Data Analysis
Week 7 Task Solutions

Further Tasks

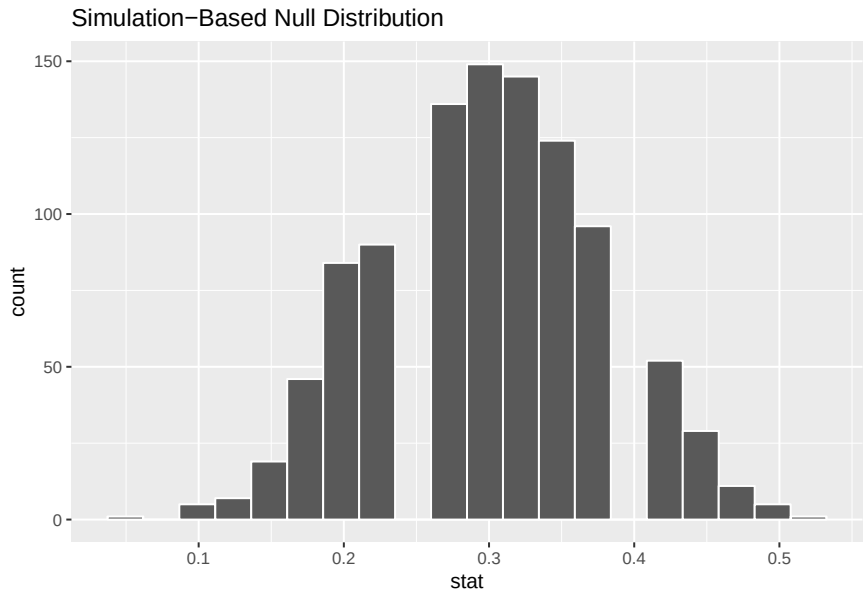
You are encouraged to complete the following tasks by using RMarkdown to produce a single document which summarises all your work, i.e. the original questions, your R code, your comments and reflections, etc.

1. In the last section, we constructed a confidence interval for the difference in the proportion of people who yawned between the "seeded" group and the "control" group (Scenario 3).

   By modifying the code in the last section in light of how we constructed a confidence interval for the age of pennies in Section on "Constructing confidence intervals" (Scenario 2), use the `mythbusters_yawn` data to constuct a confidence interval for the proportion of people who yawn when they see someone else yawn (Scenario 1). Does this overlap with the confidence interval for the proportion of people who yawn when they did not see someone else yawn (Scenario 1 again!)? Are your findings here consistent with the findings in the last section?

Solution 1. We start by focusing on the group that saw someone yawn, i.e. `group == "seed"`:

```
bootstrap_distribution <- mythbusters_yawn %>%
  filter(group=="seed") %>%
  specify(formula = yawn ~ NULL, success = "yes") %>%
  generate(reps = 1000) %>%
  calculate(stat = "prop")
bootstrap_distribution %>% visualize(bins = 20)
```



This distribution is roughly symmetric and bell-shaped but isn't quite there. Let's use the percentile-based method to compute a 95% confidence interval for the true proportion of those that yawn with a seed presented. The arguments are explicitly listed here but remember they are the defaults and simply `get_ci()` can be used.

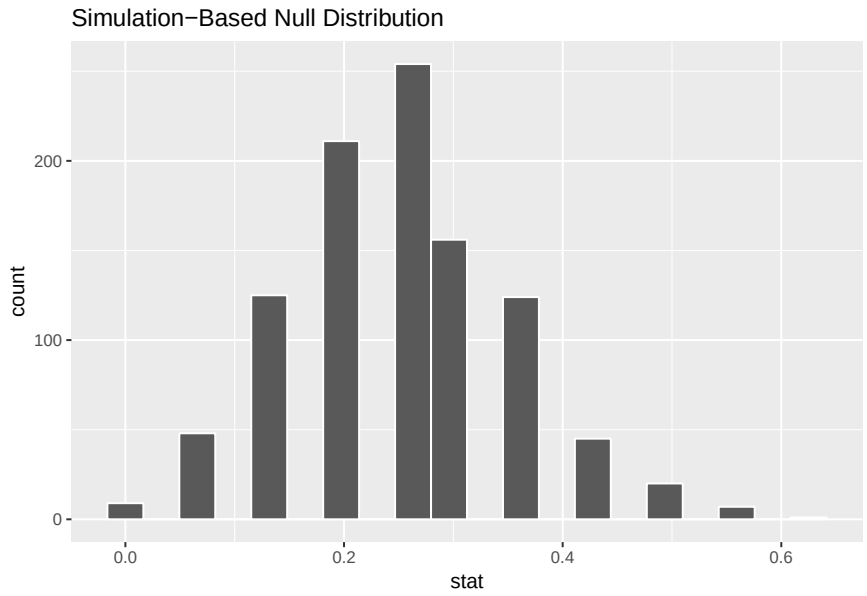
```
bootstrap_distribution %>%
  get_ci(type = "percentile", level = 0.95)
```

```
# A tibble: 1 x 2
  `2.5%` `97.5%`
   <dbl>   <dbl>
1  0.147   0.441
```

The confidence interval shown here is (0.15, 0.44). The range of plausible values for the proportion of people that yawned with after seeing someone yawn is therefore between 0.15 and 0.44.

We now repeat the exercise but for the group that didn't see someone yawn, i.e. `group == "control"`

```
bootstrap_distribution <- mythbusters_yawn %>%
  filter(group=="control") %>%
  specify(formula = yawn ~ NULL, success = "yes") %>%
  generate(reps = 1000) %>%
  calculate(stat = "prop")
bootstrap_distribution %>% visualize(bins = 20)
```




```
bootstrap_distribution %>%
  get_ci(type = "percentile", level = 0.95)
```

```
# A tibble: 1 x 2
  `2.5%` `97.5%`
   <dbl>   <dbl>
1 0.0625     0.5
```

```
Setting `type = "bootstrap"` in `generate()`.
```

The confidence interval shown here is (0.06, 0.5). The range of plausible values for the proportion of people that yawned without seeing someone yawn is therefore between 0.06 and 0.5.

Comparing the two CIs for the proportion of those that saw someone yawn (0.15, 0.44) and those that didn't (0.06, 0.5), we note that these two CIs overlap, which is consistent with the findings from the CI for the difference in the two proptions (-0.23, 0.31) in the last section, i.e. since they **do overlap** its plausible that they could each take the **same value** and therefore its plausible that their **difference is zero**, which is exactly what the CI for the difference in proportions tells us.

2. Recall the data on 144 domestic male and female adult cats that we first saw in Week 4. Each cat had their heart weight in grams (`Hwt`) and body weight in kilograms (`Bwt`) measured, and interest lies in exploring difference between females and males.

    a. Construct a bootstrap confidence intervals for the avearge heart weight of female and male cats separately? Interpret your results.

    b. Construct a bootstrap confidence interval for the difference in the avearge heart weights of female and male cats. Interpret your result.

    c. Repeat a. and b. for the body weight of cats.

Hint: You need to read in the `cats` data and remind yourself how it is organised, e.g.

```
cats <- read.csv("cats.csv")
glimpse(cats)
```

Solution 2.a.

a. Construct a bootstrap confidence intervals for the avearge heart weight of female and male cats separately? Interpret your results.

We start by focusing on female cats, i.e. `Sex == "F"`:

```
cats <- read.csv("cats.csv")
#glimpse(cats)
bootstrap_distribution <- cats %>%
  filter(Sex == "F") %>%
  specify(response = Hwt) %>%
  generate(reps = 1000) %>%
  calculate(stat = "mean")
#bootstrap_distribution %>% visualize()
percentile_ci <- bootstrap_distribution %>%
  get_ci()
#percentile_ci
```

Using the percentile method, our range of plausible values for the mean heart weight of female adult cats is 8.82 grams to 9.58 grams.

Now repeat the analysis for male cats, i.e. `Sex == "M"`:

```
bootstrap_distribution <- cats %>%
  filter(Sex == "M") %>%
  specify(response = Hwt) %>%
  generate(reps = 1000) %>%
  calculate(stat = "mean")
#bootstrap_distribution %>% visualize()
percentile_ci <- bootstrap_distribution %>%
  get_ci()
#percentile_ci
```

Using the percentile method, our range of plausible values for the mean heart weight of male adult cats is 10.83 grams to 11.79 grams.

Solution 2.b.

b. Construct a bootstrap confidence interval for the difference in the avearge heart weights of female and male cats. Interpret your result.

```
bootstrap_distribution <- cats %>%
  specify(Hwt~Sex) %>%
  generate(reps = 1000) %>%
  calculate(stat = "diff in means", order = c("F", "M"))
percentile_ci <- bootstrap_distribution %>%
  get_ci()
```

Using the percentile method, our range of plausible values for the difference in the mean heart weight between female and male adult cats is -2.77 grams to -1.45 grams. That is to say that on average female adult cats' hearts weigh between 1.45 and 2.77 grams **less** than adult male cats' hearts. (Note: the fact that the individual CIs in part a. didn't overlap told us that zero wouldn't be in the interval for the difference in the population means)

Solution 2.c.

c. Repeat a. and b. for the body weight of cats.

RMarkdown makes it very easy to repeat the analysis in parts a. and b. on a different variable.

```
cats <- read.csv("cats.csv")
bootstrap_distribution <- cats %>%
  filter(Sex == "F") %>%
  specify(response = Bwt) %>%
  generate(reps = 1000) %>%
  calculate(stat = "mean")
percentile_ci <- bootstrap_distribution %>%
  get_ci()
```

Using the percentile method, our range of plausible values for the mean body weight of female adult cats is 2.28 kilograms to 2.43 kilograms.

Now repeat the analysis for male cats, i.e. `Sex == "M"`:

```
bootstrap_distribution <- cats %>%
  filter(Sex == "M") %>%
  specify(response = Bwt) %>%
  generate(reps = 1000) %>%
  calculate(stat = "mean")
percentile_ci <- bootstrap_distribution %>%
  get_ci()
```

Using the percentile method, our range of plausible values for the mean body weight of male adult cats is 2.82 kilograms to 2.99 kilograms.

```
bootstrap_distribution <- cats %>%
  specify(Bwt~Sex) %>%
  generate(reps = 1000) %>%
  calculate(stat = "diff in means", order = c("F", "M"))
percentile_ci <- bootstrap_distribution %>%
  get_ci()
```

Using the percentile method, our range of plausible values for the difference in the mean body weight between female and male adult cats is -0.67 kilograms to -0.41 kilograms. That is to say that on average female adult cats' bodies weigh between 0.41 and 0.67 kilograms **less** than adult male cats' bodies. (Note: the fact that the individual CIs in part a. didn't overlap told us that zero wouldn't be in the interval for the difference in the population means)