PUBL0055: Introduction to Quantitative Methods

Lecture 8: Sampling, Uncertainty, and Confidence Intervals

Michal Ovádek and Indraneel Sircar

Motivation

The Motherhood Wage Penalty

Do women who have children tend to earn less? The data for this exercise came from the National Longitudinal Survey of Youth in the US which **samples** young people from the population to find out about their employment situations. We will use the data from this example to illustrate the importance of sampling in quantitative analysis.

- Unit of analysis: 2261 women aged between 19 and 30
- Dependent variable (Y): Hourly wage (measured in $s)
- Independent variable (X): 1 if the woman has at least 1 child, 0 otherwise

Difference in means recap

Let's calculate the difference in mean wages between mothers and non-mothers:

```
wage_mothers <- mean(motherhood$wage[motherhood$isMother == 1])
wage_not_mothers <- mean(motherhood$wage[motherhood$isMother == 0])
wage_mothers - wage_not_mothers
```

```
## [1] -0.4889398
```

Question: Is this a meaningful difference?

Is this a "meaningful" difference?

Two distinct approaches to answering this question:

1. Substantive answer: Consider the units in which it is measured.
    - Wage is measured in $ per hour, so mothers earn about 50 cents per hour less than non-mothers. Is that an important difference?

Is this a "meaningful" difference?

Two distinct approaches to answering this question:

1. Substantive answer: Consider the units in which it is measured.
    - Wage is measured in $ per hour, so mothers earn about 50 cents per hour less than non-mothers. Is that an important difference?
2. Statistical answer: Consider how likely it is that the difference observed in the *sample* is close to that which exists in the *population*.
    - Did we gather enough data to be confident that what we observed in the sample also applies to the population?

Is this a "meaningful" difference?

Two distinct approaches to answering this question:

1. Substantive answer: Consider the units in which it is measured.
    - Wage is measured in $ per hour, so mothers earn about 50 cents per hour less than non-mothers. Is that an important difference?
2. Statistical answer: Consider how likely it is that the difference observed in the *sample* is close to that which exists in the *population*.
    - Did we gather enough data to be confident that what we observed in the sample also applies to the population?

We must consider **both** criteria when interpreting quantitative results.

Lecture Outline

Sampling and sampling distributions

The Central Limit Theorem

Confidence intervals

Conclusion

Sampling and sampling distributions

Samples and sampling variation

Our motherhood data, like much data in political sciences, comes from a sample.

- **simple random sample** $\rightarrow$ n units are drawn at random from a population and each unit is *equally likely* to be drawn.
- **probability sample** $\rightarrow$ n units are drawn at random from a population and each unit has a *predefined probability* of being drawn.

Samples and sampling variation

Our motherhood data, like much data in political sciences, comes from a sample.

- **simple random sample** $\rightarrow$ n units are drawn at random from a population and each unit is *equally likely* to be drawn.
- **probability sample** $\rightarrow$ n units are drawn at random from a population and each unit has a *predefined probability* of being drawn.

Random sampling is useful because it helps to ensure that sample units are, *on average*, **representative** of the broader population of units.

Sampling variation

Samples help us to overcome the problem that it is often impossible to observe the entire population.

However, this comes at a cost:

Sampling variation

Samples help us to overcome the problem that it is often impossible to observe the entire population.

However, this comes at a cost:

- When you have a sample, quantities that you calculate using the sample will *not exactly match* the value of those quantities for the entire population.

Functions of samples

Most statistical methods involve calculating some quantity for a sample of data. E.g.

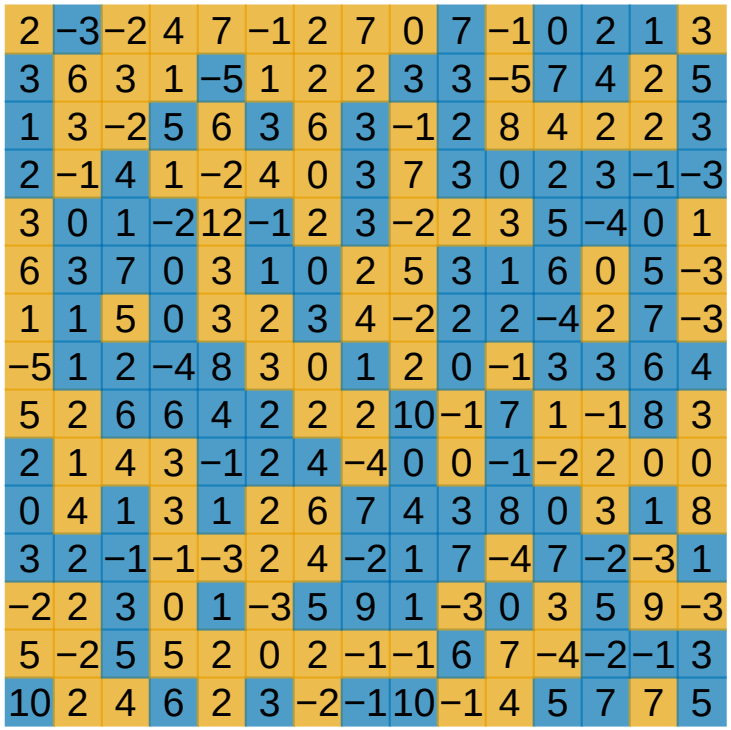
- Mean, median, mode
- The difference in means
- A β coefficient of a regression

However, any "quantity of interest" like these will also vary each time you change the sample.

Implication: We need to know something about how much our quantities of interest vary across multiple samples if we want to understand how much a single sample tells us about the population we care about.

Population vs samples example

The treatment group is blue, and the control group is orange.

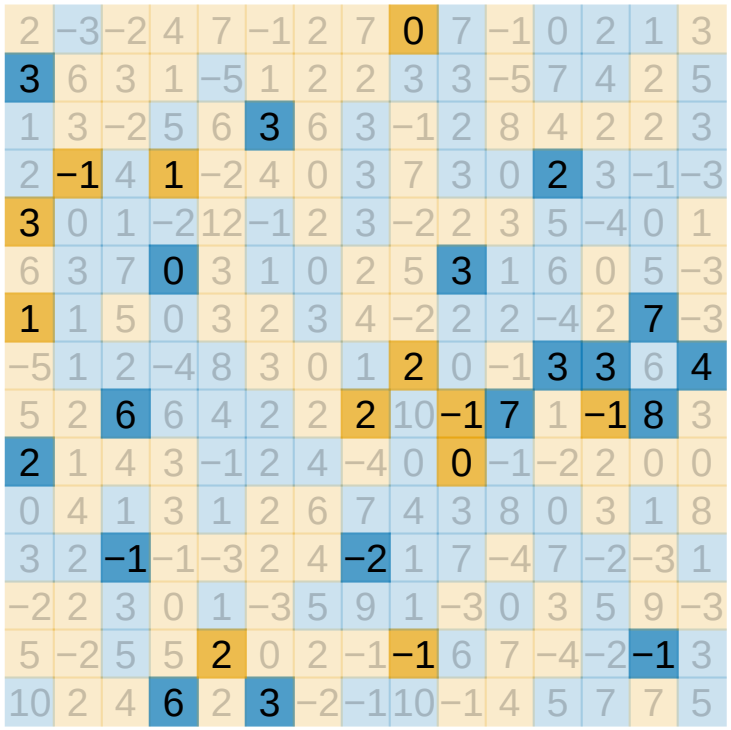


Population values:

- $\bar{Y}_{X=0} = 1.64$
- $\bar{Y}_{X=1} = 2.59$
- $\bar{Y}_{X=1} - \bar{Y}_{X=0} = 0.95$

Population vs samples example

The treatment group is blue, and the control group is orange.



Population values:

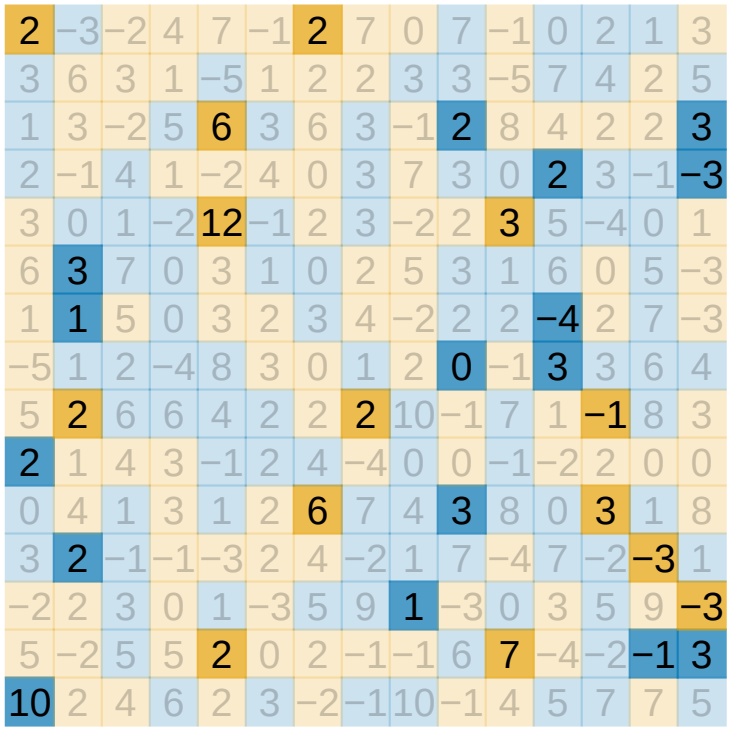
- $\bar{Y}_{X=0} = 1.64$
- $\bar{Y}_{X=1} = 2.59$
- $\bar{Y}_{X=1} - \bar{Y}_{X=0} = 0.95$

Sample values:

- $\bar{Y}_{X=0} = 0.58$
- $\bar{Y}_{X=1} = 3.11$
- $\bar{Y}_{X=1} - \bar{Y}_{X=0} = 2.53$

Population vs samples example

The treatment group is blue, and the control group is orange.



Population values:

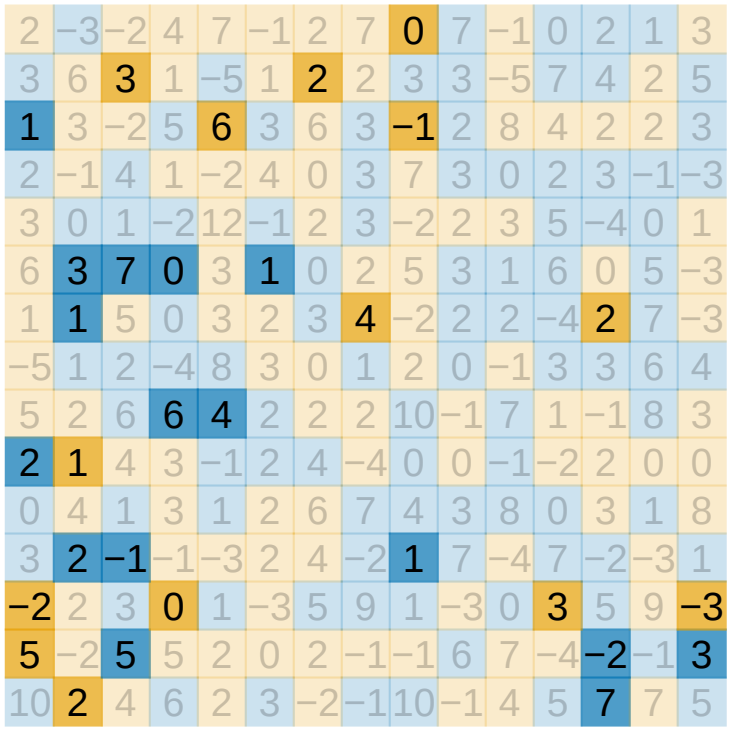
- $\bar{Y}_{X=0} = 1.64$
- $\bar{Y}_{X=1} = 2.59$
- $\bar{Y}_{X=1} - \bar{Y}_{X=0} = 0.95$

Sample values:

- $\bar{Y}_{X=0} = 2.86$
- $\bar{Y}_{X=1} = 1.69$
- $\bar{Y}_{X=1} - \bar{Y}_{X=0} = -1.17$

Population vs samples example

The treatment group is blue, and the control group is orange.



Population values:

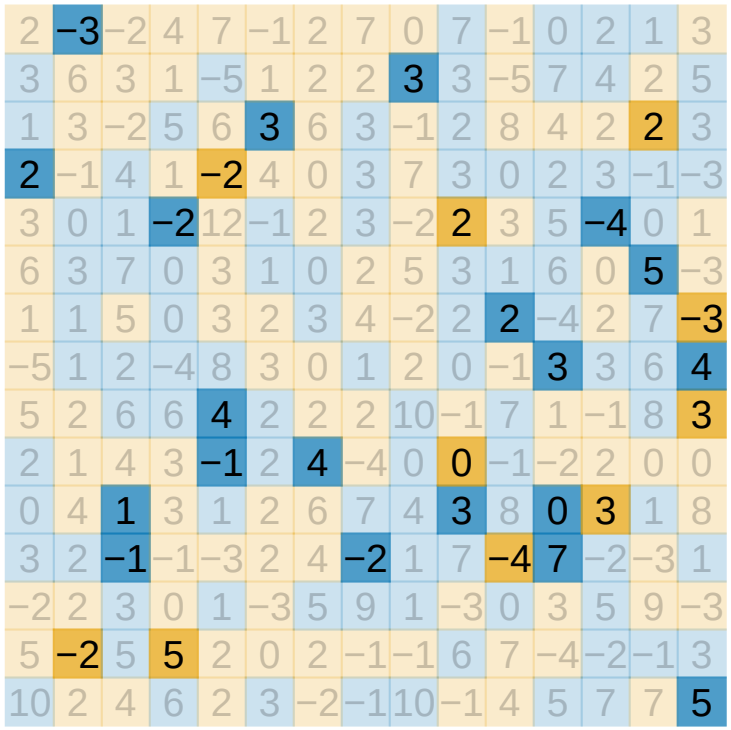
- $\bar{Y}_{X=0} = 1.64$
- $\bar{Y}_{X=1} = 2.59$
- $\bar{Y}_{X=1} - \bar{Y}_{X=0} = 0.95$

Sample values:

- $\bar{Y}_{X=0} = 1.57$
- $\bar{Y}_{X=1} = 2.5$
- $\bar{Y}_{X=1} - \bar{Y}_{X=0} = 0.93$

Population vs samples example

The treatment group is blue, and the control group is orange.



Population values:

- $\bar{Y}_{X=0} = 1.64$
- $\bar{Y}_{X=1} = 2.59$
- $\bar{Y}_{X=1} - \bar{Y}_{X=0} = 0.95$

Sample values:

- $\bar{Y}_{X=0} = 0.4$
- $\bar{Y}_{X=1} = 1.65$
- $\bar{Y}_{X=1} - \bar{Y}_{X=0} = 1.25$

Sampling variation

Intuition: if we randomly sample our observations $(Y_1, \ldots, Y_n)$ from a broader population, then

- $\bar{Y}_{X=0}$ will differ from one sample to the next
- $\bar{Y}_{X=1}$ will differ from one sample to the next

Sampling variation

Intuition: if we randomly sample our observations $(Y_1, \ldots, Y_n)$ from a broader population, then

- $\bar{Y}_{X=0}$ will differ from one sample to the next
- $\bar{Y}_{X=1}$ will differ from one sample to the next
- $\bar{Y}_{X=1} - \bar{Y}_{X=0}$ will be different from one sample to the next

Sampling variation

Intuition: if we randomly sample our observations $(Y_1, \ldots, Y_n)$ from a broader population, then

- $\bar{Y}_{X=0}$ will differ from one sample to the next
- $\bar{Y}_{X=1}$ will differ from one sample to the next
- $\bar{Y}_{X=1} - \bar{Y}_{X=0}$ will be different from one sample to the next

Implication: Without sampling many times, we do not know whether the particular *sample difference in means* we find in our data is close to or far from the *population difference in means*.

Sampling variation in potential outcomes

Sampling variation is more intuitive in the context of individuals drawn from a broader population.

What about the examples we have had when we observe the entire population?

1. Experiments where we see all treatment and control units
2. Observational studies where we observe the whole population of units

Sampling variation in potential outcomes

Sampling variation is more intuitive in the context of individuals drawn from a broader population.

What about the examples we have had when we observe the entire population?

1. Experiments where we see all treatment and control units
2. Observational studies where we observe the whole population of units

While we observe all units, we only observe a sample of the ***potential outcomes*** for all units.

Samples of potential outcomes example

True ATE = 1.13

Potential outcomes under treatment

−1	2	1	−2	2	4	1	6	2	2	3	1	5	4	7
6	−1	7	3	1	2	1	4	4	1	7	0	2	−1	3
5	6	2	1	3	0	1	1	5	3	7	2	−3	1	0
5	6	−2	−2	6	3	0	−6	−1	−1	10	0	1	4	3
−1	0	6	5	0	2	4	9	−1	6	4	4	1	−1	3
0	4	2	2	3	1	4	2	0	0	5	3	2	4	0
3	0	3	−3	5	−1	2	3	6	4	0	−1	5	−2	2
2	3	5	7	8	8	6	5	9	3	2	5	0	−3	5
9	−4	5	5	−1	6	−9	7	5	2	2	4	2	4	3
3	1	3	−2	0	−3	5	8	−2	1	−4	0	−1	3	0
3	2	1	−1	−2	6	2	4	−2	5	8	5	3	0	3
6	2	5	3	0	1	2	0	0	−1	5	1	2	1	2
6	5	2	5	3	−2	7	0	3	4	6	1	4	2	5
4	4	2	4	1	2	6	9	−3	6	2	4	3	1	1
0	1	4	4	4	1	−1	2	2	−1	0	0	1	5	3

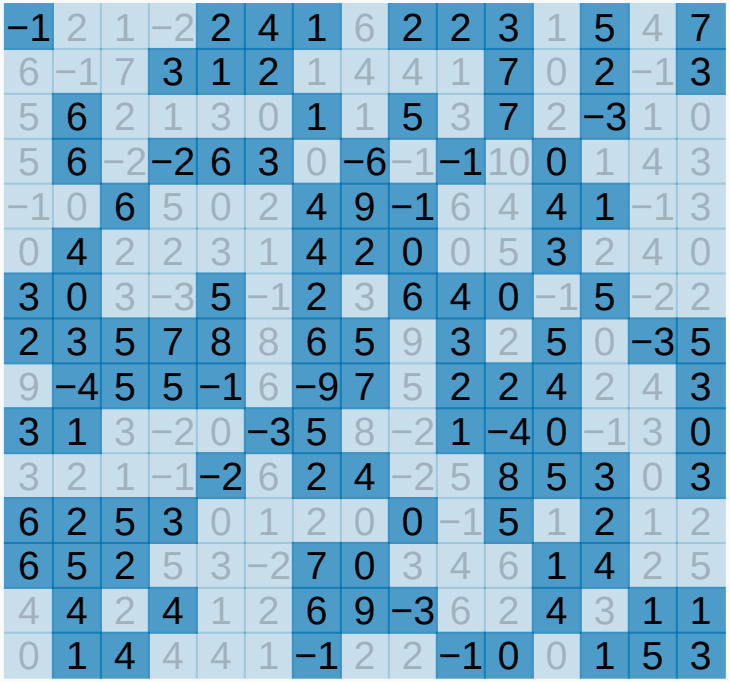
Potential outcomes under control

2	0	2	−3	1	6	−1	−2	−2	2	−5	4	3	−2	−2
4	−1	1	3	4	6	1	5	−4	2	3	4	2	3	−4
2	−2	1	4	2	1	2	−1	0	1	−1	0	5	1	0
−6	3	3	4	2	−1	0	2	−3	6	−3	5	0	1	−2
−3	−5	1	2	8	5	−1	−2	0	−4	4	−1	4	5	3
2	4	0	−3	3	−4	3	0	5	1	8	0	4	−2	2
−1	1	2	2	4	−1	4	6	3	−2	0	0	1	−1	3
3	1	3	0	0	3	−3	0	−1	−3	3	0	4	−2	7
1	8	3	6	−2	6	2	1	3	1	−2	−2	2	4	2
−2	2	3	−2	4	6	4	3	5	−4	2	1	4	4	−1
2	−1	−3	−2	2	4	1	4	−1	1	3	1	2	2	−4
−2	0	2	5	5	0	−3	4	1	0	−4	1	1	4	5
−4	3	2	2	0	6	−1	−2	3	4	−1	−2	8	3	1
4	4	3	1	−2	0	−1	−4	2	2	2	6	3	1	0
4	−6	4	7	5	0	6	−3	1	−3	1	4	−5	3	0

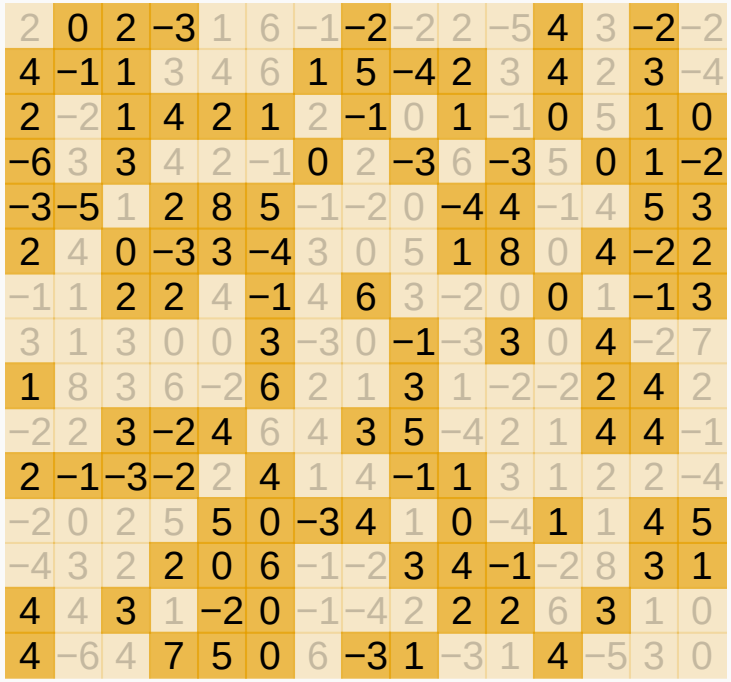
Samples of potential outcomes example

True ATE = 1.13

Potential outcomes under treatment



Potential outcomes under control

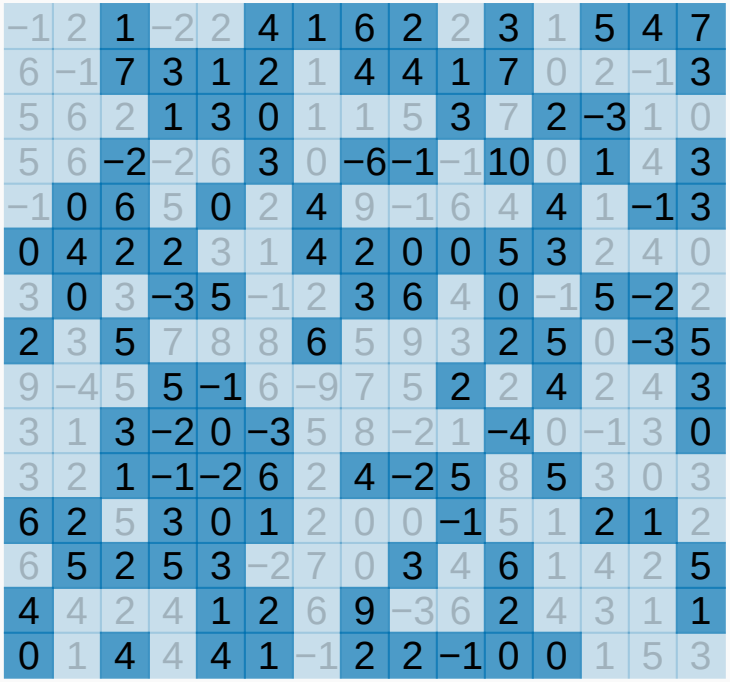


Estimated ATE = 1.21

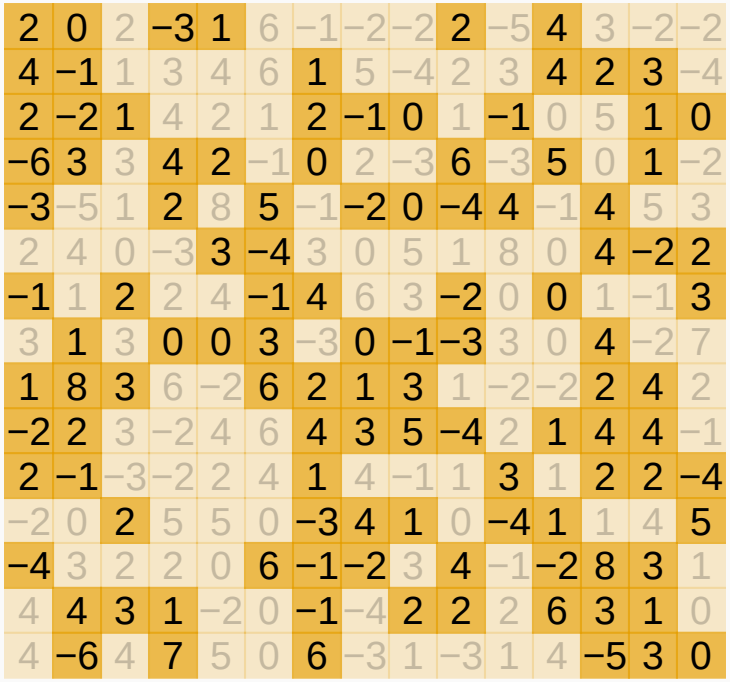
Samples of potential outcomes example

True ATE = 1.13

Potential outcomes under treatment



Potential outcomes under control

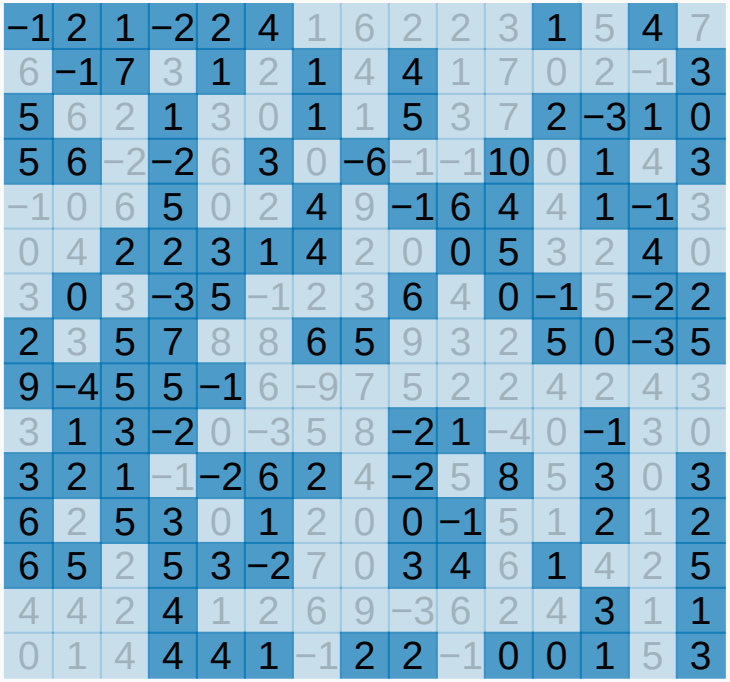


Estimated ATE = 0.89

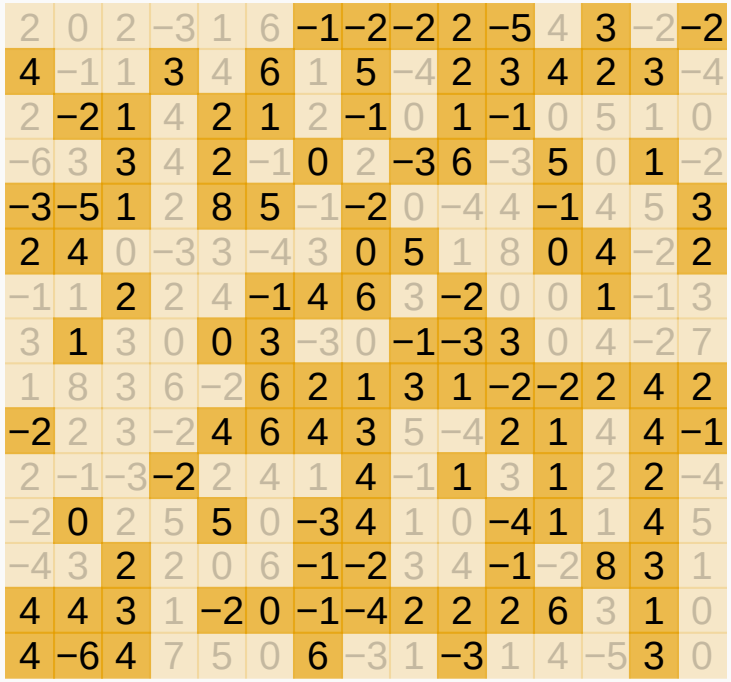
Samples of potential outcomes example

True ATE = 1.13

Potential outcomes under treatment



Potential outcomes under control



Estimated ATE = 0.77

Sampling variation in potential outcomes

Intuition: if we randomly assign our observations to treatment and control

- $\bar{Y}_{X=0}$ will differ from one sample to the next
- $\bar{Y}_{X=1}$ will differ from one sample to the next
- the estimated ATE will be different from one sample to the next

Implication: Even when we observe the entire population, we do not know whether the estimated ATE is close to or far from the true ATE.

Sampling variation

Key Question: *How much* should we expect the difference in means to vary across samples?

How much will the difference in means vary across samples?

Let's pretend that our motherhood data includes the full population, and we will sample from that population using R.

```
## Define the sample size
n_sample <- 300

## Sample once from the data
### sample 300 row numbers
sampled_rows <- sample(1:nrow(motherhood), n_sample, replace = T)
### subset of sampled rows
mother_sample <- motherhood[sampled_rows,]

## Difference in means for the sample
mean(mother_sample$wage[mother_sample$isMother == 1]) -
  mean(mother_sample$wage[mother_sample$isMother == 0])

## [1] -0.7165097
```

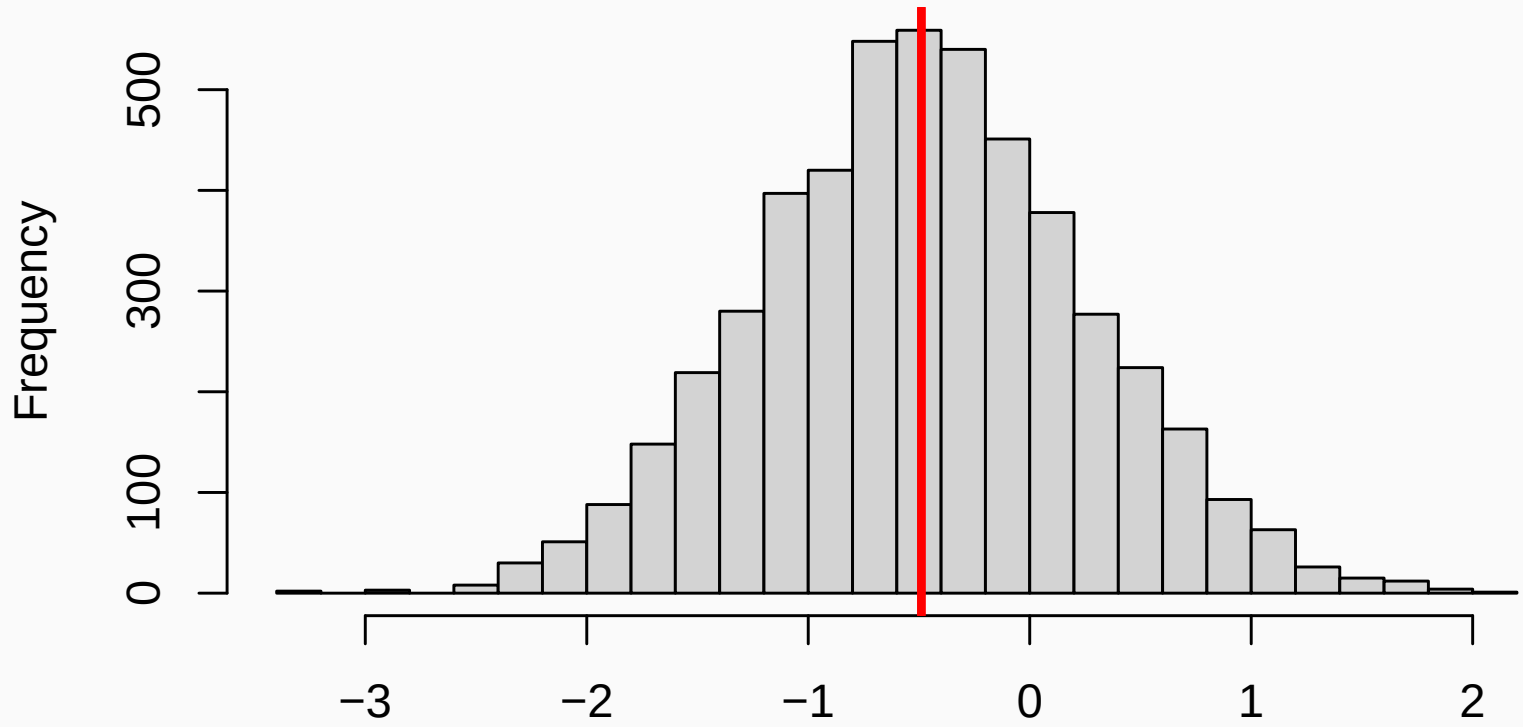
How much will the difference in means vary across samples?

If we repeatedly sample from the population, and calculate the difference in means, we will end up with a ***distribution.***

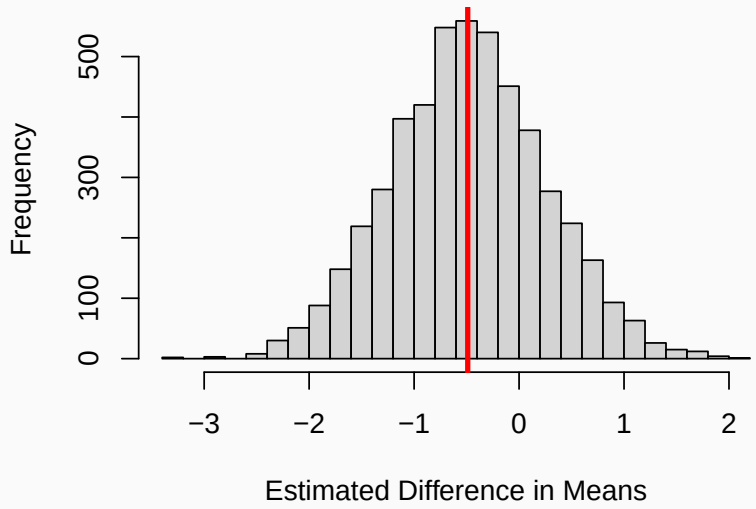
```
# Define a function to do this
diff_means <- function(){

  ## sample rows and subset
  sampled_rows <- sample(1:nrow(motherhood), n_sample, replace = T)
  mother_sample <- motherhood[sampled_rows,]

  ## calculate difference in means
  mean(mother_sample$wage[mother_sample$isMother == 1]) -
    mean(mother_sample$wage[mother_sample$isMother == 0])
}

# Replicate the sampling process 5000 times
diff_in_means_dist <- replicate(5000, diff_means())
```

How much will the difference in means vary across samples?

```
hist(diff_in_means_dist, breaks = 30,
     main = "",xlab = "Estimated Difference in Means")
abline(v = wage_mothers - wage_not_mothers, col = "red", lwd = 3)
```



Sampling distribution

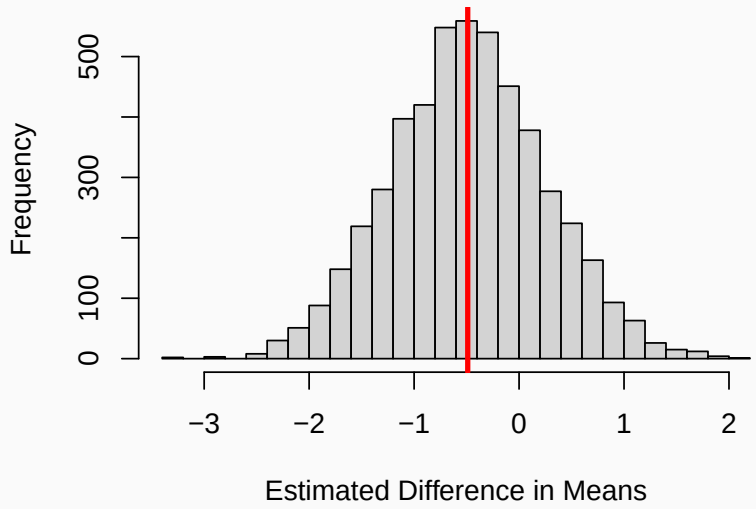


1. The distribution is centered around the true "population" difference in means

2. There is variability from sample to sample

3. This distribution takes a distinctive "bell" shape (more on this later)

Sampling distribution



1. The distribution is centered around the true "population" difference in means

2. There is variability from sample to sample

3. This distribution takes a distinctive "bell" shape (more on this later)

4. This distribution is called the **sampling distribution**

Sampling distribution

Sampling distribution

The sampling distribution is the distribution of values that results from calculating the difference in means for many samples taken from the population.

- The sampling distribution is a *hypothetical* concept, in most applications we just observe one sample, not many.
- The same is true of *potential outcomes*, some of which we cannot observe, but which help us to think about the logic of causal inference.

Standard deviation of the sampling distribution

Critical question we still haven't answered: *how much* does the difference in means vary over these hypothetical samples?

Standard deviation of the sampling distribution

Critical question we still haven't answered: *how much* does the difference in means vary over these hypothetical samples?

You already know how to calculate the spread of a distribution!

```
## standard deviation
sd(diff_in_means_dist)
```

```
## [1] 0.7382555
```

Implication: The estimates of the difference in means across samples have a standard deviation of 0.74 dollars around the mean (of the population).

Standard deviation of the sampling distribution

Critical question we still haven't answered: *how much* does the difference in means vary over these hypothetical samples?

You already know how to calculate the spread of a distribution!

```
## standard deviation
sd(diff_in_means_dist)
```

```
## [1] 0.7382555
```

Implication: The estimates of the difference in means across samples have a standard deviation of 0.74 dollars around the mean (of the population).

The *standard deviation of the sampling distribution* has a special name: the **standard error**.

Standard error

Even though we do not observe the sampling distribution, we can nonetheless calculate an estimate of the standard error of the difference in means from information found a single sample:

$$SE(\hat{Y}_{X=1} - \hat{Y}_{X=0}) = \sqrt{\frac{Var(Y_{X=1})}{n_{X=1}} + \frac{Var(Y_{X=0})}{n_{X=0}}}$$

Standard error

The standard error is the *estimated* standard deviation of the sampling distribution. It describes how much we expect the difference in means in our samples to differ from the true population difference in means, on average.

Standard error

Even though we do not observe the sampling distribution, we can nonetheless calculate an estimate of the standard error of the difference in means from information found a single sample:

$$SE(\hat{Y}_{X=1} - \hat{Y}_{X=0}) = \sqrt{\frac{Var(Y_{X=1})}{n_{X=1}} + \frac{Var(Y_{X=0})}{n_{X=0}}}$$

Intuition:

1. The standard error increases when the variance of the outcome variable – $Var(Y_{X=1})$ and $Var(Y_{X=0})$ – increases

2. The standard error decreases when the number of observations in each group – $n_{X=1}$ and $n_{X=0}$ – increases

Standard error

Even though we do not observe the sampling distribution, we can nonetheless calculate an estimate of the standard error of the difference in means from information found a single sample:

$$SE(\hat{Y}_{X=1} - \hat{Y}_{X=0}) = \sqrt{\frac{Var(Y_{X=1})}{n_{X=1}} + \frac{Var(Y_{X=0})}{n_{X=0}}}$$

For a single sample from the motherhood data:

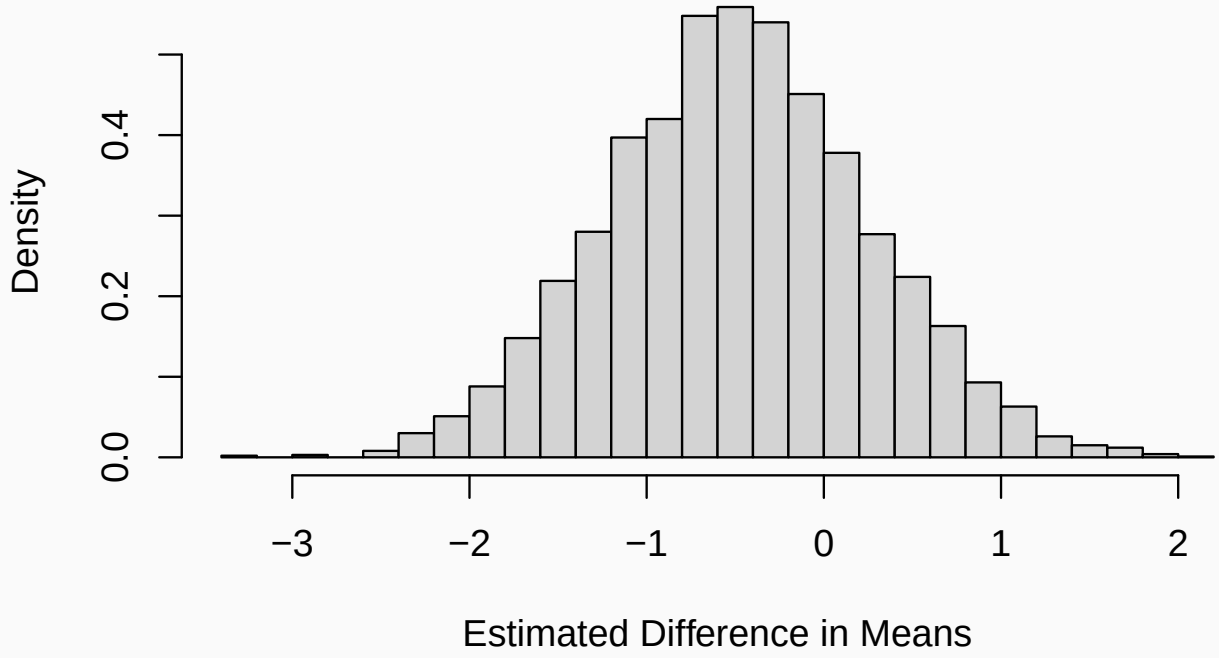
```
var_x_1 <- var(mother_sample$wage[mother_sample$isMother == 1])
var_x_0 <- var(mother_sample$wage[mother_sample$isMother == 0])

n_x_1 <- sum(mother_sample$isMother == 1)
n_x_0 <- sum(mother_sample$isMother == 0)

sqrt((var_x_1/n_x_1) + (var_x_0/n_x_0))
```

```
## [1] 0.6499978
```

The Central Limit Theorem

Example

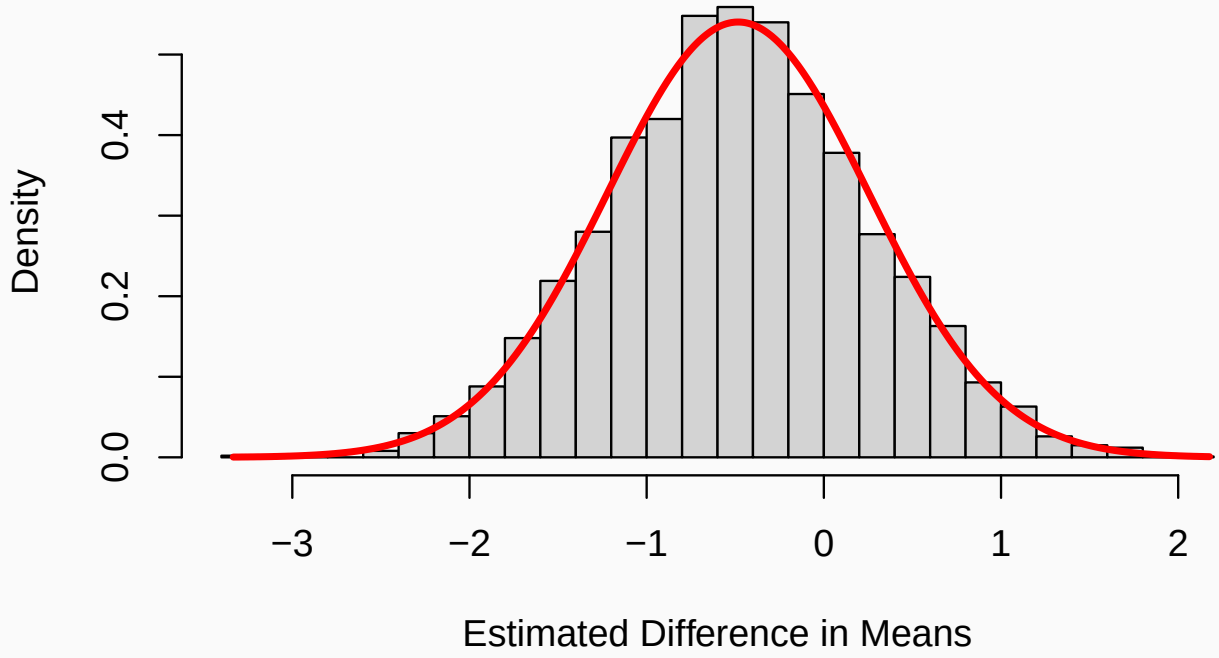
Let's look at the sampling distribution again:



Does the shape seem familiar?

Example

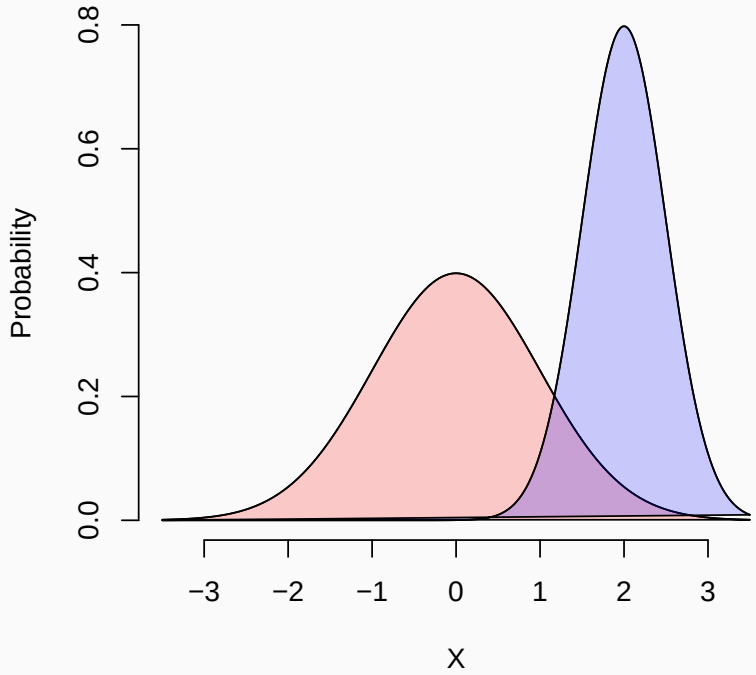
Let's look at the sampling distribution again:



Does the shape seem familiar? It closely resembles a normal distribution.

Normal distribution

The normal distribution is a probability distribution described by two parameters: the mean (μ), and the variance (σ^2)[1]



The red distribution has:

- $\mu = 0$
- $\sigma^2 = 1$

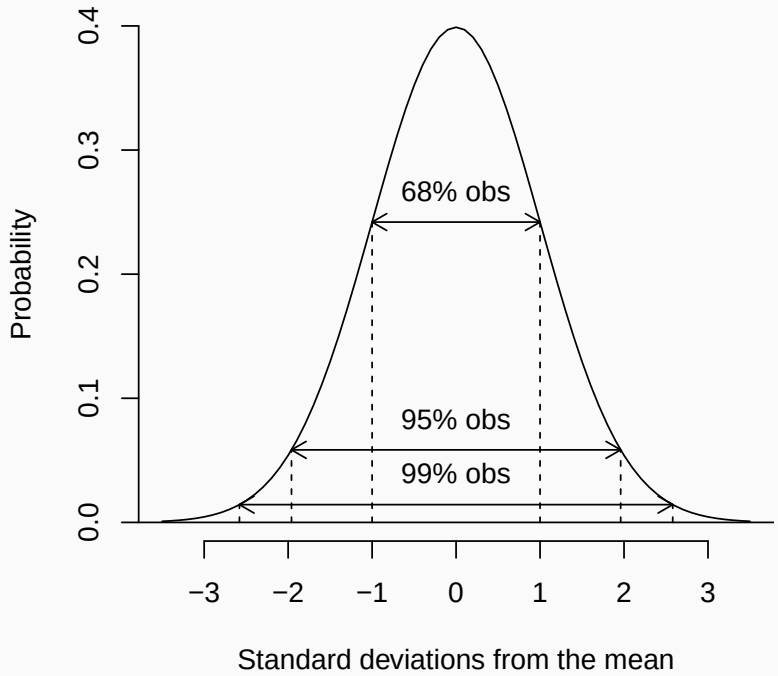
The blue distribution has:

- $\mu = 2$
- $\sigma^2 = 0.71$

[1]Note: the standard deviation is just the square root of the variance: $\sqrt{\sigma^2} = \sigma$

Normal distribution

The normal distribution is a probability distribution described by two parameters: the mean (μ), and the variance (σ^2)[1]

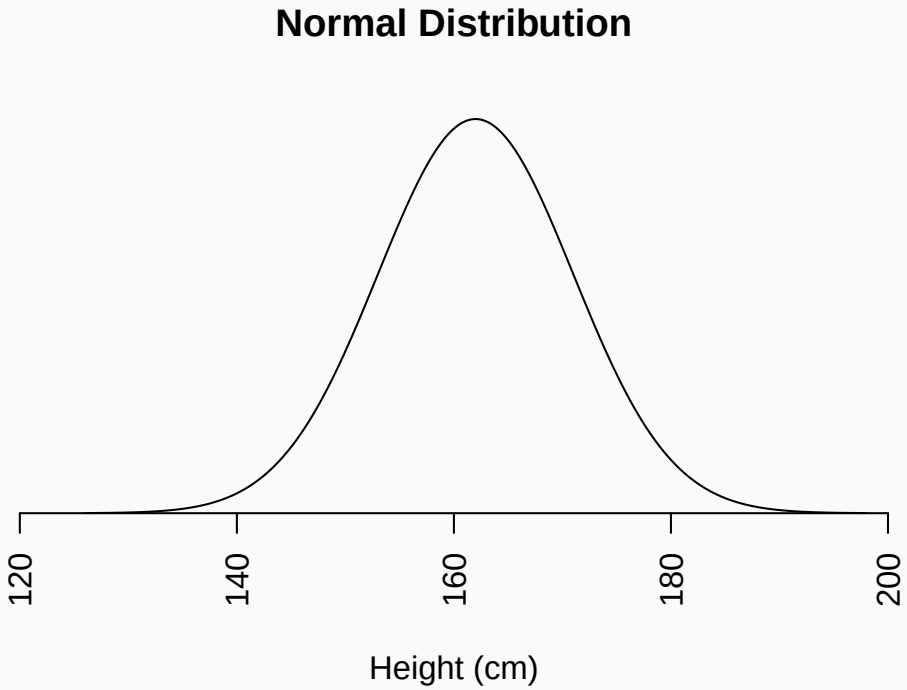


Any normally distributed variable has:

- 68% of observations within 1 sd of the mean.
- 95% of observations within 1.96 sd of the mean.
- 99% of observations within 2.58 sd of the mean.

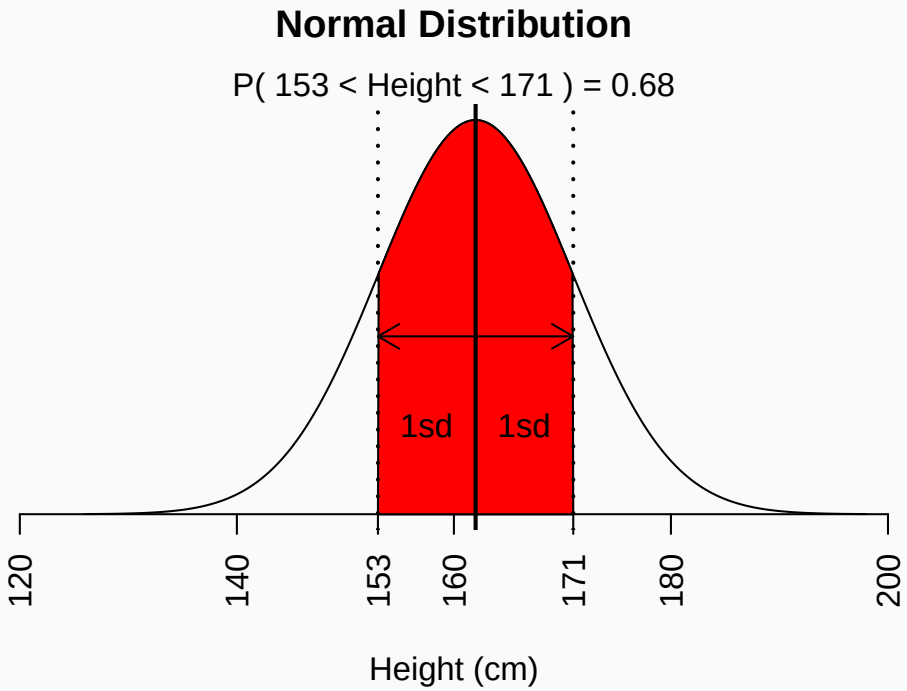
[1]Note: the standard deviation is just the square root of the variance: $\sqrt{\sigma^2} = \sigma$

Normal distribution example



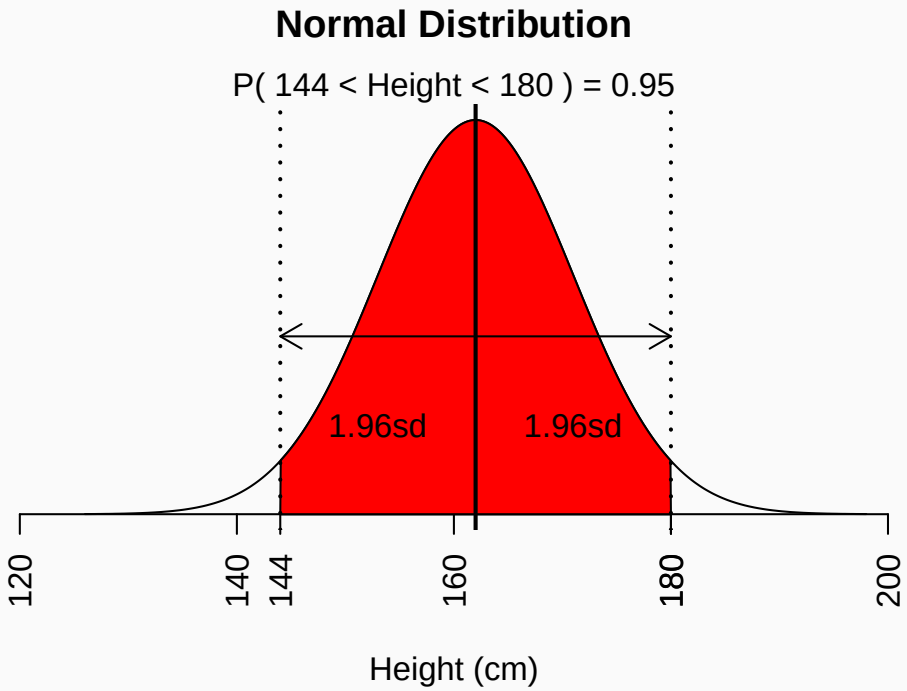
- Human height is very close to normally distributed
- The mean female height in the UK (μ) is 162cm
- The standard deviation of female heights (σ) is 9cm

Normal distribution example



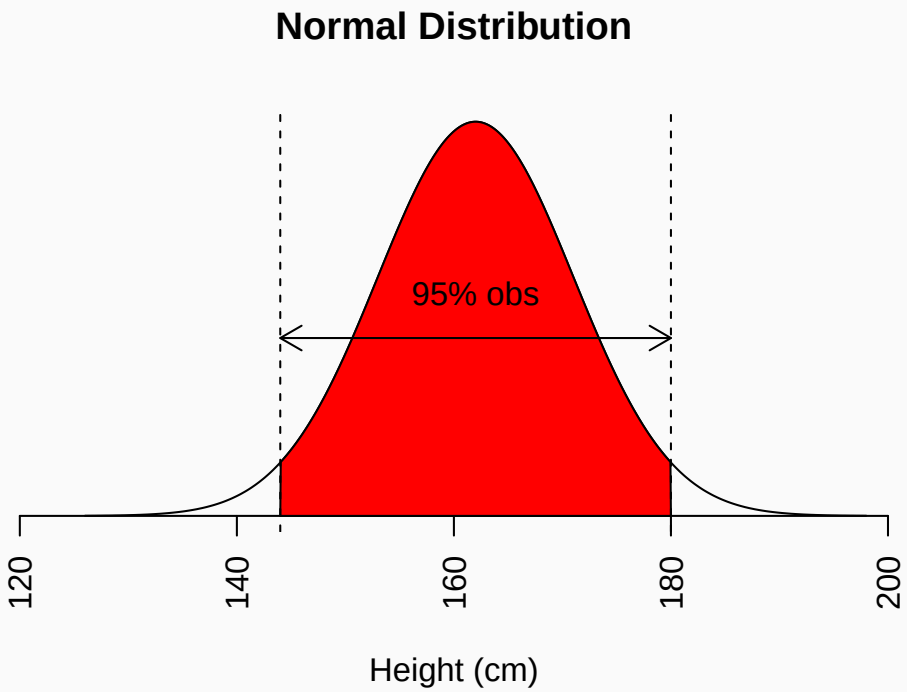
- For a normal distribution, 68% of data is within 1 sd of the mean.
- Here, 68% of women are between 153 and 171 cm.

Normal distribution example



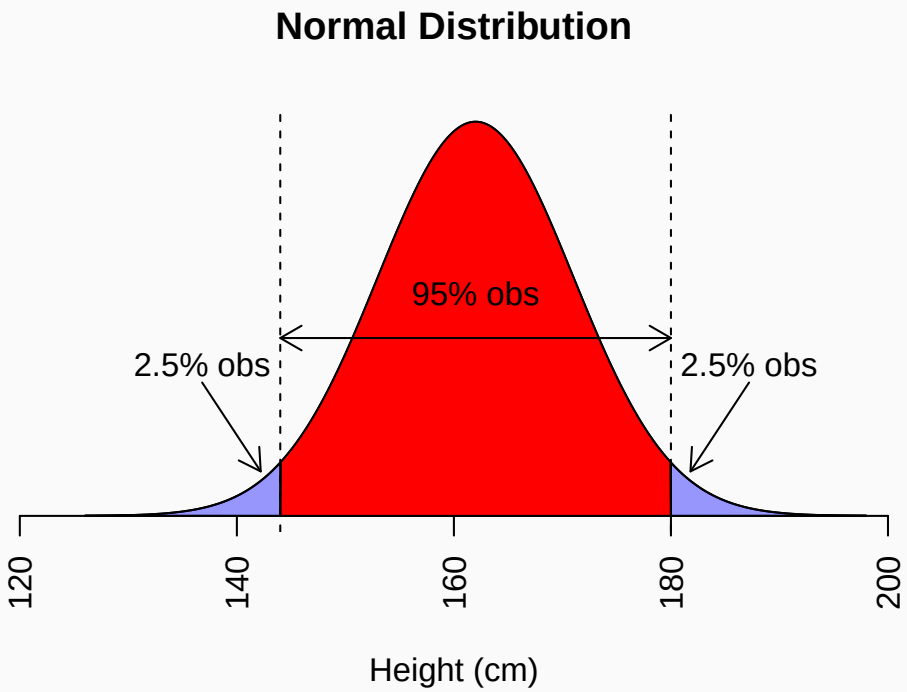
- For a normal distribution, 95% of data is within 1.96 sd of the mean.
- Here, 95% of women are between 144 and 180 cm.

Normal distribution example



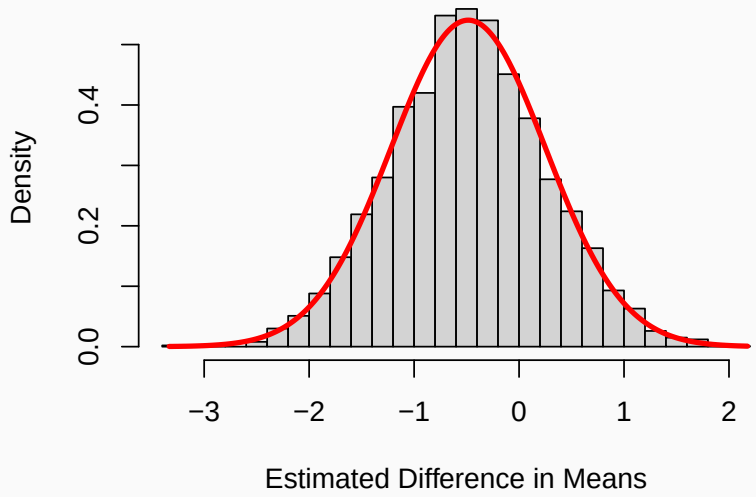
- What is the probability of observing a woman taller than 180cm?

Normal distribution example



- What is the probability of observing a woman taller than 180cm?
- Approximately 2.5%

A “normal” sampling distribution



- If our sampling distribution is normal, we will be able to use these features to calculate the probabilities of certain values
- It is not a coincidence that our sampling distribution looks normal!

The Central Limit Theorem

The Central Limit Theorem

The central limit theorem (CLT) says that, when the size of the sample (n) is large, the distribution of $\bar{Y}$ (the sample average) is approximately normal.

The sampling distribution of $\bar{Y}$ is:

- *exactly* normal when the sample is drawn from a population with the normal distribution
- *approximately* normal when the sample is drawn from a population with **any** distribution, so long as n is sufficiently large

How large is "sufficiently large" depends on the underlying Y_i distribution ($n > 30$ at a minimum).

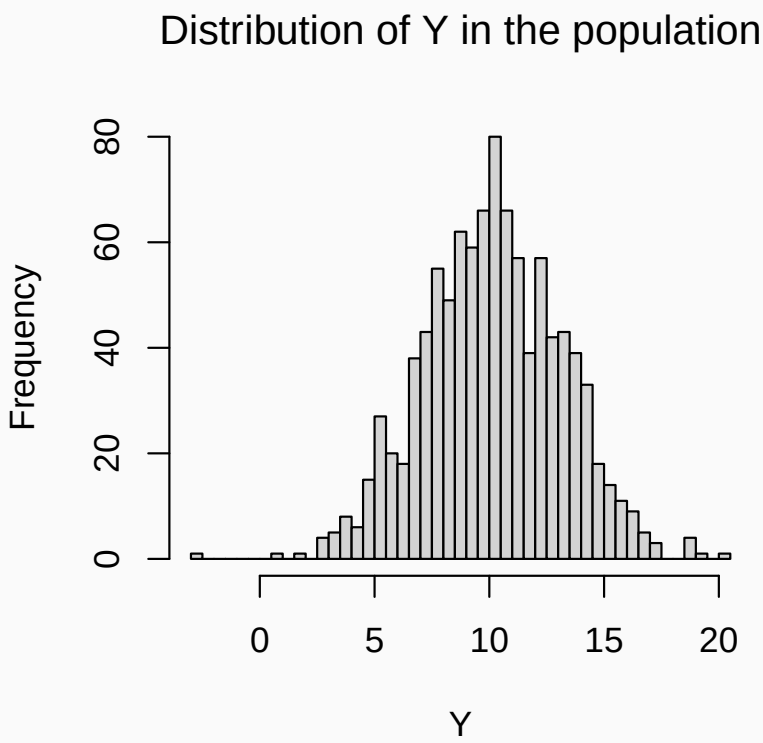
The Central Limit Theorem simulation

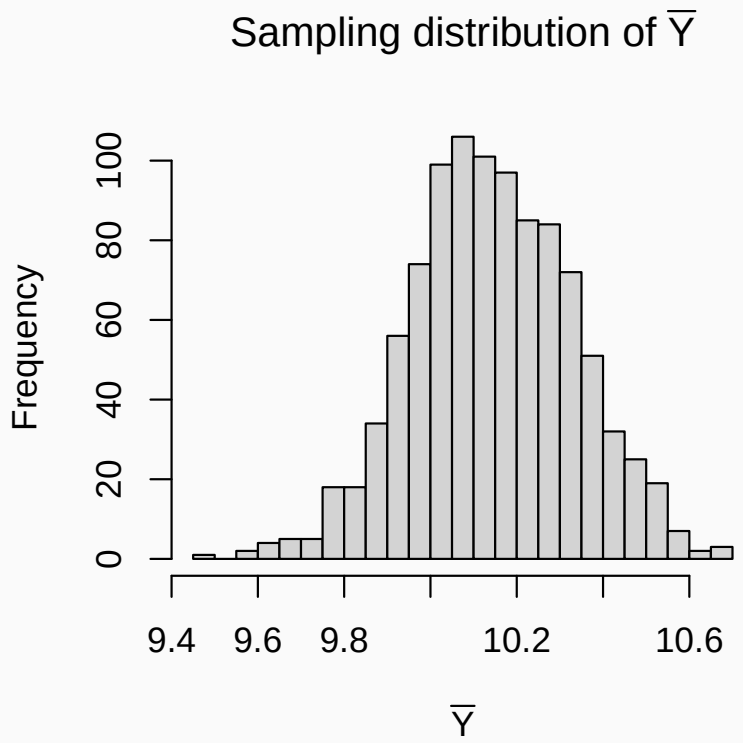
To demonstrate the CLT, we conduct the following simulation in R:

1. Create 10000 observations of Y which is our population. E.g.
   - `pop <- rnorm(n = 10000, mean = 10, sd = 3)`

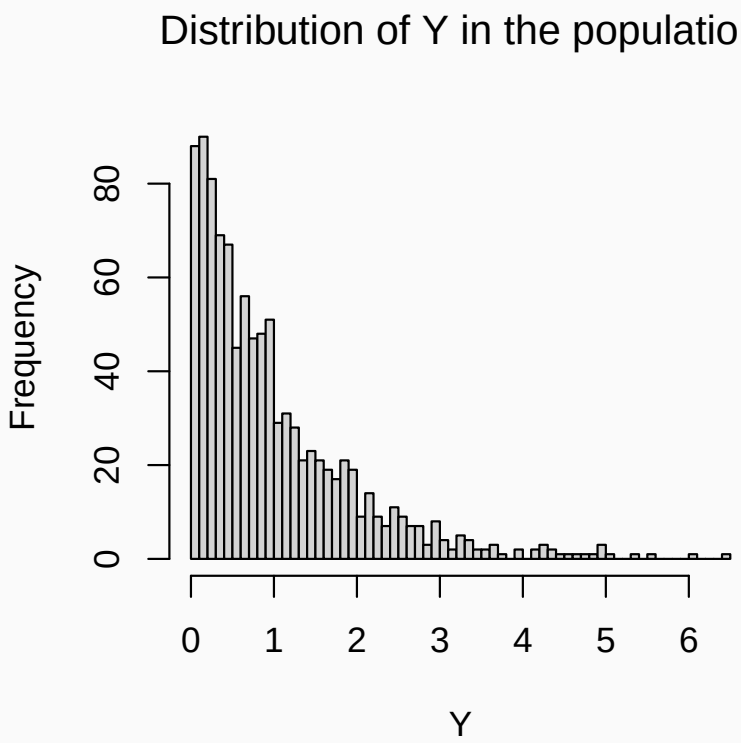
   for a normally distributed Y
2. Sample 200 observations from Y, and calculate $\bar{Y}$
   - `mean(sample(pop, n = 200))`
3. Replicate step 2 1000 times, and plot the sampling distribution
   - `replicate(1000, mean(sample(pop, n = 200)))`

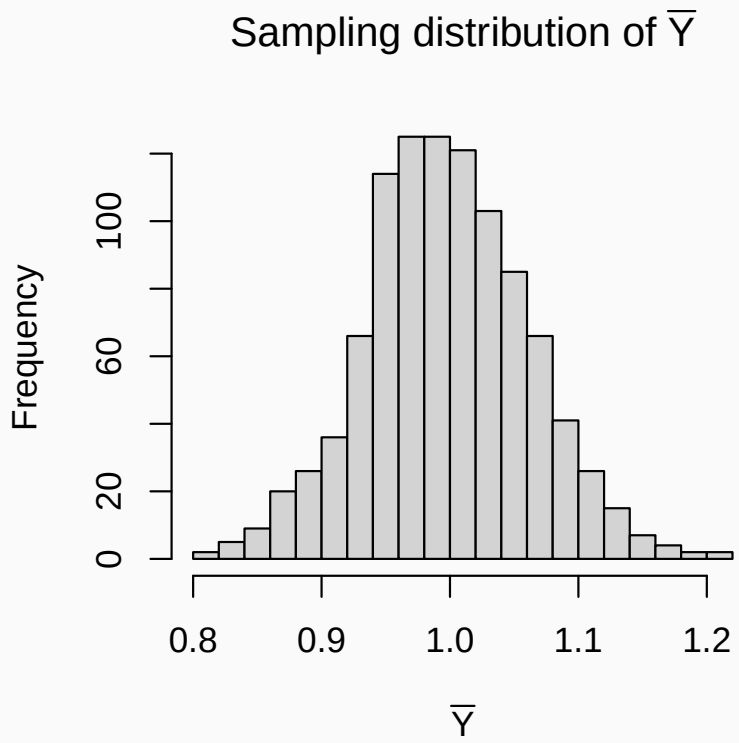
The Central Limit Theorem simulation



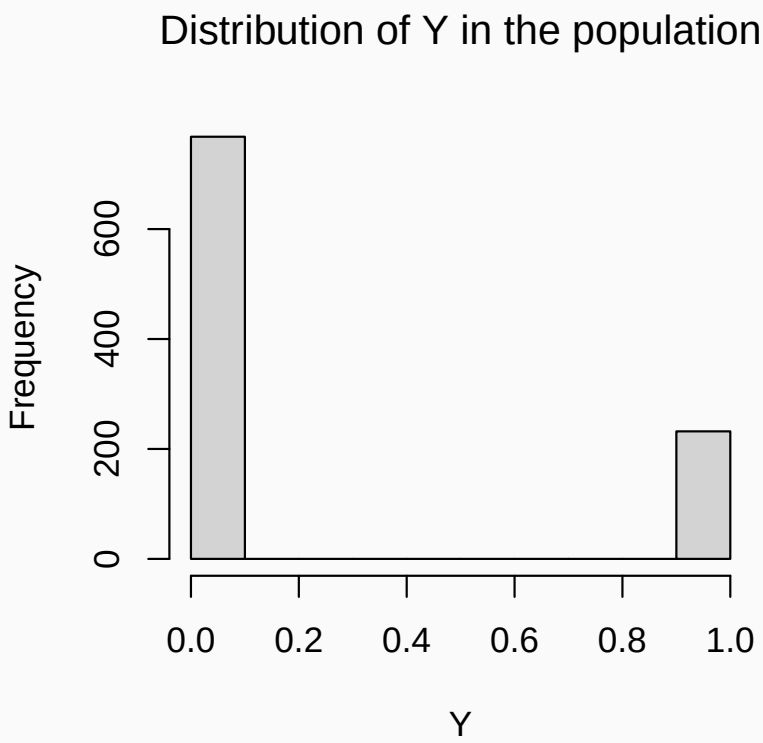


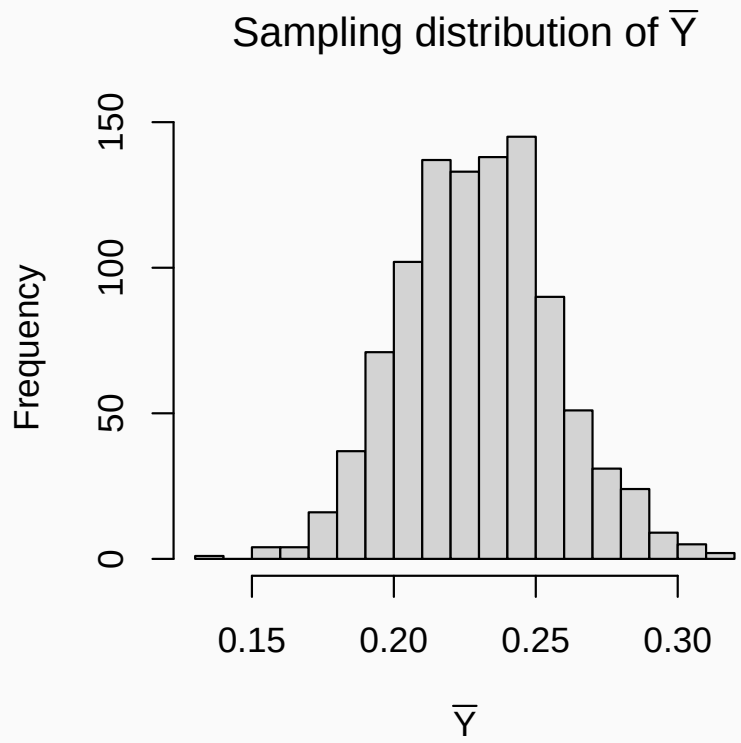
The Central Limit Theorem simulation





The Central Limit Theorem simulation





The Central Limit Theorem

Key finding: Regardless of the shape of the underlying population distribution, the sampling distribution of *sample averages* will be approximately normally distributed so long as the sample size is large enough.

- The same applies for the sampling distribution of the difference in means
- This is useful because we can apply the properties of the normal distribution to our estimated difference in means
- We will do this in the second part of the lecture.

Confidence intervals

Summarizing the variability in the difference in means

Sampling variation means we can't be sure that our sample difference in means $\bar{Y}_{X=1} - \bar{Y}_{X=0}$ is equal to the population difference in means.

While we cannot observe the sampling distribution directly, we can estimate the standard error, which summarises the variability in our estimates.

We can then use the standard error to construct confidence intervals.

Summarizing the variability in the difference in means

Sampling variation means we can't be sure that our sample difference in means $\bar{Y}_{X=1} - \bar{Y}_{X=0}$ is equal to the population difference in means.

While we cannot observe the sampling distribution directly, we can estimate the standard error, which summarises the variability in our estimates.

We can then use the standard error to construct confidence intervals.

- Confidence intervals are another way of quantifying the uncertainty about the population associated with the fact that we only observed a sample of data.

Confidence intervals

Confidence interval

A confidence interval is a range of numbers that we believe is likely to contain the true difference in means. Confidence intervals are constructed so that they contain the true difference in means in a fixed proportion of samples. This is called the confidence level, which we must select before computing the interval.

Implication: Confidence intervals quantify our uncertainty by giving a range of values that are *likely* to *include* the true population difference in means.

Calculating confidence intervals

1. Select a confidence level (typically 95% or 99%)
2. Calculate the difference in means
3. Calculate the standard error
4. Select the **critical value** of the standard normal distribution that corresponds to the confidence level[2]
   - The critical value for the 95% confidence level is 1.96
   - The critical value for the 99% confidence level is 2.58
5. Compute the upper and lower ends of the confidence interval:
   - Upper: $\bar{Y}_{X=1} - \bar{Y}_{X=0} + 1.96 * SE(\bar{Y}_{X=1} - \bar{Y}_{X=0})$
   - Lower: $\bar{Y}_{X=1} - \bar{Y}_{X=0} - 1.96 * SE(\bar{Y}_{X=1} - \bar{Y}_{X=0})$

[2] You can select any critical value, although these are the most commonly used.

Confidence intervals and the sampling distribution

Question: What is the connection between the confidence interval and the sampling distribution?

Confidence intervals and the sampling distribution

Question: What is the connection between the confidence interval and the sampling distribution?

- The sampling distribution tells us how the statistic we are calculating – here, the difference in means – will vary around the population value for that statistic across different samples.
- The confidence interval is constructed to *include* the population value with some probability across different samples.
- The fact that the sampling distribution is always approximately normal means that the standard error is all we need to calculate a confidence interval around our estimate of the difference in means.

Confidence intervals (coverage)

When repeatedly sampling from the population, confidence intervals constructed for each sample will contain the true value with a pre-specified probability.

- A 95% interval will include the true difference in 0.95 of our samples
- A 99% interval will include the true difference in 0.99 of our samples

We can simulate this by treating our motherhood data as the population:

1. Draw a sample from the motherhood data
2. Calculate the difference in means and standard error for that sample
3. Calculate the 95% confidence interval for that sample
4. What proportion of intervals include the true difference in means?

Interpretation

Interpretation confidence intervals

- a 95% CI has a .95 probability of bracketing the true population mean

Interpretation

Interpretation confidence intervals

- **a 95% CI has a .95 probability of bracketing the true population mean**
  - NOT: the true value has a .95 probablity of falling within the brackets of a given 95% CI

Interpretation

Interpretation confidence intervals

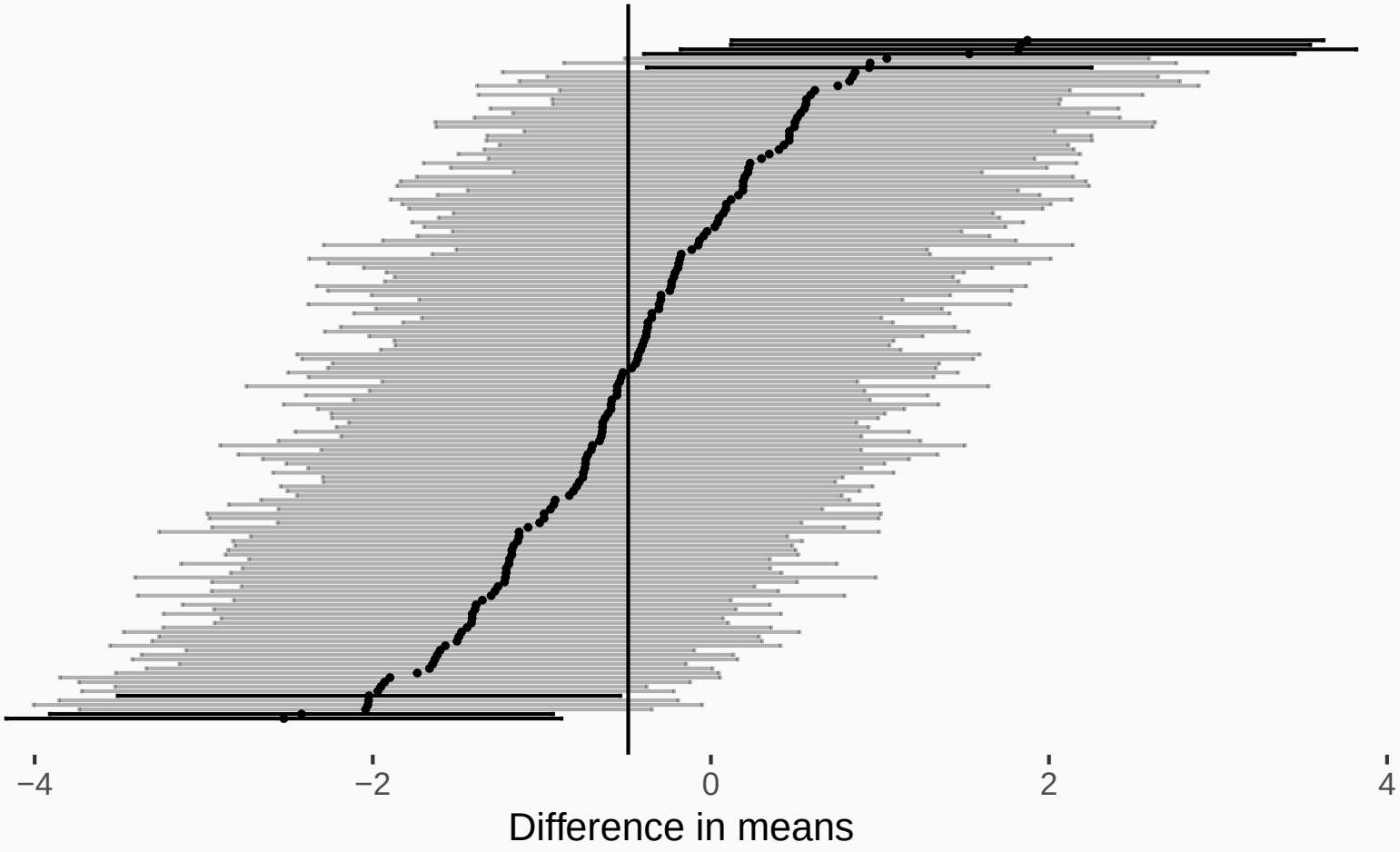
- **a 95% CI has a .95 probability of bracketing the true population mean**
  - NOT: the true value has a .95 probablity of falling within the brackets of a given 95% CI
- **a 99% CI has a .99 probability of bracketing the true population mean**

Interpretation

Interpretation confidence intervals

- **a 95% CI has a .95 probability of bracketing the true population mean**
  - NOT: the true value has a .95 probablity of falling within the brackets of a given 95% CI
- **a 99% CI has a .99 probability of bracketing the true population mean**
  - NOT: the true value has a .99 probablity of falling within the brackets of a given 99% CI

Confidence interval simulation



Calculating the confidence interval for the motherhood wage penalty

What is the confidence interval for the difference in means in the full sample of the `motherhood` data?

```
## Difference in means
wage_mothers <- mean(motherhood$wage[motherhood$isMother == 1])
wage_not_mothers <- mean(motherhood$wage[motherhood$isMother == 0])
diff_mother <- wage_mothers - wage_not_mothers
diff_mother
```

```
## [1] -0.4889398
```

Calculating the confidence interval for the motherhood wage penalty

What is the confidence interval for the difference in means in the full sample of the `motherhood` data?

```
## Standard error
var_x_1 <- var(motherhood$wage[motherhood$isMother == 1])
var_x_0 <- var(motherhood$wage[motherhood$isMother == 0])

n_x_1 <- sum(motherhood$isMother == 1)
n_x_0 <- sum(motherhood$isMother == 0)

se_motherhood <- sqrt((var_x_1/n_x_1) + (var_x_0/n_x_0))
se_motherhood
```

```
## [1] 0.2691242
```

Calculating the confidence interval for the motherhood wage penalty

What is the confidence interval for the difference in means in the full sample of the `motherhood` data?

```
## 95% Confidence interval
diff_mother - 1.96 * se_motherhood
```

```
## [1] -1.016423
```

```
diff_mother + 1.96 * se_motherhood
```

```
## [1] 0.03854365
```

Interpretation: The 95% CI for the mean difference ranges from -1.02 to 0.04

Calculating the confidence interval for the motherhood wage penalty

What is the confidence interval for the difference in means in the full sample of the `motherhood` data?

```
## 99% confidence interval
diff_mother - 2.58 * se_motherhood
```

```
## [1] -1.18328
```

```
diff_mother + 2.58 * se_motherhood
```

```
## [1] 0.2054007
```

Interpretation: The 99% CI for the mean difference ranges from -1.18 to 0.21

Note: The greater confidence level yields a wider interval

Calculating the confidence interval for the motherhood wage penalty

What is the confidence interval for the difference in means in the full sample of the `motherhood` data?

```
## 99% confidence interval
diff_mother - 2.58 * se_motherhood
```

```
## [1] -1.18328
```

```
diff_mother + 2.58 * se_motherhood
```

```
## [1] 0.2054007
```

Conclusion: It is important to note that this interval includes the value of 0. Therefore it is plausible given the size of the sample and the difference we observe in the **sample**, that there is in fact **no difference** between the hourly wages of mothers and non-mothers

Calculating the confidence interval for the motherhood wage penalty

What is the confidence interval for the difference in means in the full sample of the `motherhood` data?

An easier way:

```
t.test(x = motherhood$wage[motherhood$isMother == 1],
       y = motherhood$wage[motherhood$isMother == 0],
       conf.level = .95)
```

```
...
## t = -1.8168, df = 2140.3, p-value = 0.06939
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.01671199  0.03883242
## sample estimates:
## mean of x mean of y
##  11.32977  11.81871
```

t-distribution

- The “t-test” on the previous slide assumes that the sampling distribution follows a “t-distribution” rather than the normal distribution.
- The t-distribution is very close to a normal distribution, unless the sample size you are using is very small (eg less than 30)
- The confidence intervals calculated assuming that the sampling distribution is a t-distribution will be somewhat wider with very small sample sizes, but are indistinguishable for most data sets.

What determines the width of the confidence interval?

$$\bar{Y}_{X=1} - \bar{Y}_{X=0} \pm 1.96 * SE(\bar{Y}_{X=1} - \bar{Y}_{X=0})$$

Note that the confidence interval's width is determined by

1. The critical value
    - Larger critical values (higher confidence levels) result in wider intervals
2. The standard error
    - Larger standard errors result in wider intervals

What determines the width of the confidence interval?

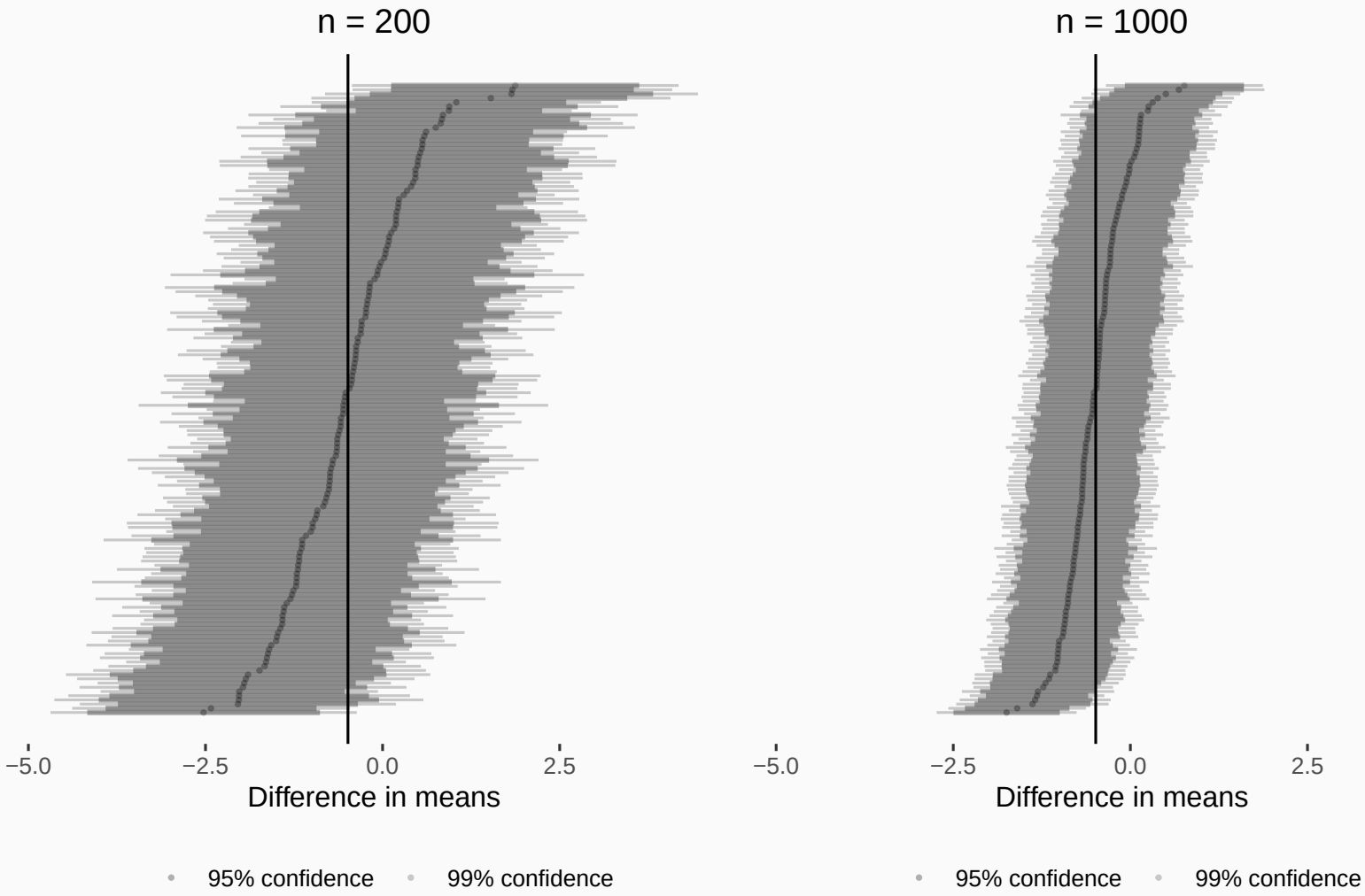
$$\bar{Y}_{X=1} - \bar{Y}_{X=0} \pm 1.96 * SE(\bar{Y}_{X=1} - \bar{Y}_{X=0})$$

Note that the confidence interval's width is determined by

1. The critical value
    - Larger critical values (higher confidence levels) result in wider intervals
2. The standard error
    - Larger standard errors result in wider intervals

Implication: the same factors that affect the size of the standard error (sample size, variance of Y) will also affect the width of the confidence interval.

Sample size and the width of the confidence interval



Confidence intervals for the RAND experiment

Does health insurance improve health outcomes?

```
## Mean health level for insured
mean_health_insured <- mean(rand$health[rand$insured == TRUE])

## Mean health level for uninsured
mean_health_uninsured <- mean(rand$health[rand$insured == FALSE])

mean_health_insured - mean_health_uninsured
```

```
## [1] -0.01895885
```

Confidence intervals for the RAND experiment

Does health insurance improve health outcomes?

```
t.test(x = rand$health[rand$insured == TRUE],
       y = rand$health[rand$insured == FALSE],
       conf.level = .95)
```

```
...
## t = -0.47235, df = 691.9, p-value = 0.6368
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.09776506  0.05984736
## sample estimates:
## mean of x mean of y
##  3.388057  3.407016
...
```

Confidence intervals for the RAND experiment

Does health insurance improve health outcomes?

```
...
## 95 percent confidence interval:
##  -0.09776506  0.05984736
...
```

Implications:

- We cannot be confident that the treatment effect was negative *in the population of potential outcomes*, from which the experiment yielded a sample.
- $\rightarrow$ We *cannot* be confident that the treatment had a negative average treatment effect.

Conclusion

What have we learned?

- Sampling variation means the quantities of interest estimated from a *sample* will not be exactly equal to those quantities in the *population*
  - Applies to sampling units from a larger population of units
  - Applies to sampling potential outcomes (under treatment vs control) for units via an experiment.
- We can conceptualise sampling uncertainty via the *sampling distribution*, which describes how much estimates will vary across samples

What have we learned?

- The *central limit theorem* says that, when n is large, the sampling distribution will be approximately normally distributed
- Confidence intervals are one way of summarising the uncertainty we have about *populations* when we do data analysis on *samples* from those populations.

Seminar

In seminars this week, you will learn to ...

1. ...calculate standard errors.
2. ...estimate and interpret confidence intervals.