Inference For Numerical Data

Alex Friedrichsen

August 31, 2021

Load and Inspect

```
download.file("http://www.openintro.org/stat/data/nc.RData", destfile = "nc.RData")
load("nc.RData")

View(nc)
head(nc)
```

```
##   fage mage      mature weeks    premie visits marital gained weight
## 1   NA   13 younger mom    39 full term     10 married     38   7.63
## 2   NA   14 younger mom    42 full term     15 married     20   7.88
## 3   19   15 younger mom    37 full term     11 married     38   6.63
## 4   21   15 younger mom    41 full term      6 married     34   8.00
## 5   NA   15 younger mom    39 full term      9 married     27   6.38
## 6   NA   15 younger mom    38 full term     19 married     22   5.38
##   lowbirthweight gender     habit  whitemom
## 1        not low   male nonsmoker not white
## 2        not low   male nonsmoker not white
## 3        not low female nonsmoker     white
## 4        not low   male nonsmoker     white
## 5        not low female nonsmoker not white
## 6            low   male nonsmoker not white
```

```
summary(nc)
```

```
##       fage            mage            mature        weeks            premie
##  Min.   :14.00   Min.   :13   mature mom :133   Min.   :20.00   full term:846
##  1st Qu.:25.00   1st Qu.:22   younger mom:867   1st Qu.:37.00   premie   :152
##  Median :30.00   Median :27                     Median :39.00   NA's     :  2
##  Mean   :30.26   Mean   :27                     Mean   :38.33
##  3rd Qu.:35.00   3rd Qu.:32                     3rd Qu.:40.00
##  Max.   :55.00   Max.   :50                     Max.   :45.00
##  NA's   :171                                    NA's   :2
##      visits           marital        gained          weight
##  Min.   : 0.0   married    :386   Min.   : 0.00   Min.   : 1.000
##  1st Qu.:10.0   not married:613   1st Qu.:20.00   1st Qu.: 6.380
##  Median :12.0   NA's       :  1   Median :30.00   Median : 7.310
##  Mean   :12.1                     Mean   :30.33   Mean   : 7.101
##  3rd Qu.:15.0                     3rd Qu.:38.00   3rd Qu.: 8.060
##  Max.   :30.0                     Max.   :85.00   Max.   :11.750
##  NA's   :9                        NA's   :27
##  lowbirthweight    gender          habit          whitemom
##  low    :111    female:503   nonsmoker:873   not white:284
```

```
##  not low:889    male  :497   smoker   :126   white    :714
##                              NA's     :  1   NA's     :  2
##
##
##
##
```

```
dim(nc)
```

```
## [1] 1000   13
```

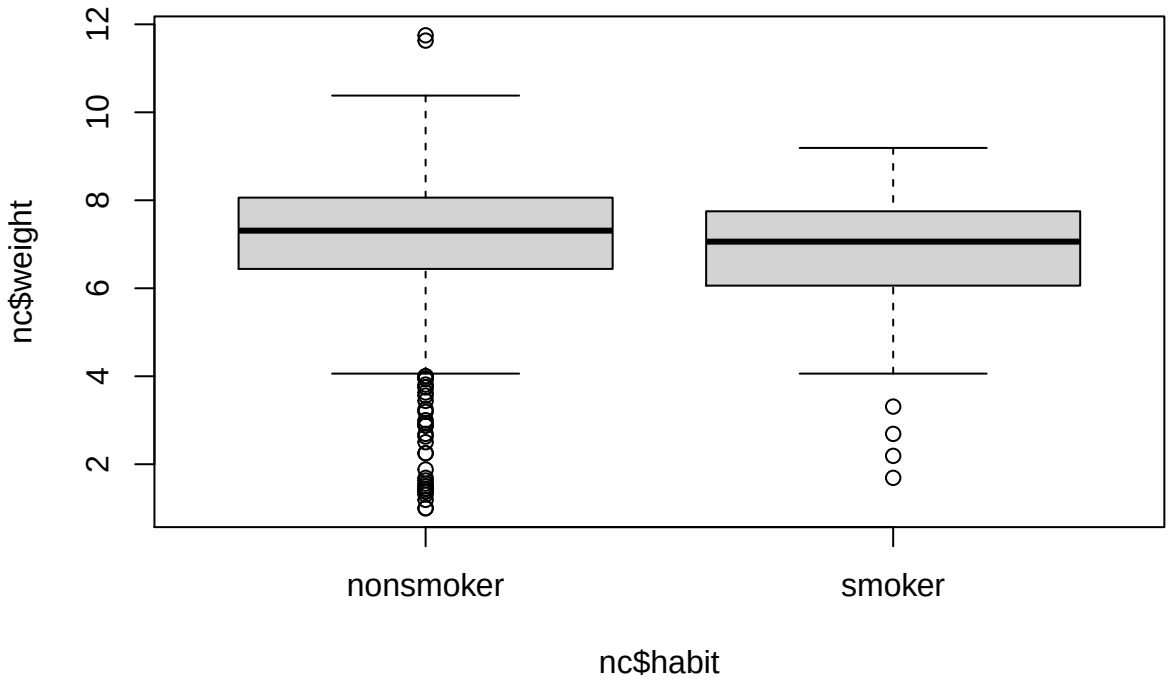
##Exercise 1 What are the cases in this data set? - Each case is one birth in the state of North Carolina.

How many cases are there in our sample? - There are 1000 cases in our sample.

Exercise 2

Make a side-by-side boxplot of habit and weight. What does the plot highlight about the relationship between these two variables? - the weight of babies of smokers appears to be less than that of non-smokers. - the median weight is less for smokers

```
boxplot(nc$weight ~ nc$habit)
```




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

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

```
## [1] 6.82873
```

#Inference ##Exercise 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.

- We must check if the data is:

(1) random: Our data is of births, the weight of the babies was up to chance.
(2) normal: our groups are size 873 and 126 for nonsmoker, smoker respectively. By the preliminary boxplots, they appear to be approximately normally distributed with no extreme, parameter warping outliers.
(3) independent: each birth is independent of every other. In addition, the number of births per year in NC is well over 10000.

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

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

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

- Null Hypothesis: There is no difference in the average weights of babies born to smoking and non-smoking mothers
- Alternative Hypothesis: There is a difference in the average weight of babies born to smoking and non-smoking mothers. Shorthand: Ho: mu1 = mu2 Ha: mu1 != mu2

```
t.test(weight ~ habit, data=nc, alternative="two.sided", var.equal=FALSE,
conf.level=0.95, paired=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  weight by habit
## t = 2.359, df = 171.32, p-value = 0.01945
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.05151165 0.57957328
## sample estimates:
## mean in group nonsmoker    mean in group smoker
##                7.144273                6.828730
```

```
#t.test(nc$weight ~ nc$habit, alternative="two.sided", var.equal=FALSE,
#conf.level=0.95, mu=0, paired=FALSE)
#t.test(nc$weight[nc$habit=="nonsmoker"], nc$weight[nc$habit=="smoker"], ... )
```

##Exercise 5 Construct and record a confidence interval for the difference between the weights of babies born to smoking and non-smoking mothers. - 95 percent confidence interval: we are 95% confident the true difference in means is between 0.05151165 and 0.57957328

```
var.test( weight ~ habit, data=nc)
```

```
##
##  F test to compare two variances
##
```

```
## data:  weight by habit
## F = 1.2003, num df = 872, denom df = 125, p-value = 0.1989
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.9076565 1.5453411
## sample estimates:
## ratio of variances
##           1.200311
```

```
#var.test(nc$weight ~ nc$habit)
t.test(weight ~ habit, data=nc, alternative="two.sided", var.equal=TRUE,
conf.level=0.95, paired=FALSE)
```

```
##
##  Two Sample t-test
##
## data:  weight by habit
## t = 2.2034, df = 997, p-value = 0.02779
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.03452013 0.59656480
## sample estimates:
## mean in group nonsmoker    mean in group smoker
##                7.144273                6.828730
```

##Exercise 6 Test for equal variances for the weights of babies born to smoking and non-smoking mothers at the .05 level. Repeat Exercise 5 and note the differences in the output from t.test() for the equal- and unequal- variance T-test. Which is more powerful in this example, the equal or unequal variance test? Why?

- At an alpha of 0.5, we do not reject the null and continue to assume the ratio of variances is equal to one.
- Conducted with assumed equal variances, a 95 percent confidence interval: we are 95% confident the true difference in means is between 0.03452013 and 0.59656480
- the second t-test with assumed equal variances will be more powerful, as there is less variability assumed in the data, giving more likely to be accurate (reflecting reality) results.

On your own (at most one page) 1. Calculate a 90% confidence interval for the average length of pregnancies (weeks) and interpret it in the context of this study. Note that since you're doing inference on a single population parameter, there is no grouping variable [i.e., habit in the example call to t.test()], so you should only list one variable in the call to t.test() leaving off a '~' and any variable that followed.

```
t.test(nc$weeks, data=nc, alternative="two.sided", var.equal=FALSE,
conf.level=0.90, paired=FALSE) #idk why but i couldnt remove the nc$ from weeks, it wouldn't run
```

```
##
##  One Sample t-test
##
## data:  nc$weeks
## t = 413.1, df = 997, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 90 percent confidence interval:
##  38.18189 38.48745
## sample estimates:
## mean of x
##  38.33467
```

- In this study, the true mean length of pregnancy is with 90% confidence between 38.18189 and 38.48745

2. Conduct a hypothesis test evaluating whether the variance of weight gained by younger mothers is different than that of mature mothers. Report a p-value and conclusion at the .05 level.

- Ho: mu1 = mu2
- Ha: mu1 != mu2

```
var.test(gained ~ mature, data=nc)
```

```
##
##  F test to compare two variances
##
## data:  gained by mature
## F = 0.88312, num df = 128, denom df = 843, p-value = 0.3795
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6871417 1.1647948
## sample estimates:
## ratio of variances
##          0.8831209
```

- p-value = 0.3795
- Conclusion: At the alpha level of .05 we choose not to reject the null hypothesis; there is no observed difference in the true mean weight gained between mature and younger mothers.

3. Conduct a hypothesis test evaluating whether the average weight gained by younger mothers is different than the average weight gained by mature mothers. Report the statistic, df, p-value, and conclusion at the .05 level.

- Ho: mu1 = mu2
- Ha: mu1 != mu2

```
t.test(gained ~ mature, data=nc, alternative="two.sided", var.equal=FALSE,
conf.level=0.95, paired=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  gained by mature
## t = -1.3765, df = 175.34, p-value = 0.1704
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.3071463  0.7676886
## sample estimates:
##  mean in group mature mom mean in group younger mom
##                  28.79070                  30.56043
```

- p-value = 0.1704
- df = 175.34
- t = -1.3765
- Conclusion: With an alpha level of 0.05, we choose not to reject the null hypothesis; there is no observable difference in the true average weights gained by mature and younger mothers.

4. Quantify the difference in weight gained by younger and mature mothers using a 95% confidence interval. State your interpretation in plain language.

- There is a 95% chance the difference in the true average weight gained by younger and mature mothers is between 28.79070 and 30.56043.

5. [A non-inference task] Determine the age cutoff that was used for younger and mature mothers.

- The age cutoff used was 35 years old and higher for mature categorization (I looked at the View of nc).