Lab 7: Inference for Means Using R

STAT 630, Fall 2021

We can use the function `t.test()` to perform two-sample t-tests and compute confidence intervals in R. Type `help(t.test)` into the console to read the documentation on this function. Some important arguments:

- `x`, `y`: numeric vectors of data values
- `alternative`: specifies the alternative hypothesis as `"two.sided"` (default), `"greater"`, or `"less"`
- `mu`: a number indicating the value of the mean, or difference in means, under the null hypothesis; default is 0
- `conf.level`: confidence level of the interval (default is 0.95)

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.

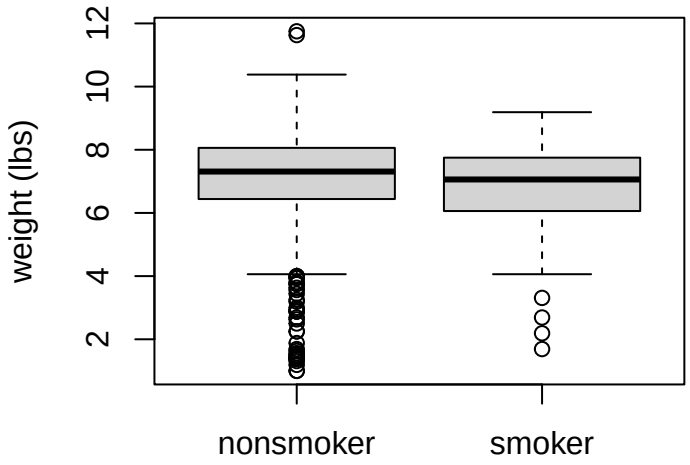
The data set is called `ncbirths` and it can be accessed from the `openintro` library. Type `help(ncbirths)` to read the documentation on this data set in the help menu.

```
library(openintro)
```

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.

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



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

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

```
nc_smoke <- subset(ncbirths, habit == "smoker")
nc_nosmoke <- subset(ncbirths, habit == "nonsmoker")
summary(nc_smoke$weight)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.690   6.077   7.060   6.829   7.735   9.190
```

```
summary(nc_nosmoke$weight)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   6.440   7.310   7.144   8.060  11.750
```

Two sample t-test

The box plot and summary statistics indicate that the babies from mothers that smoke tend to weigh less than babies from mother that do not smoke. But is the difference statistically significant? That is, are the differences in weight due to an actual effect from smoking or random sampling variability? We can use the `t.test()` function to conduct the hypothesis test and calculate a confidence interval for the difference between the two means. Specifically, we test

$H_0 : \mu_{ns} - \mu_s = 0$

$H_A : \mu_{ns} - \mu_s \neq 0$

where μ_{ns} is the population mean weight of babies for nonsmoking mothers, and μ_s is the population mean weight of babies for smoking mothers.

The conditions for the test are satisfied since the sample sizes are large (873 nonsmokers, and 126 smokers); the two groups are also independent and the sample was randomly collected.

```
t.test(nc_nosmoke$weight, nc_smoke$weight)
```

```
##
##  Welch Two Sample t-test
##
## data:  nc_nosmoke$weight and nc_smoke$weight
## 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 of x mean of y
##  7.144273  6.828730
```

Since the p-value < 0.05 we reject H_0. We are 95% confident that the population mean difference in weight $\mu_{ns} - \mu_s$ is between 0.052 and 0.580 pounds. This indicates that, on average, babies from nonsmoking mothers weigh between 0.052 and 0.58 pounds more than babies from smoking mothers.

Alternatively, we could of used the following command to conduct the t-test without using the `subset()` function.

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

This uses a formula notation in R of the form `y ~ x` where `y` is a numeric variable giving the data values, and `x` is a categorical variable specifying the two groups. Categorical variables in R are represented with factors. When using this formula notation be careful about the ordering of the levels of the factor when interpreting the results.

Exercises (in class):

1. Conduct a hypothesis test evaluating whether the average weight gained during pregnancy by younger mothers is significantly different than the average weight gained during pregnancy by mature mothers. Use $\alpha = 0.05$. (Hint: type `help(ncbirths)` to read the variable descriptions and determine which variables to use)
2. Now, a non-inference task: Determine the age cutoff for younger and mature mothers.

Paired t-test

Going back to the example from lecture, the data set `husbands_wives` (also from the `openintro` package) contains data on the ages of a random sample of married couples in Britain.

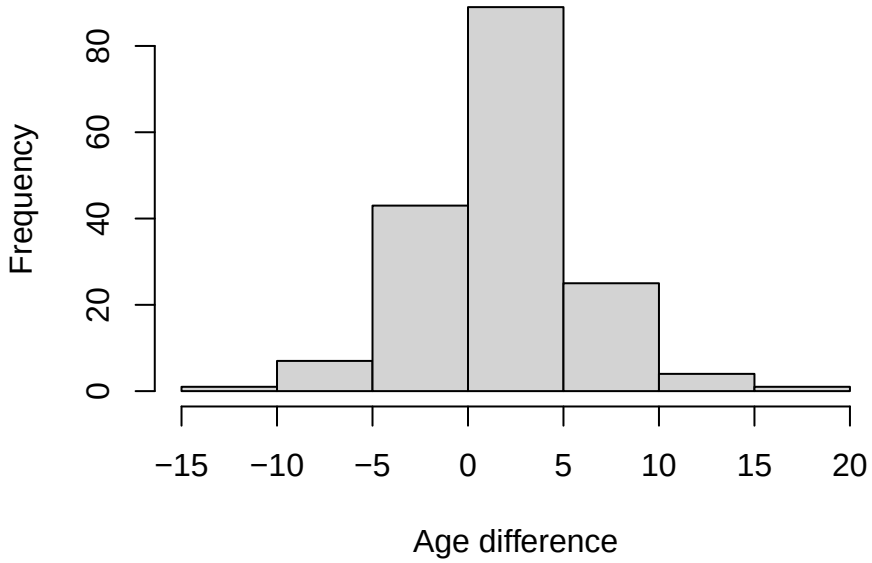
```
age_diff <- husbands_wives$age_husband - husbands_wives$age_wife
age_diff <- age_diff[!is.na(age_diff)] # remove missing entries (NA values)
length(age_diff)
```

```
## [1] 170
```

```
summary(age_diff)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -12.000   0.000   2.000   2.235   4.000  20.000
```

```
hist(age_diff, xlab="Age difference", main='')
```



The histogram shows that the differences (husband's age - wife's age) are symmetric and bell-curve shaped. Also, the sample size is large ($n = 170$), so the conditions for the t-test are well satisfied.

```
t.test(age_diff)
```

```
##
##  One Sample t-test
##
## data:  age_diff
## t = 7.1518, df = 169, p-value = 2.474e-11
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  1.618286 2.852302
## sample estimates:
## mean of x
##  2.235294
```

Since the p-value ≈ 0 we reject the null hypothesis, and conclude that the population mean difference in age is significantly different than 0. Additionally, we are 95% confident that British husbands are between 1.62 and 2.85 years older than their wives, on average.