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 condidence 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

This worksheet is to understand the phrase such as "95% OF MEANS PROJECTED TO FALL IN THIS RANGE." In a nutshell, we want to guess the `mean of the population` from the `mean of a sample` with, often, 95% confidence.

In the last worksheet, we created a population of size 50,000 consisting of "successes" and "fails" and took 2,000 samples of size 1,000 each. The task was to estimate the population proportion $p = 0.66028$ based on the sample proportions (often denoted by $\hat{p}$.

Summary: - Population proportion $p = 0.66028$ - Samples need not have the same proportion or the population proportion. - However, their distribution follows a normal distribution: - This is the consequence of the central limit theorem. - The mean of this distribution is close to the population proportion

Last time, we saw a question of "will the mean be closer to the population proportion and the standard deviation will be smaller if we take more samples". Be careful, there are two sizes:

- Size of each sample
- Number of samples we take

We took 2,000 samples each of size 1,000. So the distribution of the sample proportions consist of 2,000 observations and its mean was an excellent estimate of the population proportion.

In the following, we vary the number of samples we take with 100, 500, 1,000, and 5,000 to support our answer experimentally.

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

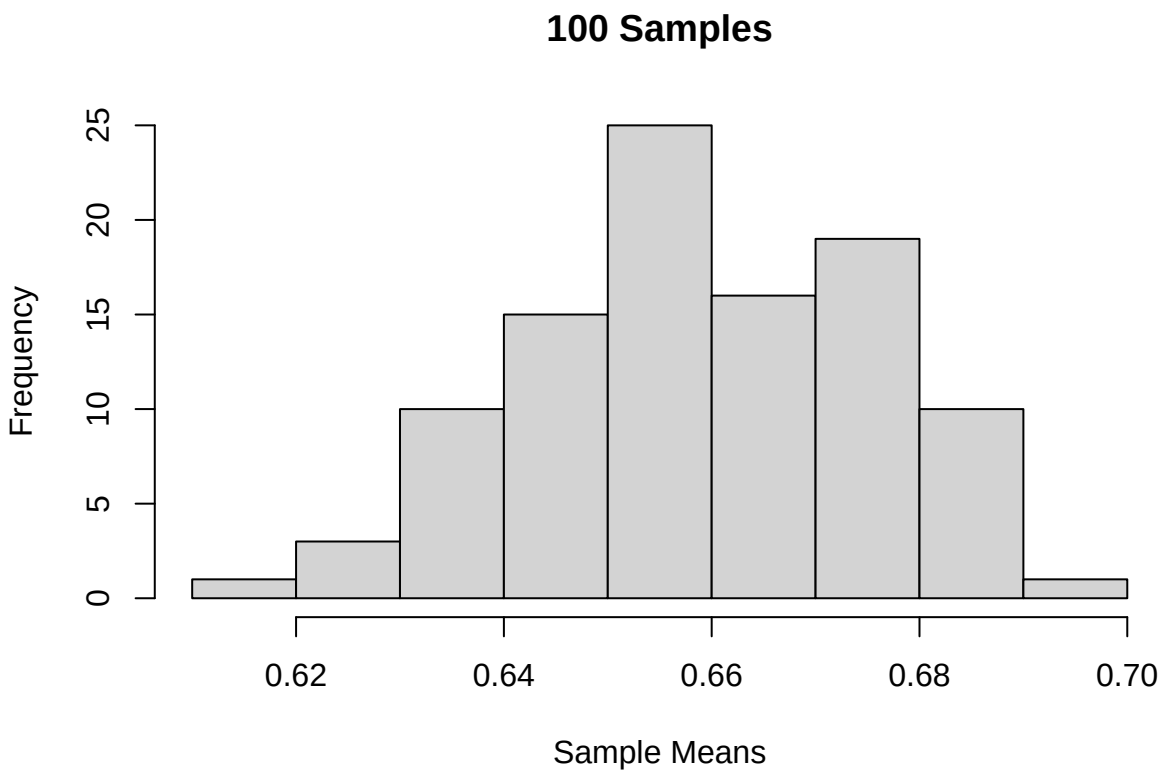
```
# the first two lines randomize the probability
set.seed(1105)
probabilities <- runif(2)
probabilities <- probabilities / sum(probabilities)

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

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

Varying the number of sample

100 Samples

```
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")
```

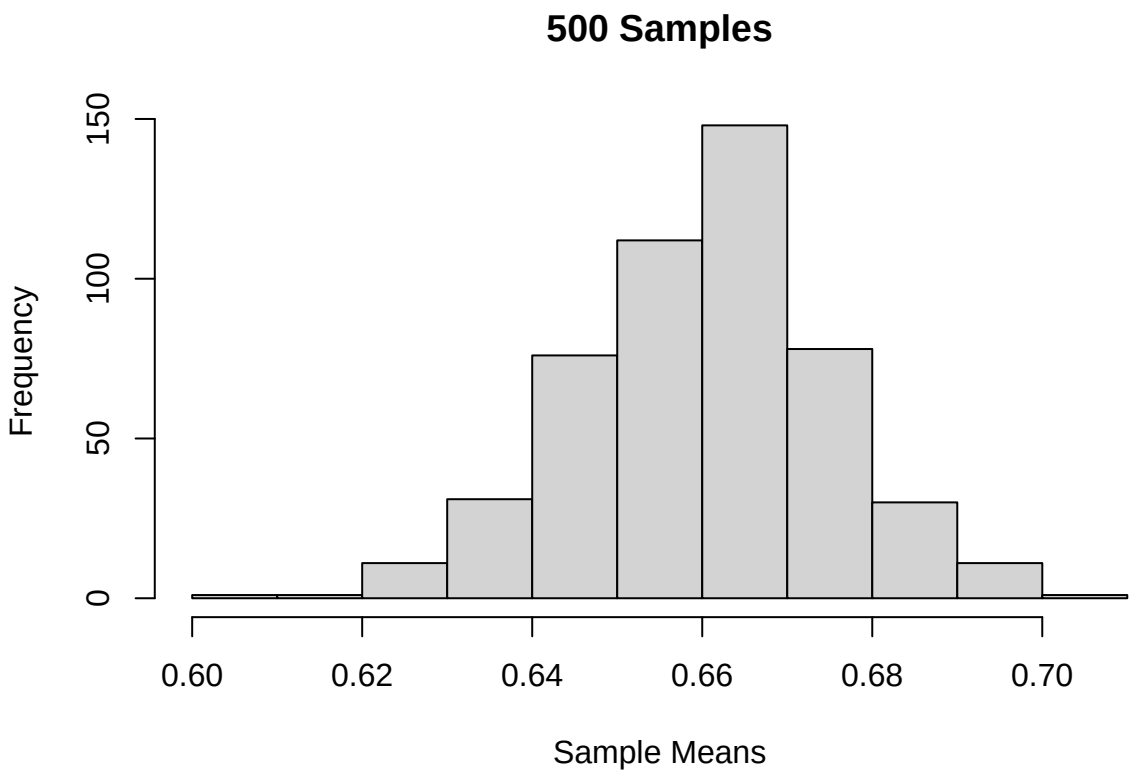



```
print(paste("Mean with", number_of_samples, "is", mean(sample_means)))
```

```
## [1] "Mean with 100 is 0.65976"
```

500 Samples

```
number_of_samples <- 500
sample_means <- replicate(number_of_samples, table(sample(population, 1000))["success"])
sample_means <- sample_means / 1000
hist(sample_means, main="500 Samples", xlab="Sample Means")
```

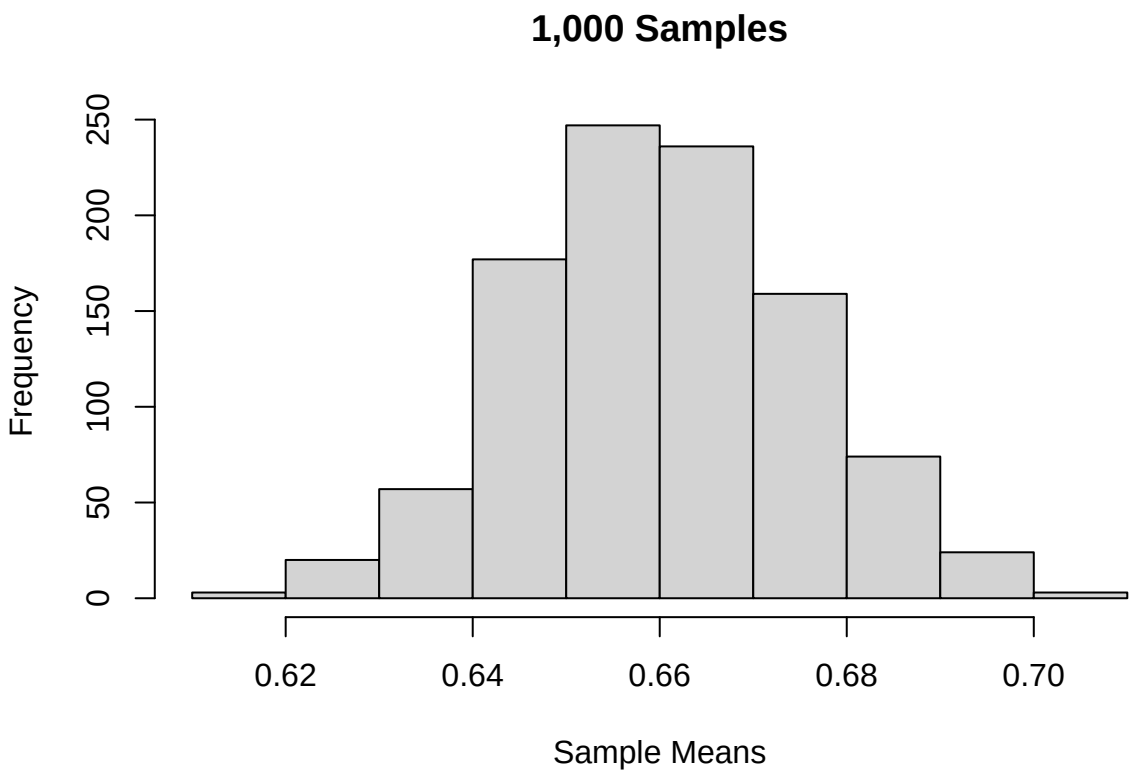



```
print(paste("Mean with", number_of_samples, "is", mean(sample_means)))
```

```
## [1] "Mean with 500 is 0.660816"
```

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")
```

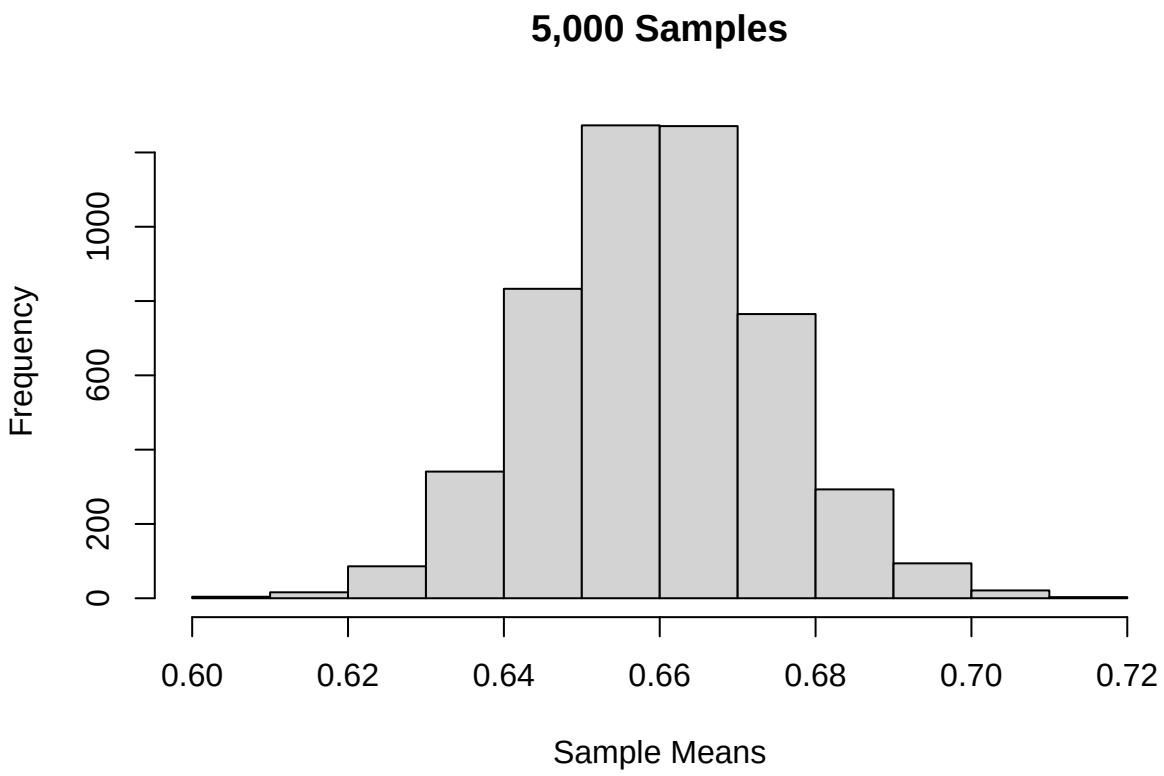



```
print(paste("Mean with", number_of_samples, "is", mean(sample_means)))
```

```
## [1] "Mean with 1000 is 0.660665"
```

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("Mean with", number_of_samples, "is", mean(sample_means)))
```

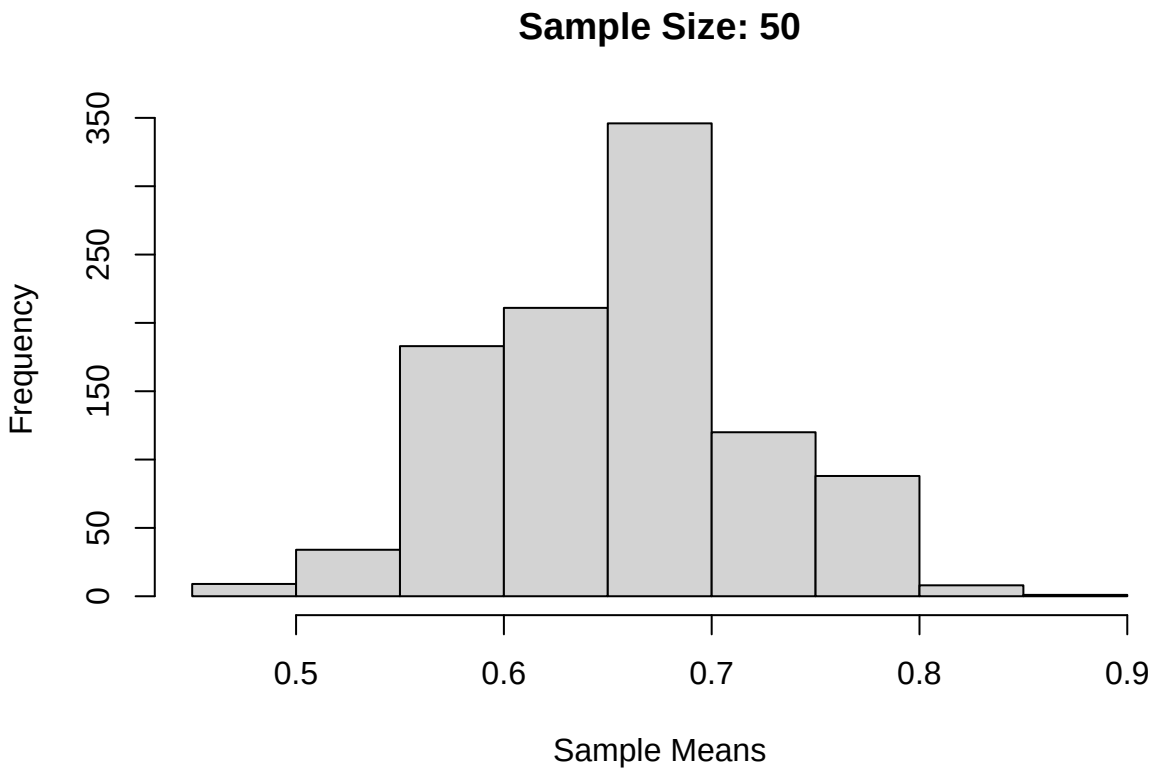
```
## [1] "Mean with 5000 is 0.6600696"
```

Varying the sample size

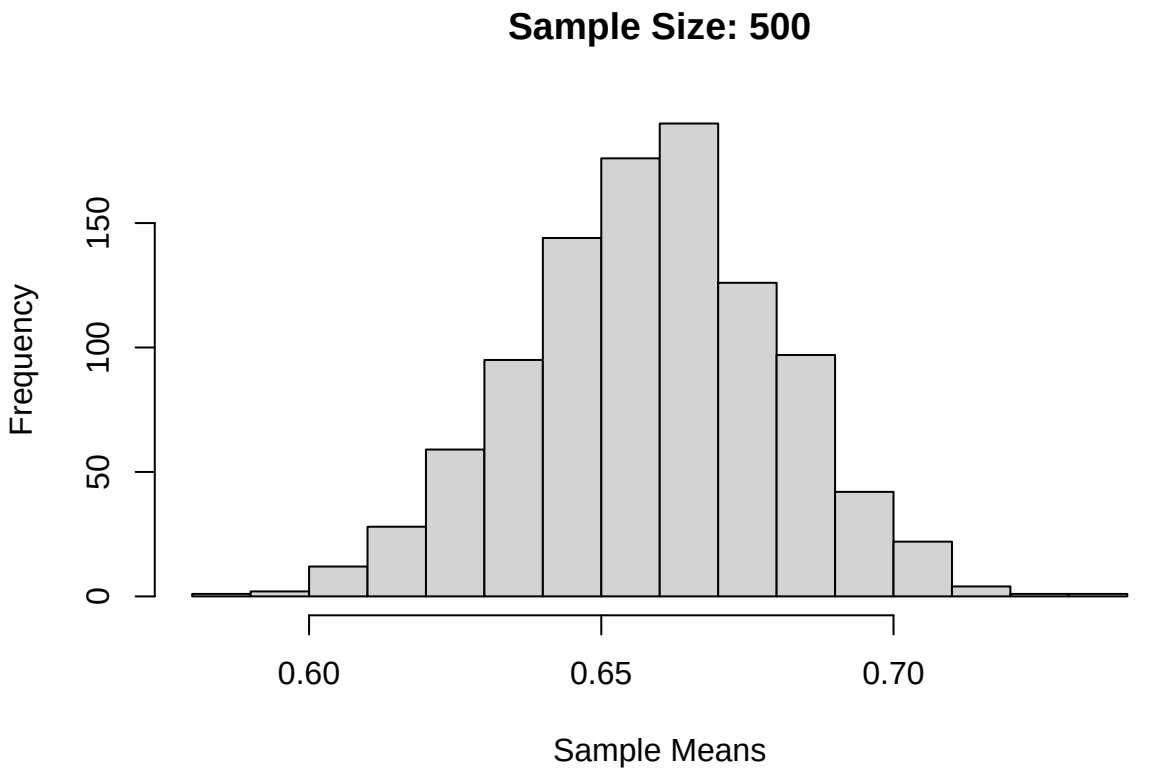
Now, we fix the number of samples at 1,000 and vary the observations in each sample to 50, 500, and 2,000.

50 observations in each sample

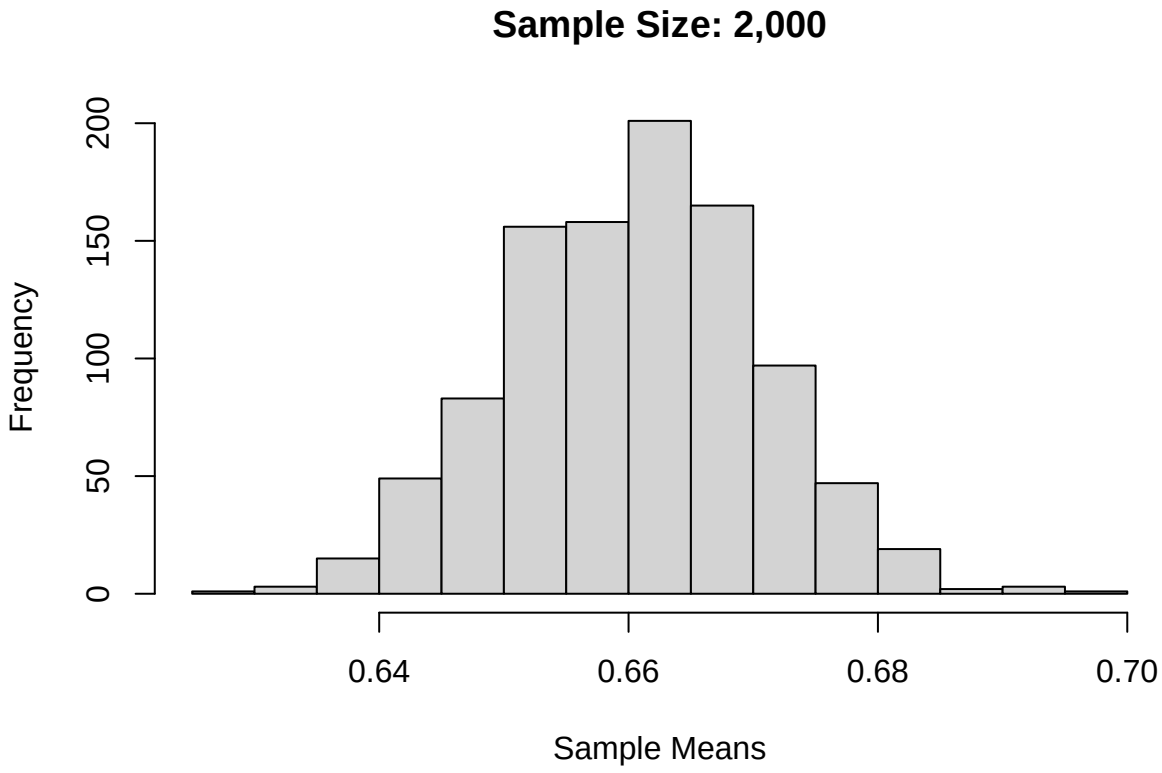
```
number_of_samples <- 1000
number_of_observations_in_sample <- 50
sample_means <- replicate(number_of_samples, table(sample(population, number_of_observations_in_sample)))
sample_means <- sample_means / number_of_observations_in_sample
hist(sample_means, main="Sample Size: 50", xlab="Sample Means")
```




```
number_of_samples <- 1000
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")
```



```r
number_of_samples <- 1000
number_of_observations_in_sample <- 2000
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: 2,000", xlab="Sample Means")
```



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

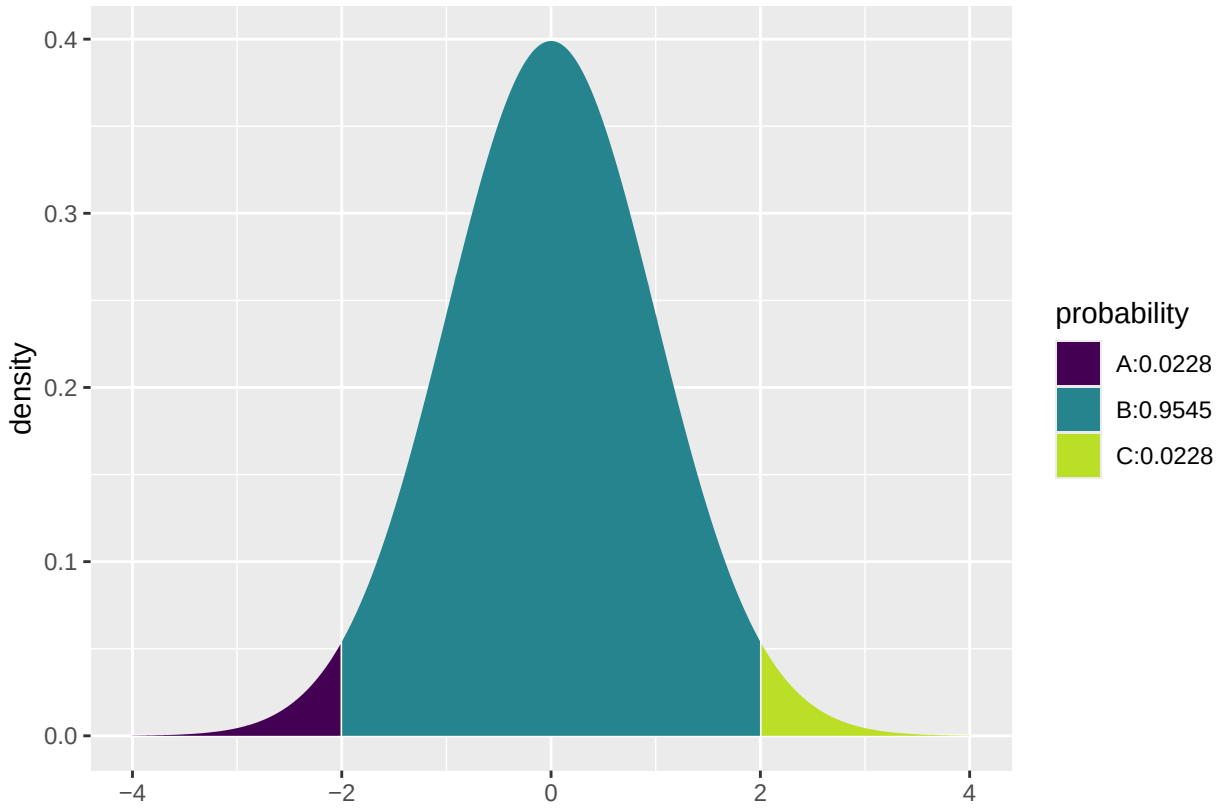
By the central limit theorem, we can estimate the population parameter by sample statistics. In reality, it is not easy to take several samples (for instance, conducting surveys cost a lot of money and effort).
Often, we rely on few samples if not one.

Point estimates are sample statistics. In the example above, the sample proportion $\hat{p}$ is a point estimate. Point estimates vary and almost never be the population parameter 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 often 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.

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

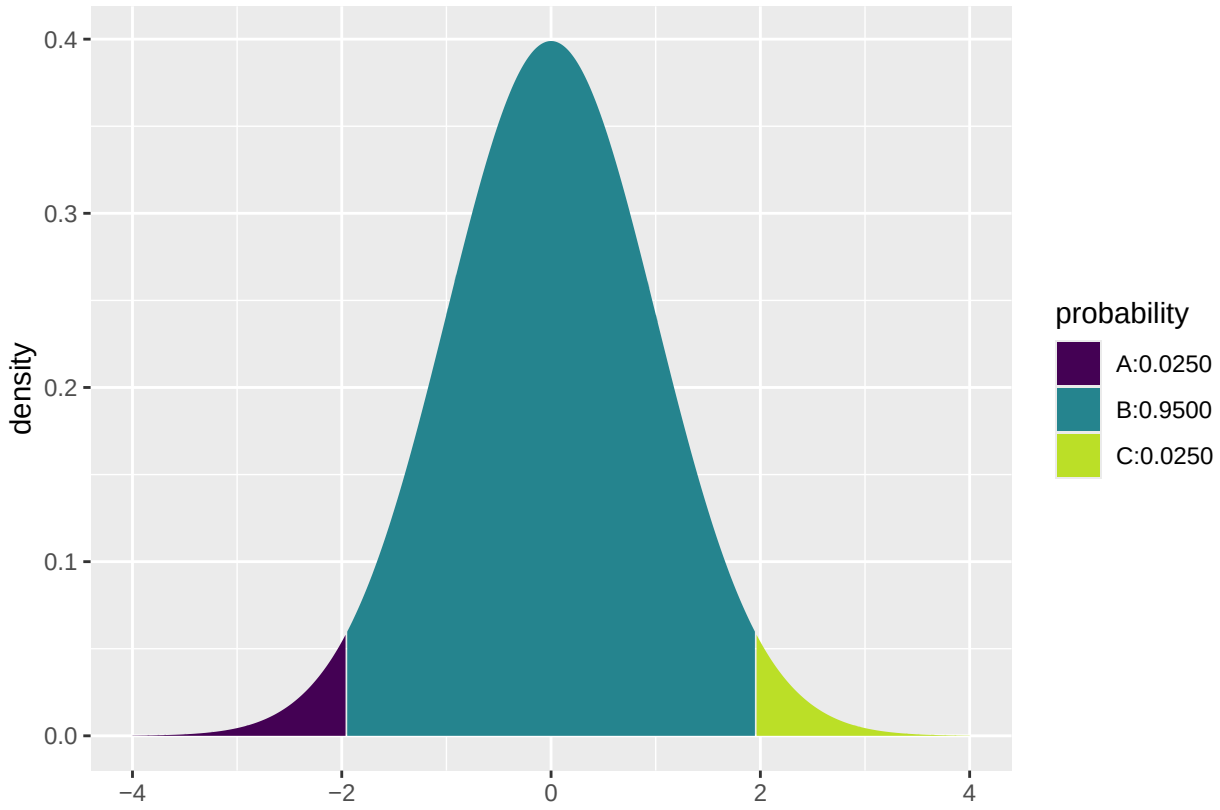
```
##
## If X ~ N(0, 1), then
##  P(X <= -2) = P(Z <= -2) = 0.02275   P(X <=  2) = P(Z <=  2) = 0.97725
##  P(X >  -2) = P(Z >  -2) = 0.97725   P(X >   2) = P(Z >   2) = 0.02275
##
```




```
## [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
```

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

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

Since we often do not know p, we use the point estimate $\hat{p}$. This is called the **plug-in principle**. That is,

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