STA130 Rstudio Homework

Problem Set 5

[Xuanqi Wei] ([1009353209]), with Josh Speagle & Scott Schwartz

Instructions

Complete the exercises in this `.Rmd` file and submit your `.Rmd` and knitted `.pdf` output through Quercus by 11:59 pm E.T. on Thursday, February 16.

```
library(tidyverse)
```

```
student_num = 1009353209
student_num_last3 = 209
```

Question 1: Old Auto Claims, New Statistical Tricks

Let's now look at some data stored in the `auto_claims_population.csv` data set, which includes claims paid (in USD) to a sample of auto insurance claimants 50 years of age and older in a specific year. Let's assume that this dataset defines the entire population of $n = 1000$ payouts that year.

```
autoclaimspop <- read_csv("auto_claims_population.csv")
glimpse(autoclaimspop)
```

```
## Rows: 1,000
## Columns: 5
## $ STATE  <chr> "STATE 15", "STATE 15", "STATE 02", "STATE 15", "STATE 04", "ST~
## $ CLASS  <chr> "F6", "F6", "C11", "C11", "C6", "C11", "C6", "C6", "C1", "C11",~
## $ GENDER <chr> "F", "M", "F", "M", "M", "M", "F", "F", "F", "M", "F", "F", "F"~
## $ AGE    <dbl> 95, 95, 92, 91, 91, 90, 90, 90, 90, 88, 88, 88, 88, 88, 88, 88,~
## $ PAID   <dbl> 2384.67, 650.00, 654.00, 3890.07, 295.99, 11756.34, 2402.00, 29~
```

(a) Create a single random sample of 20 car insurance claims from the population of claims stored in the `auto_claims_population.csv` data set and store these 20 observations in a `tibble` called `ages20`. Compute the `min()`, `mean()`, `median()`, `max()`, and `sd()` of the `AGE` variable using the `summarise()` function and include the `n=n()` summary as well to confirm the overall sample size. Then, create a histogram of the sample.

Hint: Check to ensure that the random sample is taken ***without replacement****.*

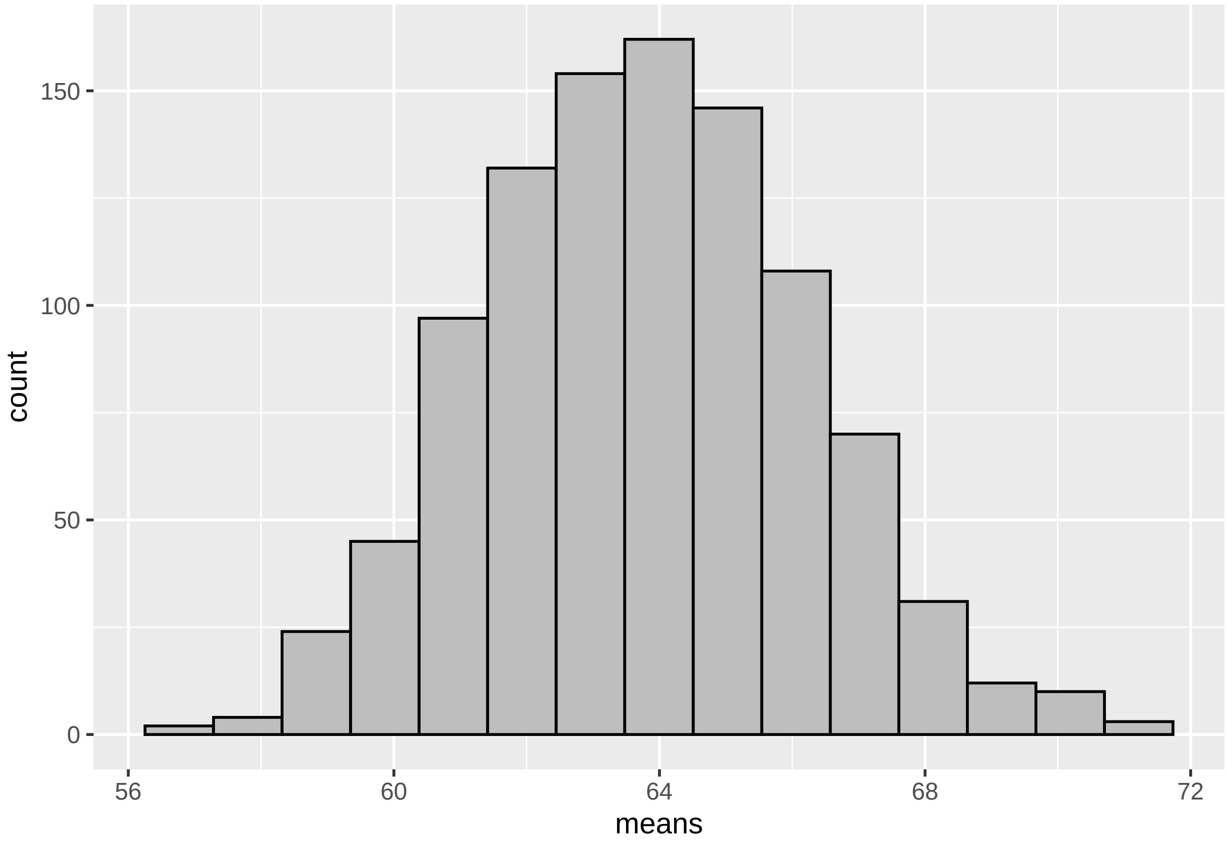
```
set.seed(student_num_last3 + 10)  # REQUIRED so the result is reproducible!
ages20 <- sample_n(autoclaimspop, 20)
ages20 %>% summarise(min=min(AGE), mean=mean(AGE), median=median(AGE), max=max(AGE), sd=sd(AGE), n=n())
```

```
## # A tibble: 1 x 6
##     min  mean median   max    sd     n
##   <dbl> <dbl>  <dbl> <dbl> <dbl> <int>
## 1    51  64.8     68    88  11.2    20
```

(b) Create $m = 1000$ random samples of size $n = 20$ from the population of claims stored in the `auto_claims_population.csv` data set. For each random sample i, compute the sample mean $\bar{x}_i$ of the `age` of the 20 claimants and make a histogram of their values.

*Hint: Check to ensure that each random sample is taken **without replacement**.*

```
set.seed(student_num_last3 + 11)  # REQUIRED so the result is reproducible!
n <- 20
repeatitions <- 1000
stat20 <- rep(NA, repeatitions)
for (i in 1:repeatitions)
{
  ages20 <- sample_n(autoclaimspop, 20)
  mean_age20 <- ages20 %>% summarise(mean=mean(AGE)) %>% as.numeric()
  stat20[i] <- mean_age20
}
sta20_t <- tibble(means=stat20)
sta20_t %>% ggplot(aes(x=means)) +
  geom_histogram(bins=15, colour="black", fill='grey')
```



(c) Calculate the overall mean ($\langle\bar{x}_i\rangle = \text{mean}(\bar{x}_1 + \bar{x}_2 + \cdots + \bar{x}_m)$) and standard deviation ($s_{\bar{x}_i} = \text{sd}(\bar{x}_1 + \bar{x}_2 + \cdots + \bar{x}_m)$) of the sample means.

```
summarise(sta20_t, mean=mean(means), sd=sd(means))
```

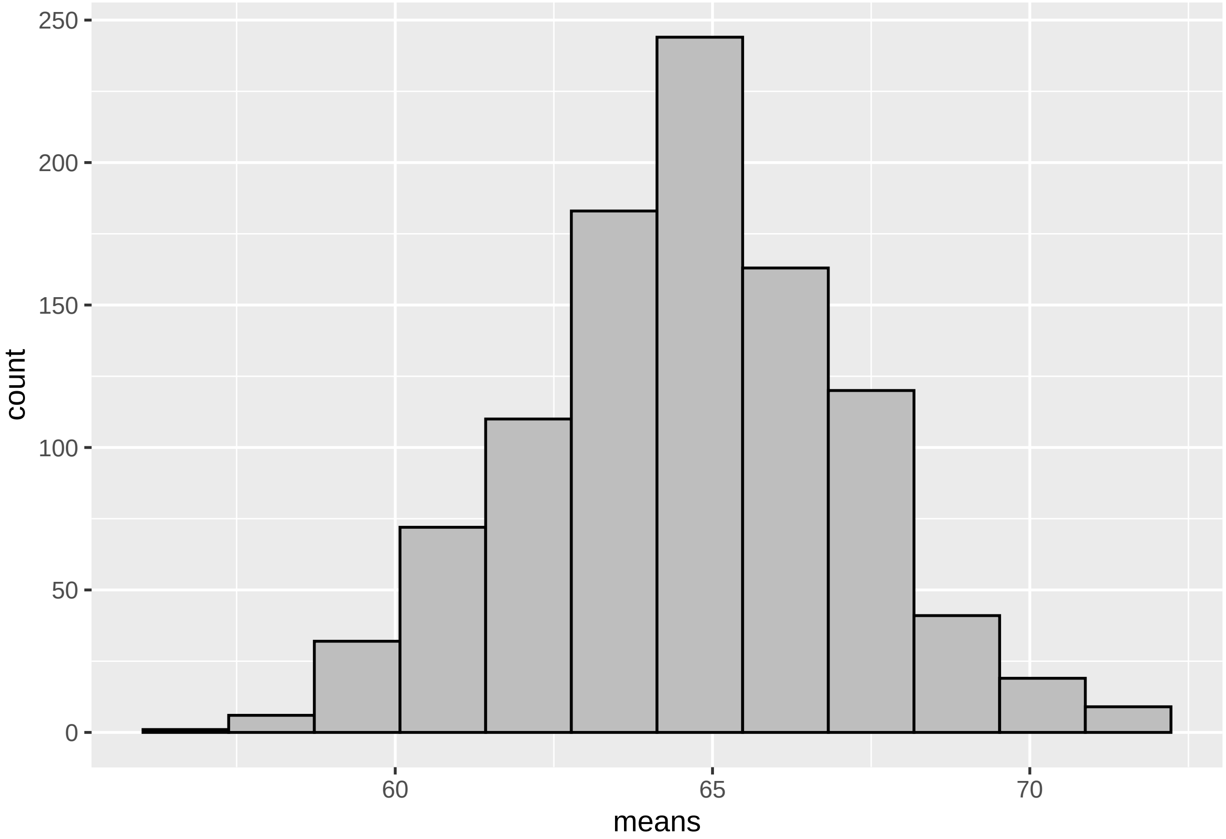
```
## # A tibble: 1 x 2
##    mean    sd
##   <dbl> <dbl>
## 1  63.8  2.41
```

Compare the mean value to the population mean μ from the entire data set of $n = 1000$ observations. Are they similar relative to the standard deviation? Justify your answer in 1-2 sentences.

> *The mean value are similar. As for eht estandard deviation, they are relatively smaller as the latter is much smaller.*

(d) Compute $m = 1000$ bootstrap samples from `ages20`. Then, repeat the analysis above by computing the mean age of claimants $\bar{x}_i$ for each bootstrap sample i, make a histogram of their values, and calculate the overall mean $\langle \bar{x}_i \rangle$ and standard deviation $s_{\bar{x}_i}$ of the bootstrap sample means.

```
set.seed(student_num_last3 + 12)  # REQUIRED so the result is reproducible!
reput <- 1000
boot_means <- rep(NA, reput)
sample_size <- 20
for (i in 1:reput)
{
  boot_stat <- ages20 %>% sample_n(size=sample_size, replace=TRUE)
  boot_means[i] <- boot_stat %>% summarize(mean(AGE)) %>% as.numeric()
}
boot_mean <- tibble(means=boot_means)
boot_mean %>% ggplot(aes(x=means)) + geom_histogram(bins=12, color="black", fill="grey")
```



```
summarise(boot_mean, mean=mean(means), sd=sd(means))
```

```
## # A tibble: 1 x 2
##    mean    sd
##   <dbl> <dbl>
## 1  64.6  2.50
```

Compare their values to the population mean μ and your previously computed values for mean of the sample means $\langle \bar{x}_i \rangle$ and the standard deviation of the sample means $s_{\bar{x}_i}$. Are they similar? Describe your observations in 1-2 sentences.

> *Yes, they are similar. A slight difference occurs on the bootstrap one which is a little greater.*

(e) What distribution do the distributions we simulated in (b) and (d) both estimate? Based on this, should we expect (portions of) our results from (b) and (d) to be similar or different? Why or why not? Please describe your thoughts on these questions in 2-3 sentences.

> *The distribution we simulated in (b) and (d) both estimate the distribution of the sample age. We should expect our results from (b) and (d) to be similar due to they have the same population and just different in the way of simulation.*

Question 2: Driving on the "Right" Side of the Road

In this question, you will explore data about cars drive on the right or left side of the road in different countries. World Standards' list of left driving countries shows that 86 of all 270 countries in the world drive on the left side of the road.

```
roaddata <- tibble(road_side = c(rep("left", 86), rep("right", 270-86)))
glimpse(roaddata)
```

```
## Rows: 270
## Columns: 1
## $ road_side <chr> "left", "left", "left", "left", "left", "left", "left", "lef~
```

(a) Pipe the `roaddata` `tibble` into the `slice_sample(n=100)` function to select a random sample of 100 countries. Call this new data `road_100`.

```
set.seed(student_num_last3 + 20)  # REQUIRED so the result is reproducible!
road_100 <- roaddata %>% slice_sample(n=100)
```

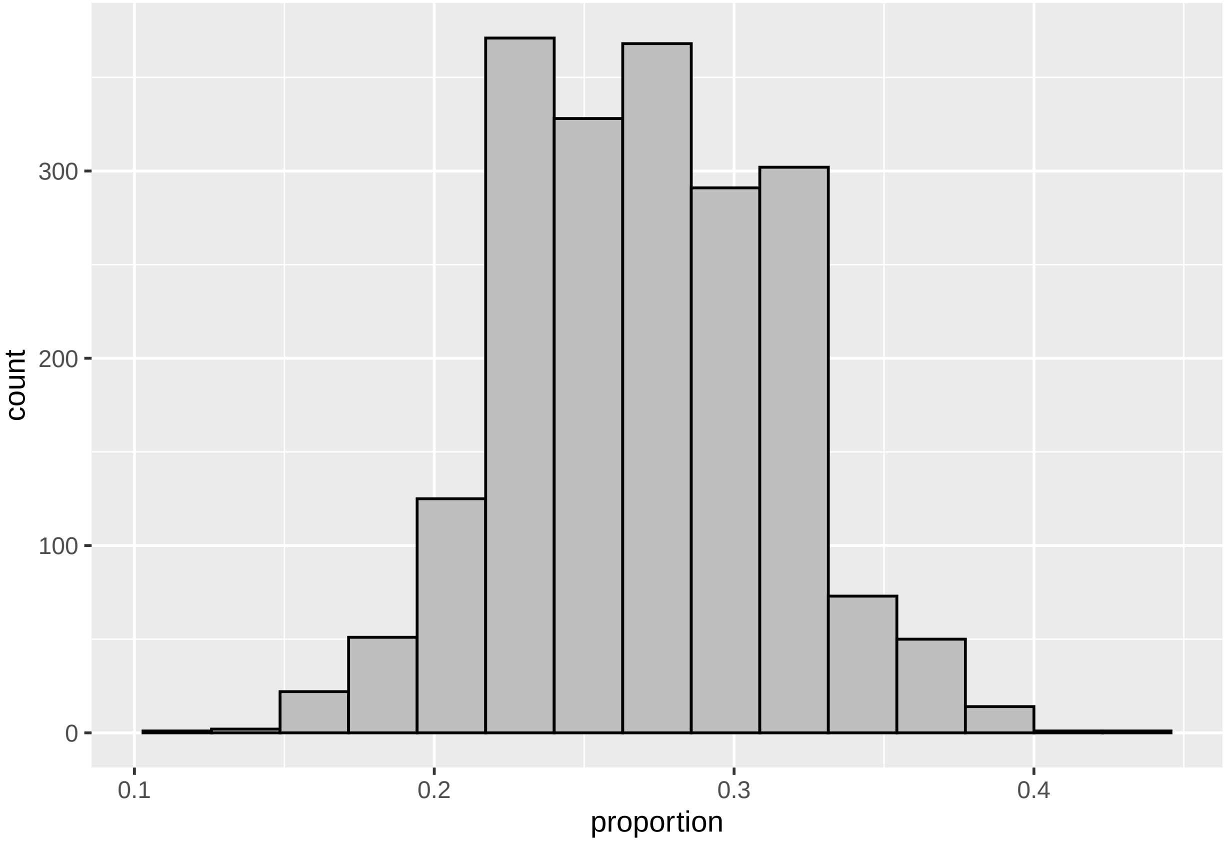
Are the observations in `roaddata` the *entire population*, or a *sample* from a population? How about `road_100`?

> *`roaddata` is the entire population, `road_100` is a sample from population.*

(b) Define f_{left} as the fraction of countries globally where cars drive on the left side of the road (86 out of 270), and $f_{\text{left},100}$ as the equivalent proportion in your random sample of 100 countries. Using `road_100`, simulate $m = 2000$ bootstrap samples and calculate the fraction of countries $\hat{f}_{\text{left},100}$ who drive on the left in each of these bootstrap samples (i.e. your simulated values). Then produce a histogram of $\hat{f}_{\text{left},100}$ to show the bootstrap sampling distribution for $f_{\text{left},100}$ (the fraction of regions that drive on the left side in a random sample of 100 countries).

```
set.seed(student_num_last3 + 21)  # REQUIRED so the result is reproducible!
reputation <- 2000
boot_proportions <- rep(NA, reputation)
sample_size <- 100
```

```
for (i in 1:reputation)
{
  boot_propor <- road_100 %>% sample_n(size = sample_size, replace=TRUE)
  boot_proportions[i] <- boot_propor %>% summarize(prop_left=mean(road_side=="left")) %>% as.numeric()
}
boot_prop <- tibble(proportion=boot_proportions)
boot_prop %>% ggplot(aes(x = proportion)) +
  geom_histogram(bins = 15, colour = "black", fill = "grey")
```



(c) Calculate a 90% bootstrap confidence interval for f_{left}, the true fraction of countries/regions which have cars drive on the left, using the bootstrap sampling distribution you generated in (c).

```
quantile(boot_proportions, c(0.05, 0.95))
```

```
##   5%  95%
## 0.20 0.34
```

(d) Assume for the moment that your 90% bootstrap confidence interval was (0.27, 0.44). Indicate whether or not each of the following statements is a correct interpretation of the confidence interval constructed in part (c) and justify your answers in 1-2 sentences.

Note: Your own confidence interval is almost certainly different from this based on your random seed.

(A) We are 90% confident that between 27% and 44% of countries/regions in our `road_100` sample from (b) drive on the left side.

Correct, due to the definition.

(B) There is a 90% chance that between 27% and 44% of *all* countries in the population drive on the left side.

> *Incorrect. Since the `chance` of something in an interval can only be 100% or 0%, it's wrong saying that `There is a 90% chance that between 27% and 44% of all countries in the population drive on the left side.`*

(C) If we consider many random samples of 100 countries/regions, and we calculate 90% bootstrap confidence intervals for each sample, approximately 90% of these confidence intervals will include the true proportion of countries/regions in the population who drive on the left side of the road.

> *Correct. to make approximately 90% of these confidence intervals including the true proportion of countries/regions in the population who drive on the left side of the road, we make the first half of the part to be the result from the simulation and the second half part to be the aim of the simulation combining provides the correct answer for approximation.*

(e) If we want to be *more* confident about capturing the proportion of all countries who drive on the left side, should we use a *wider* confidence level (e.g., 95%) or a *narrower* confidence level (e.g., 50%)? Justify your answer in 1-3 sentences.

> *If we want to be more confident, it's suggested that we should use wider confidence level because the definition of confident interval. After choosing 95% as the new confidence level, we can ensure that we are 95% confident regarding something which is more confident then saying we are 50% confident regarding something.*

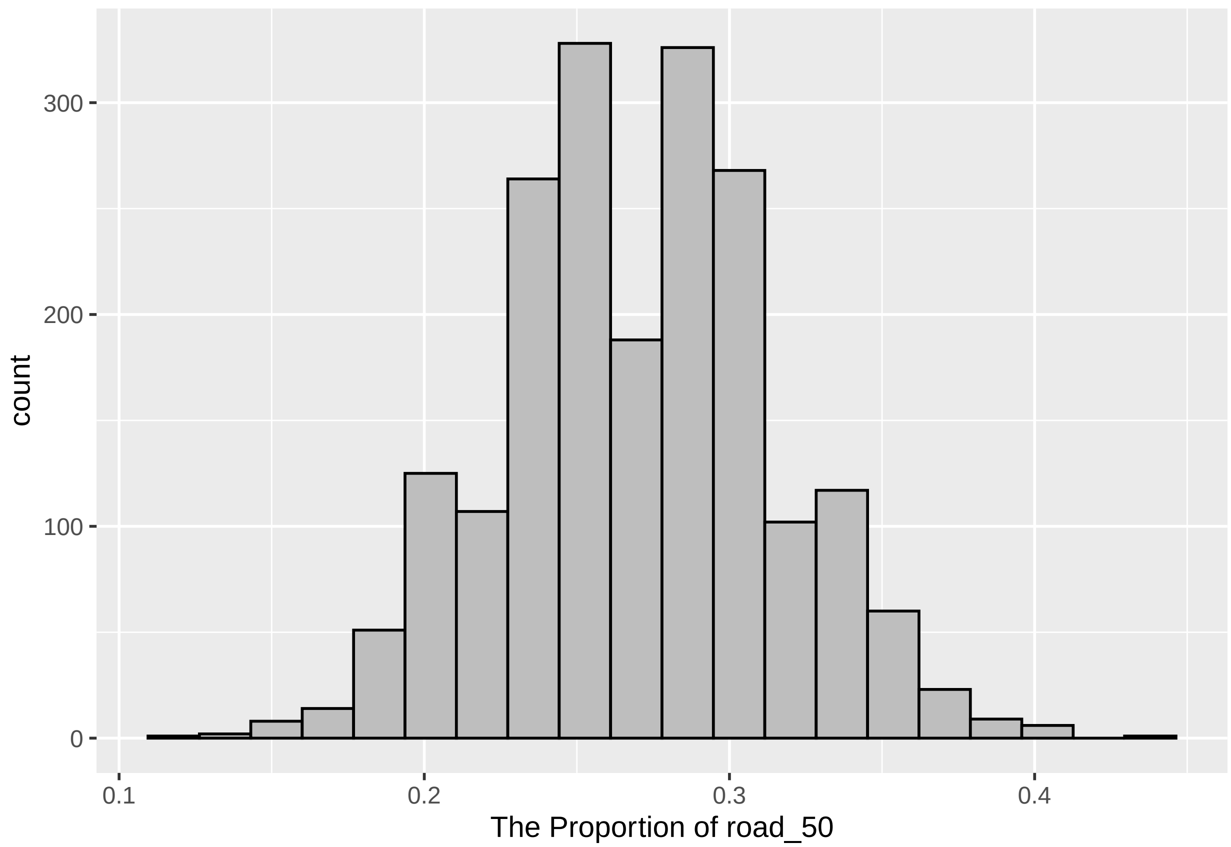
(f) If we increase the size of the sample, do we expect the size of the 90% confidence interval to increase or decrease? Why? Describe your reasoning in 1-3 sentences.

> *If we increase the size of the sample. we do expect the size of the 90% confidence interval to decrease. Becuase the more the data, the more the accuracy of sample distribution, which has narrower spread.*

(g) Define two new samples, `road_50` and `road_150` with $n = 50$ and $n = 150$ observations, respectively, and re-compute the 90% bootstrap confidence interval for each sample.

```r
set.seed(student_num_last3 + 22)  # REQUIRED so the result is reproducible!
road_50 <- roaddata %>% slice_sample(n=50)
road_150 <- roaddata %>% slice_sample(n=150)
```

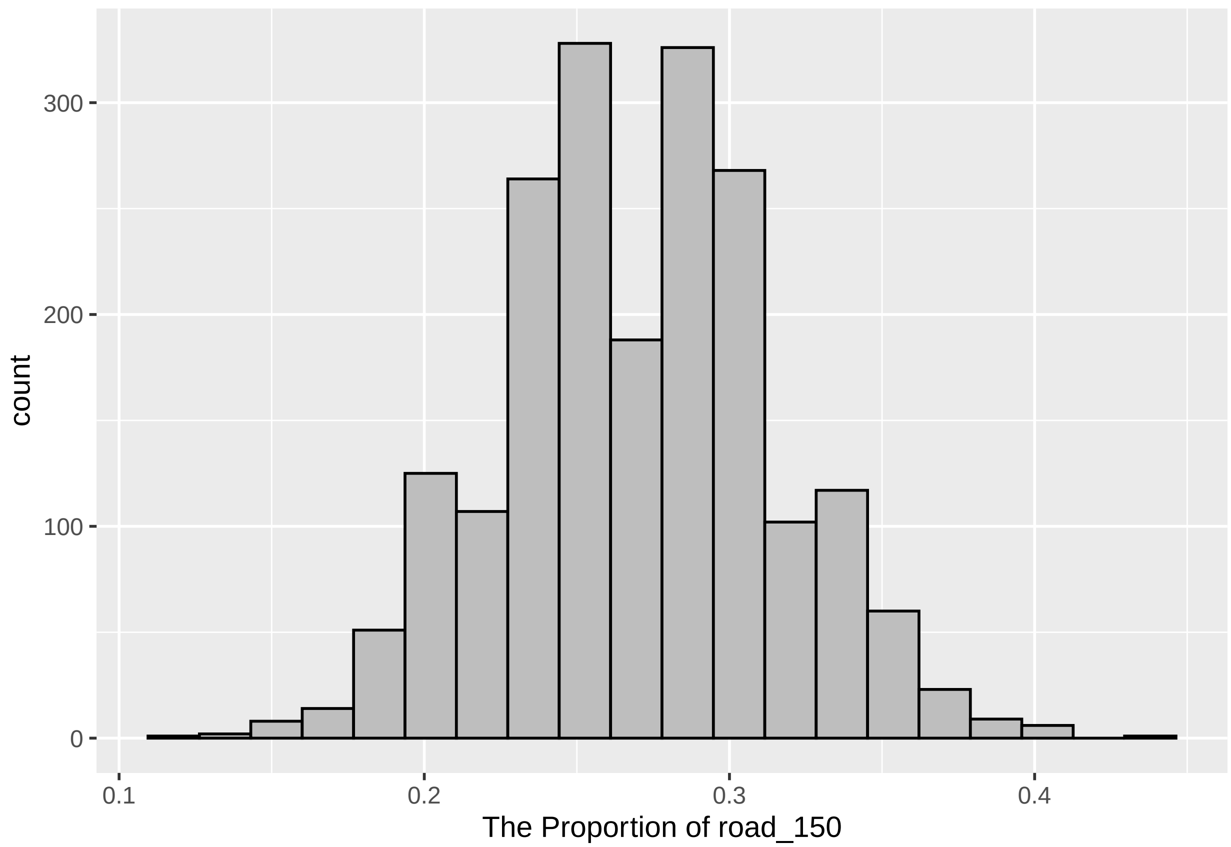
```r
set.seed(student_num_last3 + 23)  # REQUIRED so the result is reproducible!
reputation <- 2000
boot_proportions <- rep(NA, reputation)
sample_size <- 100
for (i in 1:reputation)
{
  boot_propor <- road_50 %>% sample_n(size = sample_size, replace=TRUE)
  boot_proportions[i] <- boot_propor %>% summarize(prop_left=mean(road_side=="left")) %>% as.numeric()
}
boot_prop_50 <- tibble(proportion=boot_proportions)
boot_prop %>% ggplot(aes(x = proportion)) +
geom_histogram(bins = 20, colour = "black", fill = "grey") + labs(x="The Proportion of road_50")
```



```
quantile(boot_proportions, c(0.05, 0.95))
```

```
##   5%  95%
## 0.30 0.46
```

```
reputation <- 2000
boot_proportions <- rep(NA, reputation)
sample_size <- 100
for (i in 1:reputation)
{
  boot_propor <- road_150 %>% sample_n(size = sample_size, replace=TRUE)
  boot_proportions[i] <- boot_propor %>% summarize(prop_left=mean(road_side=="left")) %>% as.numeric()
}
boot_prop_150 <- tibble(proportion=boot_proportions)
boot_prop %>% ggplot(aes(x = proportion)) +
geom_histogram(bins = 20, colour = "black", fill = "grey")+labs(x="The Proportion of road_150")
```



```
quantile(boot_proportions, c(0.05, 0.95))
```

```
##   5%  95%
## 0.29 0.44
```

Did the results match your expectations from (f)? Why or why not?

> *Yes, the result matches the expectation from (f) because CI_150 decrease and CI_50 decrease a lot as well.*

(h) Is there a minimum sample size where bootstrapping would cease to be very effective? How about a maximum sample size? Justify your answers in 1-2 sentences.

> *Yes, there's a minimum and maximum sample size where bootstrapping would cease to be very effective but we need to cover from all types from the population.*

(i) Instead of computing confidence intervals, we could instead carry out a hypothesis test to investigate whether or not countries are equally likely to drive on the right or to the left side of the road. State the null and alternative hypotheses for such a test.

> *Null Hypothesis: The possibility of Deive on the right and the left side of the road are same. Alternative Hypothesis: The possibility of drive on the right and the left side of the road are different.*

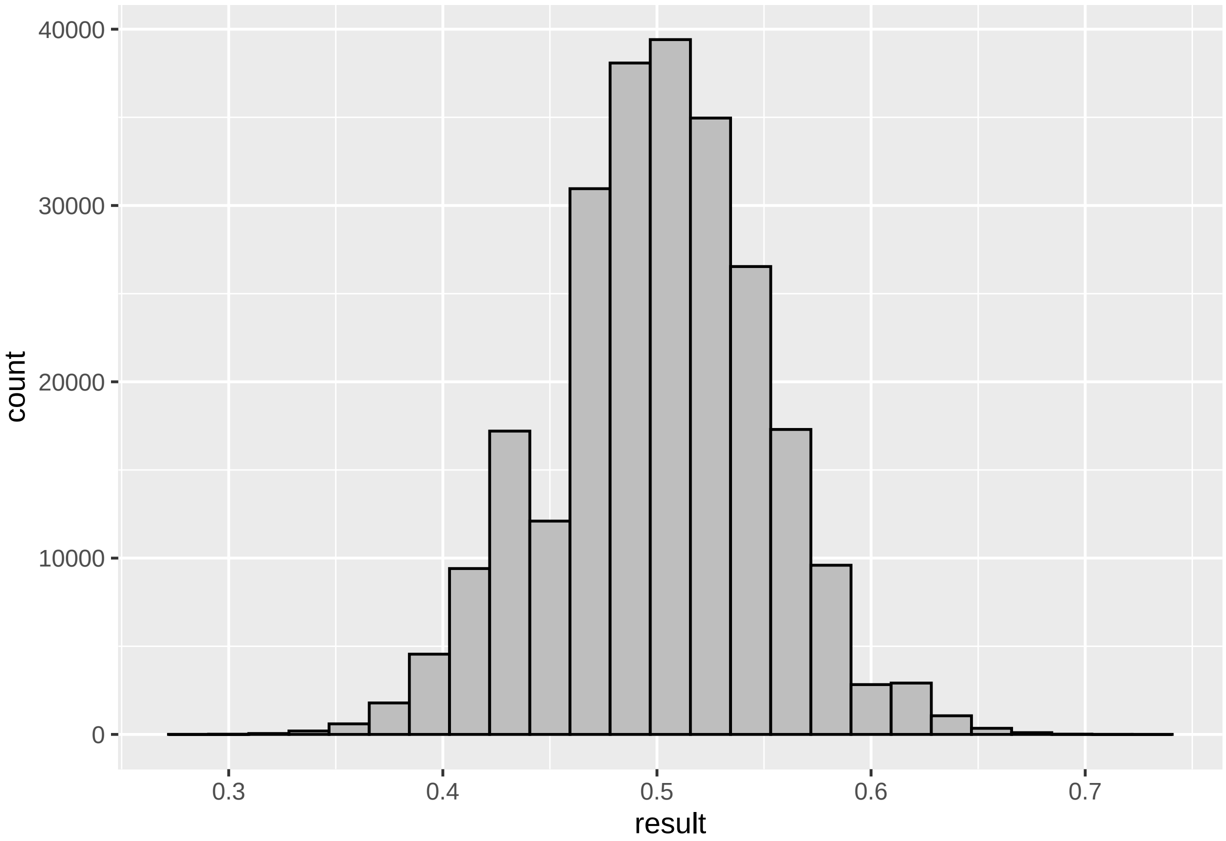
(j) Simulate $m = 250000$ values of $\hat{f}_{\text{left},100}$ you would get under the null hypothesis using direct simulation (i.e. *not bootstrapping*) and plot a histogram to highlight the associated sampling distribution. Then, compute the associated 2-sided p-value.

Important note: We have three potential quantities of interest here: the population parameter f_{left}, the sampled value $f_{\text{left},100}$ involving $n = 100$ countries/regions, and the simulated values $\hat{f}_{\text{left},100}$. Simulated values of $\hat{f}_{\text{left},100}$ give us a sampling distribution for $f_{\text{left},100}$. In general, if the simulated values are computed using bootstrapping we can use them to compute confidence intervals but cannot use them to compute p-values while if the simulated values are computed assuming a particular null hypothesis, then we can use them to compute p-values but cannot use them to compute confidence intervals.

Hint: You may be able to re-use some code from earlier answers as well as some of the code from Problem Set 4 here.

```
set.seed(student_num_last3 + 24)  # REQUIRED so the result is reproducible!
n <- 100
trails <- 250000
p1 = 0.5
responses <- rep(NA, trails)
for (i in 1:trails)
{
response <- sample(c("same","different"),
size=n, prob=c(p1, 1-p1), replace=TRUE)
responses[i] <- mean(response =="same")
}

tibble(result = responses) %>% ggplot(aes(x = result))+ geom_histogram(bins = 25, fill = "grey", colour
```



```
mean <- 86/270
2 * sum(responses > mean) / trails
```

```
## [1] 1.999752
```

(k) Decide whether you would accept or reject the null using a rejection rule based on a significance level of $\alpha = 0.1$. Is this conclusion “consistent” with the 90% confidence interval you computed earlier? Why or why not? Describe your reasoning in 2-3 sentences.

I would reject the null hypothesis. Base on the p value which is 2.0 not in the confidence interval.

Question 3: Gestation Data

In this question we will look at data from Child Health and Development Studies. Our data are adapted from the `Gestation` data set in the `mosaicData` package. Birth weight, date, and gestational period were collected as part of the Child Health and Development Studies in 1961 and 1962 for a sample of 400 mothers who had babies in these two years. Information about the baby’s parents, including age, education, height, weight, and whether the mother smoked, was also recorded.

We will find confidence intervals for parameters related to the distribution of the mother’s age, which for this sample is stored in the variable `age`.

```
gestation <- read_csv("gestation.csv")
glimpse(gestation)
```

```
## Rows: 400
## Columns: 23
## $ id        <dbl> 72, 100, 148, 164, 217, 239, 337, 365, 477, 539, 563, 591, 6~
## $ plurality <dbl> 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, ~
## $ outcome   <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ date      <dbl> 1425, 1673, 1568, 1554, 1605, 1431, 1589, 1464, 1669, 1521, ~
## $ gestation <dbl> 282, 286, 299, 351, 261, 261, 261, 266, 275, 283, 288, 267, ~
## $ sex       <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ wt        <dbl> 108, 136, 143, 140, 115, 144, 103, 114, 119, 134, 134, 122, ~
## $ parity    <dbl> 1, 4, 3, 2, 3, 2, 6, 2, 1, 1, 2, 1, 4, 10, 6, 1, 4, 2, 1, 4,~
## $ race      <dbl> 0, 0, 0, 0, 3, 0, 4, 0, 10, 0, 0, 0, 5, 0, 7, 0, 10, 0, 6, 5~
## $ age       <dbl> 23, 25, 30, 27, 33, 33, 27, 20, 23, 22, 23, 27, 26, 37, 32, ~
## $ ed        <dbl> 5, 2, 5, 5, 2, 2, 1, 2, 2, 2, 2, 4, 1, 1, 2, 5, 4, 4, 2, 2, ~
## $ ht        <dbl> 67, 62, 66, 68, 60, 68, 65, 65, 60, 67, 63, 65, NA, 65, 66, ~
## $ wt.1      <dbl> 125, 93, 136, 120, 125, 170, 112, 175, 105, 130, 92, 101, NA~
## $ drace     <chr> "0", "3", "0", "5", "3", "5", "4", "0", NA, "1", "0", "1", "~
## $ dage      <dbl> 24, 28, 34, 28, 33, 35, 29, 25, 25, 22, 26, 29, 41, 40, 35, ~
## $ ded       <dbl> 5, 2, 5, 4, 2, 4, 1, 2, 2, 2, 1, 5, NA, 2, 2, 5, 4, 2, 2, 5,~
## $ dht       <dbl> NA, 64, NA, NA, 70, NA, NA, NA, 70, 71, 71, 71, 65, 72, 69, ~
## $ dwt       <dbl> NA, 130, NA, NA, 140, NA, NA, NA, 190, 185, 178, 200, 168, 1~
## $ marital   <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ inc       <dbl> 1, 4, 2, NA, 4, 7, 6, 2, 2, 2, 1, 7, NA, 8, 1, 1, 4, 3, 1, 8~
## $ smoke     <dbl> 1, 2, 1, 3, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 3, 0, 1, 1, 0, 1, ~
## $ time      <dbl> 1, 2, 1, 4, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 3, 0, 1, 1, 0, 1, ~
## $ number    <dbl> 5, 2, 4, 2, 5, 0, 5, 2, 0, 0, 5, 1, 5, 5, 1, 0, 1, 5, 0, 3, ~
```

(a) Suppose we are interested in how the mean ages $\bar{a}_{400}$ of random samples of $n = 400$ mothers might vary across all possible samples of 400 mothers we could take from the population. Explain in 1-3 sentences why it is not possible to use these data (i.e., `gestation`) to estimate this like we did in Question 1.

According to Q1(e), all the data is from a sample randomly selected from the population, which is not the same as what this question is listed - data is from a picked data set„ and doens’t pont to any population.

REPLACE THIS TEXT WITH YOUR ANSWER

(b) Using $m = 5000$ bootstrap samples, estimate the 99% confidence interval for the true *mean* age of a mother across the entire population at the time this sample was taken.

```
set.seed(student_num_last3 + 30)  # REQUIRED so the result is reproducible!
reputation <- 5000
age_means <- rep(NA, reputation)
sample_size <- 400
for (i in 1:reputation)
{
  mo_age <- gestation %>% sample_n(size = sample_size, replace=TRUE)
  age_means[i] <- mo_age %>% summarize(mean(age)) %>% as.numeric()
}
age_mean <- tibble(means=age_means)

quantile(age_means, c(0.005, 0.995))
```

```
##     0.5%    99.5%
## 26.72997 28.17750
```

(c) Explain in 2-5 sentences why the interpretation *"There is a 99% chance that the true mean age of a mother across the entire population at the time this sample was taken is between X and Y years."* would be *INCORRECT* and what the correct interpretation should actually be.

> *As mentioned before in Q2(d)B, since the* `chance` *of something in an interval can only be 100% or 0%, it's wrong saying that 'There is a 99% chance that the true mean age of a mother across the entire population at the time this sample was taken is between X and Y years.*

(d) Using $m = 2000$ bootstrap samples, find an 80% bootstrap confidence interval for the true *median* age of a mother across the entire population at the time this sample was taken.

```
set.seed(student_num_last3 + 31)  # REQUIRED so the result is reproducible!
reputation <- 2000
age_medians <- rep(NA, reputation)
sample_size <- 400
for (i in 1:reputation)
{
  mom_age <- gestation %>% sample_n(size = sample_size, replace=TRUE)
  age_medians[i] <- mo_age %>% summarize(median(age)) %>% as.numeric()
}

quantile(age_medians, c(0.1, 0.9))
```

```
## 10% 90%
##  26  26
```

Write 1-2 sentences interpreting this interval.

> *We have 80% confident to say, the true median is in interval [26, 27].*

(e) (OPTIONAL BUT STRONGLY RECOMMENDED) Now it's time to bring everything together! It's time to verify the accuracy of your 80% bootstrap confidence intervals. First, perform the following procedure $k = 100$ times:

1. Create a random sample (i.e. sampling without replacement) containing $n = 100$ observations from the total data set of $n = 400$ observations.

2. Compute the 80% bootstrap confidence interval for the true median age based on that random sample using $m = 1000$ bootstraps.
3. Check if the true median age is within the 80% confidence interval, storing a value in a vector that takes the value `TRUE` if so and `FALSE` if not.

```
# code your answer here
set.seed(student_num_last3 + 32)  # REQUIRED so the result is reproducible!
```

Visualize the resulting outcomes in a barplot below. Then, compute the total fraction of times f_{80} the true median age falls inside the 80% confidence interval.

```
# code your answer here
```

Finally perform a hypothesis test assuming a null hypothesis of $H_0 : f_{80} = 0.8$ and an alternative hypothesis of $H_1 : f_{80} \neq 0.8$ using $m = 1000$ simulated values (for a sample size of $k = 100$).

```
# code your answer here
set.seed(student_num_last3 + 33)  # REQUIRED so the result is reproducible!
```

Based on an $\alpha = 0.05$ significance level, what conclusion should we reach about the accuracy of our confidence intervals?

> *REPLACE THIS TEXT WITH YOUR ANSWER*