tutorial_11

December 9, 2021

1 Tutorial 11 - Introduction to Statistical Inference

1.0.1 Lecture and Tutorial Learning Goals:

After completing this week's lecture and tutorial work, you will be able to: - Describe real world examples of questions that can be answered with the statistical inference methods. - Name common population parameters (e.g., mean, proportion, median, variance, standard deviation) that are often estimated using sample data, and use computation to estimate these. - Define the following statistical sampling terms (population, sample, population parameter, point estimate, sampling distribution). - Explain the difference between a population parameter and sample point estimate. - Use computation to draw random samples from a finite population. - Use computation to create a sampling distribution from a finite population. - Describe how sample size influences the sampling distribution.

[1]:
```
### Run this cell before continuing.
library(tidyverse)
library(repr)
library(digest)
library(infer)
options(repr.matrix.max.rows = 6)
source('tests_tutorial_11.R')
source('cleanup_tutorial_11.R')
```

```
── Attaching packages ─────────────────────────────────────── tidyverse
1.3.0 ──

✔ ggplot2 3.3.2     ✔ purrr   0.3.4
✔ tibble  3.0.3     ✔ dplyr   1.0.2
✔ tidyr   1.1.2     ✔ stringr 1.4.0
✔ readr   1.3.1     ✔ forcats 0.5.0

Warning message:
"package 'ggplot2' was built under R version 4.0.1"
Warning message:
"package 'tibble' was built under R version 4.0.2"
Warning message:
"package 'tidyr' was built under R version 4.0.2"
Warning message:
"package 'dplyr' was built under R version 4.0.2"
```

```
  Conflicts
tidyverse_conflicts()
  dplyr::filter() masks stats::filter()
  dplyr::lag()    masks stats::lag()

Warning message:
"package 'infer' was built under R version 4.0.3"

Attaching package: 'testthat'


The following object is masked from 'package:dplyr':

    matches


The following object is masked from 'package:purrr':

    is_null


The following object is masked from 'package:tidyr':

    matches
```

1.0.2 Virtual sampling simulation

In this tutorial you will study samples and sample means generated from different distributions. In real life, we rarely, if ever, have measurements for our entire population. Here, however, we will make simulated datasets so we can understand the behaviour of sample means.

Suppose we had the data science final grades for a large population of students.

[2]:
```
# run this cell to simulate a finite population
set.seed(20201) # DO NOT CHANGE
students_pop <- tibble(grade = (rnorm(mean = 70, sd = 8, n = 10000)))
students_pop
```

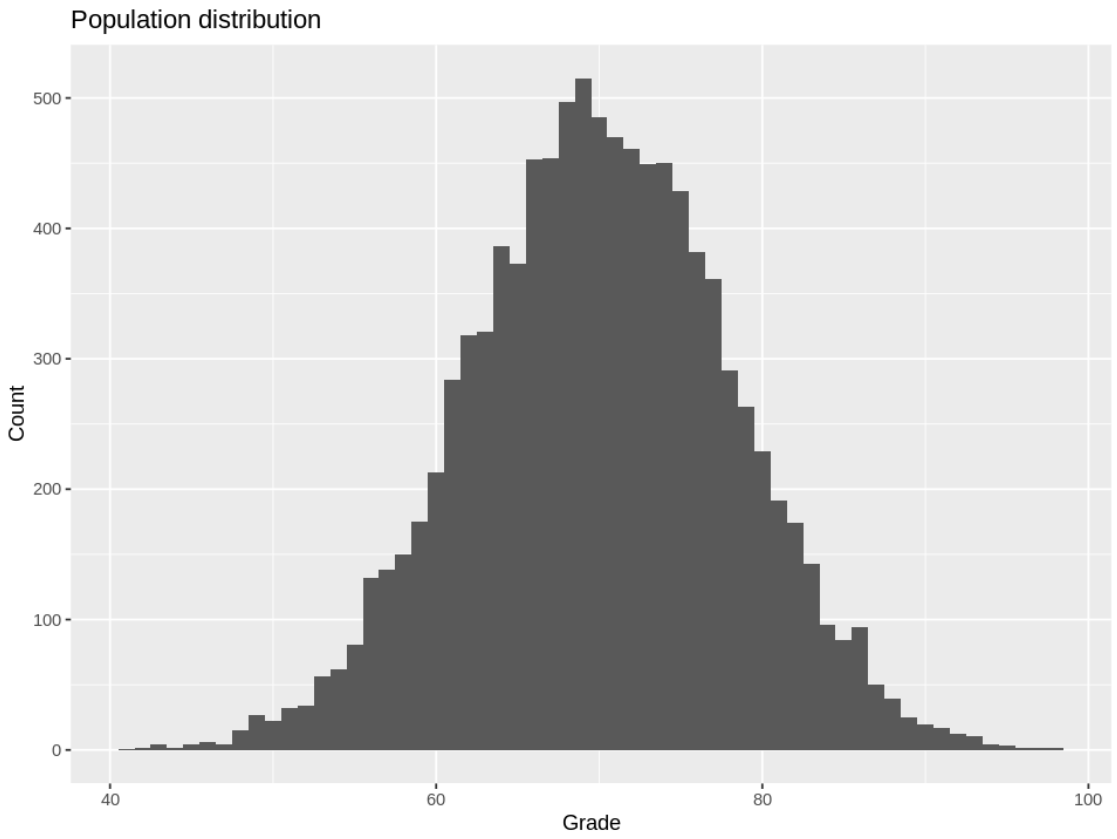
A tibble: 10000 × 1

grade
<dbl>
82.47102
73.29210
72.24852
72.65407
83.26502
68.36166

Question 1.0 {points: 1}

Visualize the distribution of the population (`students_pop`) that was just created by plotting a histogram using `binwidth = 1` in the `geom_histogram` argument. Name the plot `pop_dist` and give x-axis a descriptive label.

[3]:
```
options(repr.plot.width = 8, repr.plot.height = 6)
# ... <- ggplot(..., ...) +
#    geom_...(...) +
#    ... +
#    ggtitle("Population distribution")

# your code here
pop_dist <- ggplot(students_pop, aes(x=grade)) +
   geom_histogram(binwidth = 1) +
   labs(x = "Grade", y = "Count") +
   ggtitle("Population distribution")
pop_dist
```



[4]:
```
test_1.0()
```

```
[1] "Success!"
```

Question 1.1 {points: 3}

Describe in words the distribution above, comment on the shape, center and how spread out the distribution is.

The distribution above looks like a normal distribution and shape looks like it is bell-shaped. The shape of it is like a 2D pyramid where it peaks at around grade=70, the shape looks like a ladder up from grade=40 to grade=70 and ladder down from grade=70 to grade=100. The center of the distribution is around 70s as its mean. If we look at the x values (grades), we can see that it is quite spread out, the range of grades we encounter is between 40 and 100, and therefore it is a quite spread out distribution. We can also comment on the symmetrical skewness or symmetrical distribution, which means that both the left and right hand side of the mean are symmetrically distributed.

Question 1.2 {points: 1}

Use `summarise` to calculate the following population parameters from the `students_pop` population: - mean (use the `mean` function) - median (use the `median` function) - standard deviation (use the `sd` function)

Name this data frame `pop_parameters` which has the column names `pop_mean`, `pop_med` and `pop_sd`.

[5]:
```
# your code here
pop_parameters <- summarize(students_pop,
                            pop_mean = mean(grade), pop_med = median(grade),␣
 ↪pop_sd = sd(grade))
pop_parameters
```

A tibble: 1 × 3

pop_mean	pop_med	pop_sd
<dbl>	<dbl>	<dbl>
70.03288	70.01299	8.05165

[6]:
```
test_1.2()
```

```
[1] "Success!"
```

Question 1.2.1 {points: 1}

Draw one random sample of 5 students from our population of students (`students_pop`). Use `summarize` to calculate the mean, median, and standard deviation for these 5 students.

Name this data frame `ests_5` which should have column names `mean_5`, `med_5` and `sd_5`. Use the seed 4321.

[7]:
```
set.seed(4321) # DO NOT CHANGE!
# your code here
ests_5 <- students_pop %>%
   rep_sample_n(size = 5) %>%
   summarize(mean_5 = mean(grade), med_5 = median(grade), sd_5 = sd(grade))
ests_5
```

`summarise()` ungrouping output (override with `.groups` argument)

A tibble: 1 × 4

replicate	mean_5	med_5	sd_5
<int>	<dbl>	<dbl>	<dbl>
1	69.76367	73.5182	16.88339

[8]:
```
test_1.2.1()
```

```
[1] "Success!"
```

Question 1.2.2 Multiple Choice: {points: 1}

Which of the following is the point estimate for the average final grade for the population of data science students (rounded to two decimal places)?

A. 70.03

B. 69.76

C. 73.52

D. 8.05

Assign your answer to an object called ***answer1.2.2****. Your answer should be a single character surrounded by quotes.*

[9]:
```
# your code here
answer1.2.2 <- "B"
```

[10]:
```
test_1.2.2()
```

```
[1] "Success!"
```

Question 1.2.3 {points: 1}

Draw one random sample of 100 students from our population of students (`students_pop`). Use `summarize` to calculate the mean, median and standard deviation for these 100 students.

Name this data frame `ests_100` which has the column names `mean_100`, `med_100` and `sd_100`. Use the seed `4321`.

[11]:
```
set.seed(4321) # DO NOT CHANGE!
# your code here
ests_100 <- students_pop %>%
   rep_sample_n(size = 100) %>%
   summarize(mean_100 = mean(grade), med_100 = median(grade), sd_100 =␣
 ↪sd(grade))
ests_100
```

`summarise()` ungrouping output (override with `.groups` argument)

A tibble: 1 × 4

replicate	mean_100	med_100	sd_100
<int>	<dbl>	<dbl>	<dbl>
1	71.57394	73.53689	7.998619

[12]:
```
test_1.2.3()
```

```
[1] "Success!"
```

1.0.3 Exploring the sampling distribution of the sample mean for different populations

We will create the sampling distribution of the sample mean by taking 1500 random samples of size 5 from this population and visualize the distribution of the sample means.

Question 1.3 {points: 1}

Draw 1500 random samples from our population of students (`students_pop`). Each sample should have 5 observations. Name the data frame `samples` and use the seed `4321`.

[13]:
```
# ... <- rep_sample_n(..., size = ..., reps = ...)
set.seed(4321) # DO NOT CHANGE!
# your code here
samples <- rep_sample_n(students_pop, size = 5, reps = 1500)
head(samples)
tail(samples)
dim(samples)
```

A grouped_df: 6 × 2

replicate	grade
<int>	<dbl>
1	59.23913
1	76.06602
1	73.51820
1	92.08262
1	47.91240
2	85.42029

A grouped_df: 6 × 2

replicate	grade
<int>	<dbl>
1499	67.25662
1500	71.57016
1500	62.44389
1500	76.56512
1500	71.66241
1500	76.21743

1. 7500 2. 2

[14]:
```
test_1.3()
```

```
[1] "Success!"
```

Question 1.4 {points: 1}

Group by the sample replicate number, and then for each sample, calculate the mean. Name the data frame `sample_estimates`. The data frame should have the column names `replicate` and `sample_mean`.

[15]:
```
# your code here
sample_estimates <- samples %>%
                group_by(replicate) %>%
                summarize(sample_mean = mean(grade) )
head(sample_estimates)
tail(sample_estimates)
```

```
`summarise()` ungrouping output (override with `.groups` argument)
```

A tibble: 6 × 2

replicate <int>	sample_mean <dbl>
1	69.76367
2	75.22476
3	69.90881
4	66.11345
5	75.26159
6	63.41461

A tibble: 6 × 2

replicate <int>	sample_mean <dbl>
1495	65.80124
1496	66.44409
1497	67.86570
1498	69.39292
1499	70.64449
1500	71.69180

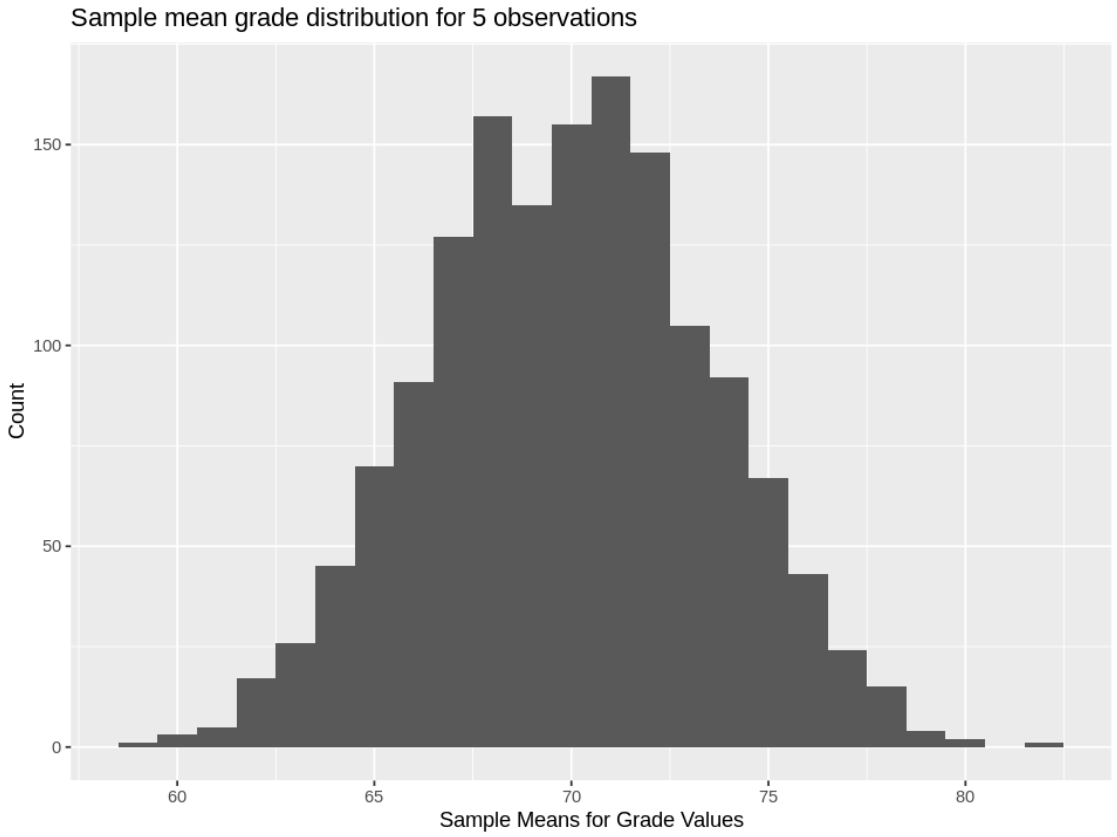
[16]:
```
test_1.4()
```

```
[1] "Success!"
```

Question 1.5 {points: 1}

Visualize the distribution of the sample estimates (`sample_estimates`) you just calculated by plotting a histogram using `binwidth = 1` in the `geom_histogram` argument. Name the plot `sampling_distribution` and give the plot (using `ggtitle`) and the x axis a descriptive label.

[17]:
```
options(repr.plot.width = 8, repr.plot.height = 6)
# your code here
sampling_distribution_5 <- ggplot(sample_estimates, aes(x= sample_mean)) +
   geom_histogram(binwidth = 1) +
   labs(x = "Sample Means for Grade Values", y = "Count") +
   ggtitle("Sample mean grade distribution for 5 observations")
#sampling_distribution
sampling_distribution_5
```



[18]:
```
test_1.5()
```

```
[1] "Success!"
```

Question 1.6 {points: 3}

Describe in words the distribution above, comment on the shape, center and how spread out the distribution is. Compare this sampling distribution to the population distribution of students' grades above.

The distribution above looks like a normal distribution, the shape of it is like a 2D pyramid where it peaks at around grade=70, the shape looks like a ladder up from grade close to 60s to grade around 70s and ladder down from grade close to 70s to grade 80s, so the shape looks like a bell-shaped distribution. The center of the distribution is around 70s as it is the mean. Also, the skewness of the distribution is symmetrical. If we look at the x values (grades), we can see that it is less spread out than the distribution we looked at earlier in the tutorial, because we used 1500 random samples of size 5 (takes the average of these replicates), and that decreased the spread of grade values, which means that we have more reliable distribution or data.

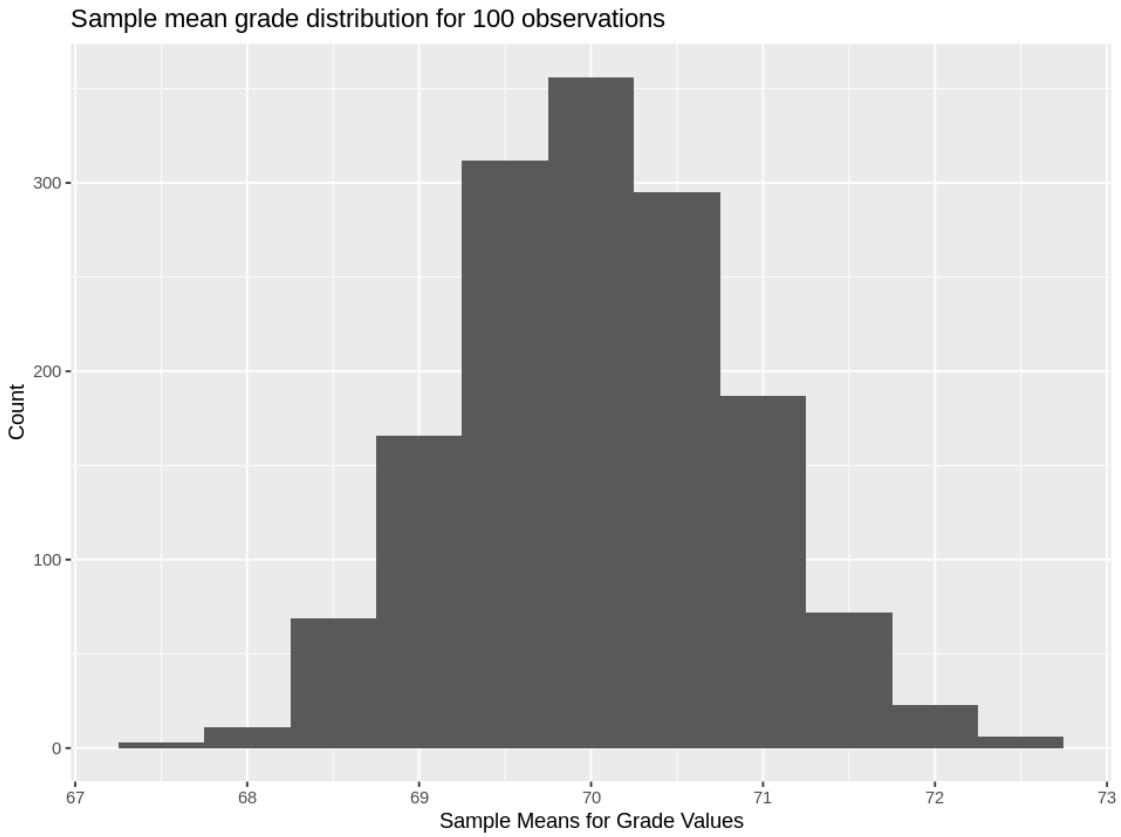
Question 1.6.1 {points: 3}

Repeat **Q1.3 - 1.5**, but now for 100 observations:
1. Draw 1500 random samples from our population of students (`students_pop`). Each sample

should have 100 observations. Use the seed `4321`. 2. Group by the sample replicate number, and then for each sample, calculate the mean (call this column `sample_mean_100`). 3. Visualize the distribution of the sample estimates you calculated by plotting a histogram using `binwidth = 0.5` in the `geom_histogram` argument. Name the plot `sampling_distribution_100` and give the plot title (using `ggtitle`) and the x axis a descriptive label.

[19]:
```
set.seed(4321) # DO NOT CHANGE!
# your code here
sample_estimates_100 <- rep_sample_n(students_pop, size = 100, reps = 1500) %>%
                group_by(replicate) %>%
                summarize(sample_mean_100 = mean(grade) )

sampling_distribution_100 <- ggplot(sample_estimates_100, aes(x=␣
↪sample_mean_100)) +
  geom_histogram(binwidth = 0.5) +
  labs(x = "Sample Means for Grade Values", y = "Count") +
  ggtitle("Sample mean grade distribution for 100 observations")

sampling_distribution_100
```

```
`summarise()` ungrouping output (override with `.groups` argument)
```



[20]:
```
set.seed(4321) # DO NOT CHANGE!

# We check that you've created objects with the right names below
# But all other tests were intentionally hidden so that you can practice␣
 ↪deciding
# when you have the correct answer.
test_that('Did not create objects named sampling_distribution_100', {
    expect_true(exists("sampling_distribution_100"))
    })
```

Question 1.6.2 {points: 3}

Suppose we do not know the parameter value for the population of data science students (as is usually the case in real life). Compare your point estimates for the population mean from **Q1.2.1 and 1.2.3** above. Which of the two point estimates is more likely to be closer to the actual value of the average final grade of the population of data science students? Briefly explain. (Hint: look at the sampling distributions for your samples of size 5 and size 100 to help you answer this question).

If we don't know the population parameter values such as mean, standard deviation etc, then we can look at the distribution from the histograms above for both distributions with size 5 and size 100. In Q1.2.1, we can see that the mean is 69.76 and standard deviation is 16.88, that means that sampling distributions with size 5 had number of people who were spread across a large range. In Q1.2.3, we can see that the mean is 71.57 and standard deviation is 7.99, this indicates that sampling distributions with size 100 had number of people who were less spread out (+/-7.99 of the mean). This is around the half of the standard deviation we got for distribution with size 5. If we look at the means they are 69.76 and 71.57, which means that they are not very far off each other.

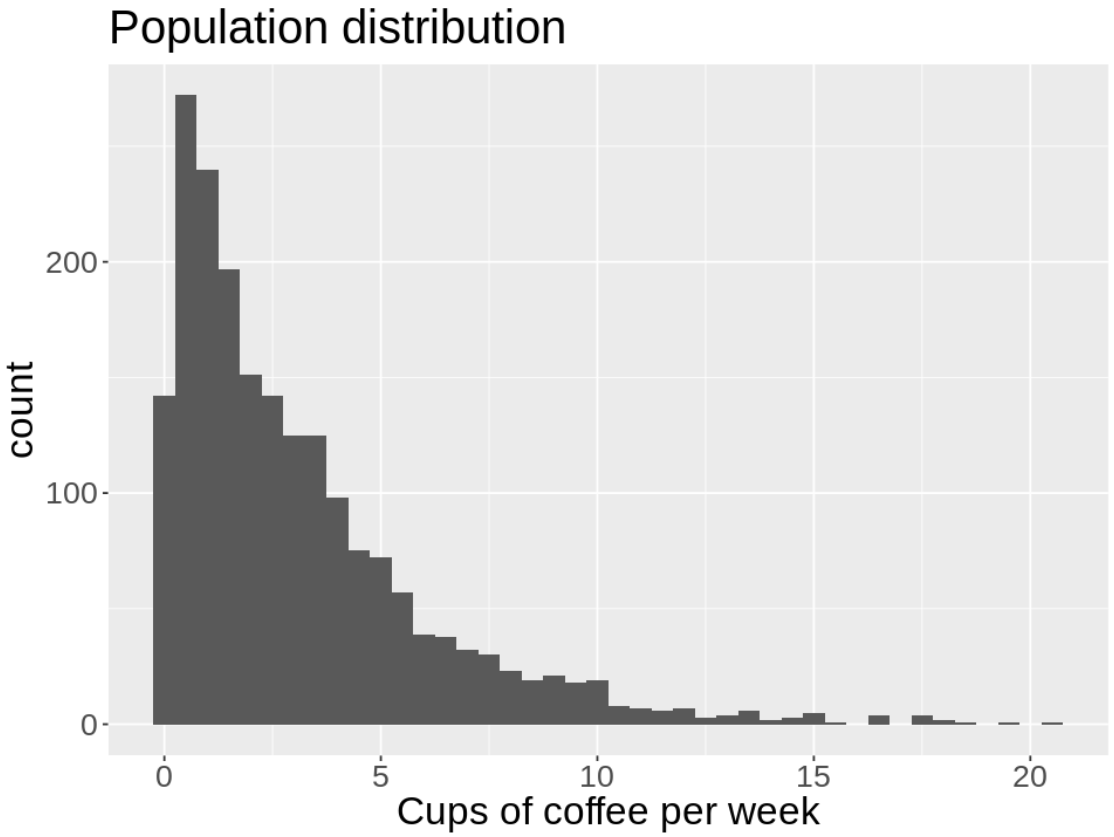
If we look at the two point estimates we can't really say which one is closer to the actual value of the average final grade, because we assume that we don't know the actual parameter value (eg. mean). Usually, the larger the n (size of the samples) is the better it is, and we can say that when the n is larger, we expect to have it closer to actual values (but no guarantees), because we get more reliable results. However, we can say that the sample distribution with more observations (size=100 in this case) does produce more reliable results for us, because the grades are much less spread out than the grades distribution for size 5.

Question 1.7 {points: 1}

Let's create a simulated dataset of the number of cups of coffee drunk per week for our population of students. Describe in words the distribution, comment on the shape, center and how spread out the distribution is.

[21]:
```
# run this cell to simulate a finite population
set.seed(2020) # DO NOT REMOVE
coffee_data = tibble(cups = rexp(n = 2000, rate = 0.34))

coffee_dist <- ggplot(coffee_data, aes(cups)) +
    geom_histogram(binwidth = 0.5) +
```

```
    xlab("Cups of coffee per week") +
    ggtitle("Population distribution") +
    theme(text = element_text(size = 20))
coffee_dist
```



The distribution above looks like a non-symmetrical distribution, and the shape doesn't look like bell-shaped this time around. The distribution peaks at around cups of coffee per week close to 1, which means the mode is that number. However, if look at the shape, we can see that the distribution has positive skewness (or left-modal), that means that the center or the mean of the plot is on the right side of the mode (where the plot peaks). And if we look at the range we see, it starts from 0 to 20 cups of coffee per week, which indicates that the distribution is quite spread out in this case.

Question 1.8 {points: 1}

Draw 1500 random samples from `coffee_data`. Each sample should have 5 observations. Assign this data frame to an object called `coffee_samples_5`.

Group by the sample replicate number, and then for each sample, calculate the mean. Name the data frame `coffee_sample_estimates_5`. The data frame should have the column names `replicate` and `coffee_sample_mean_5`.

Finally, create a plot of the sampling distribution called `coffee_sampling_distribution_5`.

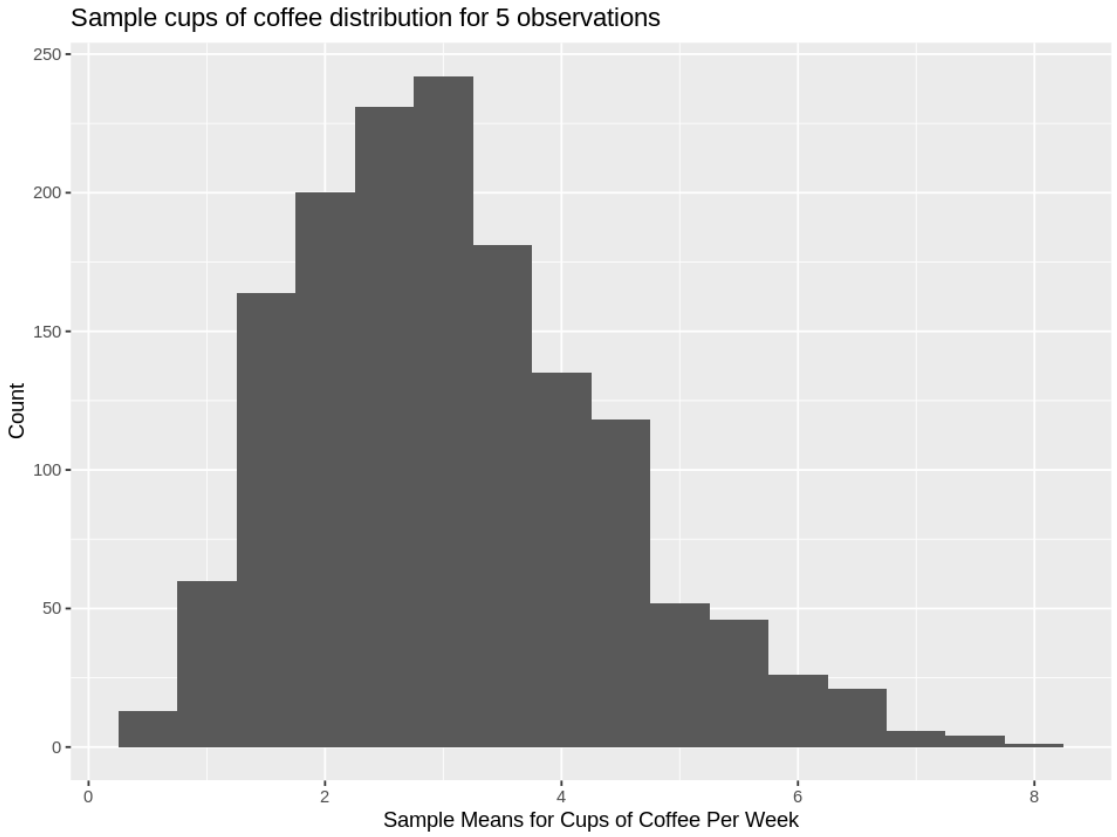
Hint: a bindwidth of 1 is a little too big for this data, try a bindwidth of 0.5 instead.

[22]:
```
set.seed(4321) # DO NOT CHANGE!

# your code here
coffee_samples_5 <- rep_sample_n(coffee_data, size = 5, reps = 1500)
coffee_sample_estimates_5 <-  coffee_samples_5 %>%
                              group_by(replicate) %>%
                              summarize(coffee_sample_mean_5 = mean(cups))
coffee_sample_estimates_5
coffee_sampling_distribution_5 <- ggplot(coffee_sample_estimates_5, aes(x=␣
 ↪coffee_sample_mean_5)) +
   geom_histogram(binwidth = 0.5) +
   labs(x = "Sample Means for Cups of Coffee Per Week", y = "Count") +
   ggtitle("Sample cups of coffee distribution for 5 observations")
coffee_sampling_distribution_5

# some extra analysis
# coffee_sum <- summarize(coffee_data,
#                              cup_mean = mean(cups), cup_med = median(cups),␣
 ↪cup_sd = sd(cups))
# coffee_sum # mean 3.12, sd 3.04, med 2.24

# coffee_estimates_sum <- summarize(coffee_sample_estimates_5,
#                              cup_mean2 = mean(coffee_sample_mean_5), cup_med2 =␣
 ↪median(coffee_sample_mean_5),
#                              cup_sd2 = sd(coffee_sample_mean_5))
# coffee_estimates_sum # mean 3.07 and sd 1.29 and med 2.88
```

`summarise()` ungrouping output (override with `.groups` argument)

A tibble: 1500 × 2

replicate <int>	coffee_sample_mean_5 <dbl>
1	2.556328
2	4.067040
3	3.030184
⋮	⋮
1498	2.826998
1499	1.978601
1500	1.636511



[23]:
```
test_1.8()
```

```
[1] "Success!"
```

Question 1.9 {points: 3}

Describe in words the distribution above, comment on the shape, center and how spread out the distribution is. Compare this sampling distribution to the population distribution above.

The distribution for 5 observations from the coffee data is still a non-symmetrical distribution, so it doesn't look like a bell-shaped distribution as we have seen in grades distribution. We can also say that the skewness is positive once again, and the center seems to be very similar to the previous plot at around 3 cups of coffee per week. However, we can clearly see that the standard deviation decreased significantly in this sampling distribution of size 5, and the distribution range is between 0 to 8, and therefore, we can say that the distribution is actually less spread out than the previous case.

Question 2.0 {points: 1}

Draw 1500 random samples from `coffee_data`. Each sample should have 5 observations. Assign this data frame to an object called `coffee_samples_30`.

Group by the sample replicate number, and then for each sample, calculate the mean. Name the data frame `coffee_sample_estimates_30`. The data frame should have the column names

`replicate` and `coffee_sample_mean_30`.

Finally, create a plot of the sampling distribution called `coffee_sampling_distribution_30`.

> Hint: use `xlim` to control the x-axis limits so that they are similar to those in the histogram above. This will make it easier to compare this histogram with that one.

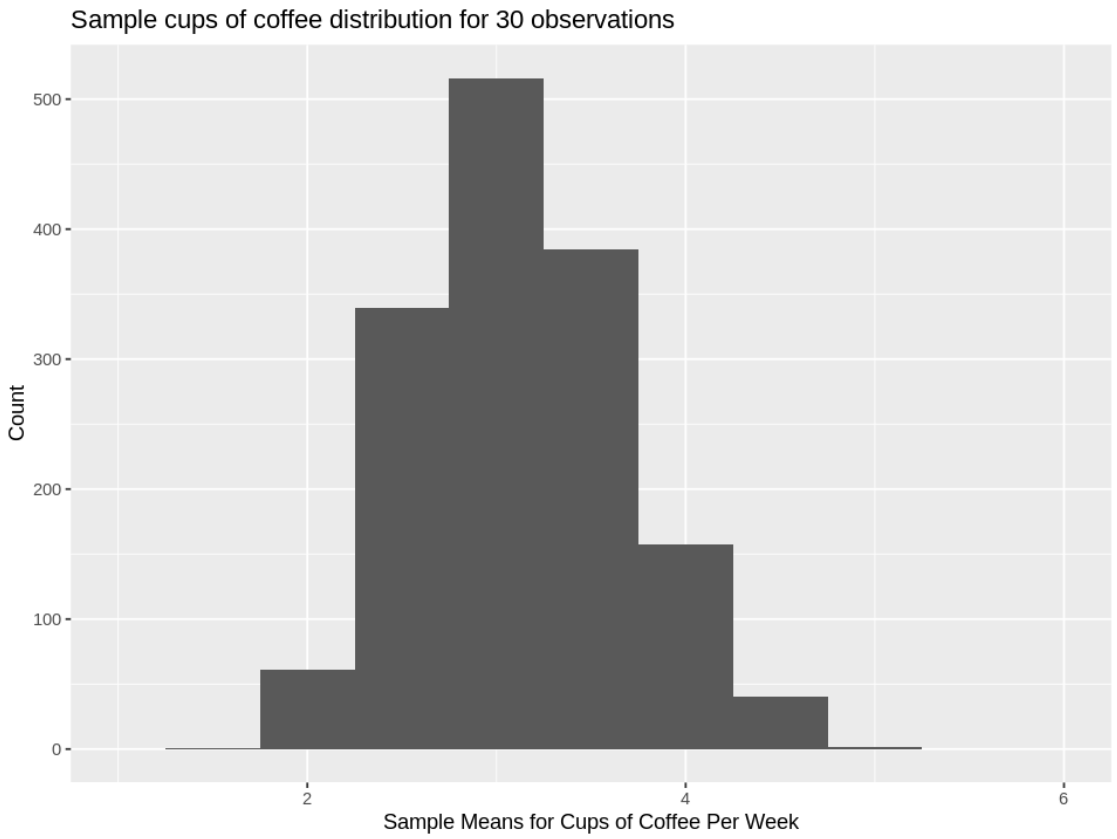
[24]:
```
set.seed(4321) # DO NOT CHANGE!

# your code here
coffee_samples_30 <- rep_sample_n(coffee_data, size = 30, reps = 1500)
coffee_sample_estimates_30 <-  coffee_samples_30 %>%
                            group_by(replicate) %>%
                            summarize(coffee_sample_mean_30 = mean(cups))
coffee_sample_estimates_30
coffee_sampling_distribution_30 <- ggplot(coffee_sample_estimates_30, aes(x=␣
 ↪coffee_sample_mean_30)) +
   geom_histogram(binwidth = 0.5) +
   labs(x = "Sample Means for Cups of Coffee Per Week", y = "Count") +
   ggtitle("Sample cups of coffee distribution for 30 observations")+
   xlim(c(1,6))
coffee_sampling_distribution_30
# extra
# coffee_estimates_sum30 <- summarize(coffee_sample_estimates_30,
#                             cup_mean2 = mean(coffee_sample_mean_30), cup_med2␣
 ↪= median(coffee_sample_mean_30),
#                             cup_sd2 = sd(coffee_sample_mean_30))
#coffee_estimates_sum30 # mean 3.12 and sd 0.54 and med 3.09
```

```
`summarise()` ungrouping output (override with `.groups` argument)
```

A tibble: 1500 × 2

replicate <int>	coffee_sample_mean_30 <dbl>
1	3.073687
2	3.753924
3	3.312196
...	...
1498	2.763284
1499	2.594383
1500	3.224412

```
Warning message:
"Removed 2 rows containing missing values (geom_bar)."
```



[25]:
```
test_2.0()
```

```
[1] "Success!"
```

Question 2.1 {points: 3}

Describe in words the distribution above, comment on the shape, center and how spread out the distribution is. Compare this sampling distribution with samples of size 30 to the sampling distribution with samples of size 5.

The distribution for 30 observations from the coffee data looks a lot more symmetrical than the previous plots, so it looks more of a bell-shaped distribution than the previous ones. We can also say that the skewness is normal, and therefore it looks a lot more like a normal distribution with size 30 than it does with size 5.

The center or mean seems to be very similar to the previous plot at around 3 cups of coffee per week. However, we can clearly see that the standard deviation decreased even more in this sampling distribution of size 30 than the size 5, and the distribution range is between 1 to 5, and therefore, we can say that the distribution is actually a lot less spread out than the samples of size 5. The standard deviation is around 0.5-0.6 in samples of size 30, that means the distribution is a lot more reliable than the size of 5.

[26]:
```
source('cleanup_tutorial_11.R')
```