - 0)/se
t
```

```
## [1] -2.709901
```

Now we will use t-score = -2.71 and degrees of freedom = 5 to calculate the p-value.

```
pt(-abs(t), df)
```

```
## [1] 0.0211406
```

We have that the p-value (0.021) is less than the siginificant value (0.05), so we will reject the null Hypothesis. In other words, the data provides sufficient evidence to suggest that there is a difference between the average numbers of traffic accident related emergency room admissions between Friday the 6th and Friday the 13th.

(b) We will calculate a t-score for a 95% confidence interval with degrees of freedom = 5.

```
#given
n <- 6
xbar <- -3.33
sd <- 3.01

#calculate critical value
t5 <- qt(c(.025, .975), df)

#standard error of difference
se = sd/sqrt(n)

#confidence interval
lower <- xbar - t5[2] * se
upper <- xbar + t5[2] * se
c(lower, upper)
```

```
## [1] -6.4888013 -0.1711987
```

(c) We must disagree with this statement because we cannot infer this increase of risk based off of observational data. While we can reject the null hypothesis, we cannot say that anything else is caused based off of this hypothesis test.

Question 4: Problem 5.50

(a) $H_0 : \mu_p = \mu_l = \mu_u = \mu_t = \mu_c$ The average number of hours spent on child care is the same across all educational attainment levels. H_A : At least one educational attainmnent level has a different average number of hours spent on child care.

(b) First, we will apply a Bonferri correction for a confidence level of 95%.

```
k = 5*4/2

sv = .05/k
sv
```

```
## [1] 0.005
```

And so, we have that the F statistic is 0.2846 which is greater than our stringent significance level of 0.005. Hence, there is not sufficient evidence to reject the null Hypothesis, the observed differences in sample means are attributable to sampling variability. In other words, there is not sufficient evidence to suggest that at least one educational attainment level has a different average number of hours spent on child care than any other level.

Question 5

This question is based on North Carolina births data, available on Canvas. In 2004, the state of North Carolina released a large data set containing information on births recorded in this state. This data set is useful to researchers studying the relation between habits and practices of expectant mothers and the birth of their children. We will work with a random sample of observations from this data set.

We have observations on 13 different variables, some categorical and some numerical. The meaning of each variable is as follows.

```
fage - father's age in years.
mage - mother's age in years.
mature - maturity status of mother.
weeks - length of pregnancy in weeks.
premie - whether the birth was classified as premature (premie) or full-term.
visits - number of hospital visits during pregnancy.
marital - whether mother is married or not married at birth.
gained - weight gained by mother during pregnancy in pounds.
weight - weight of the baby at birth in pounds.
lowbirthweight - whether baby was classified as low birthweight ( low ) or not ( not low ).
gender - gender of the baby, female or male.
habit - status of the mother as a nonsmoker or a smoker.
whitemom - whether mom is white or not white.
```

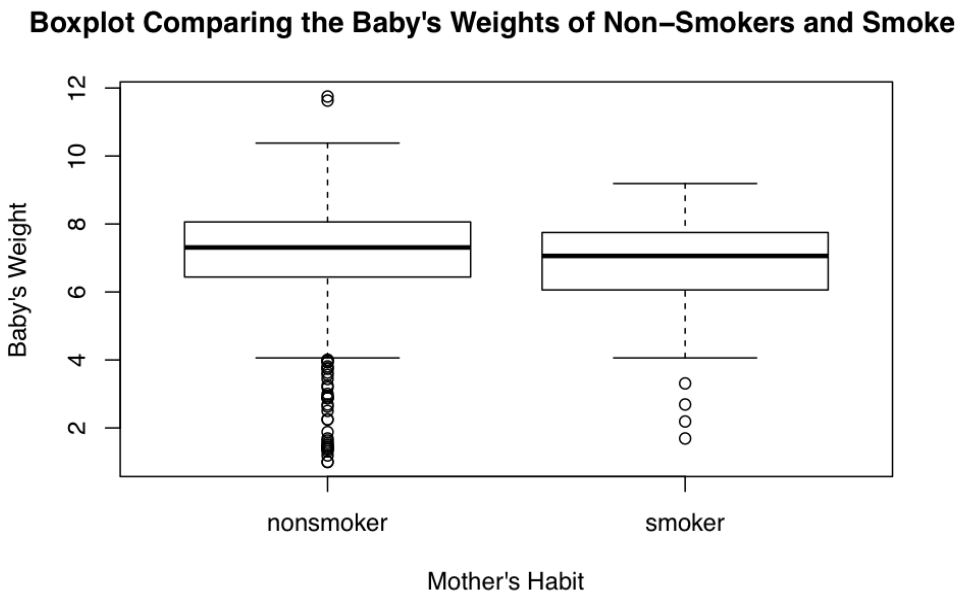
```
nc = read.csv("/Users/osamples/Desktop/MSDS5043/nc.csv", header = TRUE, sep = ",")

summary(nc$habit)
```

```
## nonsmoker    smoker      NA's
##       873       126         1
```

(a) 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, main = "Boxplot Comparing the Baby's Weights of Non-Smokers and Smokers",
```



The box plot shows that the range for nonsmoking mothers is larger than that of smoking mothers. There are also a lot more outlers represented for nonsmoking mothers. Also, the box plot shows that nonsmoking mothers represent the largest baby weights. In terms of average baby weights, we can see the difference

in median for nonsmoking and smoking mothers; however, we can not determine the difference in mean or significance of this difference from the box plot.

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

$H_0 : \mu_s - \mu_{ns} = 0$ There is no difference in the average (mean) weights of babies born to smoking and non-smoking mothers. $H_A : \mu_s - \mu_{ns} \neq 0$ There is a difference in the average (mean) weights of babies born to smoking and non-smoking mothers.

(c) Conduct a hypothesis test evaluating whether the average weight gained by younger mothers is different than the average weight gained by mature mothers.

```
summary(nc$weight)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   6.380   7.310   7.101   8.060  11.750
```

```
sd(nc$weight)
```

```
## [1] 1.50886
```

```
length(nc$weight)
```

```
## [1] 1000
```

First of all, based on these summary statistics, we have that the data is not strongly skewed and is reflective of a normal distribution. We also have that the data is a random sample with only 1000 observations which is less than 10% of the total population so that our data is independent as well. Hence, we have met our conditions for a hypothesis test.

```
by(nc$weight, nc$habit, mean)
```

```
## nc$habit: nonsmoker
## [1] 7.144273
## ------------------------------------------------------------
## nc$habit: smoker
## [1] 6.82873
```

```
rm(sd)
by(nc$weight, nc$habit, sd)
```

```
## nc$habit: nonsmoker
## [1] 1.518681
## ------------------------------------------------------------
## nc$habit: smoker
## [1] 1.38618
```

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

```
## nc$habit: nonsmoker
## [1] 873
## ------------------------------------------------------------
## nc$habit: smoker
## [1] 126
```

Now we will conduct the following hypothesis test: $H_0 : \mu_{diff} = 0$
$H_A : \mu_{diff} \neq 0$ with a 95% confidence interval such that the significance level is $\alpha = 0.05$ because we have a large n.

```r
#given differences
nns = 873
ns = 126
xbar = 7.144273 - 6.82873
sdns = 1.518681
sds = 1.38618
z = 1.96

#standard error of difference
se = sqrt(((sdns^2)/nns) + ((sds^2)/ns))

#calculate confidence intervale
lower <- xbar - z * se
upper <- xbar + z * se

c(lower, upper)
```

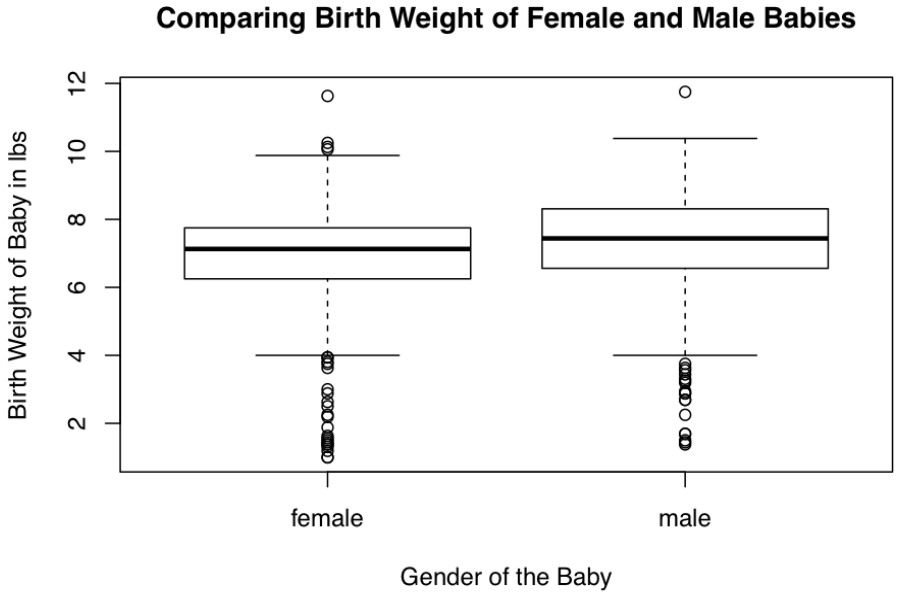
```
## [1] 0.05337239 0.57771361
```

And so, we have a 95% confidence interval of (0.0534, 0.5777) that does not include 0. This means there is sufficient evidence to reject the null hypothesis. In other words, there is sufficient evidence to suggest that there is a difference in the average (mean) weights of babies born to smoking and non-smoking mothers.

(d) 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.

I would like to establish if a baby's gender affects the average birth weight of a baby.

```r
boxplot(nc$weight ~ nc$gender, main = "Comparing Birth Weight of Female and Male Babies", ylab = "Birth
```



The

box plot shows that the range for male babies is larger than that of female babies. In terms of average baby weight, we can see the difference in median for female and male babies; however, we can not determine the difference in mean or significance of this difference from the box plot.

Rather, the difference in mean can be calculated as follows:

```
by(nc$weight, nc$gender, summary)
```

```
## nc$gender: female
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   6.250   7.130   6.903   7.750  11.630
## ------------------------------------------------------------
## nc$gender: male
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.380   6.560   7.440   7.302   8.310  11.750
```

And so, there is a difference in mean. We will test to see if this is a significant difference.

We have the following hypothesis:

$H_0 : \mu_s - \mu_{ns} = 0$ There is no difference in the average (mean) baby weight of female children and male children. $H_A : \mu_s - \mu_{ns} \neq 0$ There is a difference in the average (mean) baby weight of female children and male children.

```
by(nc$weight, nc$gender, sd)
```

```
## nc$gender: female
## [1] 1.475869
## ------------------------------------------------------------
## nc$gender: male
## [1] 1.516846
```

```
by(nc$weight, nc$gender, length)
```

```
## nc$gender: female
## [1] 503
## ------------------------------------------------------------
## nc$gender: male
## [1] 497
```

```
by(nc$weight, nc$gender, mean)
```

```
## nc$gender: female
## [1] 6.902883
## ------------------------------------------------------------
## nc$gender: male
## [1] 7.301509
```

First of all, based on these summary statistics, we have that the data is not strongly skewed and is reflective of a normal distribution. We also have that the data is a random sample with only 1000 observations which is less than 10% of the total population so that our data is independent as well. Hence, we have met our conditions for a hypothesis test.

Now we will conduct the following hypothesis test: $H_0 : \mu_{diff} = 0$
$H_A : \mu_{diff} \neq 0$ with a 95% confidence interval such that the significance level is $\alpha = 0.05$ because we have a large n.

```
#given differences
nf = 503
nm = 497
```

```r
xbar = 6.902883 - 7.301509
sdf = 1.475869
sdm = 1.516846
z = 1.96

#standard error of difference
se = sqrt(((sdf^2)/nf) + ((sdm^2)/nm))

#calculate confidence intervale
lower <- xbar - z * se
upper <- xbar + z * se

c(lower, upper)
```

```
## [1] -0.5841524 -0.2130996
```

And so, we have a 95% confidence interval of (-0.5841, -0.2131) that does not include 0. This means there is sufficient evidence to reject the null hypothesis. In other words, there is sufficient evidence to suggest that there is a difference in the average (mean) weights of babies born to female and male babies.