Intro to inference using SAS (lab)

North Carolina Births

In 2004, the state of North Carolina released a large data set containing information about births recorded in the state. This data set is useful for 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.[1]

Exploratory Analysis

Load the **nc** data set into SAS by importing a raw data file formatted as comma-separated values (CSV). The data set contains observations on 13 variables, some categorical and some numerical.

```
filename file1 url
'http://www.uvm.edu/~rsingle/stat231/data/other/nc_sas.csv';

proc import datafile=file1
            out=nc
            dbms=csv
            replace;
   getnames=yes;
   guessingrows=max;
run;
```

[1] This lab was modified from an OpenIntro lab that was released under Creative Commons Attribution-ShareAlike 3.0 Unported (*http://creativecommons.org/licenses/by-sa/3.0/*). The original lab was adapted for OpenIntro by Mine Çetinkaya-Rundel from a lab written by the faculty and TAs of UCLA Statistics. It was then modified by SAS Institute Inc.

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 or full-term
visits	Number of hospital visits during pregnancy
marital	Whether mother is married at the time of giving 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 birth weight (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 characteristics of the babies represented 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. We can use the MEANS procedure for numerical variables and the FREQ procedure for categorical variables.

The MAXDEC= option in the PROC MEANS statement specifies the number of decimal places for all summary statistics reported by the procedure. We specify the numeric variables to be summarized using the VAR statement in PROC MEANS. Frequency tables for categorical variables are created for variables listed in the TABLES statement in PROC FREQ.

```
proc means data=nc maxdec=2;
   var fage mage weeks visits weight gained;
run;

proc freq data=nc;
   tables marital lowbirthweight*habit;
run;
```

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 the smoking habits of mothers and the weights of their babies. Plotting the data is a useful first step because it helps us quickly visualize trends, identify strong associations, and develop research questions.

Exercise 2: Create a side-by-side box plot of **habit** and **weight**. What does the plot highlight about the relationship between these two variables?

```
proc sgplot data=nc;
   vbox weight / category=habit;
run;
```

The box plots show how the medians (indicated by the horizontal line inside each box) and the means (indicated by the diamonds) of the two distributions compare. We can obtain the sample means for the two smoking groups by using PROC MEANS with a CLASS statement, indicating that **habit** is the variable that should be used to group the observations.

```
proc means data=nc mean maxdec=2;
   class habit;
   var weight;
run;
```

There is a difference between the two sample means, but is this difference statistically significant? In order to answer this question, we will conduct a hypothesis test.

Inference

Exercise 3: Write the hypotheses for testing if the average weights of babies born to smoking and non-smoking mothers are different. State the conclusion of the hypothesis test and report a 95% confidence interval for the difference between the weights of babies born to smoking and non-smoking mothers.

We will use the TTEST procedure in SAS for conducting hypothesis tests and constructing confidence intervals.

```
proc ttest data=nc sides=2 H0=0;
   class habit;
   var weight;
   title "Two-Sample t-test Comparing Birthweights by Smoking Status";
run;
```

Let's go through the statements and options in PROC TTEST:

- The CLASS and VAR statements play the same roles that they did for the MEANS procedure: the CLASS statement indicates the grouping variable, and the VAR statement indicates the analysis variable.
- When performing a hypothesis test, we can supply the null value using the H0= option in the PROC TTEST statement. In this case, the null value is *0*, because the null hypothesis sets the two population means equal to each other. Because *0* is the default value, we could have omitted the H0= option for this analysis.
- We can specify the type of hypothesis test with the SIDES= option in the PROC TTEST statement. The default is a two-sided test (SIDES=2). Left- and right-tailed tests can be specified using SIDES=L and SIDES=R, respectively.

PROC TTEST produces four tables of output along with two sets of graphs.

- First, look at the last output table, which is titled "Equality of Variances." It reports the result of the folded *F* test for the null hypothesis that the two groups have the same variance for birth weight. If this test has a significant *p*-value ($p < 0.05$), there is enough evidence to conclude that the variances for the two groups are unequal.
- Estimates and confidence intervals for the mean birth weight within each group, as well as the difference between groups, are shown in the second table. The third table presents results for the hypothesis test for equality of the mean birth weights. If the folded *F* test *p*-value from the Equality of Variances table is significant, use the Satterthwaite rows in those tables. Otherwise, use the Pooled rows.

On Your Own

1. Calculate a 90% confidence interval for the average length of pregnancies (**weeks**) and interpret it in context. Note that because you're performing inference on a single population parameter, there is no grouping variable (i.e., no CLASS statement) in PROC TTEST. You can change the confidence level from the default of 95% (ALPHA=.05) by adding the ALPHA= option to the PROC TTEST statement.
2. Determine the age cutoff that the state used to define younger vs. mature mothers. Explain how your method of determining this works.
3. Conduct a hypothesis test evaluating whether the average weight gained by younger mothers is different than the average weight gained by mature mothers.
4. Repeat the hypothesis testing #3, but using ANOVA via PROC GLM. Show the correspondence between p-values and test statistics in the two PROCedures.

use the variable 'gained' here