Introduction to Statistical Inference (QTM 100 Lab)

Lecture 8: Sampling Distribution of the Mean and Inference for a Single Mean

Justin Eloriaga — Emory University

Fall 2024

Gameplan

Why should we care?

Data Preliminaries

Population vs Sampling Distribution of a Numeric Variable

One Sample t-test

The t distribution

Why should we care?

Recall

- Population Distribution
- Data Distribution
- **Sampling Distribution**
  - A probability distribution of a statistic obtained from a large number of samples drawn from a specific population

Recall

- Population Distribution
- Data Distribution
- **Sampling Distribution**
  - A probability distribution of a statistic obtained from a large number of samples drawn from a specific population

In lab 5, we looked at a categorical variable. Here, we will look at a numerical variable.

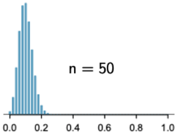
Why do we care about the Sampling Distribution

Why do we care about the Sampling Distribution

The sampling distribution *illustrates the importance of the sample size*

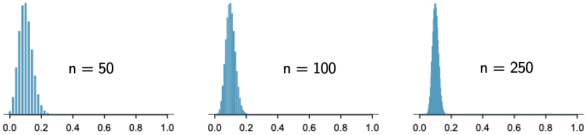
Why do we care about the Sampling Distribution

The sampling distribution *illustrates the importance of the sample size*



Why do we care about the Sampling Distribution

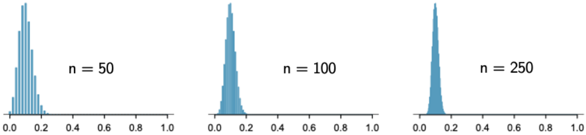
The sampling distribution *illustrates the importance of the sample size*



In real life, we don't observe the sampling distribution (we are lucky if we get one sample!) **BUT** we know that hypothetically, the sampling distribution of our statistic exists and our statistic falls somewhere within this distribution.

Why do we care about the Sampling Distribution

The sampling distribution *illustrates the importance of the sample size*



In real life, we don't observe the sampling distribution (we are lucky if we get one sample!) **BUT** we know that hypothetically, the sampling distribution of our statistic exists and our statistic falls somewhere within this distribution.

When the sample n is sufficiently large, the sampling distribution of the proportion is approximately normal.

Data Preliminaries

Using `yrbss2013` again

We revisit the Youth Risk and Morbidity survey data that we used in Week 5.

- Like before, we treat the *whole* dataset as the population (even if this is not really an ideal assumption).
- Use the usual `setwd()` and `read.csv()` functions to load the dataset

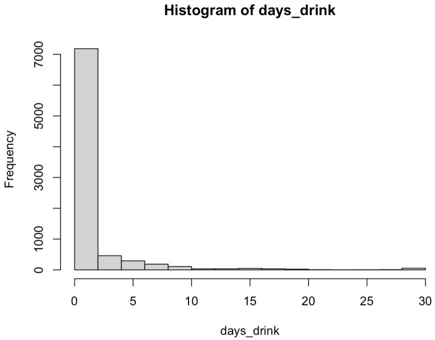
Population vs Sampling Distribution of a Numeric Variable

Looking at days_drink

Let's look at the distribution of days the students drank at least one drink of alcohol. This is done using the days_drink variable

```
days_drink <-
yrbss$days_drinks
summary(days_drink)
hist(days_drink)
```

What do you notice about the population distribution?



Creating a sample

We can obtain estimates of parameters such as the mean based on random samples.

```
samp_dd1 <- sample(x = days_drink, size = 50)
mean(samp_dd1)
```

Compare this to the population distribution. What do you see?

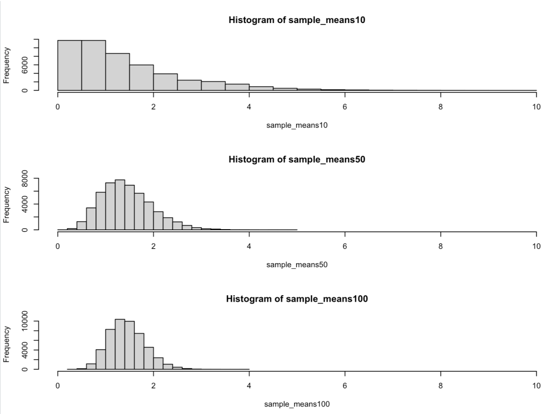
Doing sample loops again

We do the loop as we did before

```
 sample_means50 <- rep(NA, 50000)
#Creates an empty vector of 50000 lines
for(i in 1:50000){
samp_d <- sample(days_drink, 50)
#Creates a vector with 50 values from the
"days_drink" vector
sample_means50[i] <- mean(samp_d)
#Adds the mean of samp to the sample_means vector
}
hist(sample_means50)
```

More Loops

Try and do this for a sample size of 10 and 100. Try to do it in the same loop!



What can you tell about the distribution and the spread?

One Sample t-test

Do you do your course evals?

- Researchers are concerned about the validity of course evals because most students don't respond.
- Claim: There is an 80% response rate in course evaluations.
- Let's use the `cls_perc_eval` variable in the Course Evals dataset to answer this question

One-sample t-test

Our hypothesis are as follows:

$$H_0 : \mu = 80$$

$$H_a : \mu \neq 80$$

Use the `t.test()` function to implement this test

```
t.test(evals$cls_perc_eval, mu = 80)
```

```
	One Sample t-test

data:  evals$cls_perc_eval
t = -7.1555, df = 462, p-value = 3.294e-12
alternative hypothesis: true mean is not equal to 80
95 percent confidence interval:
 72.89749 75.95808
sample estimates:
mean of x
 74.42779
```

What do we conclude?

Modifications to the t-test

We can specify the confidence interval using the `conf.level` option

```
t.test(evals$cls_perc_eval, mu = 80, conf.level =
0.90)
```

Alternatively, we can do a one-sided test using the `alternative` option.

```
t.test(evals$cls_perc_eval, mu = 80, alternative =
"less")
```

The t distribution

Overview

- We used `pnorm()` and `qnorm()` to calculate probabilities from the normal.
- The probabilities can be used to calculate p-values based on a test statistic, and the quantiles can be used to identify t-scores for confidence intervals of a certain level.
- By default, the `t` functions utilize lower tail areas.

Using `pt()` for a two sided test

Suppose we performed a one sample t-test with a two sided H_a with 50 degrees of freedom and a test-statistic of $t = -2$.

```
2*pt(-2,df = 50)
```

The p-value for this test would be given by twice the lower tail area under the curve.

- Area under the curve less than -2 for a t distribution with 50 degrees of freedom (then multiplies by 2 to yield a p-value for a two-sided H_a of 0.0509).
- Try `2*(1-pt(2,df = 50))` and `2*pt(2,df = 50, lower.tail = F)`. All should yield the same thing by **symmetry**!

Using `qt()` to get t-scores for a confidence interval

- For a 95% confidence interval, this would correspond to a lower tail area under the curve of 0.025.
- In general, for a specified α, use $\alpha/2$ as the lower tail area under the curve to calculate the t-score

```
qt(0.025,df = 50)
```

- Given $df = 50$, the quantile that corresponds to the 2.5th percentile is -2.01
- Equivalently, you could also calculate the 97.5th percentile to yield the positive t-score using `qt(0.975,df = 50)`