Analysis of Psychological Data

Lab 5. Warming Up for Statistical Inference: Sampling Distribution of the Mean

Ihnwhi Heo (iheo2@ucmerced.edu)
Quantitative Methods, Measurement, and Statistics
Website: https://ihnwhiheo.github.io
Office: https://ucmerced.zoom.us/j/2093557522 (Thursday 3:30 - 5:30 pm)

Some announcements

Homework 3 is due on March 8 (Tuesday)

Don't forget to submit it on CatCourses

Exam 2 is on March 17 (Thursday)

What are we going to do?

Recap to give you a big picture

Sampling distribution of the mean

Do it together

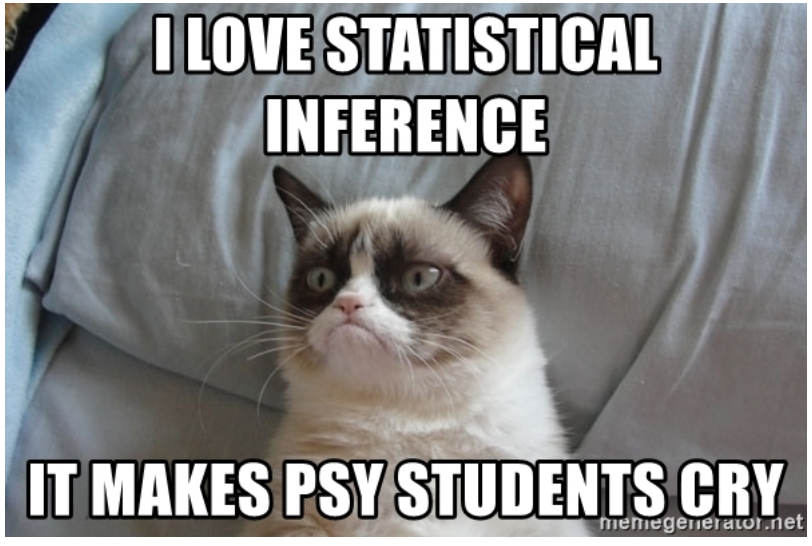
Q&A session for homework 3

Prelude

From now onwards, we do **inferential statistics** (*aka.* **statistical inference**)

Logic is crucial, so do not hesitate to ask me questions if anything is unclear!

Let's prove ourselves against this grumpy cat!



Statistical inference

Main idea

Let's make a best guess and test if that guess is true
$\longrightarrow$ Estimation and hypothesis testing

Where do we start?

Let's assume we are interested in one sample statistic (e.g., a sample mean)
and have the distribution of all the possible sample statistics
$\longrightarrow$ Sampling distribution

In this course, we are mainly interested in **a sample mean**
$\longrightarrow$ Sampling distribution of the mean

Statistical inference

Statement of probability

We speak of statistical inference in the statement of probability
$\longrightarrow$ If we have a normal distribution, we can calculate the area under the curve (i.e., proportion, probability) using z-scores

A closer look at the probability

When you flip a coin, what is the probability of the head (or the front side)?
What if you saw three heads in a row during the first three tosses?

Statistical inference

Repeated sampling

Probability is about after a large number of times

As the number of trials (infinitely) increases, the average outcome converges to the expected probability ~~(frequentist statistics)~~

Why important to take this into account?

Helpful to understand the precise idea of frequentist statistical inference
(e.g., confidence interval, *p*-value)

Statistical inference

Remember...

Let's make a best guess and test if that guess is true
$\longrightarrow$ Estimation and hypothesis testing

What are we interested in?

We are interested in one sample mean
$\longrightarrow$ We want to know the population mean using sample means
$\longrightarrow$ Sampling distribution of the mean

Central limit theorem

Why important?

We can apply probabilistic and statistical methods **that work for normal distributions** to problems involving other types of distributions

Characteristics

No matter what the original distribution looks like, the sampling distribution approximates a normal distribution if the sample sizes are larger enough

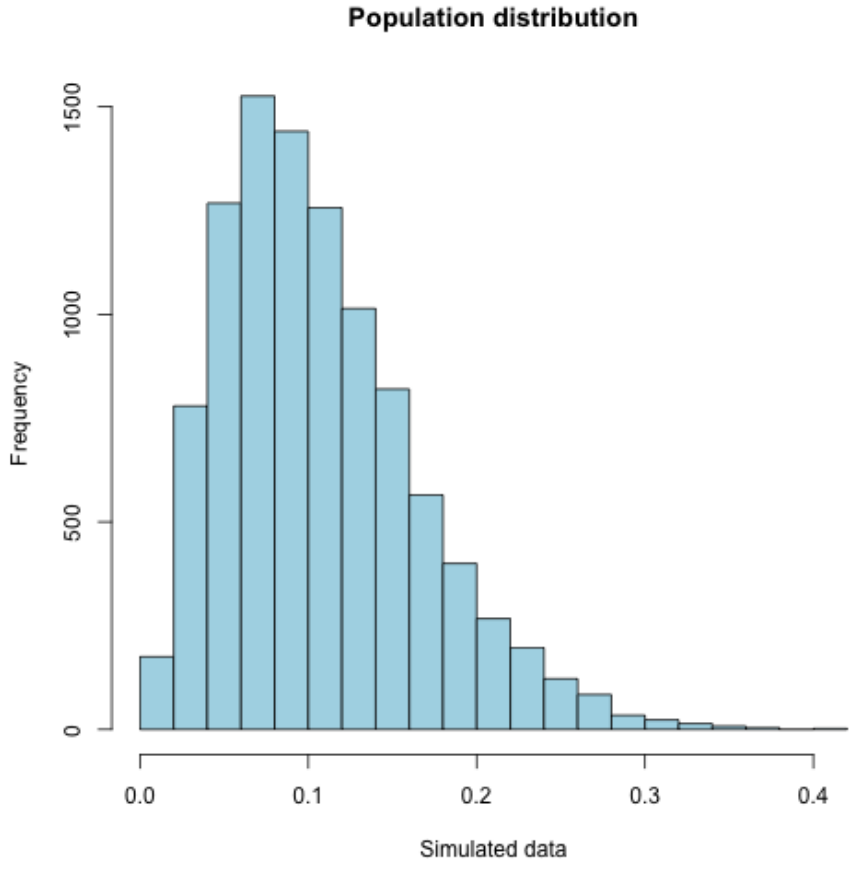
Code from https://www.analyticsvidhya.com/blog/2019/05/statistics-101-introduction-central-limit-theorem/

Central limit theorem

Your population is distributed as

```
set.seed(322) # Replicability
dat <- rbeta(10000, 3, 25) # Data simulation
hist(dat, col ="lightblue", main="Population distri
```

The population mean is 0.107.

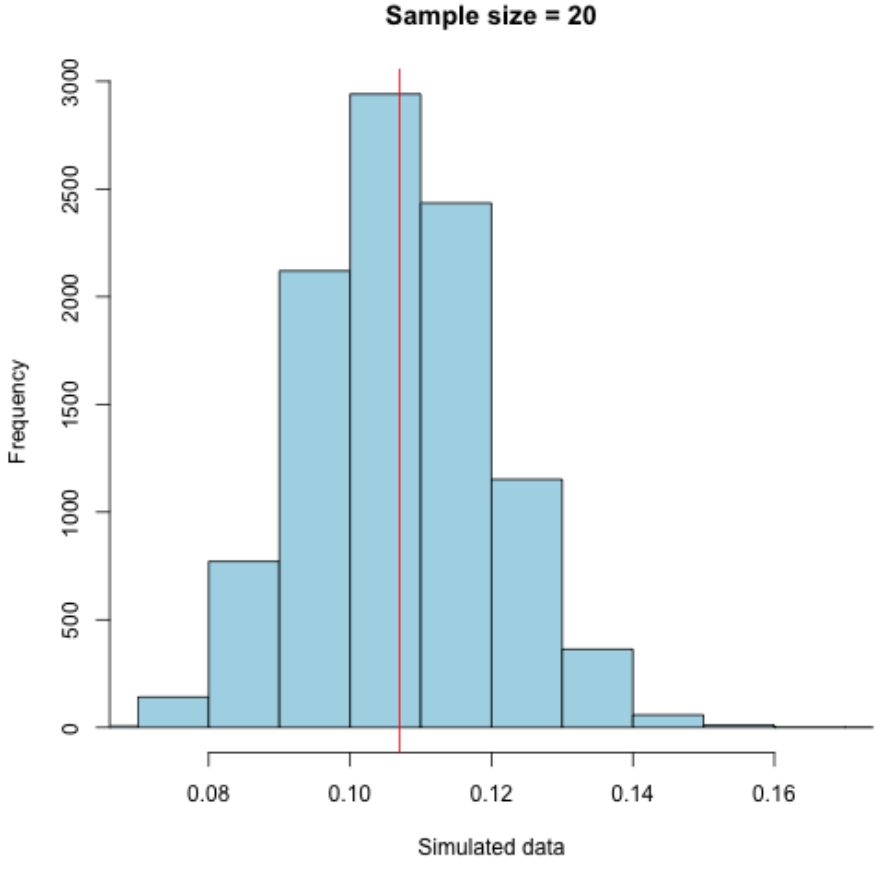


Central limit theorem

Collect a sample of 20 values, calculate the mean, and plot it... We repeat this 10000 times

```
set.seed(20210923) # Replicability
sample.20 <- c() # Empty vector
n=10000 # The number of iterations
for (i in 1:n) {
  sample.20[i] = mean(sample(dat,20, replace = TRUE
hist(sample.20, col ="lightblue", main="Sample size
abline(v = mean(sample.20), col = "Red")
```

The mean is 0.107.

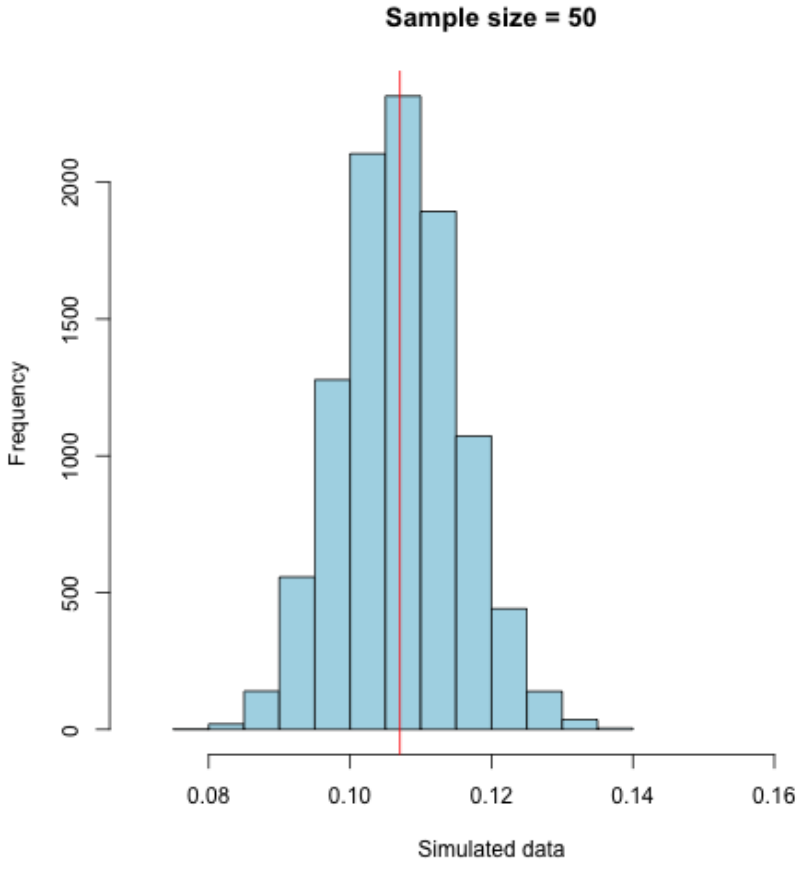


Central limit theorem

Collect a sample of 50 values, calculate the mean, and plot it... We repeat this 10000 times

```
set.seed(20210923) # Replicability
sample.50 <- c() # Empty vector
n=10000 # The number of iterations
for (i in 1:n) {
  sample.50[i] = mean(sample(dat,50, replace = TRUE
hist(sample.50, col ="lightblue", main="Sample size
abline(v = mean(sample.50), col = "Red")
```

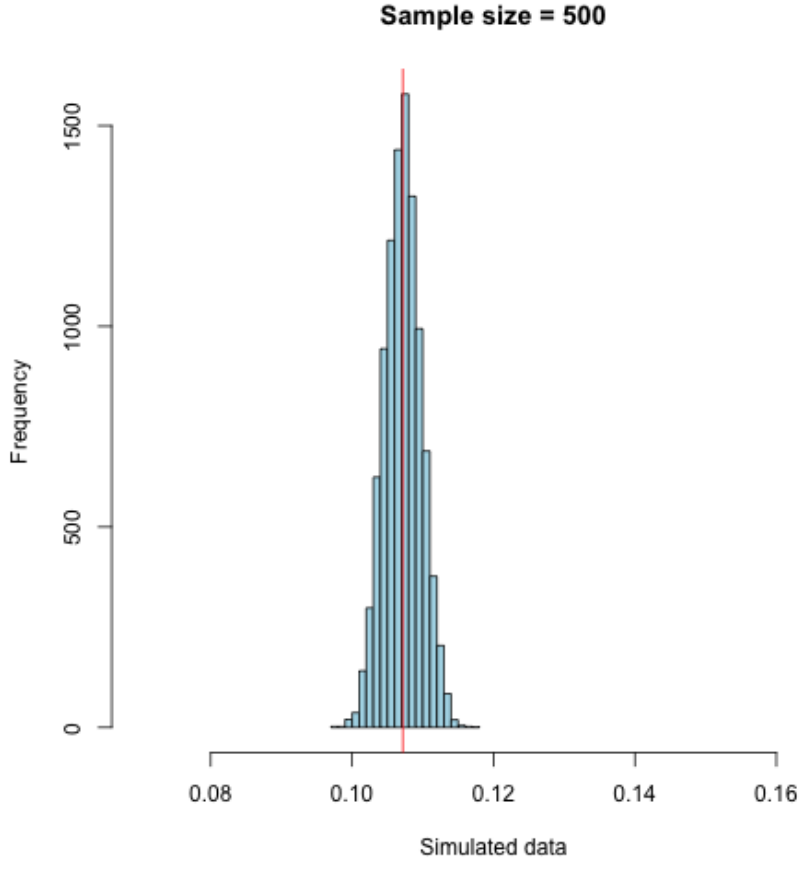
The mean is 0.107.



Central limit theorem

Collect a sample of 500 values, calculate the mean, and plot it... We repeat this 10000 times

```
set.seed(20210923) # Replicability
sample.500 <- c() # Empty vector
n=10000 # The number of iterations
for (i in 1:n) {
  sample.500[i] = mean(sample(dat,500, replace = TR
hist(sample.500, col ="lightblue", main="Sample siz
abline(v = mean(sample.500), col = "Red")
```



The mean is 0.107.

Sampling distribution of the mean

What is it?

A distribution of the sample means (i,e., collection of the means of possible samples)

What can we do?

Use the z-table since the sampling distribution is **normally distributed**

Calculate the probability of getting any specific mean from a random sample

Sampling distribution of the mean

Sampling error

Deviation of the (each) sample mean from the population mean

$$\mu_{\bar{X}} - \mu$$

Sampling distribution of the mean

Standard error of the mean

Standard deviation of the sampling distribution of the mean (spread of sampling error)

Typical distance that a sample mean deviates from the population mean

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Sampling distribution of the mean

Is normally distributed with...

Mean

$$\mu_{\bar{X}} = \mu$$

Standard deviation

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Sampling distribution of the mean

Some tips

Be careful which area under the curve (i.e., proportion, probability) is needed before looking at the z-table

Be careful whether you are using a raw score or a sample mean

Do it together

Q13 in HW3

A researcher records the following response times (in seconds) to a visual stimulus. Assuming these data are normally distributed, N(23, 3.1), what is the probability that participants responded in 20 seconds or less? (Note. round z score to 2 decimal places.)

- .1660
- .6660
- The probability is not listed in the unit normal table.
- .3340

Do it together

Q19 in HW3

A researcher selects a sample of 49 participants from a population with a mean of 12 and a standard deviation of 3.5. What is the probability of selecting a sample mean of 13 or larger from this population?

- equal to the probability of selecting a score above the mean
- greater than .31
- about one standard deviation below the mean
- less than .03

Q&A session for homework 3

Before you go home...

Lab materials are available at

https://github.com/lhnwhiHeo/PSY010

Any questions or comments?

Office hours or my email

Thanks! Have a good one!