STA130H1F – Winter 2020

Week 6 Practice Problems - Sample Answers

N. Moon & L. Bolton [INSERT YOUR NAME HERE]

Instructions

How do I hand in these problems for the February 13 deadline?

Your complete .Rmd file that you create for *QUESTIONS 1 & 2* of these practice problems and the resulting pdf (i.e., the one you 'Knit to PDF' from your .Rmd file) must be uploaded into a Quercus assignment (link: https://q.utoronto.ca/courses/138992/assignments/284430) by 11:59PM, on February 13. Late problem sets or problems submitted another way (e.g., by email) are *not* accepted.

What should I bring to tutorial on February 14?

R output (e.g., output and explanations) for *QUESTIONS 1 & 2* only. You can either bring a hardcopy or bring your laptop with the output.

Tutorial Grading

Tutorial grades will be assigned according to the following marking scheme.

	Mark
Completion of required problems (due on Quercus the day before your tutorial)	1
Attendance for the entire tutorial	1
In-class exercises	4
Total	6

Practice Problems

Question 1

In this question, you will explore data about whether couples observed kissing in an airport tilt their heads to the left or the right. The data is adapted from a real study, Gunturkun (2003) (link: https://www.nature.com/articles/421711a).

This is the opening of the article:

> "I observed kissing couples in public places (international airports, large railway stations, beaches and parks) in the United States, Germany and Turkey. The head-turning behaviour of each couple was recorded for a single kiss, with only the first being counted in instances of multiple kissing. The following criteria had to be met to qualify: lip contact, face-to-face positioning, no hand-held objects (as these might induce a side preference), and an obvious head-turning direction during kissing. Subjects' ages ranged from about 13–70 years.
>
> Of 124 kissing pairs, 80 (64.5%) turned their heads to the right and 44 (35.5%) turned to the left."

Here is a data frame with the data from the kissing study:

```r
# Create a data frame
direction <- c( rep("right", 80), rep("left", 124-80) )
kissdata <- tibble(direction)
```

(a) Are the observations in `kissdata` the entire population or a sample from a population?

The observations in `kissdata` are the 124 couples selected to study. The `kissdata` data frame does not *include data all kissing couples (i.e., a population).*

(b) Simulate 1000 bootstrap samples and calculate the proportion of couples who kiss to the left in each of these bootstrap samples. Produce a histogram of the bootstrap sampling distribution of the proportion of people who kiss to the left. Set the seed as the last *three* digits of your student number.

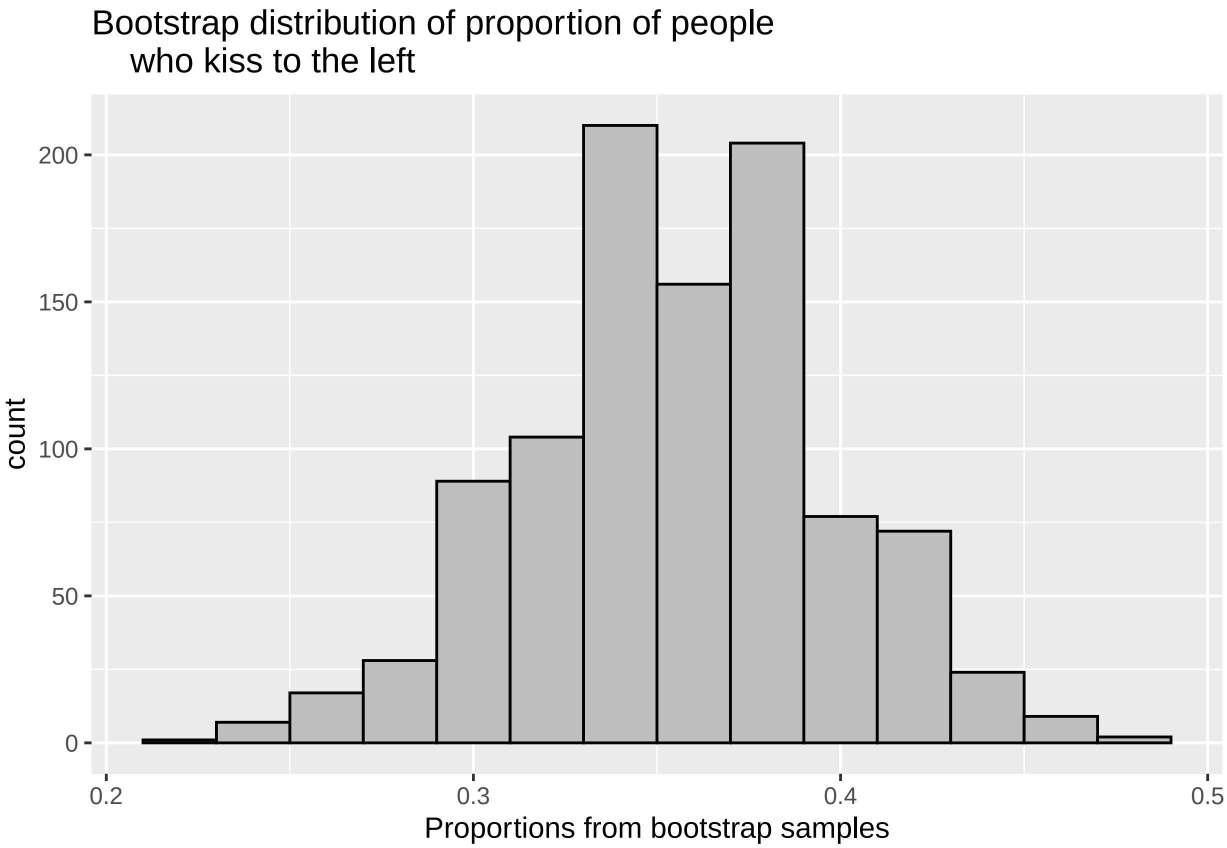
```r
set.seed(333) # change to the last three digits of your student number

boot_p <- rep(NA, 1000)  # where we'll store the bootstrap proportions

for (i in 1:1000)
{
  boot_samp <- kissdata %>% sample_n(size = 124, replace=TRUE)
  boot_p[i] <- as.numeric(boot_samp %>%
                            filter(direction == "left") %>%
                            summarize(n()))/124
}

boot_p <- tibble(boot_p)
```

```
ggplot(boot_p, aes(x=boot_p)) + geom_histogram(binwidth=0.02, fill="gray", color="black") +
  labs(x="Proportions from bootstrap samples",
    title="Bootstrap distribution of proportion of people
    who kiss to the left")
```



(c) Calculate a 95% confidence interval for the proportion of people who kiss to the left based on the Bootstrap distribution you generated in (b).

```
quantile(boot_p$boot_p, c(0.025, 0.975))
```

```
##      2.5%     97.5%
## 0.2739919 0.4354839
```

Based on the above, we see that the confidence interval is (0.27, 0.44).

(d) 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.

(i) We are 95% confident that between 27% and 44% of kissing couples in this sample tilt their head to the left when they kiss.

Incorrect. We know in this sample that 35.5% of kissing couples tilt their head to the left ($\hat{p} = 0.355$) *so we're 100% sure that* $\hat{p}$ *is between 0.27 and 0.44.*

(ii) There is a 95% chance that between 27% and 44% of all kissing couples in the population tilt their head to the left when they kiss.

Incorrect.This specific interval either does or doesn't include p, the proportion of all couples who tilt their heads to the left when kissing (i.e., there is a 0% chance or a 100% chance - we just don't know which). We cannot conclude we are 95% sure (nor that there is a 95% chance or 0.95 probability) that p is in this particular interval. Rather, we are confident about the method used to produce this interval: for 95% of the possible samples of n=124 couples we could end up with, intervals constructed this way will include p

(iii) We are 95% confident that between 27% and 44% of all kissing couples in the population tilt their head to the left when they kiss.

Correct. Someone reading this interpretation of a confidence interval would have to know what "95% confident" means though! This does not *mean we are 95% sure (nor that there is a 95% chance or 0.95 probability) that p is in this interval. Rather, we are confident about the method used to produce this interval: for 95% of the possible samples of n=124 couples we could end up with, intervals constructed this way will include p.*

(iv) If we considered many random samples of 124 couples, and we calculated 95% confidence intervals for each sample, 95% of these confidence intervals will include the true proportion of kissing couples in the population who tilt their heads to the left when they kiss.

Correct. Although this doesn't report the confidence interval computed, this is a valid description about how confidence intervals behave.

(e) If we want to be *more* confident about capturing the proportion of all couples who tilt their heads to the left when kissing, should we use a *wider* confidence level or a *narrower* confidence level? Explain your answer.

A wider confidence interval would capture the population parameter in more samples. We would get a wider confidence interval if we set a higher confidence level (e.g., instead of 95%, use 98%). We would be extending to more of the bootstrap sampling distribution.

(f) We could carry out an hypothesis test to investigate whether or not couples are equally likely to tilt their heads to the right or to the left when they kiss. Our hypotheses would be:

$$H_0 : p = 0.5$$

versus

$$H_A : p \neq 0.5$$

where p is the proportion of couples who tilt their heads to the left when they kiss. Using Gunturkun's data, we would get a P-value of 0.003. Do this hypothesis test and the confidence interval you produced in (c) tell a similar story? Why or why not?

From the hypothesis test, we have strong evidence against the hypothesis that the proportion of couples who tilt their head to the left is 0.5 because our p-value is so small (p-value=0.003). Our 95% confidence interval for this proportion is (0.27, 0.44) so does not include 0.5. This suggests that p=0.5 is not a plausible value of the proportion of all couples who tilt their heads to the left when kissing based on these data. So the hypothesis test and confidence interval both suggest that couples are not *equally likely to tilt their heads to the left or right when they kiss.*

Question 2

A few weeks ago in lecture, we considered data on car insurance claims paid by an insurer over a certain time period in data set `AutoClaims`. Assume the data set `auto_claims_population.csv` includes *ALL* claims paid (in USD) to claimants 50 years of age and older in a specific year. In other words, it represents a 'population' of car insurance claims in that year.

(a) Produce appropriate data summaries of paid claims (`PAID`) and comment the shape, centre and spread of this distribution.

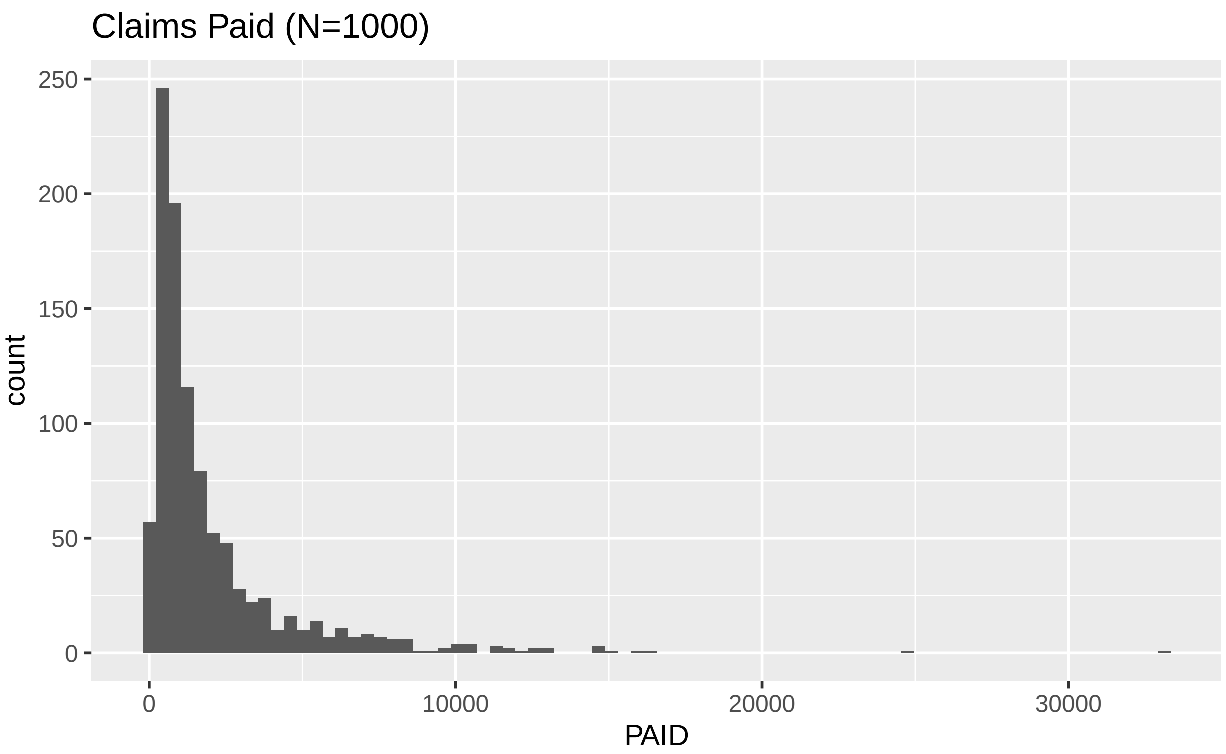
```
AutoClaimsPop <- read_csv("auto_claims_population.csv")
glimpse(AutoClaimsPop)
```

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

```
summarise(AutoClaimsPop,
          min=min(PAID),
          mean = mean(PAID),
          median = median(PAID),
          max=max(PAID),
          sd = sd(PAID),
          n=n())
```

```
## # A tibble: 1 x 6
##     min  mean median    max    sd     n
##   <dbl> <dbl>  <dbl>  <dbl> <dbl> <int>
## 1   9.5 2018.  1049. 33138. 2729.  1000
```

```
ggplot(AutoClaimsPop, aes(x=PAID)) +
  geom_histogram(bins=80) +
  labs(title="Claims Paid (N=1000)")
```

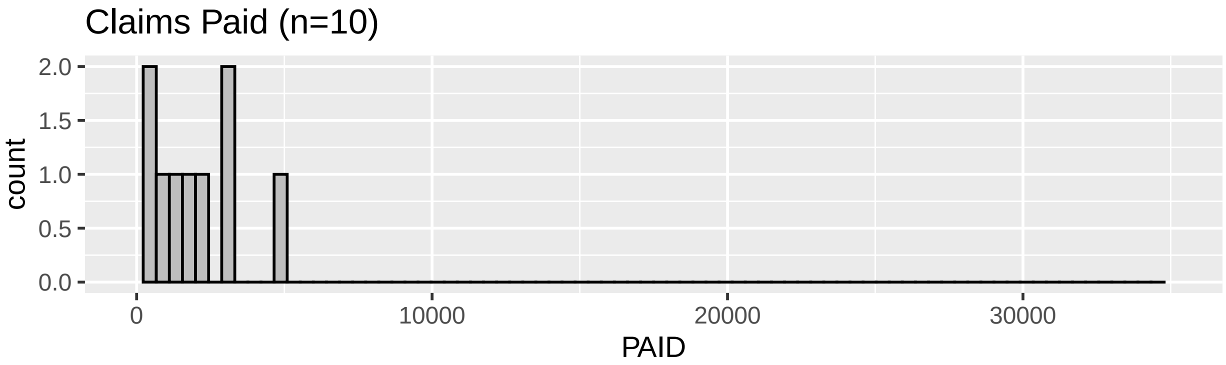


The distribution of claims paid to all claimants aged 50 and over during this year follows a right-skewed distribution with high outliers (i.e., a few very expensive claims and half the claims were less than $\$1,049$*). The mean claim was* $\$2,018$ *was almost twice as large as the median claim, of* $\$1,049$*. Overall, the claims ranged from* $\$9.50$ *US to* $\$33,137.5$ *US with a standard deviation of* $\$2,729.36$ *US.*

(b)

(i) Select a random sample of size n=10 from the population in (a) and produce appropriate data summaries of the paid claims in this sample. Set the seed as the last *three* digits of your student ID number.

```r
set.seed(420)
sample10 <- tibble(PAID=sample(AutoClaimsPop$PAID,10))
ggplot(sample10, aes(x=PAID)) +
  geom_histogram(bins=80, color="black", fill="gray") +
  xlim(0,35000) +
  labs(title="Claims Paid (n=10)")
```

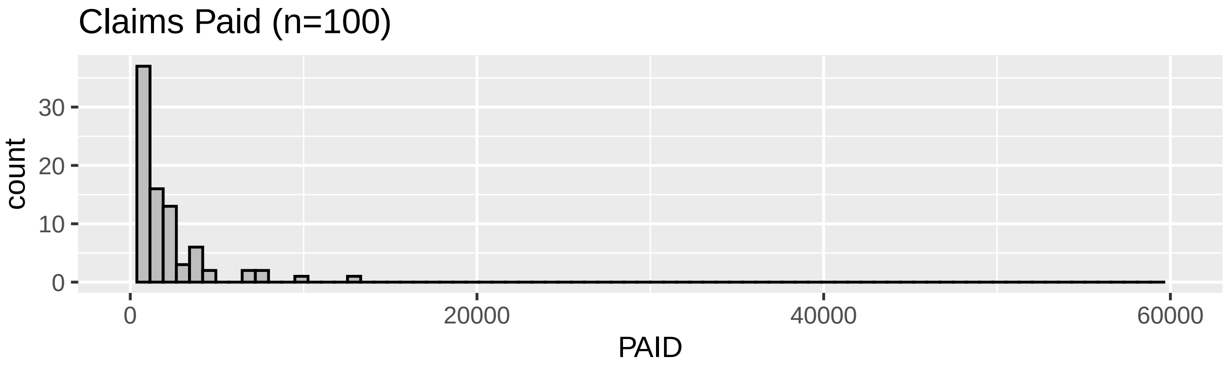


```
summarise(sample10,
          min=min(PAID),
          mean = mean(PAID),
          median = median(PAID),
          max=max(PAID),
          sd = sd(PAID),
          n=n())
```

```
## # A tibble: 1 x 6
##     min  mean median   max    sd     n
##   <dbl> <dbl>  <dbl> <dbl> <dbl> <int>
## 1  143. 1777.  1542. 4930. 1591.    10
```

(ii) Select a random sample of size n=100 from the population in (a) and produce appropriate data summaries of the paid claims in this sample. Again, set the seed as the last four digits of your student number.

```
set.seed(420)
sample100 <- tibble(PAID=sample(AutoClaimsPop$PAID,100))
ggplot(sample100, aes(x=PAID)) +
  geom_histogram(bins=80, color="black", fill="gray") +
  xlim(0,60000) +
  labs(title="Claims Paid (n=100)")
```



```
summarise(sample100,
          min=min(PAID),
          mean = mean(PAID),
          median = median(PAID),
          max=max(PAID),
          sd = sd(PAID),
          n=n())
```

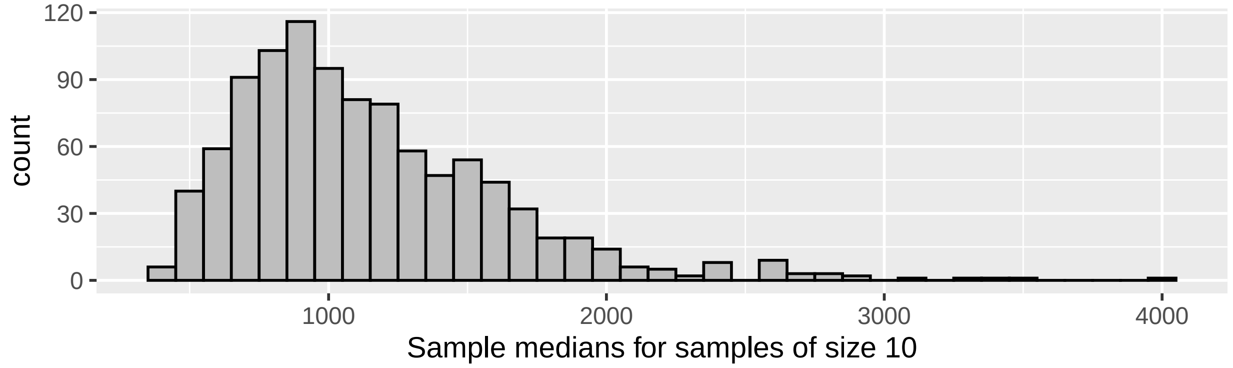
```
## # A tibble: 1 x 6
##     min  mean median    max    sd     n
##   <dbl> <dbl>  <dbl>  <dbl> <dbl> <int>
## 1   125 1778.   990. 12582. 2068.   100
```

(iii) How do the distributions in b(i) and b(ii) compare to the distribution in (a)? Which one of b(i) and b(ii) resembles the distribution in (a) more? Why is this so?

Both samples have right-skewed distributions for paid claims which resembles the distribution of all claims paid to drivers age 50 or over that year, but the summary statistics for the samples differ from the paid claims for the population. The distribution of paid claims (and associated summaries) for the sample with n=100 more closely resembles the distribution in (a) than the distribution of paid claims for the sample with n=10. This is not surprising because we would expect the larger random sample to be more representative of the population.

(c) Estimate the sampling distributions of sample *median* of paid claims by taking 1000 samples of (i) size n=10 and (ii) size n=100 from the distribution in (a) and produce appropriate data summaries. Set the seed as the last four digits of your student number for each set of simulations.

```
## (i)
n <- 10
repetitions <- 1000
set.seed(1460)
sim10 <- rep(NA, repetitions)
for (i in 1:repetitions)
{
  new_sim <- sample(AutoClaimsPop$PAID,size = n)
  sim_median <- median(new_sim)
  sim10[i] <- sim_median
}
sim10 <- tibble(median = sim10)
sim10 %>% ggplot(aes(x = median)) +
  geom_histogram(binwidth = 100, colour = "black", fill = "grey") +
  labs(x="Sample medians for samples of size 10")
```



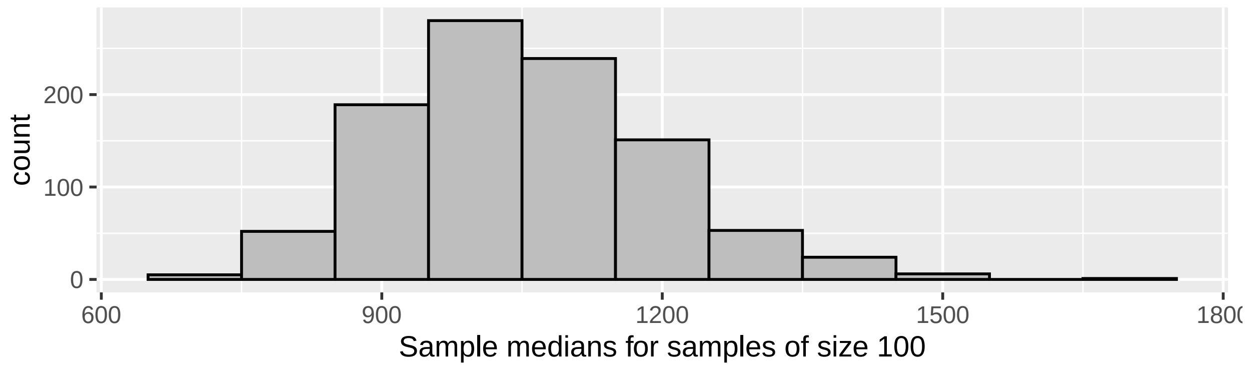
```
summary(sim10$median)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   400.5   810.0  1036.1  1144.4  1391.4  3997.1
```

```
sd(sim10$median)
```

```
## [1] 478.5171
```

```
## (ii)
n <- 100
repetitions <- 1000
sim100 <- rep(NA, repetitions)
set.seed(1460)
for (i in 1:repetitions)
{
  new_sim <- sample(AutoClaimsPop$PAID,size = n)
  sim_median <- median(new_sim)
  sim100[i] <- sim_median
}
sim100 <- tibble(median = sim100)
sim100 %>% ggplot(aes(x = median)) +
  geom_histogram(binwidth = 100, colour = "black", fill = "grey") +
  labs(x="Sample medians for samples of size 100")
```



```
summary(sim100$median)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   711.8   950.6  1043.3  1053.3  1135.8  1698.3
```

```
sd(sim100$median)
```

```
## [1] 140.1926
```

(iii) How do these two estimated sampling distributions (i.e., the one for the sample medians when the sample size is 10 versus the one for sample medians when the sample size is 100) compare?

They are both right-skewed, although the one for medians of random samples of size 100 is not as right-skewed as the one for medians of random samples of size 10. Although the medians of each of the sampling distributions are relatively close, the distribution of medians based on larger random samples (i.e., n=100) has much less variation than the distribution of means based on smaller random samples (i.e., n=10). In particular, the sample medians in the sampling distribution based on smaller samples (n=10) range from approximately $329 *to* $4135*, while the sample medians in the sample distribution for larger samples (n=100) range from* $715 *to* $1533*.*

(d) Explain how and why the distributions you estimated in part (c) are different from the distributions you estimated in part (b) above.

The estimated distributions in (b) and (c) are centred at approximately the same USD as the distribution of paid claims for the population; however the spreads and shapes differ. The distributions in (b) are data from samples which estimate of the distribution of paid claims in the population. The estimated sampling distributions of the sample medians in (c) are much less spread out (i.e., smaller standard deviation) than the distributions of paid claims in the population (and their estimates in part (b)) and the estimated sampling distribution of sample medians in (c) are not as right-skewed as the distribution of paid claims in the population, and its estimates in part (b).

These are estimates of different distributions. The distributions in part (b) are estimates of the distribution of paid claims *in the population of all claims paid to drivers aged 50 and over that year. These estimates were obtained based on the paid claims for one random sample from the population. In contrast, the distributions in part (c) are estimates of* sampling distributions of the sample median paid claims*. 1000 random samples, each of size n (n=10 or 100), were drawn from the population and the median paid claims were computed for each of these samples to estimate these two distributions of median claims based on the two different sample sizes.*

Question 3

Consider the car insurance claims paid by an insurer over a certain time period in data set from Question 2 again. Assume this data set includes *ALL* claims paid (in USD) to claimants 50 years of age and older in a specific year. In other words, it represents a 'population' of car insurance claims in that year.

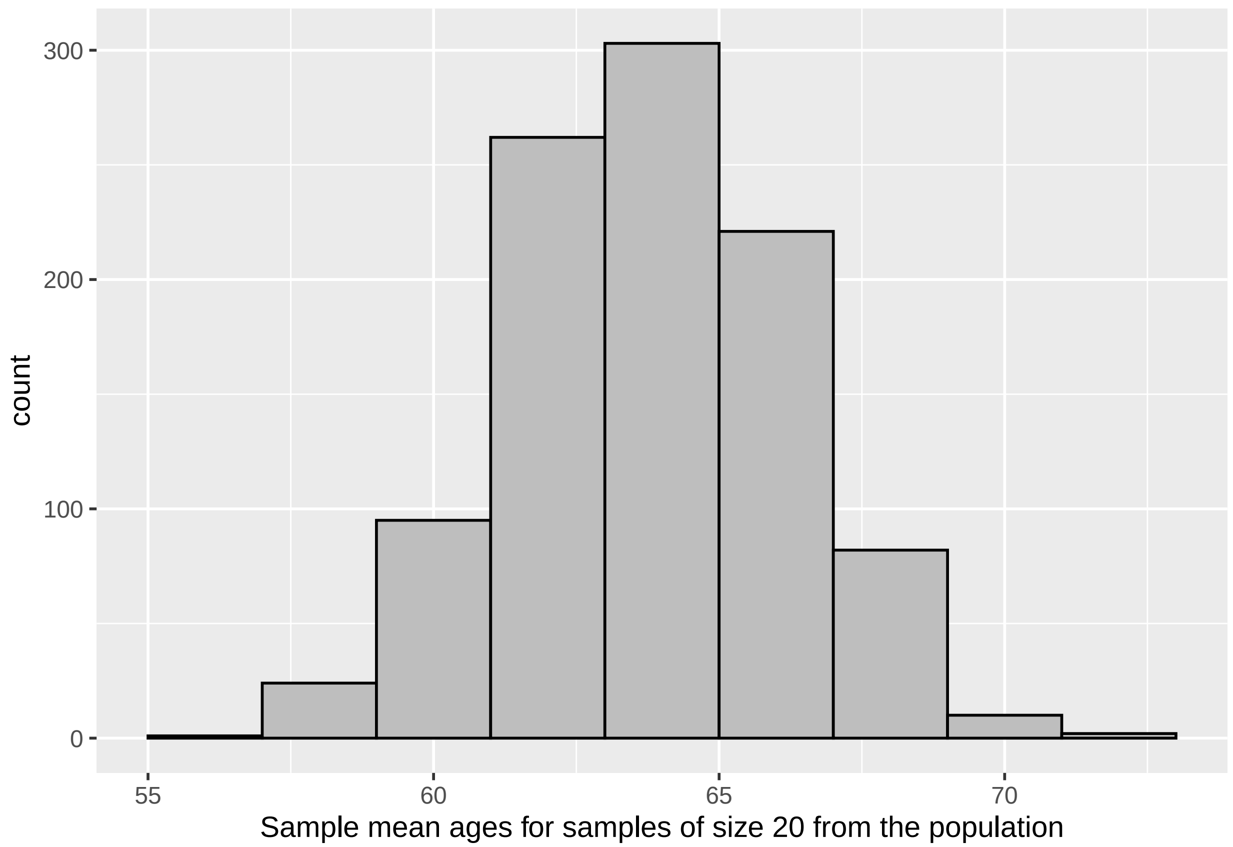
(a) Select 1000 samples of size 20 from the population of claims stored in the `auto_claims_population.csv` data set (each sample is taken without replacement, so there are no repeated observations within each sample). Compute the mean age of claimants for each sample and produce appropriate summaries of the simulated sample means. Set the seed to the last *three* digits of your student number.

```r
AutoClaimsPop <- read_csv("auto_claims_population.csv")

set.seed(246)
n <- 20
repetitions <- 1000
sim20 <- rep(NA, repetitions)

for (i in 1:repetitions)
{
  new_sim <- AutoClaimsPop %>% sample_n(size=20, replace=FALSE)
  sim_mean <- new_sim %>%
    summarize(mean(AGE)) %>%
    as.numeric()

  sim20[i] <- sim_mean
}
sim20 <- tibble(means = sim20)
sim20 %>% ggplot(aes(x = means)) +
  geom_histogram(binwidth = 2, colour = "black", fill = "grey") +
  labs(x="Sample mean ages for samples of size 20 from the population")
```



```
summarise(sim20,
          min=min(means),
          mean = mean(means),
          median = median(means),
          max=max(means),
          sd = sd(means),
          n=n())
```

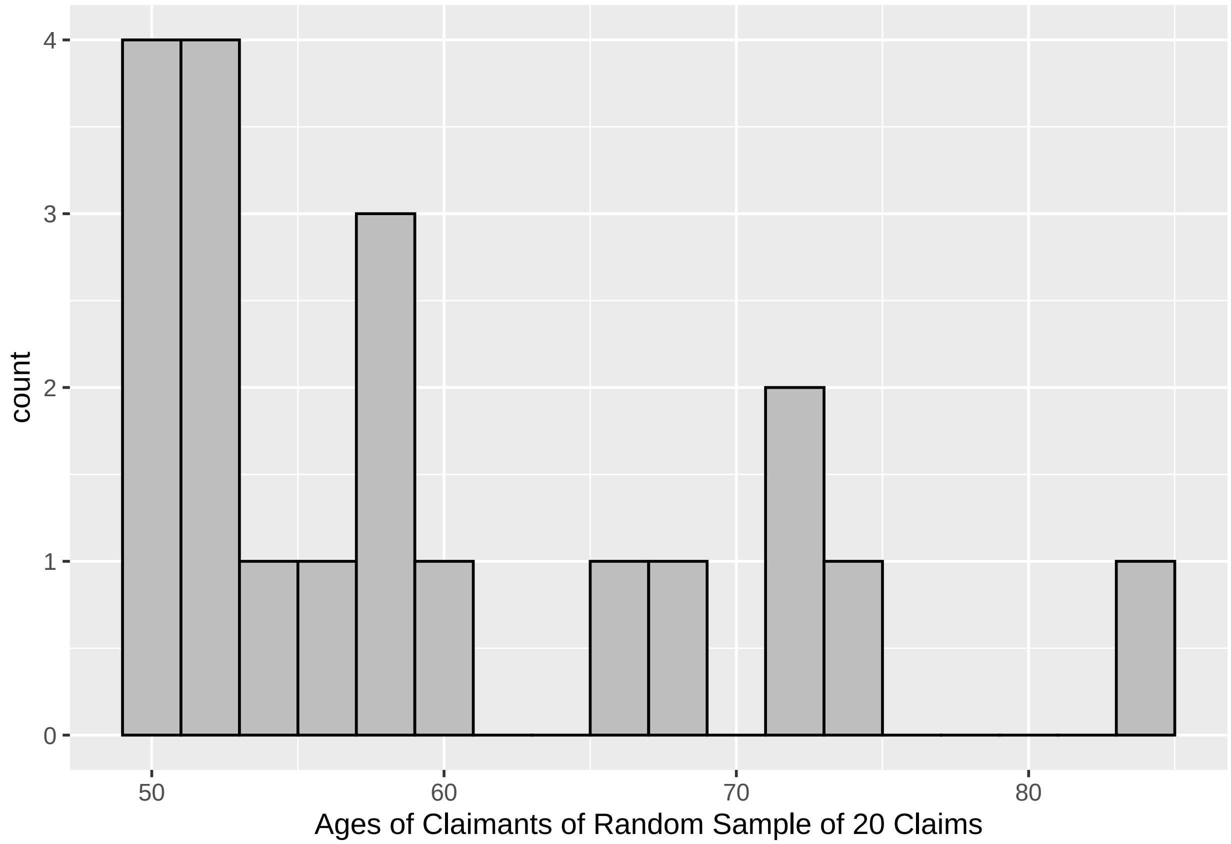
```
## # A tibble: 1 x 6
##     min  mean median   max    sd     n
##   <dbl> <dbl>  <dbl> <dbl> <dbl> <int>
## 1  56.4  63.8   63.8  71.8  2.39  1000
```

(b) Now suppose we only had data for ONE random sample of 20 car insurance claims, and that these 20 observations are stored in ages20.

```
set.seed(321)
ages20 <- tibble(age=sample(AutoClaimsPop$AGE,size = 20, replace=FALSE))
glimpse(ages20)
```

```
## Observations: 20
## Variables: 1
## $ age <dbl> 75, 72, 68, 52, 50, 59, 52, 57, 53, 73, 84, 51, 51, 51, 61, 52,...
```

```
ages20 %>% ggplot(aes(x = age)) +
  geom_histogram(binwidth = 2, colour = "black", fill = "grey") +
  labs(x="Ages of Claimants of Random Sample of 20 Claims")
```

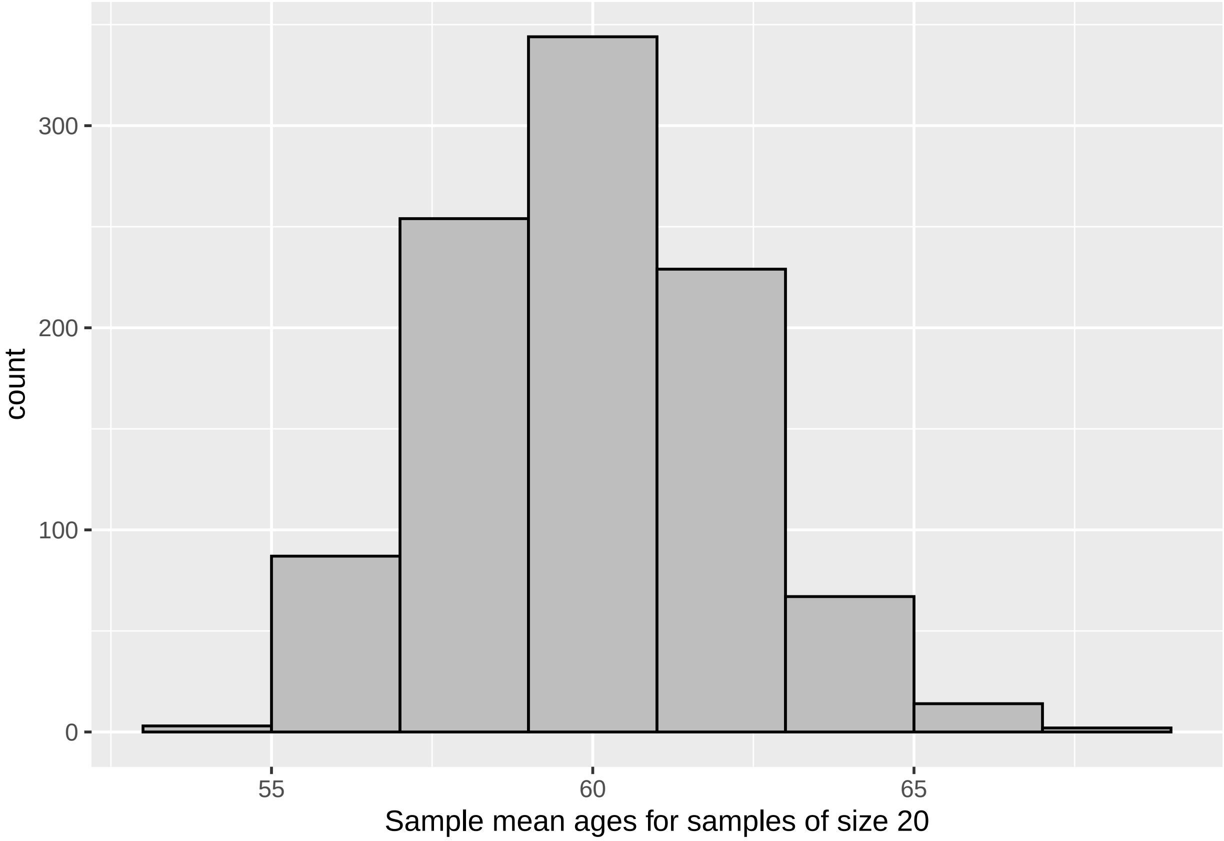


```
summarise(ages20,
          min=min(age),
          mean = mean(age),
          median = median(age),
          max=max(age),
          sd = sd(age),
          n=n())
```

```
## # A tibble: 1 x 6
##     min  mean median   max    sd     n
##   <dbl> <dbl>  <dbl> <dbl> <dbl> <int>
## 1    50  60.0   57.5    84  9.88    20
```

Use R to take 1000 bootstrap samples from the ages of the claimants of the claims sampled and stored in `ages20`. Compute the mean age of claimants for each bootstrap sample of claims and produce appropriate summaries of the bootstrap sample means. Set the seed to the last *three* digits of your student number.

```
set.seed(246)
boot_means <- rep(NA, 1000)  # where we'll store the bootstrap means
sample_size <- 20
for (i in 1:1000)
{
  boot_samp <- ages20 %>% sample_n(size = sample_size, replace=TRUE)
  boot_means[i] <- as.numeric(boot_samp %>% summarize(mean(age)))
}
boot_means <- tibble(means=boot_means)
boot_means %>% ggplot(aes(x = means)) +
  geom_histogram(binwidth = 2, colour = "black", fill = "grey") +
  labs(x="Sample mean ages for samples of size 20")
```



```
summarise(boot_means,
          min=min(means),
          mean = mean(means),
          median = median(means),
          max=max(means),
          sd = sd(means),
          n=n())
```

```
## # A tibble: 1 x 6
##     min  mean median   max    sd     n
##   <dbl> <dbl>  <dbl> <dbl> <dbl> <int>
## 1  53.8  60.0   60.0    68  2.19  1000
```

(c) What distribution do the distributions you simulated in (a) and (b) both estimate? Comment on the similarities and differences in the estimates you obtained.

Both distributions estimate the same sampling distribution - the sampling distribution of the sample mean age of claimants based on a random samples of 20 claims. *So, it is not surprising that both estimated distributions are similar in terms of shape, centre and spread. They are both approximately symmetric and unimodal, their means are relatively close (63.8 and 60 years respectively) and their standard deviations just differ a little bit (2.39 vs 2.19 years).*

The estimate of the sampling distribution in (a) was obtained by sampling directly from the population; whereas the estimate of the sampling distribution in (b) was obtained by resampling (i.e., taking bootstrap samples) from a specific random sample of 20 claims. If the sample is not representative of the population of claims, then the estimate of the sampling distribution based on bootstrap samples from that non-representative sample will not reflect the sampling distribution of mean ages very well.

Question 4

In this question we will look at data from the 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 — age, education, height, weight, and whether the mother smoked was also recorded.

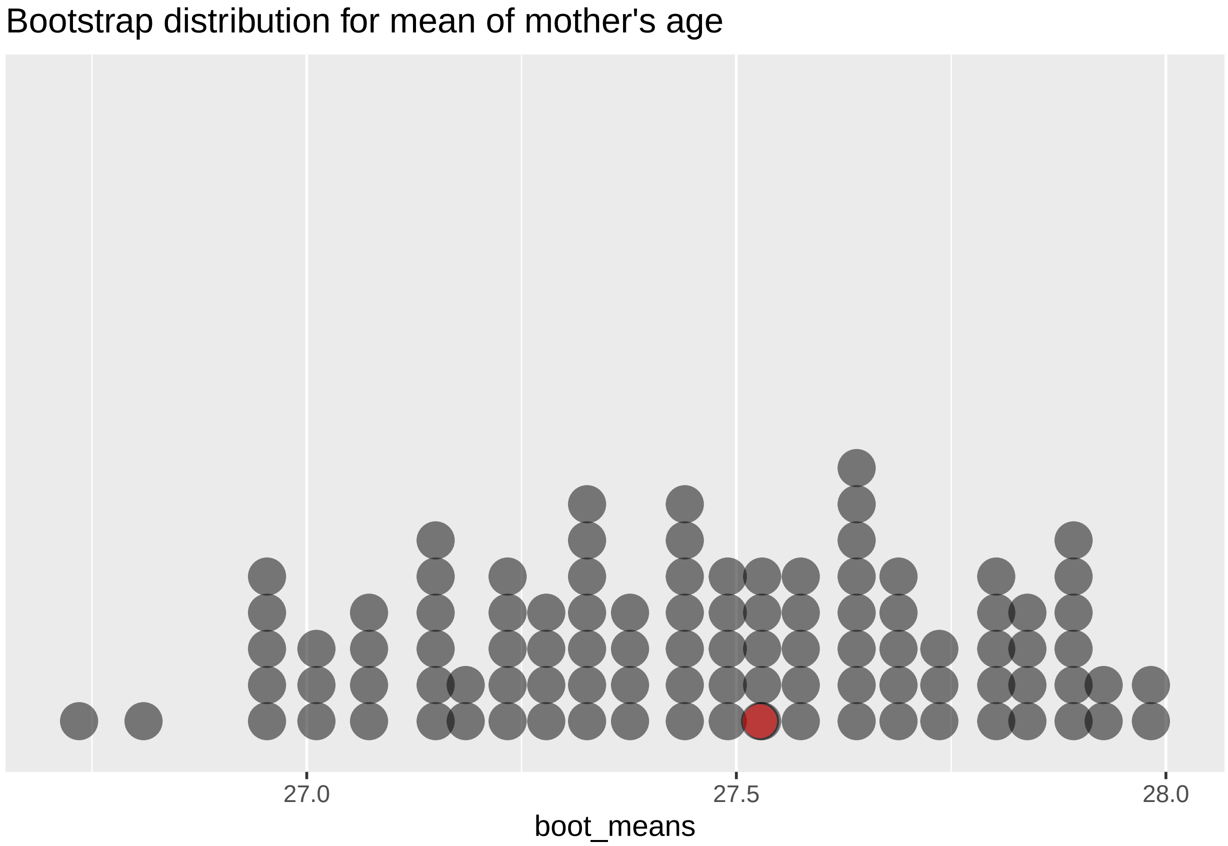
You will find confidence intervals for parameters related to the distribution of the mother's age, which for this sample is stored in the variable age.

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

(a) Suppose we are interested in how means of random samples of n=400 mothers vary across possible samples of 400 mothers we could take from the population. Explain why it is not possible to use these data (i.e., 'Gestation') to estimate this like we did in Question 3a.

In 3a we estimated the sampling distribution of the sample mean based on samples of n=20 observations by repeatedly drawing samples of size 20 from the population of claims, which were available in the 'auto_claims_population.csv' data set. The data for this question are on a sample of mothers who participated in the Child Health and Development Studies in 1961 and 1962. The n=400 ages here, then, represent ages for a sample *of mothers, not the entire population. Therefore, we cannot repeatedly take samples of 400 observations from the population. We do not have data on the entire population.*

(b) The plot below shows the bootstrap distribution for the mean of mother's age for 100 bootstrap samples. The red dot is the estimate of the mean for the first bootstrap sample, and the grey dots are the estimates of the mean for the other 99 bootstrap samples.



```
## # A tibble: 1 x 6
##     min  mean median   max    sd     n
##   <dbl> <dbl>  <dbl> <dbl> <dbl> <int>
## 1  26.7  27.5   27.5  28.0 0.299   100
```

i) Explain how the value of the red dot is calculated.

The red dot is the mean of mother's age for one bootstrap sample. The bootstrap sample is obtained by taking a random sample with replacement, from the original sample data, with the same number of observations as the original sample.

(ii) Using this plot, estimate a 90% confidence interval for the mean of mother's age.

The 90% confidence interval ranges from approximately the 5th largest data point to the 95th largest data point. This interval will be from a value a little below 27.0 to a value a little below 28.0.

(c)

(i) Use R to find a 99% bootstrap confidence interval for the mean of mother's age. Use 2000 bootstrap samples. *NOTE:* More bootstrap samples is better, but if you find this times out or takes too long in RStudio Cloud, try using 1000 bootstrap samples instead.

```r
repetitions <- 2000
boot_means <- rep(NA, repetitions)  # where we'll store the bootstrap means
sample_size <- as.numeric(Gestation %>% summarize(n()))
set.seed(50)
for (i in 1:repetitions)
{
  boot_samp <- Gestation %>% sample_n(size = sample_size, replace=TRUE)
  boot_means[i] <- as.numeric(boot_samp %>% summarize(mean(age)))
}
quantile(boot_means,c(.005,.995))
```

```
##     0.5%    99.5%
## 26.78249 28.17759
```

(ii) Suppose your confidence interval was (26.8, 28.2). Explain why the interpretation *"We are 99% sure that the true mean of a mother's age at the time this sample was taken is between 26.8 and 28.2 years."* is *INCORRECT*. What is a correct interpretation?

The true mean age of mothers in 1961/62 is unknown, but it's not random. In other words, it's a fixed, but unknown constant. Therefore it either is or isn't in this interval (i.e., the chance is either 0% or 100%). We just do not know either way.

We can conclude that we are 99% confident *that the true mean mother's age in 1961/62 was between 26 and 27 years. We are confident in this because the method we used to obtain the interval will produce intervals that do include the true value of the parameter of interest for 99% of the possible samples we could take.*

(d)

(i) Use R to find a 95% bootstrap confidence interval for the *median* of mother's age. Use 2000 bootstrap samples. *NOTE:* More bootstrap samples is better, but if you find this times out or takes too long in RStudio Cloud, try using 1000 bootstrap samples instead.

```r
repetitions <- 2000;
boot_medians <- rep(NA, repetitions)
sample_size <- as.numeric(Gestation %>% summarize(n()))
set.seed(579)
for (i in 1:repetitions)
{
  boot_samp <- Gestation %>% sample_n(size = sample_size, replace=TRUE)
  boot_medians[i] <- as.numeric(boot_samp %>% summarize(median(age)))
}
quantile(boot_medians,c(0.025,0.975))
```

```
##  2.5% 97.5%
##    26    27
```

(ii) Write an interpretation of this interval.

The 95% bootstrap confidence interval for the median of mother's age is (26, 27). We are 95% confident that the median age of all mothers in 1961/62 is between 26 and 27 years based on these data.