Lab 3: Inference for Two Means Using R

STAT 310, Spring 2023

In this lab we will go over how to perform a hypothesis test and compute a confidence interval for the difference between two means. We will use a data set on birth records to answer the question: "Is there convincing evidence that newborns from mothers who smoke have a different average birth weight than newborns from mothers who don't smoke?"

North Carolina Births Data Set

In 2004, the state of North Carolina released a large data set containing information on births recorded in this state. This data set has been of interest to medical researchers who are studying the relation between habits and practices of expectant mothers and the birth of their children. We will work with a random sample of 1,000 cases from the complete data set collected for this study.

The data set is called `ncbirths` and it can be accessed from the `openintro` package. To install this package run the following command in the R console:

```
install.packages("openintro")
```

You only need to install the package on your computer once. Next, to load the contents of the package into RStudio run the command:

```
library(openintro)
```

Note that you need to run the `library()` command each time you open RStudio and want to use a particular package.

The `ncbirths` data set should now be available in your RStudio. Type the following command to look at a scrollable, spreadsheet display of the entire data set:

```
View(ncbirths)
```

Use `dim()` to check the dimensions:

```
dim(ncbirths)
```

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

We see that the data set has 1000 rows and 13 columns (variables). Descriptions of the variables are provided in the help menu, which can be viewed by typing:

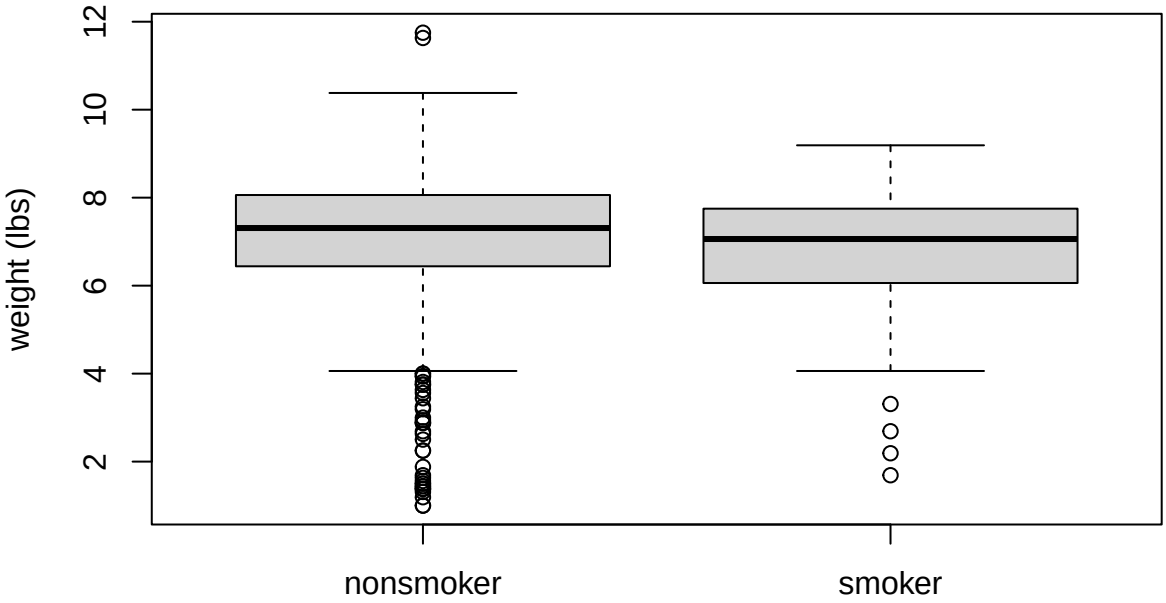
```
help(ncbirths)
```

Exploratory Analysis

As a first step in the analysis, we will look at some graphical and numerical summaries of the data. Of primary interest is the relationship between the mother's smoking status and the baby's weight.

First, we create side-by-side box plots to look at the distributions of birth weight for babies from `nonsmoker` and `smoker` mothers.

```
boxplot(weight ~ habit, data = ncbirths, xlab = "", ylab = "weight (lbs)")
```



Next, we use the `table()` function to count the number `nonsmoker` and `smoker` mothers in the data set:

```
table(ncbirths$habit)
```

```
##
## nonsmoker    smoker
##       873       126
```

We can also use the `aggregate()` function to compute the sample mean and standard deviation of birth weight for babies from `nonsmoker` and `smoker` mothers separately.

```
aggregate(weight ~ habit, data = ncbirths, FUN = mean)
```

```
##       habit   weight
## 1 nonsmoker 7.144273
## 2    smoker 6.828730
```

```
aggregate(weight ~ habit, data = ncbirths, FUN = sd)
```

```
##       habit   weight
## 1 nonsmoker 1.518681
## 2    smoker 1.386180
```

Two Sample t-test and Confidence Interval

The box plots and summary statistics indicate that babies from mothers who smoke tend to weigh less than babies from mother who do not smoke. But is the difference statistically significant? That is, is the difference in birth weight due to random chance (sampling variability), or is there an actual effect from smoking?

We can use the `t.test()` function to perform the hypothesis test and calculate a confidence interval for the difference between the two means. Specifically, the null and alternative hypothesis are

$H_0 : \mu_{ns} = \mu_s$ (the mean birth weight of babies from nonsmoking mothers *is the same* as the mean birth weight of babies from smoking mothers)

$H_A : \mu_{ns} \neq \mu_s$ (the mean birth weight of babies from nonsmoking mothers *is different* than the mean birth weight of babies from smoking mothers)

Note that the conditions for the test are satisfied since the sample sizes are large (there 873 nonsmokers, and 126 smokers); the sample was also randomly collected, and the two groups are independent (smoking and nonsmoking mothers are not related).

Now we can proceed with the hypothesis test:

```
t.test(weight ~ habit, data = ncbirths)
```

```
##
##  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 between group nonsmoker and group smoker is not equa
## 95 percent confidence interval:
##  0.05151165 0.57957328
## sample estimates:
## mean in group nonsmoker    mean in group smoker
##                7.144273                6.828730
```

Since the p-value $= 0.01945 < 0.05$, we reject H_0. Therefore, there is a statistically significant difference between the two means. The data provide convincing evidence that the mean birth weight of babies from smoking mothers is less than the mean birth weight of babies from nonsmoking mothers.

A 95% confidence interval for the difference between the two population means, $\mu_{ns} - \mu_s$, is (0.052, 0.58). So we are 95% confident that the population mean birth weight of babies from nonsmoking mothers is between 0.052 and 0.58 pounds higher than the population mean birth weight of babies from smoking mothers.

While the default confidence level is 95%, other confidence levels can be specified in the `t.test()` function. For example, run the following command:

```
t.test(weight ~ habit, data = ncbirths, conf.level = 0.9)
```

Note that similar to the `boxplot()` function, the `t.test()` function also uses the formula notation `y ~ x` where `y` is a numerical variable, and `x` is a categorical variable specifying the two groups.