SEIS 631 Foundations of Data Analysis
Assignment 5

In 2004, the state of North Carolina released a large data set containing information on births recorded in the 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 which is in the **nc.RData** file. We have observations on 13 different variables, some categorical and some numerical. The meaning of each variable is as follows.

variable	**description**
`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`.

Q1) What does each case (observation) in this data represent? (See Multiple Choice in Canvas.)

Q2) How many cases do we have?

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.

We can 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. To do this we'll use the `summarise` and **`group`**`_by` functions from the Tidyverse, specifically in the `dplyr` package. `group_by()` doesn't actually change a `data.frame` but it adds meta-data that other functions (like `summarise`) can. Use. The syntax is just `group_by(data, group1, group2, …)` where data is your `data.frame` or `tibble` to be grouped and `group1`, `group2`, `group3`, are column names containing grouping variables. Note that you can group by multiple variables. In this case we'll only need one. Once the data are grouped, `summarise` finds summary statistics for each of the groups you've specified. This function takes a `data.frame` or `tibble` (grouped or ungrouped), and summarizes based on the function(s) you specifcy. The syntax is `sumarise(data, summary_name1 = function1(args), summary_name2 = function2(args), …)` where data is your `data.frame`, and `function1`, `function2`, etc. are the summary functions (like `mean()` or `sd()`). In the results, the summaries will be labeled with `summary_name1`, `summary_name2`, etc. Here is the code to summarize the weight of the baby (average), grouping by the smoking habits of the mother. I've also included the group size, *N*, with the `N = n()` argument.

```
library(dplyr)

summarise(group_by(nc, habit), mean_wgt = mean(weight), N = n())
```

Note that I've nested the group_by and summarise commands. A cleaner and more readable way to nest functions is by "piping." The `%>%` command is a binary operation that "pipes" whatever is on the left into whatever function is no the right. So if you had a variable `x` that you wanted to find the mean of you could type

```
x %>% mean()
```

This equivalent to `mean(x)`. For something as simple as finding the mean, this is not that helpful, but if we rewrite the code above using two sets of pipes (and some line spacing to make it even more readable) we get

```
nc %>%
    group_by(habit) %>%
    summarise(mean_wgt = mean(weight), N = n())
```

Written like this, it's clear that we start with the `nc` data set, group it by the `habit` variable, and find the mean `weight` and group size for each group. Also note that although `summarise` is the default, `summarize` is equivalent.

Q3) What mean weight do you get for smokers?

Q4) Non-smokers?

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

Q5) Write the hypotheses for testing if the average weights of babies born to smoking and non-smoking mothers are different. (See multiple choice in Canvas)

We do not know the population standard deviation, but according to the CLT because the sample size is large enough we could use a z-test (provided the other assumptions are met—what are those assumptions? Are they met?). Alternatively, we could use a t-test. Here we will do both and compare the results.

Two Sample Z-test

Generate histograms, boxplots, and Normal Probability plots to check the assumptions of normality. Then calculate the z-test stastic. Recall the formula for the two-sample case is

$$Z = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\sigma_1^2/_{n_1} + \sigma_2^2/_{n_2}}}$$

And in this case we will approximate σ_1 and σ_2 with s_1 and s_2. You'll need to calculate these based on the data. Use the `group_by` and `summarise` functions we saw above. Next, use the `pnorm` function to calculate the p-value. Is this a one sided or two sided p-value?

Q6) What is the p-value you calculated for this test?

Q7) Based on the p-value (and an alpha value of 5%) what conclusion do you draw?

Because we don't know the population standard deviations, the z-test is only an approximation. A better approach is to use a t-test. In R, this is even easier than above. We'll use the `t.test` function. This function can take input in a variety of ways (see the help documentation) but in this case we perform the test using this code

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

We'll look at each argument in turn:

`weight~habit`: the variable of interest is weight and we are grouping it by habit (weight is modeled by habit)

`data = nc`: this specifices the data.frame where to R can find the variables weight and habit.

`alternative = “two.sided”`: What kind of test is this? Options are “two.sided,” “greater,” or “less.”

`var.equal = TRUE`: The traditional two-sample t-test assumes equal variances between the two samples. In this case, they are close enough. If the variances are not equal (and we set this argument to FALSE) R performs a modified version of the test called a Welch test.

`conf.level = 0.95`: specifies the confidence level. This is the same as $1 - \alpha$.

Q8) What is the p-value for this hypothesis test?

Q9) Would you draw the same conclusion from this test as we did for the two sample Z-test?

Notice that the p-value is slightly higher for the t-test vs. the z-test. The z-test provides a slight underestimate of the p-value because it is an approximation based on the CLT.

Q10) Find the confidence interval in the output from the t-test. Write a sentence interpreting those numbers both in the context of the problem (i.e. you should mention weight and smoking) and in the context of the hypothesis test (i.e. how does this support the conclusion you made above).

Submission:

Save your Rhistory and any code for your own reference or if you have questions. Take the quiz on Canvas to enter you responses to each of the questions.