Math 300 Lesson 19 Notes

Sampling

YOUR NAME HERE

June, 2022

Contents

Objectives

1. Explain the need for randomization in sampling.
2. Explain the impact of sample size on the sample distribution and number of replications on the sampling procedure.
3. Using a sampling distribution, make decisions about the population of interest.

Reading

Chapter 7 - 7.2

Lesson

Remember that you will be running this more like a lab than a lecture. You want them using `R` and answering questions. Have them open the notes rmd and work through it together.

Work through the learning checks LC7.1 - 7.7.

- Learning checks 7.3 - 7.5 are not in the book but are in the notes for students. We are not sure why they were dropped.
- Our solutions we discuss may be different from those in the back of the book. Make sure to ask questions about answers that you don't understand.

Setup

```
library(tidyverse)
library(moderndive)
```

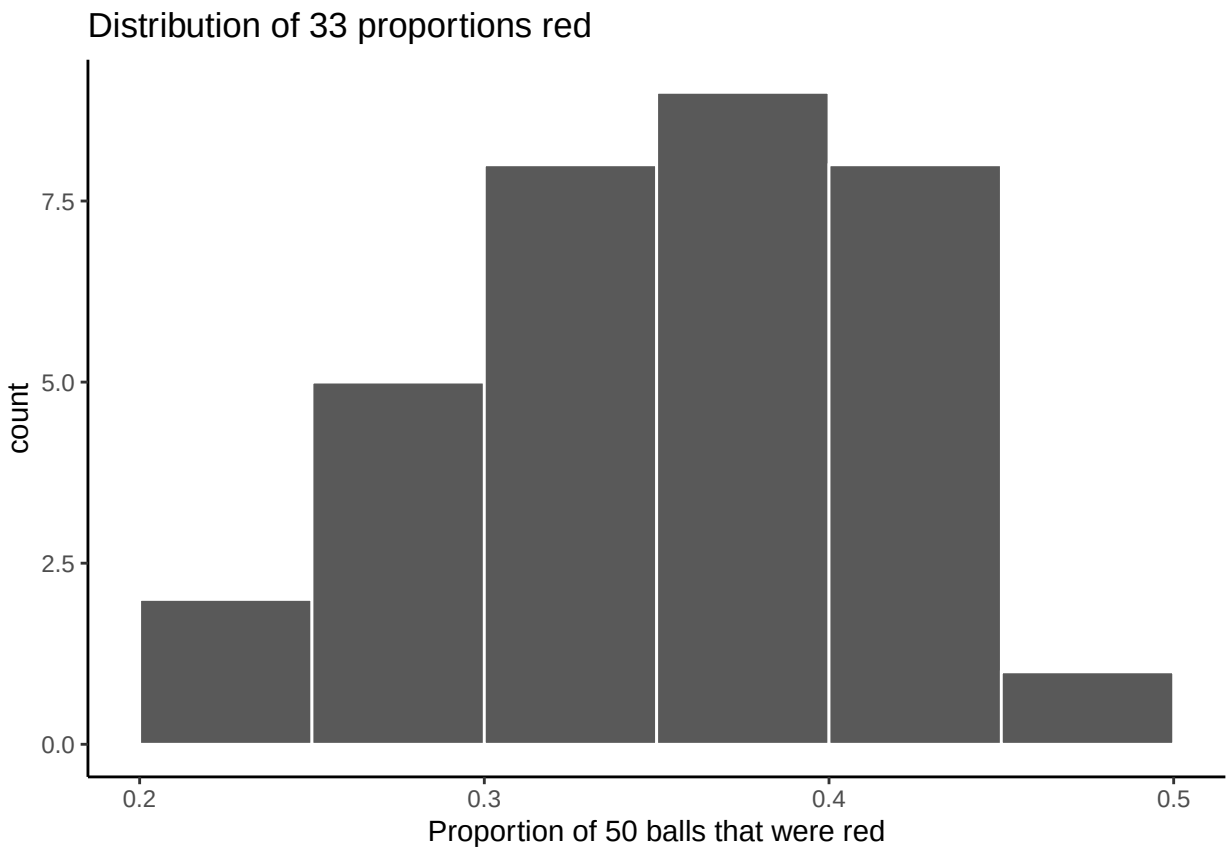
Here is the data from the reading and the histogram.

```
head(tactile_prop_red)
```

```
## # A tibble: 6 x 4
##   group            replicate red_balls prop_red
##   <chr>                <int>     <int>    <dbl>
## 1 Ilyas, Yohan             1        21     0.42
## 2 Morgan, Terrance         2        17     0.34
## 3 Martin, Thomas           3        21     0.42
## 4 Clark, Frank             4        21     0.42
## 5 Riddhi, Karina           5        18     0.36
## 6 Andrew, Tyler            6        19     0.38
```

- What is each row in the data?
- Is this data tidy?

```
ggplot(tactile_prop_red, aes(x = prop_red)) +
  geom_histogram(binwidth = 0.05, boundary = 0.4, color = "white") +
  labs(x = "Proportion of 50 balls that were red",
       title = "Distribution of 33 proportions red") +
  theme_classic()
```



- What do you think is the true proportion of red balls in the bin?

LC 7.1 (Objective 1)

(LC 7.1) Why was it important to mix the bowl before we sampled the balls?

Solution:

LC 7.2 (Objective 1)

(LC 7.2) Why is it that our 33 groups of friends did not all have the same numbers of balls that were red out of 50, and hence different proportions red?

Solution:

LC 7.3 (Objective 2)

(LC 7.3) Why couldn't we study the effects of sampling variation when we used the virtual shovel only once? Why did we need to take more than one virtual sample (in our case 33 virtual samples)?

Solution:

LC 7.4 (Objective 2)

(LC 7.4) Why did we not take 1000 "tactile" samples of 50 balls by hand?

Solution:

LC 7.5 (Objective 3)

Code for this problem. Complete the code and remove extra comment symbols

```
# Segment 2: sample size = 50 ------------------------------
# Virtually use shovel 1000 times
#set.seed(107)
# virtual_samples_50 <- bowl %>%
#  rep_sample_n(size = ____________, reps = 1000)
```

```
# Compute resulting 1000 replicates of proportion red
# virtual_prop_red_50 <- virtual_samples_50 %>%
#   group_by(_______________) %>%
#   summarize(red = sum(color == "_______________")) %>%
#   mutate(prop_red = ___________ / 50)
```

```
#summary(virtual_prop_red_50)
```

```
#Plot distribution via a histogram
# ggplot(virtual_prop_red_50, aes(x = ________)) +
#   geom_histogram(binwidth = 0.05, boundary = 0.4, color = "white") +
#   labs(x = "Proportion of 50 balls that were red", title = "50")  +
#   theme_classic()
```

(LC 7.5) Looking at the figure we just created, would you say that sampling 50 balls where 30% of them were red is likely or not? What about sampling 50 balls where 10% of them were red?

Solution:

LC 7.6 (Objective 2)

(LC 7.6) In Figure 7.9, we used shovels to take 1000 samples each, computed the resulting 1000 proportions of the shovel's balls that were red, and then visualized the distribution of these 1000 proportions in a histogram. We did this for shovels with 25, 50, and 100 slots in them. As the size of the shovels increased, the histograms got narrower. In other words, as the size of the shovels increased from 25 to 50 to 100, did the 1000 proportions

- A. vary less,
- B. vary by the same amount, or
- C. vary more?

  Solution:

LC 7.7 (Objective 2)

(LC 7.7) What summary statistic did we use to quantify how much the 1000 proportions red varied?

- A. The inter-quartile range
- B. The standard deviation
- C. The range: the largest value minus the smallest.

Solution:

Documenting software

- File creation date: 2022-06-08
- R version 4.1.3 (2022-03-10)
- `tidyverse` package version: 1.3.1
- `moderndive` package version: 0.5.4