Instructions

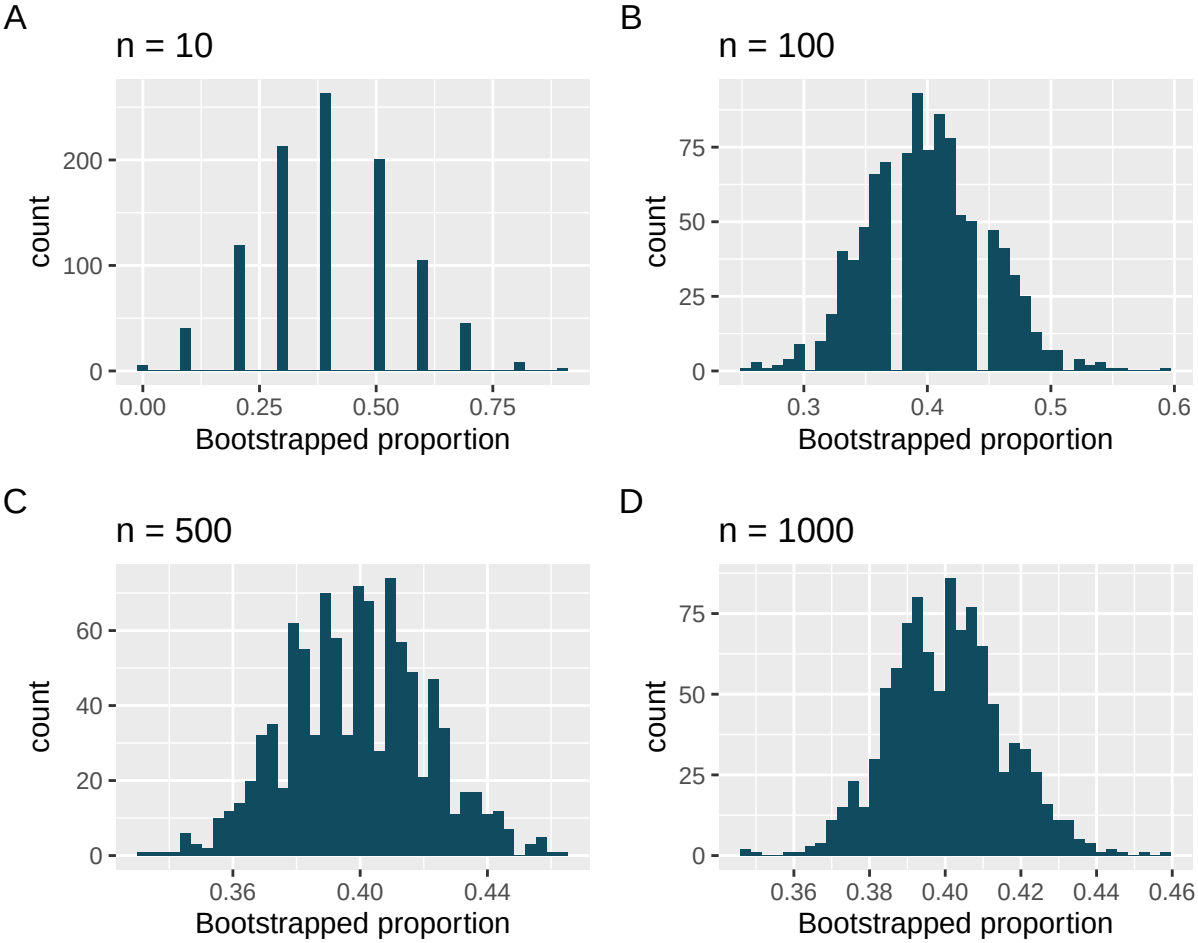
Upload a PDF file, named with your UC Davis email ID and homework number (e.g., sfrei_hw4.pdf), to Gradescope (accessible through Canvas). You will give the commands to answer each question in its own code block, which will also produce output that will be automatically embedded in the output file. All code used to answer the question must be supplied, as well as written statements where appropriate.

All code used to produce your results must be shown in your PDF file (e.g., do not use `echo = FALSE` or `include = FALSE` as options anywhere). Rmd files do not need to be submitted, but may be requested by the TA and must be available when the assignment is submitted.

Students may choose to collaborate with each other on the homework, but must clearly indicate with whom they collaborated.

Problem 1: [IMS] 12.6

1. **Bootstrap distributions of $\hat{p}$, III.** Each of the following four distributions was created using a different dataset. Each dataset had the same proportion of successes ($\hat{p} = 0.4$) but a different sample size. The four datasets were given by $n = 10, 100, 500$, and 1000.



Consider each of the following values for the true popluation p (proportion of success). Datasets A, B, C, D were bootstrapped 1000 times, with bootstrap proportions as given in the histograms provided. For each parameter value, list the datasets which could plausibly have come from that population. (Hint: there may be more than one dataset for each parameter value.)

`a. p = 0.05`

`b. p = 0.25`

`c. p = 0.45`

`d. p = 0.55`

`e. p = 0.75`

To list the datasets which could plausibly have come from that population, we need to consider two things: * the bootstrap sample (there are 1,000 repeated bootstrap samples) will be centered around the *sample proportion* of whatever dataset we have, * the *sample proportion* of the dataset will in general be centered around the population proportion. However, this will depend upon the sample size: if we only have a sample size of 3, it would be very hard to tell if, when we saw say a sample proportion of 1/3, whether this came

from a $p = 0.1$ or $p = 0.9$ population.

We will use this reasoning as we consider each of the following.

For a, p=0.05, if the dataset comes from the distribution with this parameter, the typical sample proportion seen in a dataset will be around 0.05. However, in A, B, C, D, none of them center around 0.05. It means that it's unlikely for a distribution with p=0.05 to generate the dataset similar to A, B, C, D. The only possible one would be A, since there are so few samples ($n = 10$) it is possible (but rather unlikely) that we just had a sample with this outcome, but it is most likely none of these.

For b, p=0.25, analogously, we can be very confident that none of B, C, or D had p=0.25 since there are over 100 samples and having a sample proportion of 0.40 when the true proportion is 0.25 would be very unlikely. This is more possible in the $n = 10$ setting, so A is possible, but it is also reasonable to say it us unlikely.

For c, p=0.45, similar arguments show that A and B are likely. C and D are very unlikely since we have 500+ examples and they are centered around 0.39 - 0.41 very tightly, so we can assume the proportion in the sample proportion was around 0.39-0.41, not 0.45.

For d, p=0.55, the data centers around 0.55 in A so the answer is A, and clearly not B,C, or D.

For e, p=0.75, the data centers around 0.75 in none of them. The only possible one would be A, since there are so few samples it is somewhat possible that we just had a sample with this outcome, but it is most likely none of these.

Problem 2

In this problem we work through how to do a randomization test in R. We will work with the following example dataset.

```
df <- tibble(
  group = rep(c("control", "treatment"), each=50), # 50 control and 50 treatment observations
  response = factor(c(rep("success", 20), rep("failure", 30),
                      rep("success", 29), rep("failure", 21)),
                    levels = c("success", "failure"))
)
```

Part (a) - computing sample proportion

Write a function `compute_proportions` which takes in the argument `tib`, a tibble, which we assume takes the form of the above tibble (i.e., it has two columns, "group" and "response", where "group" is a factor which takes values "control" and "treatment" while "response" is a factor which takes values "success" and "failure".)

The function `compute_proportions` returns a tibble with two variables, "group" and "proportion". Each row of the returned table corresponds to the proportion of "success" within each group of "control" and "treatment".

```
compute_proportions <- function(tib) {
  tib %>%
    group_by(group) %>%
    summarise(proportion = mean(response == "success"))
}
```

Part (b) - generating random treatment/control vectors

Write a function `randomized_treatment` which takes as its argument `tib`, a tibble in the same format as the previous part of the problem, and returns a tibble where the treatments and controls are assigned randomly. *Hint: the function `sample()` in R should be helpful. Examine what this function does to a vector of categorical variables. What happens when you repeatedly apply this function to the same tibble (e.g., the tibble `df` in the introduction to the problem?)

```
randomized_treatment <- function(tib) {
  tib %>%
    mutate(group = sample(group))
}
```

Part (c) - randomization test

Write a function `randomization_tests` which takes two arguments: `tib`, a tibble, and `n`, an integer with default value of 500. Assume `tib` has the same form as in the previous problem. It returns a named list, with one component, `actual_diff`, which returns the difference in proportion of successes in treatment vs. proportion of successes in control (this difference is positive if there are more successes in treatment). The second component of the named list, `randomized_differences`, is a vector of length `n` which computes the difference in proportions between treatment and control when the treatment and control groups are assigned *randomly*, using the `randomized_treatment()` function from the previous part of the problem.

You may wish to first initialize a vector `differences`, and to use a for loop - see below.

```
randomization_tests <- function(tib, n = 500) {
  set.seed(123) # For reproducibility
  differences <- numeric(n)
```

```
  actual_diff <- compute_proportions(tib) %>%
    summarise(diff = diff(proportion)) %>%
    pull(diff)

  for(i in 1:n) {
    shuffled_tib <- randomized_treatment(tib)
    shuffled_diff <- compute_proportions(shuffled_tib) %>%
      summarise(diff = diff(proportion)) %>%
      pull(diff)
    differences[i] <- shuffled_diff
  }

  list(actual_diff = actual_diff, differences = differences)
}
```

Part (d) - plotting the results

Write a function, `plot_randomization_results`, which takes as its argument a named list `randomization_results` which has the same format as the output of the previous part of the problem (i.e., a component "actual_diff" and a component "differences"), and returns a histogram with binwidth 0.005 of the values of the differences that come from using $n = 1000$ randomization tests. In addition, plot a dashed red line which denotes the actual difference observed in the original data. Make sure the x-axis, y-axis, and title of the plot are descriptive, and that the x-axis and y-axis labels are visible.

```
plot_randomization_results <- function(randomization_results) {
  library(ggplot2)

  ggplot(data.frame(difference = randomization_results$differences), aes(x = difference)) +
    geom_histogram(binwidth = 0.005, fill = "blue", color = "black") +
    geom_vline(xintercept = randomization_results$actual_diff,
               color = "red", linetype = "dashed", linewidth=1) +
    theme_minimal() +
    xlab("Difference in Proportions") +
    ylab("Frequency") +
    ggtitle("Randomization Test Results")
}
```

Part (e) - analysis of how extreme events are

Write a function `percent_more_extreme()`, which takes as its input `randomization_results` (a named list which is the output of the function `randomization_tests` from part (b)) and returns the proportion of the simulated differences-in-proportions in randomizations which have as extreme as a result as the observed difference in the data. For example:

- If the original data had difference in proportions of 0.03, and randomizations had difference in proportions of -0.02, 0.01, 0.04, 0.05, and 0.07, then the function would return $0.6 = 3/5$ since 0.04, 0.05, and 0.07 were more extreme (positive) than the observed difference of 0.03.
- If the original data had difference in proportions of -0.01, and randomizations had difference in proportions of -0.02, 0.01, 0.04, 0.05, and 0.07, then the function would return $0.2 = 1/5$ since only -0.02 was more extreme (negative) than the observed difference of -0.01.

For this problem you can assume that the observed differences in the data is not zero. Be sure that the reader can read every line of your code in the PDF!

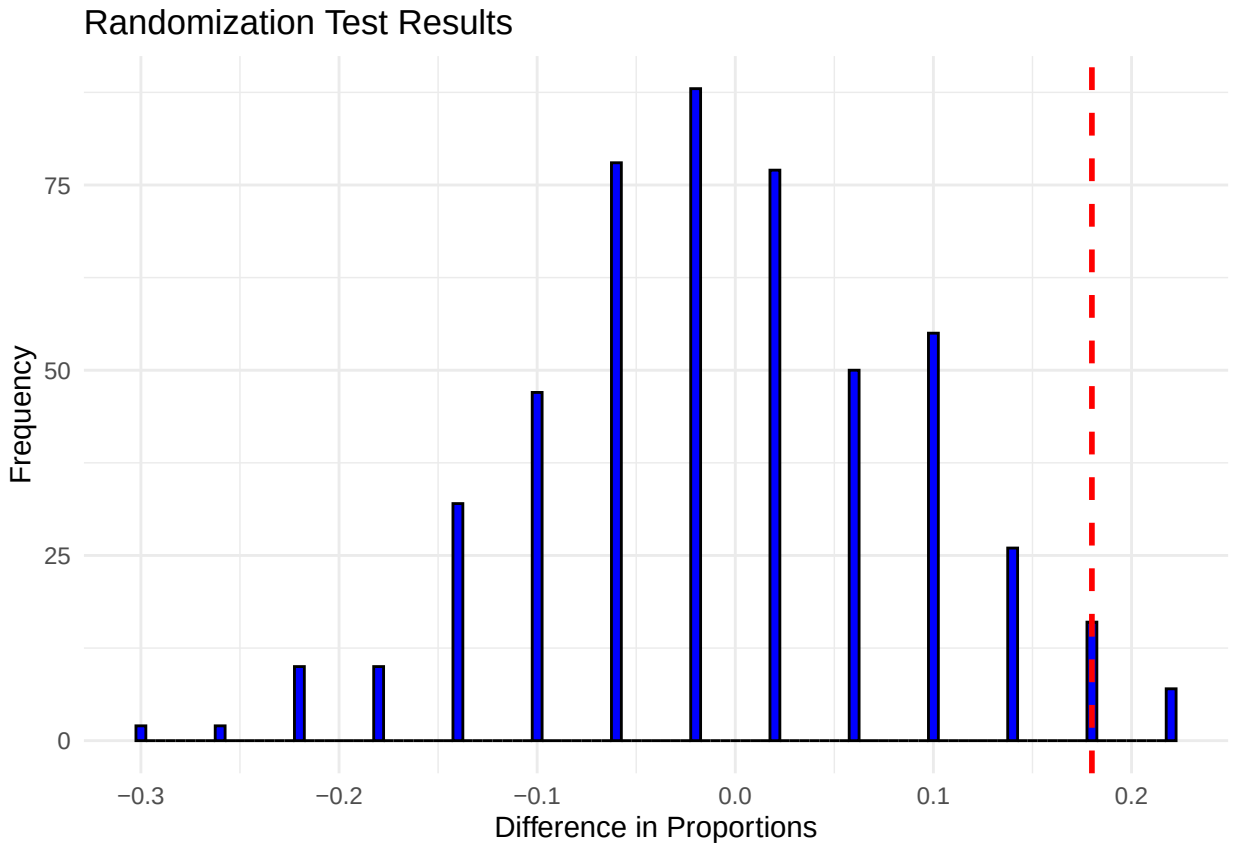
```
percent_more_extreme <- function(randomization_results) {
  differences <- randomization_results$differences
```

```
  actual_diff <- randomization_results$actual_diff
  if(actual_diff < 0) {
    result <- sum(differences < actual_diff) / length(differences)
  }
  if(actual_diff > 0) {
    result <- sum(differences >actual_diff) / length(differences)
  }
  return(result)
}
```

Part (f) - running the analysis and interpreting

Now run the code that you've built up as follows:

```
randomization_results <- randomization_tests(df)
plot_randomization_results(randomization_results)
```




```
percent_more_extreme(randomization_results)
```

```
[1] 0.014
```

Interpret the graph and the results of the function `percent_more_extreme`. How likely was it that the treatment is independent of success/failure?