Lesson 19 Companion

Research question:

For this example I'm going to use the "bowls of M&Ms" data set. The research question here is whether or not there is an association between bowl size and the number of M&Ms taken from the bowl. To conduct this study, researchers had participants attend two study sessions. In one session they were offered a small bowl and in another they were offered a large bowl.

H_0 : There is no association between bowl size and number of M&Ms taken.

H_a : When a large bowl is present, people tend to take more M&Ms.

Notice there is something a little different about the alternate hypothesis here. I don't know if this was the textbook's attempt to get you to try something new or perhaps these were really the hypothesis of the researchers. Can you identify how this is going to change your analysis?

Another very important aspect of this experimental design (and the reason for a seperate chapter) is that each participant participates in two trials of the study. We have a measurement for a given participant under both conditions: large and small bowl. This means that we have paired data for each participant. It seems to me that this would be the preferred method for almost all experiments because it helps remove the uncertainty of different participants. Unfortunately, a lot of experiments/studies can not collect paired data: you can't collect the GCI of children who were breastfed and then un-breastfeed them to collect that GCI as well.

Because we are dealing with paired data, the way we approach the problem is slightly different. Rather than our relevant statistic being the **difference of the averages of each condition** our relevant statistic is the **average difference between conditions**. That is:

$\bar{x}_d = \frac{\sum_{i=1}^{n} x_{i,a} - x_{i,b}}{n}$ **not** $\bar{x}_a - \bar{x}_b$

You will see that in the case of $\bar{x}_d$ we are summing the difference between the number of M&Ms for each condition for each participant and dividing by the total number participants (finding the average difference).

Our parameter of interest is the population average difference between conditions so we can rewrite our hypothesis in notation like this:

$H_0 : \mu_d = 0$

$H_a : \mu_d > 0$

Note: Because we are doing a one-sided test (and we wish to test whether more M&Ms are taken from a larger bowl) it's important to pick the order of your subtraction in the summation and your inequality properly. As I've decided to use a *greater than* inequality I should order my group such that my summation is:

$\bar{x}_d = \frac{\sum_{i=1}^{n} x_{i,large} - x_{i,small}}{n}$

This means that if we will have a positive average if the average person takes more from the large bowl.

```r
library(tidyverse)

m_m = read_table2("http://www.isi-stats.com/isi/data/chap7/BowlsMMs.txt")

head(m_m)
```

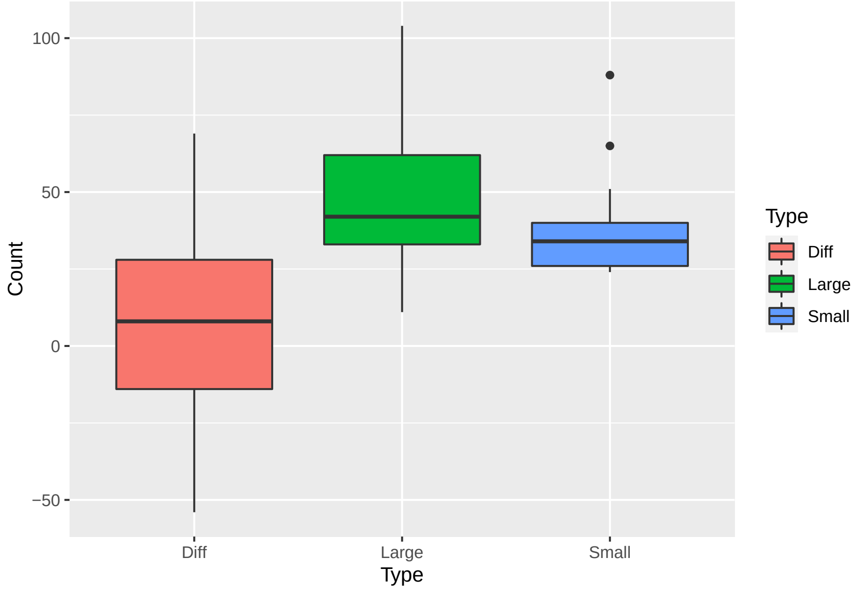
```
## # A tibble: 6 x 2
##   Small Large
##   <dbl> <dbl>
## 1    33    41
## 2    24    92
## 3    35    61
## 4    24    19
## 5    40    21
## 6    33    35
```

```
sample_stat = m_m %>%
  mutate(Diff = Large - Small) %>%
  summarise(x_d = mean(Diff))

null_value = 0
```

Data exploration:

I'm going to so a little bit fancier with this data exploration to get a box plot of our values.

```
m_m %>%
  mutate(Diff = Large - Small) %>%
  pivot_longer(c(Small, Large, Diff), names_to = "Type", values_to = "Count") %>%
  ggplot(aes(x = Type, y = Count, fill = Type)) +
  geom_boxplot()
```



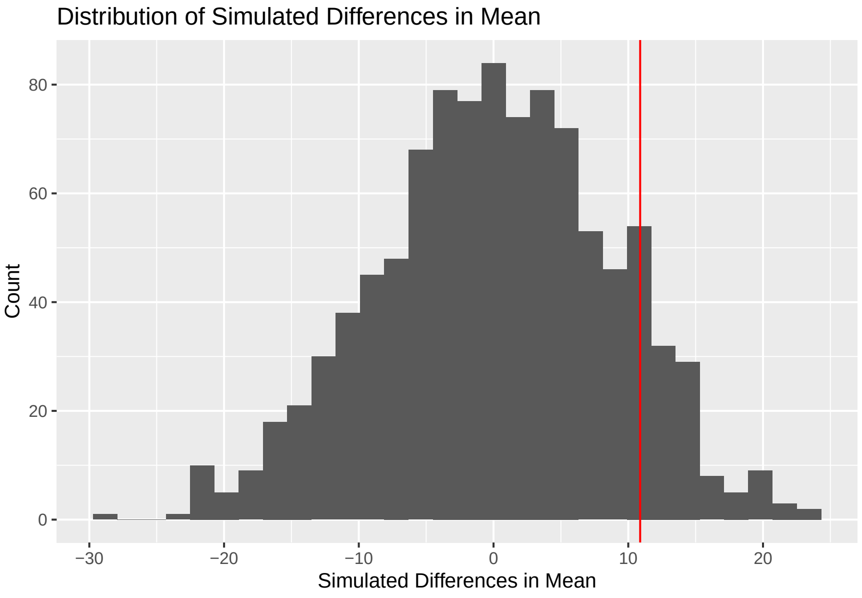
As you can see from this command, I've take "wide" data (row is participants, columns are value for that participant) and changed it to "long" data (row for each value, participant's have several rows each). This allows me to more easily produce the box plot you see here. I've included not only the box plot for the large and small bowls but also a boxplot of the difference between the number of M&Ms each partitipant took.

It appears from this boxplot that the median number of pieces taken from the larger bowl is higher than the from the smaller bowl and that the number taken from the larger bowl varies more than the smaller bowl. You can also see two outliers in the smaller bowl values at the top of the range.

Simulation-based approach:

In order to create the null distribution here we will randomly reassign the “bowl sizes” to the measurements. This is how we ensure there is no association between the bowl size and the number of M&Ms. Some participants will get their types switched, and some will not.

```
replications_dataframe = NULL

num_reps = 1000

for (i in 1:num_reps){

  trial_stat = m_m %>%
    rowwise() %>%
    mutate(Rand = runif(1)) %>%
    mutate(new_Small = ifelse(Rand > 0.5, Small, Large),
           new_Large = ifelse(Rand > 0.5, Large, Small)) %>%
    ungroup() %>%
    mutate(Diff = new_Large - new_Small) %>%
    summarise(trial_stat = mean(Diff))

  replications_dataframe = rbind(replications_dataframe, data.frame(trial_stat))

}

replications_dataframe %>%
  ggplot(aes(x = trial_stat)) +
  geom_histogram() +
  labs(x = "Simulated Differences in Mean", y = "Count",
       title = "Distribution of Simulated Differences in Mean") +
  geom_vline(xintercept = sample_stat[[1]], color = "red")
```



```
replications_dataframe %>%
  summarise(pvalue = sum(trial_stat >= sample_stat[[1]]) / n())
```

```
##   pvalue
## 1  0.116
```

A p-value of 0.116 provides little to no evidence against the null hypothesis and thus suggests (shockingly to me) that there is no statistically significant association between the size of the bowl and the number of M&Ms a person takes.

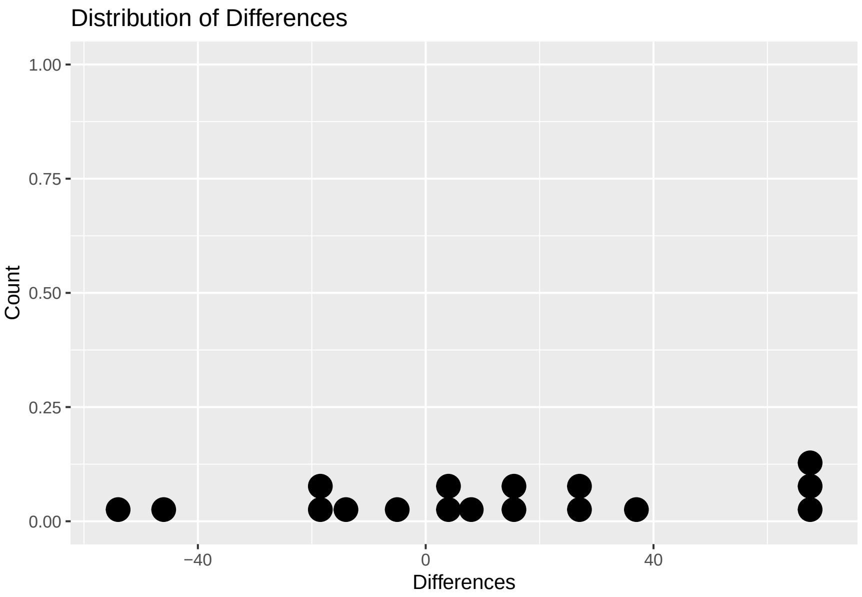
Paired t-test

The validity conditions for the paired t-test are:

1. The distribution of the differences is symmetric **or**
2. You have at least 20 pairs and the distributions are not strongly skewed.

I suppose we better hope for a symmetric distribution of the differences because we clearly don't have 20 pairs of measurements.

```
m_m %>%
  mutate(Diff = Large - Small) %>%
  ggplot(aes(x = Diff)) +
  geom_dotplot() +
  labs(x = "Differences", y = "Count",
       title = "Distribution of Differences")
```



Notice I used **geom_dotplot()** here instead of **geom_histogram()** because there are so few pairs. The book states that the distribution of these difference is symmetric enough to utilize the theory-based approach. This should give you an idea of how flexible the theory-based approach is to the symmetry of the differences. I wouldn't blame you for saying this isn't symmetrical. If you decided that, you would need to use the simulation-based approach. We can always compare the simulation-based and theory-based p-values in the end.

Here is the code to find the p-value using **pt()**:

```
n = 17

s_d = m_m %>%
  mutate(Diff = Large - Small) %>%
  summarise(s_d = sd(Diff))
```

```
stand_error = s_d / sqrt(n)

t_stat = (sample_stat - null_value) / stand_error

#One-sided p-value
1 - pt(t_stat[[1]], df = n - 1)
```

```
## [1] 0.1171413
```

Of course here is the quicker way:

```
t.test(x = m_m$Large, y = m_m$Small, paired = TRUE, alternative = "greater")
```

```
##
##  Paired t-test
##
## data:  m_m$Large and m_m$Small
## t = 1.236, df = 16, p-value = 0.1171
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -4.488747       Inf
## sample estimates:
## mean of the differences
##                10.88235
```

Our simulation-based p-value was 0.116 and the theory-based p-value was 0.117 so I suppose we were justified in accepting that the validity conditions were met.

Confidence interval:

```
lower = sample_stat + qt(0.025, df = n - 1) * stand_error

upper = sample_stat - qt(0.025, df = n - 1) * stand_error

paste("(", lower, ",", upper, ")")
```

```
## [1] "( -7.78170563915685 , 29.5464115215098 )"
```

We can interpret this interval as saying: I am 95% confident that, on average in the long run, students take up to 7.78 candies less from a large bowl or up to 29.54 candies more from a large bowl when compared to a small bowl. Clearly this interval contains zero and therefore zero is plausible value for the average difference between conditions. This agrees with our relatively high p-value.