Lab 2: Inference for numerical data

North Carolina births

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.

Exploratory analysis

Load the nc data set into our workspace.

```
source("https://www.uvm.edu/~rsingle/stat211/data/ncbirths.R")
```

We have observations on 13 different variables, some categorical and some numerical. The meaning of each variable is as follows. NOTE: the exercises use the 'weight' variable, but the "on your own" uses the 'gained' variable.

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.

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

As a first step in the analysis, we should consider summaries of the data. This can be done using the summary command:

```
summary(nc)
```

As you review the variable summaries, consider which variables are categorical and which are numerical. For numerical variables, are there outliers? If you aren't sure or want to take a closer look at the data, make a graph.

Consider the possible relationship between a mother's smoking habit and the weight of her baby. Plotting the data is a useful first step because it helps us quickly visualize trends, identify strong associations, and develop research questions.

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 box plots show how the medians of the two distributions compare, but we can also compare the means of the distributions using the following function to split the weight variable into the habit groups, then take the mean of

each using the mean function.

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

There is an observed difference, but is this difference statistically significant? In order to answer this question we will conduct a hypothesis test.

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.

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

Next, we introduce the t.test() function that we will use for conducting hypothesis tests and constructing confidence intervals. Below are 3 equivalent calls to t.test()

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

- mu the NULL value of the mean [or difference in means if doing a 2-sample test]. (default=0).
- alternative The alternative hypothesis can be "less", "greater", or "two.sided". (default="two.sided").
- var.equal a logical variable indicating whether to treat the two variances as being equal. If TRUE then the pooled variance is used to estimate the variance, otherwise the Welch (or Satterthwaite) approximation to the degrees of freedom is used. (default=FALSE)
- conf.level confidence level of the interval. (deafult=0.95)
- paired a logical indicating whether you want a paired t-test. (default=FALSE)

Exercise 5 Construct and record a confidence interval for the difference between the weights of babies born to smoking and non-smoking mothers.

By default, an unequal/unpooled/separate variance T-test is done. In order to decide between an equal-variance and unequal-variance T-test, it is necessary to check if the variances are equal. The var.test() function can be used for this (we will use the Fmax test in class). The F-test performed by the var.test() function has similar limitations as those of the Fmax test that we will(or have) discuss(ed).

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

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?

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.
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. Compare your results to what you would have from the Fmax test (by hand, as there is no Fmax test in R) – indicating the Fmax statistic, degrees of freedom, and critical value.
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.
4. Quantify the difference in weight gained by younger and mature mothers using a 95% confidence interval. State your interpretation in plain language.
5. [A non-inference task] Determine the age cutoff that was used for younger and mature mothers.