Sampling and the Central Limit Theorem Part II

your name

2024-11-07

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 points 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 the central limit theorem by example
2. Understand point estimates, standard errors, and the confidence interval

---

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 the last worksheet, we created a population of size 50,000 consisting of "successes" and "failures" and took 2,000 samples each of size 1,000. We used the mean of the distribution, consisting of the sample proportions (often denoted by $\hat{p}$, to estimate the population proportion $p = 0.66028$).

Summary

- Samples proportions vary and they are almost never be the population proportion.

- However, the distribution of the sample proportions follows a normal distribution whose mean approximates the population proportion.
  - This is a consequence of the central limit theorem.

Last time, we questioned whether the mean would be closer to the population proportion, and whether the standard deviation would be smaller if we took more samples. Be careful as there deal with two sizes:

- Size of each sample
- Number of samples we take

Experiments

The following code block creates the population of our simulations.

```r
# the first two lines randomize the probability
set.seed(1105)
probabilities <- runif(2)
probabilities <- probabilities / sum(probabilities)

# generate population of size 50,000 consisting of successes and failures
population <-sample(c("success", "fail"), 50000, replace = TRUE, prob=probabilities)
population_proportion <- table(population)["success"]/50000
print(paste("The population parameter is", population_proportion))
```

```
## [1] "The population parameter is 0.66028"
```

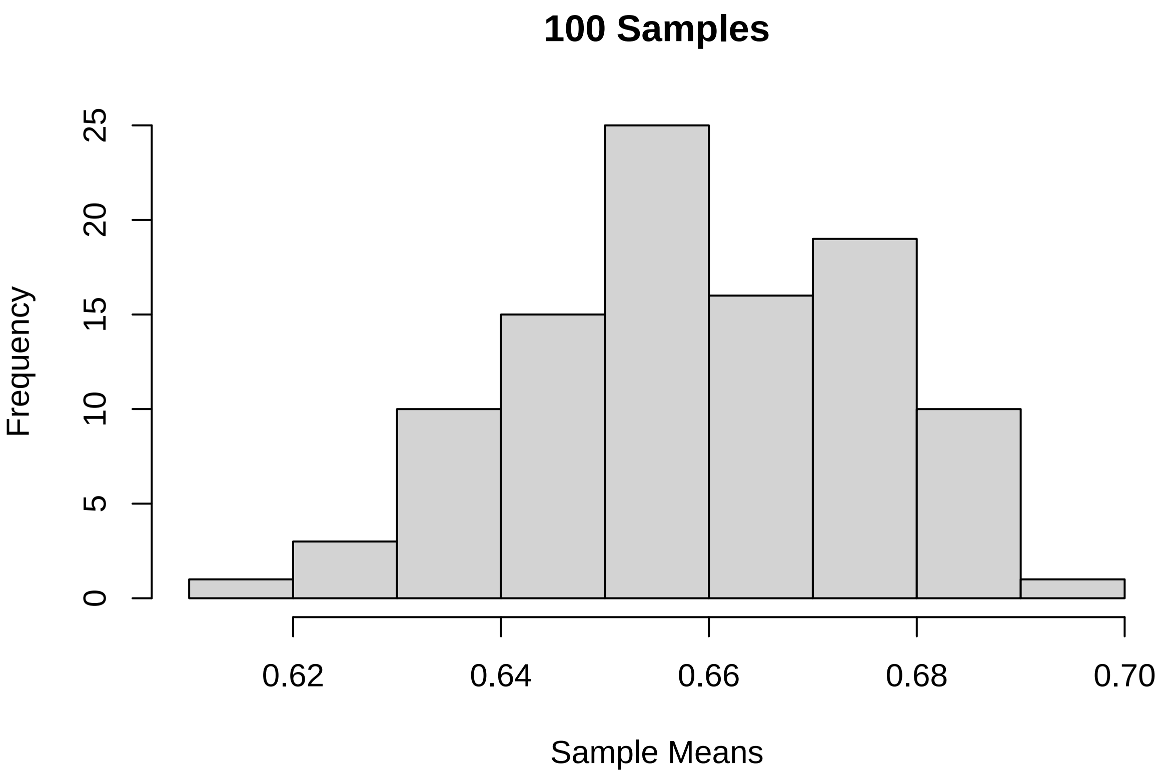
Varying the number of samples

In the following, we vary the number of samples we take with 100, 1,000, and 5,000 to support our answer experimentally.

Note: We can execute the following with a `for` loop instead of using multiple code blocks. But since it is not a programming course, we will be happy with executing it multiple times.

```r

number_of_samples <- 100
sample_means <- replicate(number_of_samples, table(sample(population, 1000))["success"])
sample_means <- sample_means / 1000
hist(sample_means, main="100 Samples", xlab="Sample Means")
```



100 Samples

```
print(paste("The mean with", number_of_samples, "is", mean(sample_means)))
```

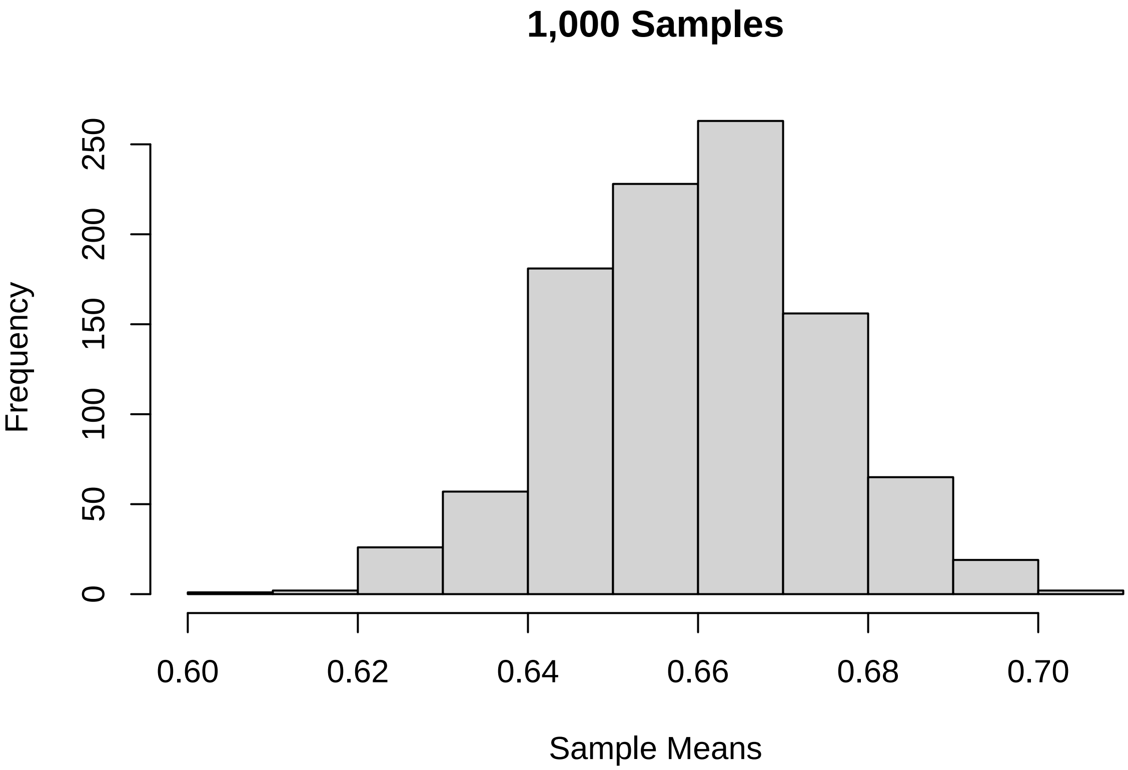
```
## [1] "The mean with 100 is 0.65976"
```

```
print(paste("The error, point estimate - population parameter is", mean(sample_means) - population_prop
```

```
## [1] "The error, point estimate - population parameter is -0.000519999999999965"
```

1,000 Samples

```
number_of_samples <- 1000
sample_means <- replicate(number_of_samples, table(sample(population, 1000))["success"])
sample_means <- sample_means / 1000
hist(sample_means, main="1,000 Samples", xlab="Sample Means")
```



```
print(paste("The mean with", number_of_samples, "is", mean(sample_means)))
```

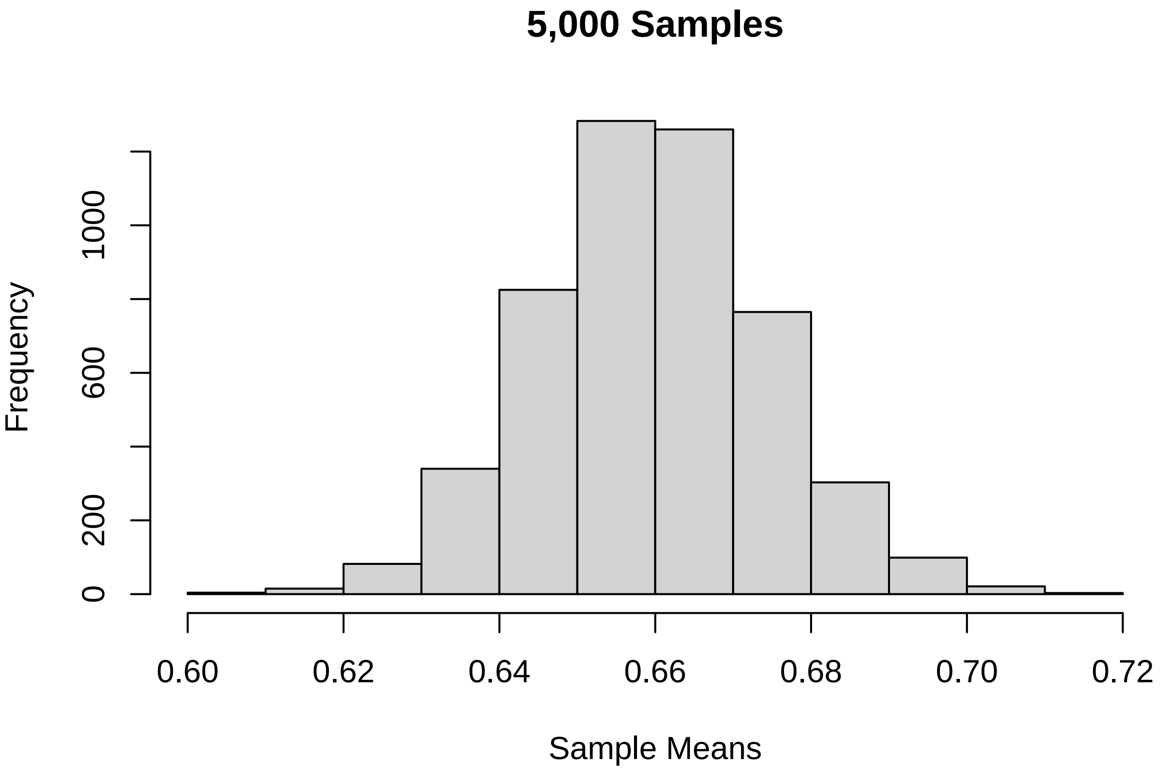
```
## [1] "The mean with 1000 is 0.660207"
```

```
print(paste("The error, point estimate - population parameter is", mean(sample_means) - population_prop
```

```
## [1] "The error, point estimate - population parameter is -7.29999999999897e-05"
```

5,000 Samples

```
number_of_samples <- 5000
sample_means <- replicate(number_of_samples, table(sample(population, 1000))["success"])
sample_means <- sample_means / 1000
hist(sample_means, main="5,000 Samples", xlab="Sample Means")
```



```
print(paste("The mean with", number_of_samples, "is", mean(sample_means)))
```

```
## [1] "The mean with 5000 is 0.6602046"
```

```
print(paste("The error, point estimate - population parameter is", mean(sample_means) - population_prop
```

```
## [1] "The error, point estimate - population parameter is -7.53999999999477e-05"
```

Question

(a) What was the error with 100 samples?
answer:

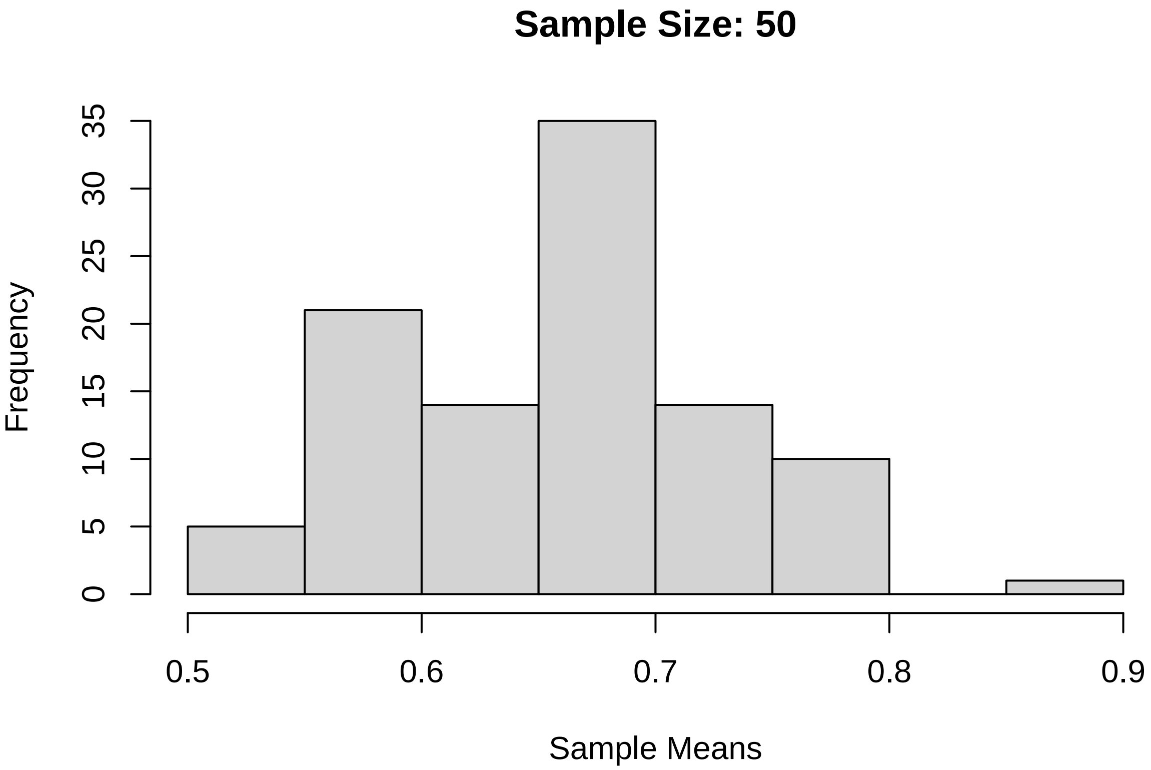
(b) Are 100 samples enough to estimate the population proportion? (open-ended question). answer:

Varying the sample size

Now, we fix the number of samples at 100 and vary the observations in each sample to 50, and 500.

50 observations in each sample

```
number_of_samples <- 100
number_of_observations_in_sample <- 50
sample_means <- replicate(number_of_samples, table(sample(population, number_of_observations_in_sample)
sample_means <- sample_means / number_of_observations_in_sample
hist(sample_means, main="Sample Size: 50", xlab="Sample Means")
```



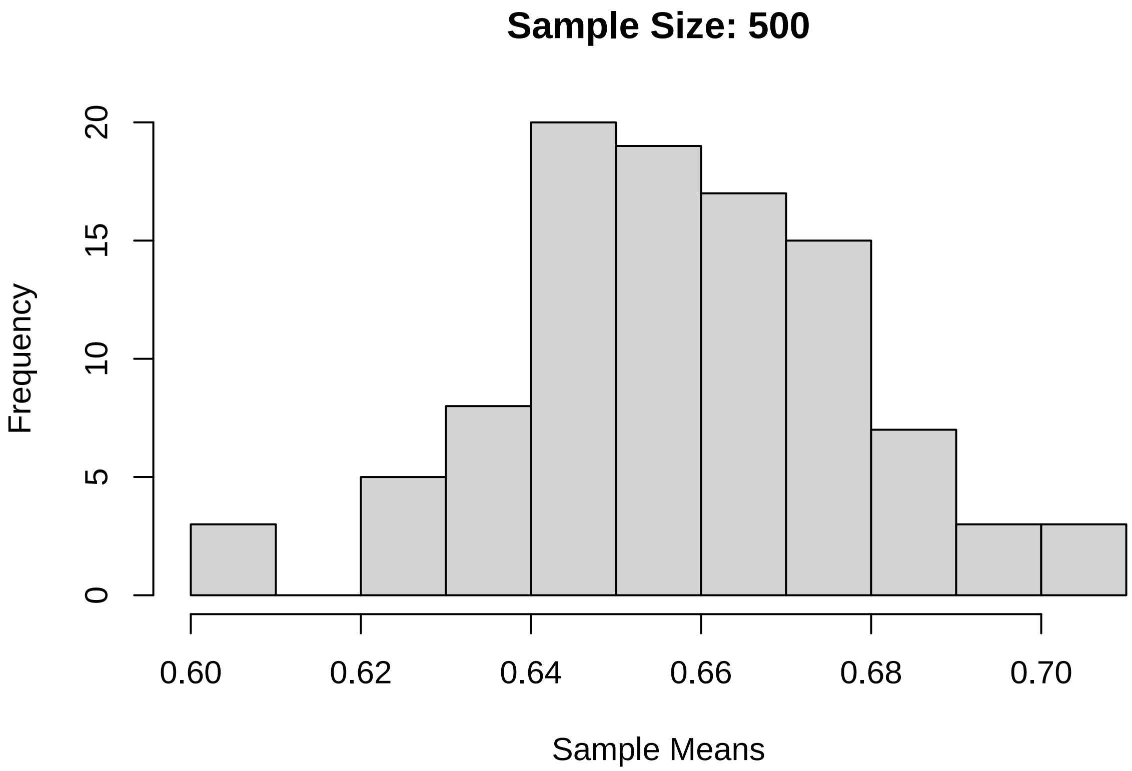
```r
print(paste("The mean with", number_of_samples, "is", mean(sample_means)))
```

```
## [1] "The mean with 100 is 0.6634"
```

```r
print(paste("The error, point estimate - population parameter is", mean(sample_means) - population_prop
```

```
## [1] "The error, point estimate - population parameter is 0.00312000000000001"
```

```r
number_of_samples <- 100
number_of_observations_in_sample <- 500
sample_means <- replicate(number_of_samples, table(sample(population, number_of_observations_in_sample)
sample_means <- sample_means / number_of_observations_in_sample
hist(sample_means, main="Sample Size: 500", xlab="Sample Means")
```



```
print(paste("The mean with", number_of_samples, "is", mean(sample_means)))
```

```
## [1] "The mean with 100 is 0.65884"
```

```
print(paste("The error, point estimate - population parameter is", mean(sample_means) - population_prop
```

```
## [1] "The error, point estimate - population parameter is -0.00144"
```

Question:

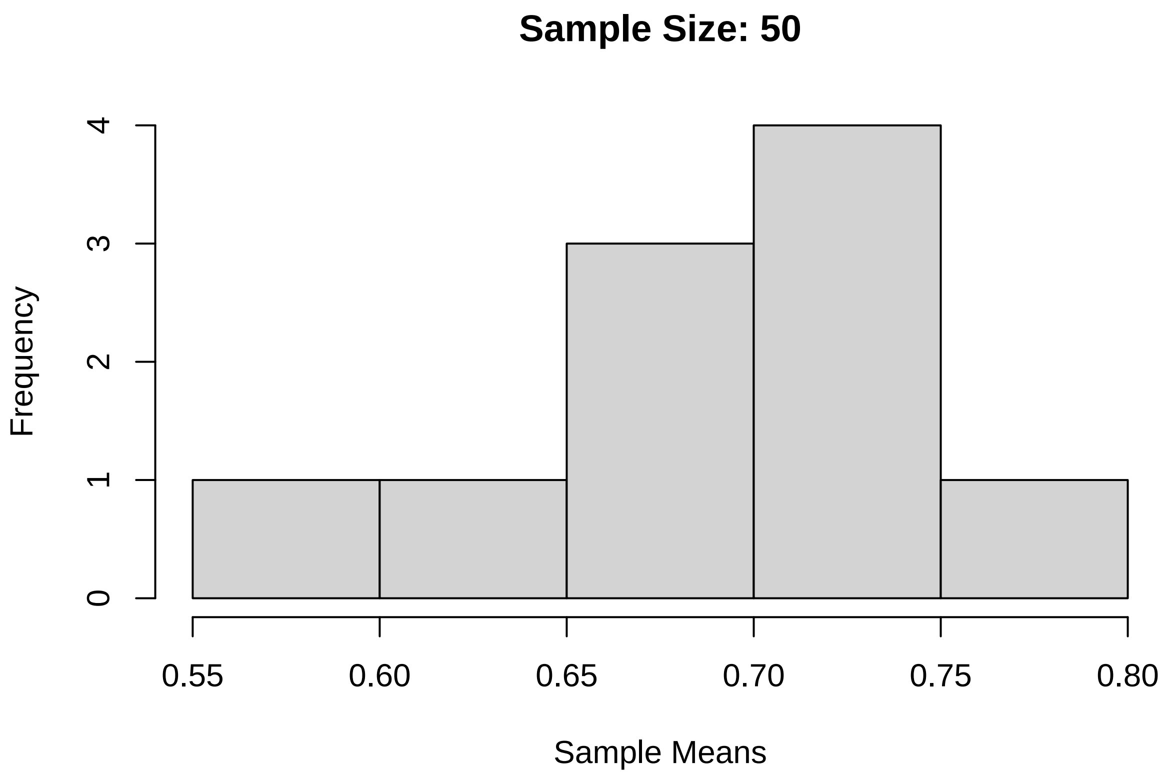
(a) Did 50 observations in the samples provide a good estimate? (open-ended question)

(b) Or did the 500-observation simulation provide a better result than the 50-observation case? (open-ended question)

Task

In the next code block, choose your own number of samples and observations in samples. The goal is to choose the smallest number for the best approximation.

```
number_of_samples <- 10                     #change this
number_of_observations_in_sample <- 50     #change this
sample_means <- replicate(number_of_samples, table(sample(population, number_of_observations_in_sample))
sample_means <- sample_means / number_of_observations_in_sample
hist(sample_means, main="Sample Size: 50", xlab="Sample Means")
```



```
print(paste("The mean with", number_of_samples, "is", mean(sample_means)))
```

```
## [1] "The mean with 10 is 0.698"
```

```
print(paste("The error, point estimate - population parameter is", mean(sample_means) - population_prop
```

```
## [1] "The error, point estimate - population parameter is 0.03772"
```

Numerical variables

The population proportion in the example above was based on a categorical variable. The central limit theorem applies to numerical variables too. Next time we will an example.

Using one sample

By the central limit theorem, we can estimate the population parameter by sample statistics. In reality, it is not easy to obtain several samples (for instance, conducting multiple surveys cost a lot of money and effort). Often, we want to deduce the best estimate based on a single sample.

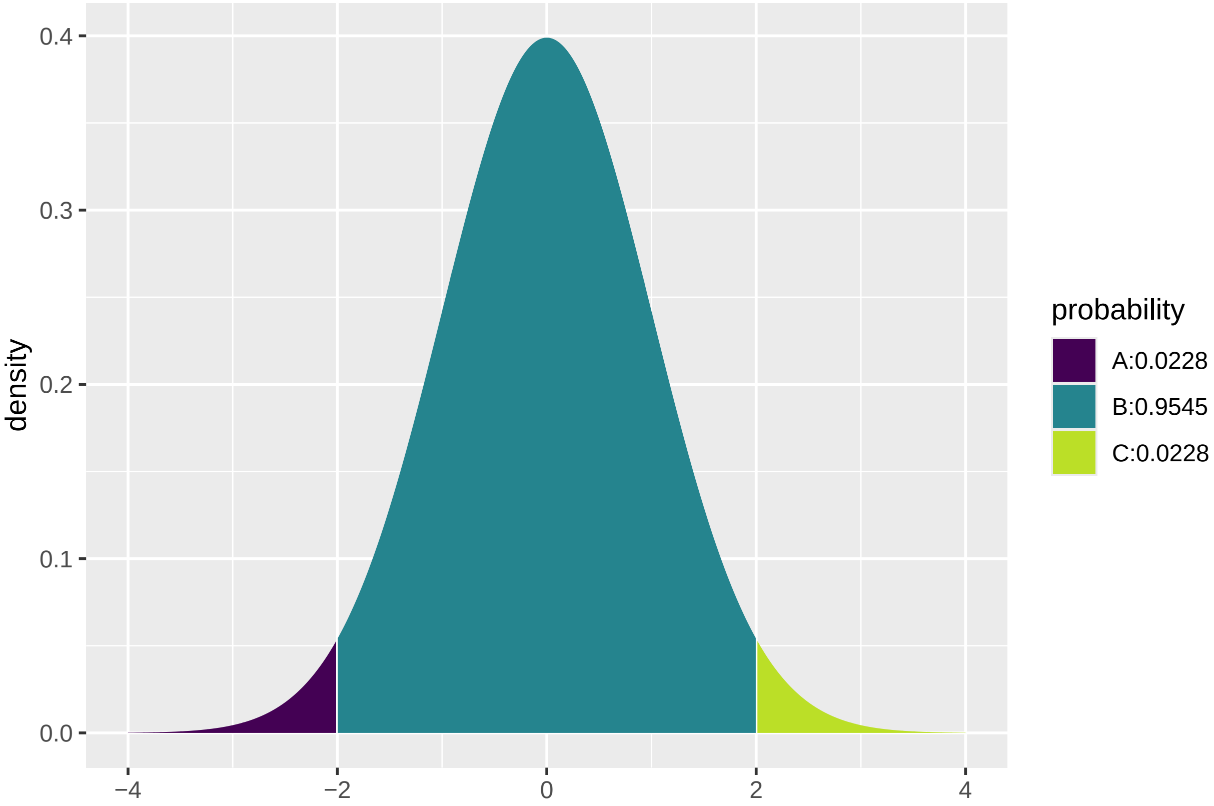
A point estimate is a sample statistic. In the example above, the sample proportion $\hat{p}$ is a point estimate. Point estimates is an estimate and almost never the population parameter which we are interested in. Hence it is often more reasonable making an interval around the point estimate. This interval is called the **confidence interval**.

The most commonly used interval is 95%. Recall for the standard normal distribution $P(-2 < Z < 2) \approx 0.9544$ which is slightly over 95%. For 95%, -1.96 and 1.96 will do the job.

```
xpnorm(c(-2,2))
```

```
##
```

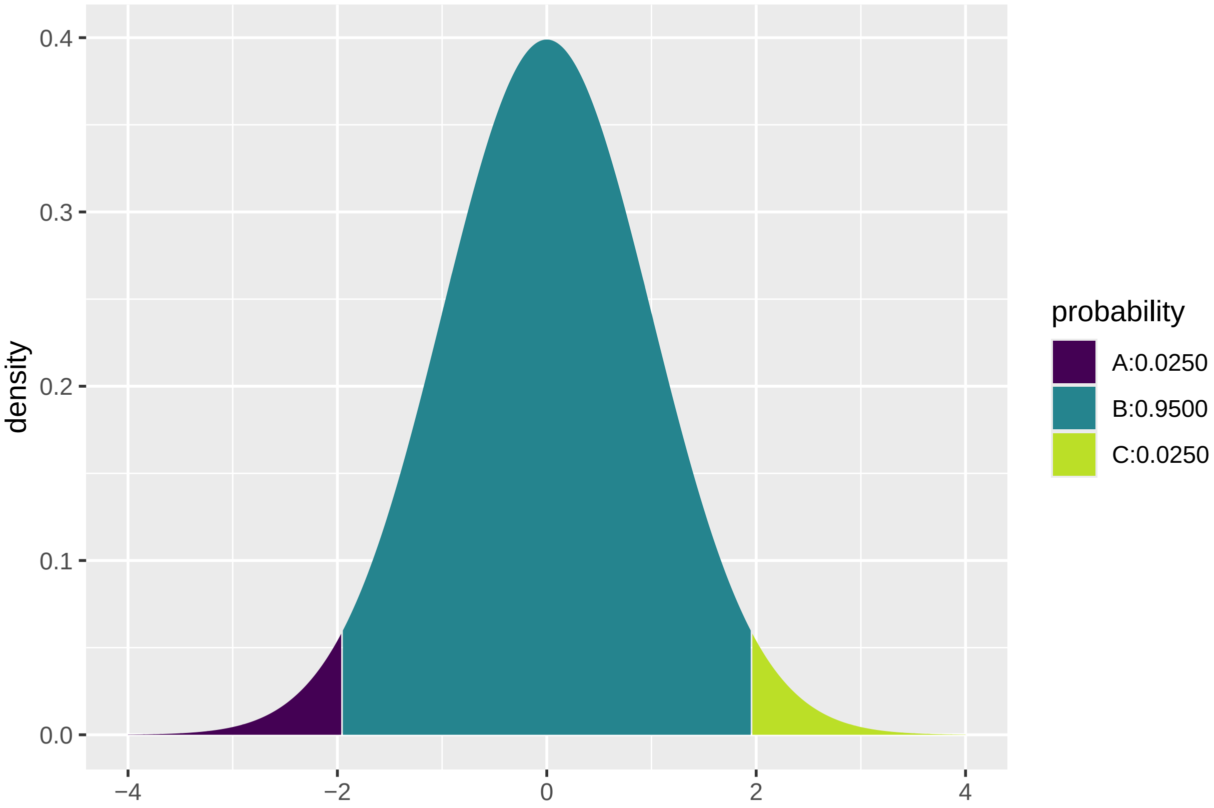
```
## If X ~ N(0, 1), then
##  P(X <= -2) = P(Z <= -2) = 0.02275   P(X <=  2) = P(Z <=  2) = 0.97725
##  P(X >  -2) = P(Z >  -2) = 0.97725   P(X >   2) = P(Z >   2) = 0.02275
##
```



```
## [1] 0.02275013 0.97724987
```

```
xpnorm(c(-1.96,1.96))
```

```
##
## If X ~ N(0, 1), then
##  P(X <= -1.96) = P(Z <= -1.96) = 0.025   P(X <=  1.96) = P(Z <=  1.96) = 0.975
##  P(X >  -1.96) = P(Z >  -1.96) = 0.975   P(X >   1.96) = P(Z >   1.96) = 0.025
##
```



```
## [1] 0.0249979 0.9750021
```

In short, the higher the confidence level, the wider the confidence interval. That is, the 99% confidence interval is wider than the 95% confidence interval. The 99% interval does not necessarily mean better than the 95% interval. For instance, 99% of time people arrive at a party within 4 hours of the starting time is less useful then 95% of time people arrive at a party within 2 hours especially when the party last for 3 hours.

Remember, when we were given data, we used the Z-score to convert it to the standard normal distribution scale.

$$Z = \frac{x-\mu}{\sigma},$$

where μ is the mean and σ is the standard deviation. In the proportion setting, we use the *standard error* with the population proportion p is defined as

$$SE_p = \sqrt{\frac{p(1-p)}{n}},$$

where n is the sample size (the number of observations in a sample). Since we often do not know p, we use the point estimate $\hat{p}$ for p in the formula. This is called the **plug-in principle**. That is,

$$SE_p = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

For proportions, the standard error plays the role of the standard deviation, and it provides the confidence interval.

Task

Find the 95% confidence interval for the point estimate defined in the next code block. Here the number of observations in a sample is 100, so $n = 100$.

```
set.seed(8201)
my_sample <- sample(population, 100)
my_point_estimate <- table(my_sample)[["success"]]
p_hat <- my_point_estimate/100
p_hat
```

```
## [1] 0.69
```

This point estimate is stored in the variable `p_hat`. Use the plug-in principle for the standard error. In the next code blocks, follow these steps to find the 95% confidence interval.

1. Use `p_hat` and the plug-in principle to compute the standard error
2. Compute the end points of the interval using the formula

$$(\hat{p} - 1.96 \cdot SE, \hat{p} + 1.96 \cdot SE)$$

Step 1.

```
# Compute the standard error; the formula is around L233. For square root, use sqrt().
my_SE <- sqrt(p_hat*(1-p_hat)/100)
my_SE
```

```
## [1] 0.04624932
```

Step 2.

```
# Find the left and right end point of the interval
p_hat - 1.96 * my_SE
```

```
## [1] 0.5993513
```

```
p_hat + 1.95 * my_SE
```

```
## [1] 0.7801862
```

Question:

(a) Did the population proportion lie in the confidence interval?

(b) Change the seed number or delete the line and execute the above three code blocks a few times to get an example where the population proportion does not fall in the confidence interval.

It may be easier to copy the code below and execute them as a single code block.

```
# Use it if needed
```