Lesson 14 Companion

Research question:

I'm going to use the parent's smoking habit vs. baby's gender example for this document. The research question is: Does a parent's smoking habit have an effect on the gender of their baby? To test this question, scientists collected data on the whether both parents smoked and the gender of their child.

H_0 : The population proportion of boys born to smoking parents is the same as the population proportion of boys born to nonsmoking parents. Put simply: There is no association between smoking status of parents and the sex of child.
H_a : The population proportion of boys born to smoking parents is different from the population proportion of boys born to nonsmoking parents. Or: There is an association between smoking status of parents and the sex of child.

Our explanatory variable is the parent's smoking status and the response variable is the sex of the baby. Hopefully, this is fairly clear. What might not be so clear is the choice of sample statistic. When we dealt with single proportion problems our relevant statistic was the sample proportion (p); however, now that we have two proportions, our sample statistic is the difference between these two proportions:

$p_{nonsmoker} - p_{smoker}$

Two things to keep track of:

- How you define success: Here we are led to defining success as "boy" for our proportions. As the father of three boys and one girl, I would definitely define it the other way around but I'll go with it for now.
- The order of your proportions: Doesn't matter which way you pick as long as you remain consistent.

We can also rewrite our hypothesis in notation:

$H_0 : \pi_{nonsmoker} - \pi_{smoker} = 0$
$H_a : \pi_{nonsmoker} - \pi_{smoker} \neq 0$

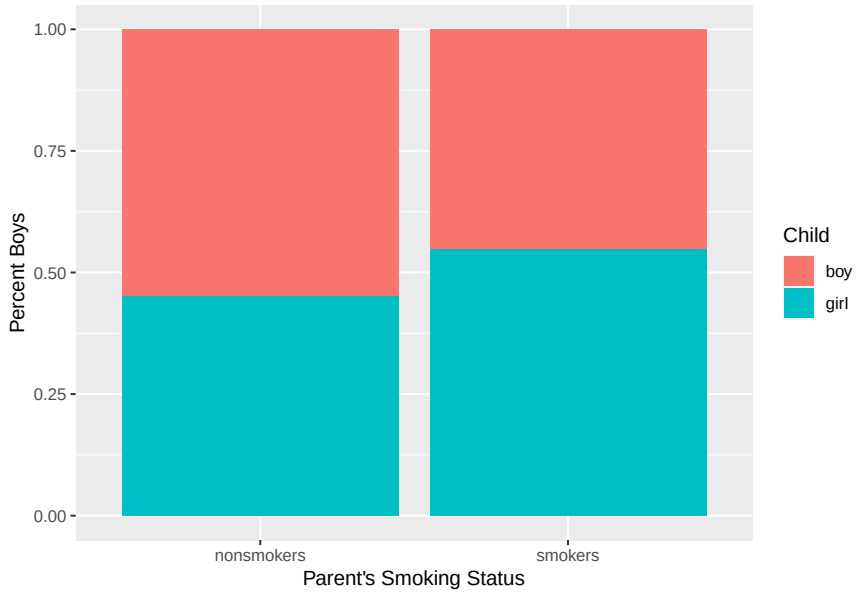
Data exploration:

Let's start with some data exploration:

```
library(tidyverse)

smoking = read_table2("http://www.isi-stats.com/isi/data/chap5/Smoking.txt")

head(smoking)
```

```
## # A tibble: 6 x 2
##   Parents Child
##   <chr>   <chr>
## 1 smokers girl
## 2 smokers girl
## 3 smokers girl
## 4 smokers girl
## 5 smokers girl
## 6 smokers girl
```

```r
smoking %>%
  count(Parents, Child) %>%
  ggplot(aes(fill = Child, y = n, x = Parents)) +
  geom_bar(position = "fill", stat = "identity") +
  labs(x = "Parent's Smoking Status", y = "Percent Boys")
```




```r
#No reason you have to program this is, you could just
# get the number from the first two lines and compute the
# proportion yourself. #Rflex
p_nonsmokers = smoking %>%
  count(Parents, Child) %>%
  filter(Parents == "nonsmokers") %>%
  mutate(p = n / (n + lead(n))) %>%
  select(p) %>%
  drop_na()

p_smokers = smoking %>%
  count(Parents, Child) %>%
  filter(Parents == "smokers") %>%
  mutate(p = n / (n + lead(n))) %>%
  select(p) %>%
  drop_na()

sample_stat = (p_nonsmokers - p_smokers)[[1]]

#For use later
null_diff = 0
```

The only new code here is the production of the stacked bar chart. I feed in our dataframe and used the **count()** function to group by the parent's smoking status and then by the baby's sex and count how many we have in each of these four categories (smokers/boy, smokers/girl,...). The rest of the bar chart code was taken from: https://www.r-graph-gallery.com/48-grouped-barplot-with-ggplot2.html. I won't explain it further because it's a lot of *R* black magic to me.

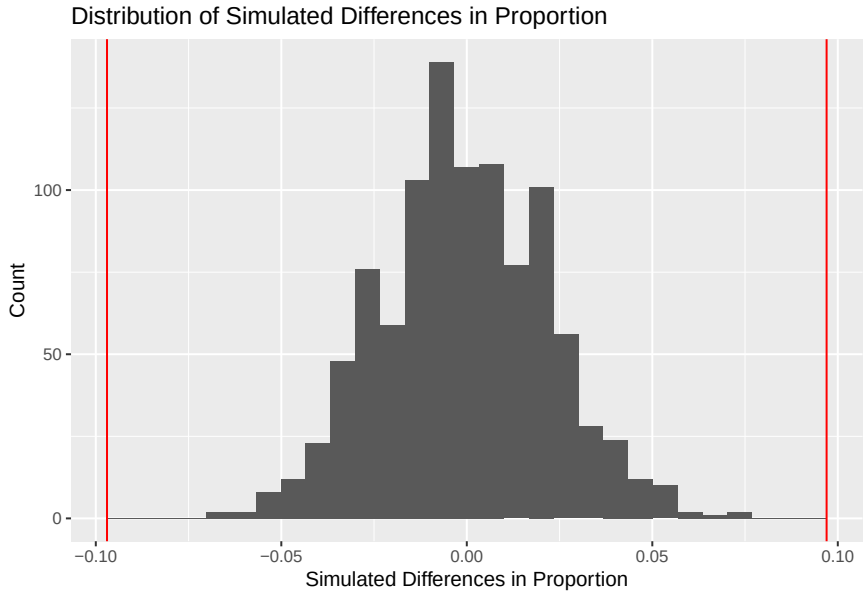
So what can we take from the segmented bar chart? I seems to me that, in our sample, the proportion of

boys is relatively close together for both status'. The proportion of boys from smoking parents is definitely higher but both proportion hover around 50%.

Simulation-based approach:

We start by building our null distribution which, of course, assumes the truth of the null hypothesis. To do this we need to devise a strategy to break the association between the explanatory and response variables. We can do this by randomly assigning whether or not a baby's parents are smokers or nonsmokers while maintaining the same number of smokers and nonsmokers.

```
replications_dataframe = NULL

num_reps = 1000

for (i in 1:num_reps){

  #Produce a new dataframe (sim_smoking) with a new column
  #that contains scrambled parents by sampling the original w/o
  #replacement.
  sim_smoking = smoking %>%
    mutate(new_Parents = sample(Parents, size = n(), replace = FALSE))

  #Calculate the difference in proportions.
  #This is where is nice to have an automatic way
  # to calculate these proportions.
  p_nonsmokers = sim_smoking %>%
    count(new_Parents, Child) %>%
    filter(new_Parents == "nonsmokers") %>%
    mutate(p = n / (n + lead(n))) %>%
    select(p) %>%
    drop_na()

  p_smokers = sim_smoking %>%
    count(new_Parents, Child) %>%
    filter(new_Parents == "smokers") %>%
    mutate(p = n / (n + lead(n))) %>%
    select(p) %>%
    drop_na()

  trial_stat = (p_nonsmokers - p_smokers)[[1]]

  #Add it to my list of simulated differences
  replications_dataframe = rbind(replications_dataframe, data.frame(trial_stat))

}

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




```
replications_dataframe %>%
  summarise(pvalue = sum(abs(trial_stat) >= abs(sample_stat)) / n())
```

```
##   pvalue
## 1      0
```

As you can see from the histogram and the p-value, this sample offers very strong evidence against the null hypothesis. This suggests there is a difference in the proportion of boys born to smokers vs. nonsmokers.

Let's take a look at the theory-based approach.

Two-sample z-test:

If we are going to put any trust into the p-value of our theory-based approach we better verify the validity conditions. For this test we need to have at least 10 observations in each group.

```
smoking %>%
  count(Parents, Child)
```

```
## # A tibble: 4 x 3
##   Parents    Child     n
##   <chr>      <chr> <int>
## 1 nonsmokers boy    1975
## 2 nonsmokers girl   1627
## 3 smokers    boy     255
## 4 smokers    girl    310
```

I think we just barely scrape by with this dataset. Just kidding, we've got plenty of observations.

Let me show you two ways of computing the p-value. The first will standardize our sample statistic and the second will not.

```
#Calculate the overall proportion of success (p_hat)
p_hat = smoking %>%
    count(Child) %>%
    mutate(p = n / (n + lead(n))) %>%
    select(p) %>%
    drop_na()
```

```r
#Calculate sample size for both groups
n_non = smoking %>%
  filter(Parents == "nonsmokers") %>%
  summarise(count = n())

n_smoke = smoking %>%
  filter(Parents == "smokers") %>%
  summarise(count = n())

#Calculate the standard error of the statistic
stand_error = sqrt( (p_hat * (1 - p_hat)) *
                      ((1 / n_non) + (1 / n_smoke)))


z = (sample_stat - null_diff) / stand_error

2 * (1 - pnorm(abs(z[[1]])))
```

```
## [1] 1.731004e-05
```

```r
2 * (1 - pnorm(abs(sample_stat), mean = null_diff, sd = stand_error[[1]]))
```

```
## [1] 1.731004e-05
```

While the second approach looks much shorter, it uses some of the same values calculated in the first so it's really a matter of dropping the standardization and using *R*'s ability to calculate the area under a custom normal curve. The p-values calculated aren't numerically the same but they are practically the same for the purposes of judging evidence against the null hypothesis. It's clear that we still have very strong evidence against the null hypothesis.

You will notice I used [[**1**]] several times in the code. This is kind of a time saving mechanism for me as I like to program in the code to pull the values I need. Doing this the way I have leaves me with a "list" for the values of **p_hat** and **stand_error** (and several other places). As what I really want are the first values in these lists, the [[**1**]] allows me to pull that first value out to use it. There is likely a better way to do this but this is what I've got right now.

Confidence interval:

We can compute a confidence interval for our population parameter ($\pi_{nonsmoker} - \pi_{smoker}$) as well.

```r
#If it seems counterintiuitive that the "lower" equation
# has a plus sign, remember that qnorm(0.025) is a negative number.
lower = sample_stat + qnorm(0.025) * stand_error

upper = sample_stat - qnorm(0.025) * stand_error

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

```
## [1] "( 0.0527449312789062 , 0.141213194246244 )"
```

Because we have positive values in our confidence interval we can interpret it by saying: babies born to nonsmoking parents are between 5.3% and 14.1% more likely (with 95% confidence) to be a boy than babies born to smoking parents. If we had negative values, we'd have to say they are less likely.

This is a little different than the other interpretation we have learned (we are 95% confident that the true difference in proportion is between 5.3% and 14.1%): however, it provides much more clarity in what we have calculated.

Practical importance:

We have a p-value that suggests there is a difference in the true proportion of boys born to nonsmokers and smokers. The confidence interval doesn't include zero and therefore agrees with this (not unexpectedly). Now we are faced with deciding if this result is practically important. Based on the lower value of 5.3%, I assess this result it practically important. Increasing the chances of a boy by more than 5% is a non-trivial difference. I don't think we would have picked up smoking but if you had offered me something that would increase my chances of having a girl after having three boys... I would taken it.