

Math 300 Lesson 20 Notes

Sampling Framework

YOUR NAME HERE

June, 2022

Contents

Objectives	1
Reading	1
Lesson	1
Documenting software	4

Objectives

1. Know and correctly use the terminology and notation for the sampling concepts such as population, census, point estimate, biased, etc.
2. Explain the impact of sample size on the sample distribution and number of replications on the sampling procedure.
3. Use and explain the terms accuracy and precision in the context of sampling.

Reading

Chapter 7.3

Lesson

Work through the learning checks LC7.8 - 7.20. Complete the code as necessary.

- This line from the book is a good summary:

What you have been performing is statistical inference. This is one of the most important concepts in all of statistics. So much so, we included this term in the title of our book: “Statistical Inference via Data Science”. More generally speaking,

– If the sampling of a sample of size n is done at random, then – the sample is unbiased and representative of the population of size N , thus – any result based on the sample can generalize to the population, thus – the point estimate is a “good guess” of the unknown population parameter, thus – instead of performing a census, we can infer about the population using sampling.

- Try to use the terminology and notation as much as possible from this point on. You can use the bin or red and white balls as an example to go back to.
- Standard error and standard deviation can get confused. The standard error is simply the standard deviation of the point estimate.

Setup

Code to setup the learning checks. Complete the code and remove extra comment symbols.

```
library(tidyverse)
library(moderndiver)

# 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 7.8 (Objective 1)

(LC 7.8) In the case of our bowl activity, what is the *population parameter*? Do we know its value?

Solution:

LC 7.9 (Objective 1)

(LC 7.9) What would performing a census in our bowl activity correspond to? Why did we not perform a census?

Solution:

LC 7.10 (Objective 1)

(LC 7.10) What purpose do *point estimates* serve in general? What is the name of the point estimate specific to our bowl activity? What is its mathematical notation?

Solution:

LC 7.11 (Objective 1)

(LC 7.11) How did we ensure that our tactile samples using the shovel were random?

Solution:

LC 7.12 (Objective 1)

(LC 7.12) Why is it important that sampling be done *at random*?

Solution:

LC 7.13 (Objective)

(LC 7.13) What are we *inferring* about the bowl based on the samples using the shovel?

Solution:

LC 7.14 (Objective)

(LC 7.14) What purpose did the *sampling distributions* serve?

Solution:

LC 7.15 (Objective 1)

(LC 7.15) What does the *standard error* of the sample proportion \hat{p} quantify?

Solution:

LC 7.16 (Objective 2)

(LC 7.16) The table that follows is a version of Table 7.3 matching sample sizes n to different *standard errors* of the sample proportion \hat{p} , but with the rows randomly re-ordered and the sample sizes removed. Fill in the table by matching the correct sample sizes to the correct standard errors. Simply list the order of sample size in your answer.

Table 1: Three standard errors of the sample proportion based on $n = 25, 50, 100$

Sample size	Standard error of \hat{p}
n =	0.067
n =	0.048
n =	0.094

Solution:

For the following four learning checks, let the *estimate* be the sample proportion \hat{p} : the proportion of a shovel's balls that were red. It estimates the population proportion p : the proportion of the bowl's balls that were red.

LC 7.17 (Objective 3)

(LC 7.17) What is the difference between an *accurate* estimate and a *precise* estimate?

Solution:

LC 7.18 (Objective 3)

(LC 7.18) How do we ensure that an estimate is *accurate*? How do we ensure that an estimate is *precise*?

Solution:

LC 7.19 (Objective 1)

(LC 7.19) In a real-life situation, we would not take 1000 different samples to infer about a population, but rather only one. Then, what was the purpose of our exercises where we took 1000 different samples?

Solution:

LC 7.20 (Objective 3)

(LC 7.20) Figure 7.13 with the targets shows four combinations of “accurate versus precise” estimates. Draw four corresponding *sampling distributions* of the sample proportion \hat{p} , like the one in the left-most plot in Figure 7-12. Complete the code.

Solution:

Documenting software

- File creation date: 2022-06-30
- R version 4.1.3 (2022-03-10)
- `tidyverse` package version: 1.3.1
- `moderndive` package version: 0.5.4