Sampling and the Central Limit Theorem Part I

your name

2024-11-02

Math 2265 Chapter 5. Foundations for Inference

Section 5.1 Point Estimates and Sampling Variability

- Work as a group!
- You will need to replace `"ans"` or `your_answer` in the source code
- Update your name in L3
- Add your group members' name below; students may lose one point if Question 0 is unanswered
- Make sure you save and `knit` your work (to html or pdf) before submitting it to Canvas
- Please only submit your work if you attended the class and worked with other students; this is not an online course

---

Goal

1. Understand how to sample in R
2. Learn the central limit theorem by example

Question 0. Who are your group members? (List their first names)

Answer:

1. `<name_1>`
2. `<name_2>`
3. `<name_3>`
4. `<name_4>`

If you need more time to get used to `Markdown`, use the `Visual` mode.

The icon is located in the upper-left corner next to `source`.

Normal distribution, Central Limit Theorem, Confidence Interval

These worksheets aim to help you understand the phrase such as "95% OF MEANS PROJECTED TO FALL IN THIS RANGE." In a nutshell, we want to estimate the `mean (or proportion) of the population` from the `mean (or proportion) of a sample` with, often, 95% confidence.

In this note, we create population proportion data and take (several) samples. Recall that the size of the population is "big," and its sample is reasonable. In our example, the population consists of 50,000 observations, and the samples are 1,000. The data set consists of "success" and "fail" though it can be

replaced by any categories with two unique values such as, head and tail, candidate A and candidate B. Here we use two to easily compute the proportion.

The R-variable `population` will serve as the population. We will select 50,000 between "success" and "fail" and save the result in `population`. Our goal is to guess the success and fail ratio of the population proportion (population parameter) via sample proportions (sample statistics). Here the success rate is

$$\text{success proportion} = \frac{\text{number of successes}}{\text{number of successes} + \text{number of fails}}$$

```
# the following three lines randomize the probability of choosing successes and fails
set.seed(1105)
probabilities <- runif(2)
probabilities <- probabilities / sum(probabilities)

# generate population of size 50,000 consisting of successes and fails
population <-sample(c("success", "fail"), 50000, replace = TRUE, prob=probabilities)
head(population)
```

```
## [1] "success" "fail"    "success" "success" "fail"    "fail"
```

Let's take a sample of size 1,000 from the population and compute its mean.

```
my_sample <- sample(population, 1000)
head(my_sample)
```

```
## [1] "success" "success" "fail"    "success" "success" "success"
```

```
table(my_sample)
```

```
## my_sample
##    fail success
##     342     658
```

This means that among the 1,000 samples there are s number of successes and f number of fails. Since s+f = 1,000, the ratio is nothing but s/1000. This is a sample statistic. Take a few more samples and compute their proportions by code.

```
my_sample <- sample(population, 1000)
print(table(my_sample))
```

```
## my_sample
##    fail success
##     340     660
```

```
print(paste("p_hat =", table(my_sample)["success"]/1000))
```

```
## [1] "p_hat = 0.66"
```

```
my_sample <- sample(population, 1000)
print(table(my_sample))
```

```
## my_sample
##    fail success
##     332     668
```

```
print(paste("p_hat =", table(my_sample)["success"]/1000))
```

```
## [1] "p_hat = 0.668"
```

```r
my_sample <- sample(population, 1000)
print(table(my_sample))
```

```
## my_sample
##    fail success
##     327     673
```

```r
print(paste("p_hat =", table(my_sample)["success"]/1000))
```

```
## [1] "p_hat = 0.673"
```

Question 1:

(a) What are your point estimates? Answer:

(b) Are they all the same? Answer:

(c) What is your guess for the population proportion? Answer:

The distribution of the sample proportions

What distribution do the sample proportions have? In other words, if we make a data set consisting of the sample proportions, what is the distribution of the success proportion? We can answer this question experimentally. The R function `replicate` is helpful in this case. For instance,

```r
# replicate executes table(sample(population, 1000)) twice
replicate(2, table(sample(population, 1000)))
```

```
##
##           [,1] [,2]
##   fail     339  330
##   success  661  670
```

Can you read the output?

Task

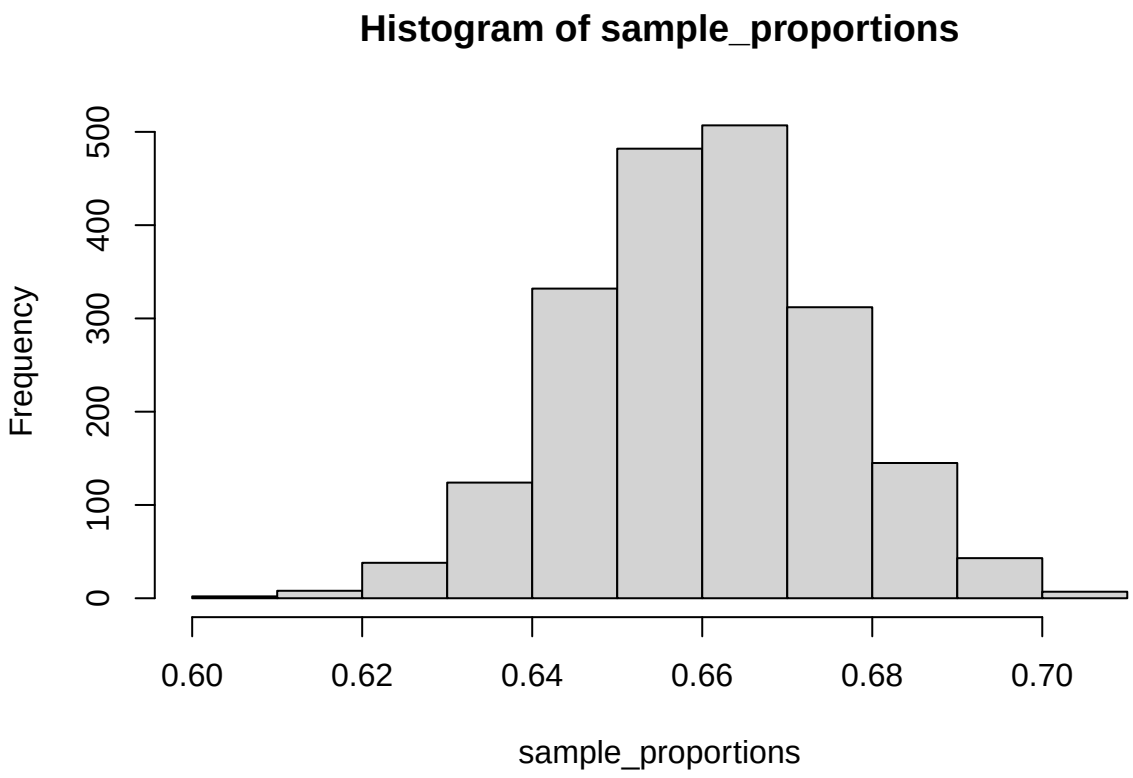
Complete the following code to take 2,000 samples.

```r
sample_proportions <- replicate(2000, table(sample(population, 1000))["success"]) # ["success"] is to c
sample_proportions <- sample_proportions/1000 # we divide by 1000 to get the proportion
head(sample_proportions)
```

```
## success success success success success success
##   0.643   0.691   0.657   0.670   0.647   0.672
```

Now, let's see its distribution. Here, we used `hist`, but you can get the same result with `ggplot` and `geom_histogram`.

```r
hist(sample_proportions)
```



Question 2

(a) What distribution does `sample_proportions` follow? Answer:

(b) What is the mean and standard deviation of `sample_proportions`?

```
# use this block to find the mean and standard deviation of `sample_proportions`
```

Here is the population proportion. Compare it with your answer above. Your answer is correct if they are close.

```
table(population)["success"]/50000
```

```
## success
## 0.66028
```

(c) If we were to take more samples, will the mean be closer to the population proportion? Answer:

(d) What is the Z-score when $X = 0.50$? Recall, the Z-score is $z = \frac{x-\bar{x}}{\sigma}$

```
# use this block to find the mean and the Z-score for X = 0.50
```

(e) If we were to take more samples, will the standard deviation be bigger or smaller? Answer:

Next time, we will go over the standard error for a point estimate which is closely related to the standard deviation.