Week 5: CLT and SE of the Sample Mean Difference

Wenhao Jiang

Department of Sociology
New York University

October 7, 2022

Central Limit Theorem (Recap)

Definition

- The distribution of the sample means will be approximately normally distributed with sufficiently large sample sizes.

Definition

- The distribution of the sample means will be approximately normally distributed with sufficiently large sample sizes.

Intuition Behind CLT

- Suppose we are interested in American's attitudes towards abortion (support v. oppose to legal abortion). We want to estimate the proportion of the Americans who support legal abortion (p) using one sample.
  - Although we almost never draw multiple samples from the population in reality, let's imagine that we have abundant research budget and we can draw random samples for an large number (K) of times.
  - We draw random samples ($n = 1000$) for K times (e.g., $K = 2000$) from the population.

Intuition Behind CLT

- Suppose we are interested in American's attitudes towards abortion (support v. oppose to legal abortion). We want to estimate the proportion of the Americans who support legal abortion (p) using one sample.
  - Although we almost never draw multiple samples from the population in reality, let's imagine that we have abundant research budget and we can draw random samples for an large number (K) of times.
  - We draw random samples ($n = 1000$) for K times (e.g., $K = 2000$) from the population.
  - In each random sample, we calculate the proportion of individuals who support legal abortion $\hat{p}$. There will be K such $\hat{p}$.

Intuition Behind CLT

- Suppose we are interested in American's attitudes towards abortion (support v. oppose to legal abortion). We want to estimate the proportion of the Americans who support legal abortion (p) using one sample.
  - Although we almost never draw multiple samples from the population in reality, let's imagine that we have abundant research budget and we can draw random samples for an large number (K) of times.
  - We draw random samples ($n = 1000$) for K times (e.g., $K = 2000$) from the population.
  - In each random sample, we calculate the proportion of individuals who support legal abortion $\hat{p}$. There will be K such $\hat{p}$.
  - We stack all these $\hat{p}$ and make a histogram. The histogram will look like a bell curve

Intuition Behind CLT

- Suppose we are interested in American's attitudes towards abortion (support v. oppose to legal abortion). We want to estimate the proportion of the Americans who support legal abortion (p) using one sample.
  - Although we almost never draw multiple samples from the population in reality, let's imagine that we have abundant research budget and we can draw random samples for an large number (K) of times.
  - We draw random samples ($n = 1000$) for K times (e.g., $K = 2000$) from the population.
  - In each random sample, we calculate the proportion of individuals who support legal abortion $\hat{p}$. There will be K such $\hat{p}$.
  - We stack all these $\hat{p}$ and make a histogram. The histogram will look like a bell curve
  - The $\hat{p}$ that corresponds to the peak of the histogram will be p
  - The standard deviation of the K number of $\hat{p}$s will be $\sqrt{\frac{p(1-p)}{n}}$ (We also call it **standard error**)

Simulation

- Suppose the true population **parameter** $p = 0.52$, *i.e.*, the proportion of the population that support legal abortion is 0.52. Note that we never observe this population parameter.

```
## create population, 1 means support, 0 means oppose to legal abortion
population <- c(rep(0,4800000),rep(1,5200000))

## sample with n=1000
sample <- sample(population,1000)

## calculate sample mean
mean(sample)
```

```
## [1] 0.538
```

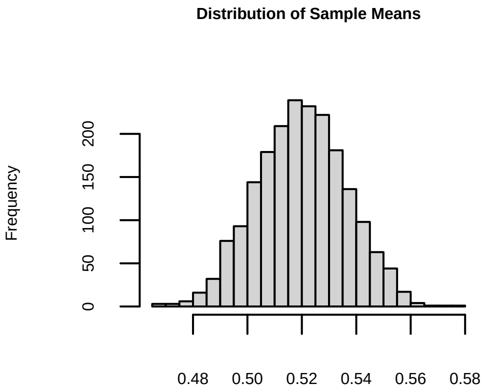
Simulation

- Now iterate this process for $K = 2000$ times

```
## empty vector to store means of sample (K times)
mean_of_sample <- c()

## iterate the process of 2000 times
for (i in 1:2000){
  ## sample with n=1000
  sample <- sample(population,1000)
  mean <- mean(sample)
  mean_of_sample <- c(mean_of_sample,mean)
}
```

Simulation - Normal Distribution of Sample Means

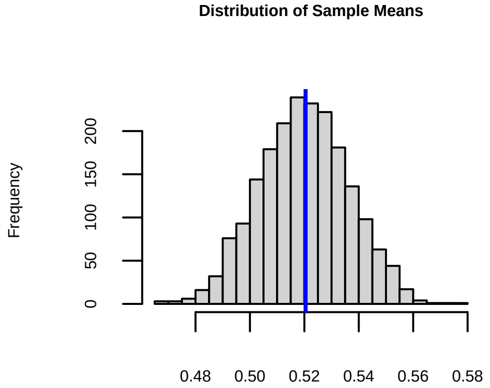
```
## plot histogram
hist(mean_of_sample,breaks=20,
     main="Distribution of Sample Means",
     xlab="Sample Mean",
     cex.lab=0.5, cex.axis=0.5, cex.main=0.5, cex.sub=0.5)
```



Simulation - Mean

- What is the mean of the K $\hat{p}$s?

```
## plot histogram
hist(mean_of_sample,breaks=20, main="Distribution of Sample Means", xlab="Sample M
     cex.lab=0.5, cex.axis=0.5, cex.main=0.5, cex.sub=0.5)
abline(v=mean(mean_of_sample),col="blue",lwd=2)
```



Simulation - Standard Error

- What is the standard deviation of K $\hat{p}$s?
- $\sqrt{\frac{p(1-p)}{n}}$

Simulation - Standard Error

- What is the standard deviation of K $\hat{p}$s?
- $\sqrt{\frac{p(1-p)}{n}}$
- We also call it **standard error** of $\hat{p}$

```
print(round(sqrt(0.52*(1-0.52)/1000),3))
```

```
## [1] 0.016
```

- Let's verify if this is correct.

```
round(sd(mean_of_sample),3)
```

```
## [1] 0.016
```

Simulation - Confidence Interval

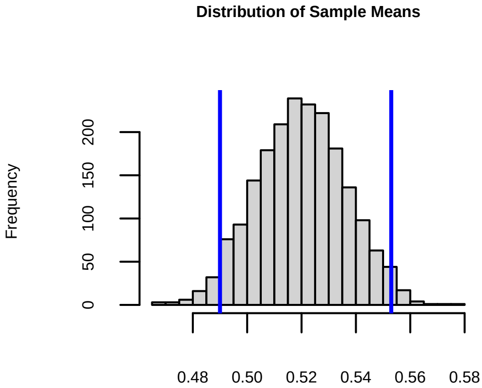
- In the normal distribution of sample means, 95% of the $\hat{p}$s (from the K samples) will fall into
- $p - 1.96 \times SE_{\hat{p}} \leq \hat{p} \leq p + 1.96 \times SE_{\hat{p}}$

Simulation - Confidence Interval

- In the normal distribution of sample means, 95% of the $\hat{p}$s (from the K samples) will fall into
- $p - 1.96 \times SE_{\hat{p}} \leq \hat{p} \leq p + 1.96 \times SE_{\hat{p}}$
- Equivalently, $\hat{p} - 1.96 \times SE_{\hat{p}} \leq p \leq \hat{p} + 1.96 \times SE_{\hat{p}}$
- This is the **Confidence Interval** for the population p

Simulation - Confidence Interval

```
## plot histogram
hist(mean_of_sample,breaks=20, main="Distribution of Sample Means", xlab="S
     cex.lab=0.5, cex.axis=0.5, cex.main=0.5, cex.sub=0.5)
abline(v=quantile(mean_of_sample,0.025),col="blue",lwd=2)
abline(v=quantile(mean_of_sample,0.975),col="blue",lwd=2)
```



More CLT

- The uncertainty of $\hat{p}$ is $SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$, where n is the sample size.
- Here p is the proportion of the **population** who support legal abortion (or other characteristics), which we never observe in the real world.

More CLT

- The uncertainty of $\hat{p}$ is $SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$, where n is the sample size.
- Here p is the proportion of the **population** who support legal abortion (or other characteristics), which we never observe in the real world.
- In real settings, instead of using p, we use $\hat{p}$ to estimate $SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

More CLT

- The uncertainty of $\hat{p}$ is $SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$, where n is the sample size.
- Here p is the proportion of the **population** who support legal abortion (or other characteristics), which we never observe in the real world.
- In real settings, instead of using p, we use $\hat{p}$ to estimate $SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.
- We then derive the **margin of error** (*e.g.*, $1.96 \times SE_{\hat{p}}$ if we want 95% confidence interval), and the 95% confidence interval of the population p (*e.g.*, $p \in [\hat{p} - 1.96 \times SE_{\hat{p}}, \hat{p} + 1.96 \times SE_{\hat{p}}]$).

Exercise

- In the 2018 GSS respondents were asked "Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if she wants one for any reason." Of 1,524 adults, 764 said "yes" and 760 said "no."
- We shall estimate the population proportion who would respond yes to this question (p).
- **What is the point estimate?**

Exercise

- In the 2018 GSS respondents were asked "Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if she wants one for any reason." Of 1,524 adults, 764 said "yes" and 760 said "no."
- We shall estimate the population proportion who would respond yes to this question (p).
- **What is the point estimate?**
  - $\hat{p} = 764/1524 = 0.501$
- **What is the SE of the point estimate?**

Exercise

- In the 2018 GSS respondents were asked "Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if she wants one for any reason." Of 1,524 adults, 764 said "yes" and 760 said "no."
- We shall estimate the population proportion who would respond yes to this question (p).
- **What is the point estimate?**
  - $\hat{p} = 764/1524 = 0.501$
- **What is the SE of the point estimate?**
  - $\sqrt{\frac{0.501\times(1-0.501)}{1524}} = 0.0128$
- **What is the margin of error for 95% Confidence Interval**

Exercise

- In the 2018 GSS respondents were asked "Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if she wants one for any reason." Of 1,524 adults, 764 said "yes" and 760 said "no."
- We shall estimate the population proportion who would respond yes to this question (p).
- **What is the point estimate?**
  - $\hat{p} = 764/1524 = 0.501$
- **What is the SE of the point estimate?**
  - $\sqrt{\frac{0.501 \times (1-0.501)}{1524}} = 0.0128$
- **What is the margin of error for 95% Confidence Interval**
  - $MOE_{\hat{p}} = 1.96 \times 0.0128 = 0.0251$
- **What is the 95% Confidence Interval for population p?**

Exercise

- In the 2018 GSS respondents were asked "Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if she wants one for any reason." Of 1,524 adults, 764 said "yes" and 760 said "no."
- We shall estimate the population proportion who would respond yes to this question (p).
- **What is the point estimate?**
  - $\hat{p} = 764/1524 = 0.501$
- **What is the SE of the point estimate?**
  - $\sqrt{\frac{0.501\times(1-0.501)}{1524}} = 0.0128$
- **What is the margin of error for 95% Confidence Interval**
  - $MOE_{\hat{p}} = 1.96 \times 0.0128 = 0.0251$
- **What is the 95% Confidence Interval for population p?**
  - $[0.501 - MOE_{\hat{p}}, 0.501 + MOE_{\hat{p}}] = [0.501 - 0.0251, 0.501 + 0.0251] = [0.4759, 0.5261]$

Difference between two Sample Means

Basics

- Now we extend the point estimate from one proportion to two proportions
- For example, we want to estimate the difference between the proportion of men and women supporting legal abortion, and how uncertain we are

Difference between two Sample Means

- Suppose the proportion of men in the population supporting abortion is p_m, the proportion of women in the population supporting abortion is p_w, and the difference is $p_m - p_w$

Difference between two Sample Means

- Suppose the proportion of men in the population supporting abortion is p_m, the proportion of women in the population supporting abortion is p_w, and the difference is $p_m - p_w$
- We draw a random sample from the population with n_m men and n_w women ($n_m + n_w$ in total).
- We get point estimates $\hat{p}_m$, $\hat{p}_w$, and $\hat{p}_m - \hat{p}_w$

Difference between two Sample Means

- Suppose the proportion of men in the population supporting abortion is p_m, the proportion of women in the population supporting abortion is p_w, and the difference is $p_m - p_w$
- We draw a random sample from the population with n_m men and n_w women ($n_m + n_w$ in total).
- We get point estimates $\hat{p}_m$, $\hat{p}_w$, and $\hat{p}_m - \hat{p}_w$
- What is the standard error of $\hat{p_m}$?
  - $SE_{\hat{p}_m} = \sqrt{\frac{p_m(1-p_m)}{n_m}}$, and we estimate it using $SE_{\hat{p}_m} = \sqrt{\frac{\hat{p}_m(1-\hat{p}_m)}{n_m}}$

Difference between two Sample Means

- What is the standard error of $\hat{p}_m - \hat{p}_w$?
  - $SE_{\hat{p}_m - \hat{p}_w} = \sqrt{\frac{p_m(1-p_m)}{n_m} + \frac{p_w(1-p_w)}{n_w}}$, and we estimate it using $SE_{\hat{p}_m - \hat{p}_w} = \sqrt{\frac{\hat{p}_m(1-\hat{p}_m)}{n_m} + \frac{\hat{p}_w(1-\hat{p}_w)}{n_w}}$
- Check textbook pp. 217

Difference between two Sample Means - Simulation

- Create a male population with $p_m = 0.48$ and a female population with $p_w = 0.56$. The "true" difference is $p_m - p_w = -0.08$

```
## create population, 1 means support, 0 means oppose to legal abortion
men <- c(rep(0,5200000),rep(1,4800000))
women <- c(rep(0,4400000),rep(1,5600000))

## draw a random sample, 1200 men and 800 women
sample_men <- sample(men, 1200)
sample_women <- sample(women, 800)

## estimate the proportion of men and women supporting abortion
mean(sample_men)
```

```
## [1] 0.4666667
```

```
mean(sample_women)
```

```
## [1] 0.53125
```

Difference between two Sample Means - Simulation

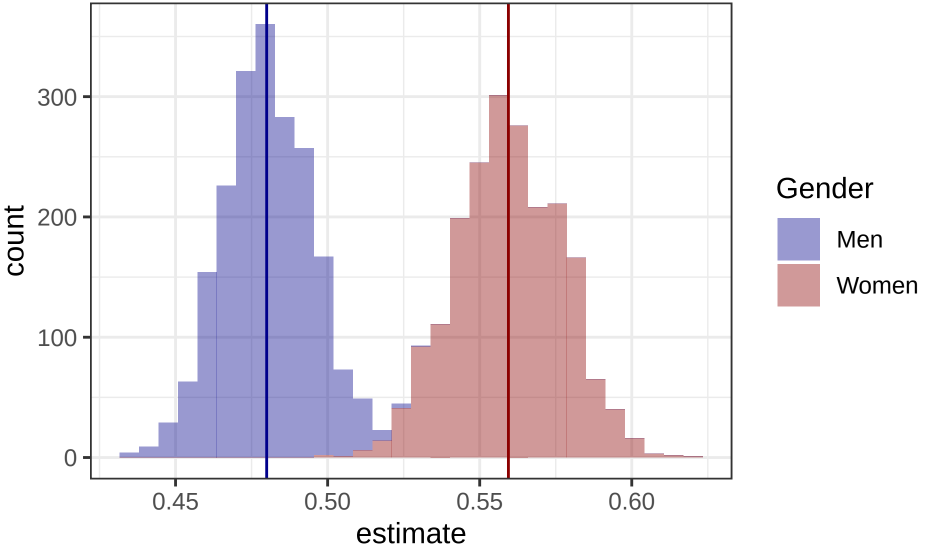
- Now iterate this process for $K = 2000$ times

```
## empty vector to store means of sample (K times)
mean_of_sample_men <- c()
mean_of_sample_women <- c()
mean_difference <- c()

## iterate the process of 2000 times
for (i in 1:2000){
  ## sample with nm=1200 and nw=800
  sample_men <- sample(men,1200)
  mean_men <- mean(sample_men)
  mean_of_sample_men <- c(mean_of_sample_men,mean_men)

  sample_women <- sample(women,800)
  mean_women <- mean(sample_women)
  mean_of_sample_women <- c(mean_of_sample_women,mean_women)

  sample_difference <- mean_men - mean_women
  mean_difference <- c(mean_difference,sample_difference)
}
```

Difference between two Sample Means - Simulation

- Now iterate this process for $K = 2000$ times

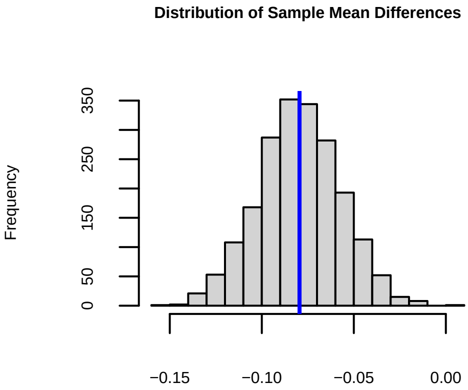
```
library(ggplot2)
dat <- data.frame(estimate = c(mean_of_sample_men, mean_of_sample_women),
                  sex = c(rep("men",2000),rep("women",2000)))
ggplot(dat,aes(x=estimate, fill = sex)) + geom_histogram(alpha = 0.4) +
  scale_fill_manual(name="Gender",values=c("darkblue","darkred"),labels=c("Men","W
  theme_bw() +
  geom_vline(xintercept = mean(mean_of_sample_men),color="darkblue") +
  geom_vline(xintercept = mean(mean_of_sample_women),color="darkred")
```


300
200
100
0
count
0.45
0.50
0.55
0.60
estimate
Gender
Men
Women

Difference between two Sample Means - Simulation

- What is the distribution of K times of $\hat{p}_m - \hat{p}_w$?

```
## plot histogram
hist(mean_difference,breaks=20, main="Distribution of Sample Mean Differences",
     xlab="Sample Mean Differences",
     cex.lab=0.5, cex.axis=0.5, cex.main=0.5, cex.sub=0.5)
abline(v=mean(mean_difference),col="blue",lwd=2)
```



Difference between two Sample Means - Simulation

- What is the standard error of $\hat{p}_m - \hat{p}_w$?

Difference between two Sample Means - Simulation

- What is the standard error of $\hat{p}_m - \hat{p}_w$?
- According to the formula:
  - $SE_{\hat{p}_m - \hat{p}_w} = \sqrt{\frac{p_m(1-p_m)}{n_m} + \frac{p_w(1-p_w)}{n_w}} = \sqrt{\frac{0.48(1-0.48)}{1200} + \frac{0.56(1-0.56)}{800}} = 0.023$
- See if the simulation returns the same result

```
round(sd(mean_difference),3)
```

```
## [1] 0.022
```

- Very close. We can expect the two to be the same if we keep sampling from the population ($K \rightarrow infinity$)

Difference between two Sample Means - Simulation

- What is the standard error of $\hat{p}_m - \hat{p}_w$?

Difference between two Sample Means - Simulation

- What is the standard error of $\hat{p}_m - \hat{p}_w$?
- According to the formula:
  - $SE_{\hat{p}_m - \hat{p}_w} = \sqrt{\frac{p_m(1-p_m)}{n_m} + \frac{p_w(1-p_w)}{n_w}} = \sqrt{\frac{0.48(1-0.48)}{1200} + \frac{0.56(1-0.56)}{800}} = 0.023$
  - Again, we never observe p_m and p_w, so we use $\hat{p}_m$ and $\hat{p}_w$ from one single sample to replace p_m and p_w in the above equation

Difference between two Sample Means - A Special Case

- What is the standard error of $\hat{p}_m - \hat{p}_w$ **under the null hypothesis** $H_0 : p_m = p_w = p$?
- In this **special case** of null hypothesis $H_0 : p_m = p_w = p$, we will replace p_m and p_w by a uniform p.
- $SE_{\hat{p}_m - \hat{p}_w} = \sqrt{\frac{p(1-p)}{n_m} + \frac{p(1-p)}{n_w}}$

Difference between two Sample Means - A Special Case

- What is the standard error of $\hat{p}_m - \hat{p}_w$ **under the null hypothesis** $H_0 : p_m = p_w = p$?
- In this **special case** of null hypothesis $H_0 : p_m = p_w = p$, we will replace p_m and p_w by a uniform p.
- $SE_{\hat{p}_m - \hat{p}_w} = \sqrt{\frac{p(1-p)}{n_m} + \frac{p(1-p)}{n_w}}$
- We do not observe p. We estimate p using the sample we drew.
- We denote this $\hat{p}$ in the case as $\hat{p}_{pooled}$, and $\hat{p}_{pooled} = \frac{\hat{p}_m \times n_m + \hat{p}_w \times n_w}{n_m + n_w}$

Difference between two Sample Means - A Special Case

- What is the standard error of $\hat{p}_m - \hat{p}_w$ **under the null hypothesis** $H_0 : p_m = p_w = p$?
- In this **special case** of null hypothesis $H_0 : p_m = p_w = p$, we will replace p_m and p_w by a uniform p.
- $SE_{\hat{p}_m - \hat{p}_w} = \sqrt{\frac{p(1-p)}{n_m} + \frac{p(1-p)}{n_w}}$
- We do not observe p. We estimate p using the sample we drew.
- We denote this $\hat{p}$ in the case as $\hat{p}_{pooled}$, and $\hat{p}_{pooled} = \frac{\hat{p}_m \times n_m + \hat{p}_w \times n_w}{n_m + n_w}$
- $SE_{\hat{p}_m - \hat{p}_w} = \sqrt{\frac{\hat{p}_{pooled}(1-\hat{p}_{pooled})}{n_m} + \frac{\hat{p}_{pooled}(1-\hat{p}_{pooled})}{n_w}}$

Logistics

- We will talk more about Standard Error, Confidence Interval, and Hypothesis Testing next week.
- If you find understanding the above concepts or process hard, please be sure to book an office hour with me next week.