# Causality and randomness

## Lecture 9

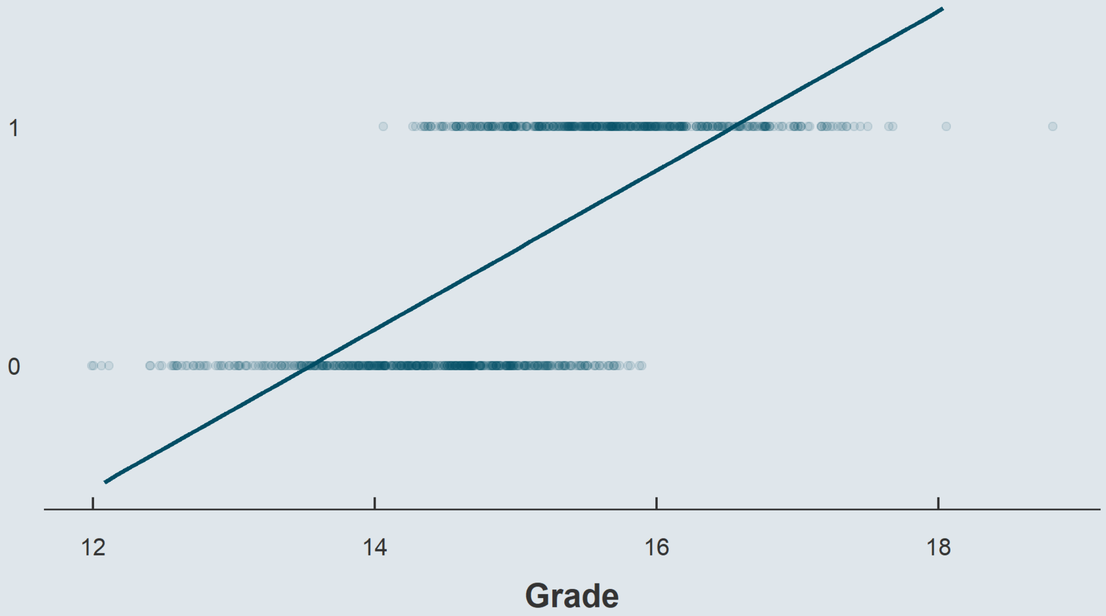
Louis SIRUGUE

11/2021

# Last time we saw

**1) An OLS regression with a binary dependant variable is called a *linear probability model***

$$1\{y_i = \text{Accepted}\} = \hat{\alpha} + \hat{\beta} \times \text{Grade}_i + \hat{\varepsilon}_i$$



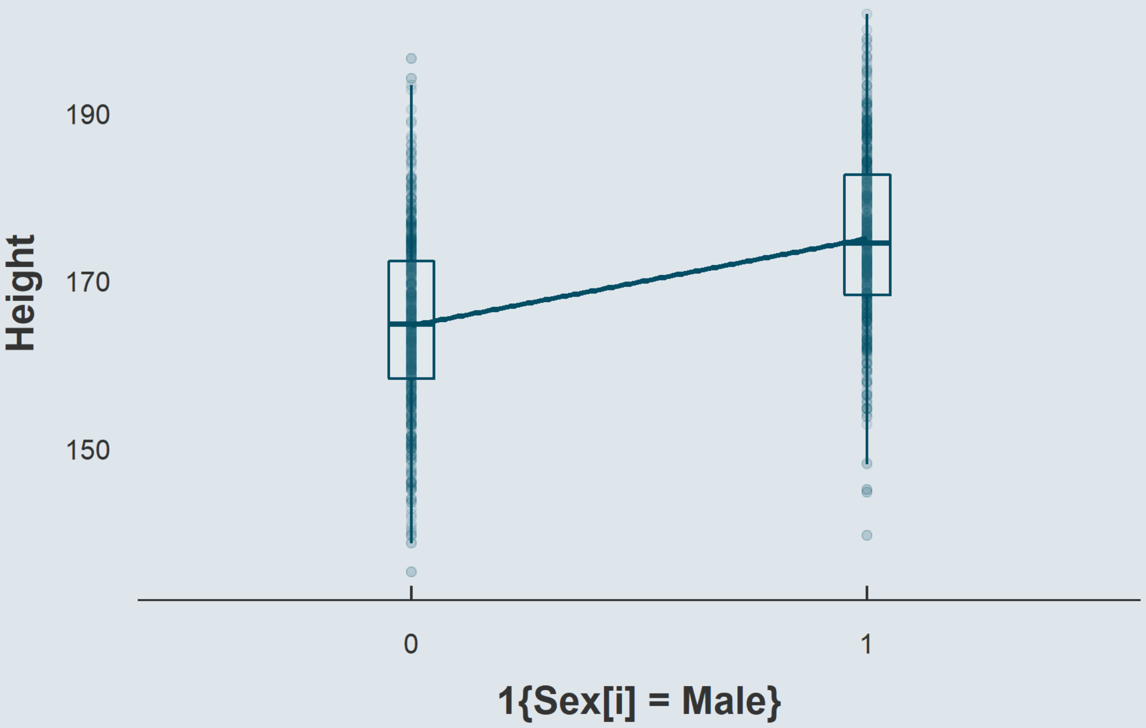
- $\hat{\beta}$ indicates by how much the probability that $y$ equals 1 would increase on expectation for a 1 point increase in $x$

- But linear probability models can lead to probabilities that are lower than 0 and greater than 1

# Last time we saw

**2) An OLS regression with a binary independent variable:**

$$\text{Height}_i = \hat{\alpha} + \hat{\beta} \times 1\{x_i = \text{Male}\} + \hat{\varepsilon}_i$$



- $\hat{\alpha}$ indicates the average value of $y$ when $x$ is equal to 0, i.e., for the reference category
- $\hat{\beta}$ indicates the difference between the average value of $y$ for the two groups

# Last time we saw

**3) An OLS regression with an independant variable with more than 2 categories**

$$\text{Earnings}_i = \hat{\alpha} + \hat{\beta}_1 1\{\text{Race}_i = \text{Black}\} + \hat{\beta}_2 1\{\text{Race}_i = \text{Asian}\} + \hat{\beta}_3 1\{\text{Race}_i = \text{Other}\} + \hat{\varepsilon}_i$$

- It should be specified as a sum of binary variables, omitting the category we want as a reference

```
summary(lm(Earnings ~ relevel(as.factor(Race), "White"), asec_2020))$coefficients[, c(1, 2, 4)]
```

```
##                                          Estimate Std. Error     Pr(>|t|)
## (Intercept)                              62880.49   344.0464 0.000000e+00
## relevel(as.factor(Race), "White")Asian  15110.29  1199.9326 2.559272e-36
## relevel(as.factor(Race), "White")Black -12302.99   996.8981 5.947231e-35
## relevel(as.factor(Race), "White")Other -13401.79  1609.0045 8.294160e-17
```

  - $\hat{\alpha}$ is the average earnings for the reference category
  - Coefficients are *relative* to the reference category

# Last time we saw

**4) Interactions allow to estimate how the coefficient of a given variable varies depending on the value of a third variable**

$$\text{Earnings}_i = \hat{\alpha} + \hat{\beta}_1\text{Hours}_i + \hat{\beta}_2 1\{\text{Sex}_i = \text{Male}\} + \hat{\beta}_3\text{Hours}_i \times 1\{\text{Sex}_i = \text{Male}\} + \hat{\varepsilon}_i$$

- **It captures the difference between males and females in their expected returns to working an additional hour per week**
  - With $\hat{\beta}_3 > 0$: working an additional hour matters more if you are a male
- **Or equivalently, by how much the effect of being a male varies for an additional hour of work per week**
  - With $\hat{\beta}_3 > 0$: being a male matters more if you work an additional hour
- Note that the expected returns to working an additional hour per week is:
  - $\hat{\beta}_1$ for females
  - $\hat{\beta}_1 + \hat{\beta}_3$ for males

# Last time we saw

### 5) Non-linearities

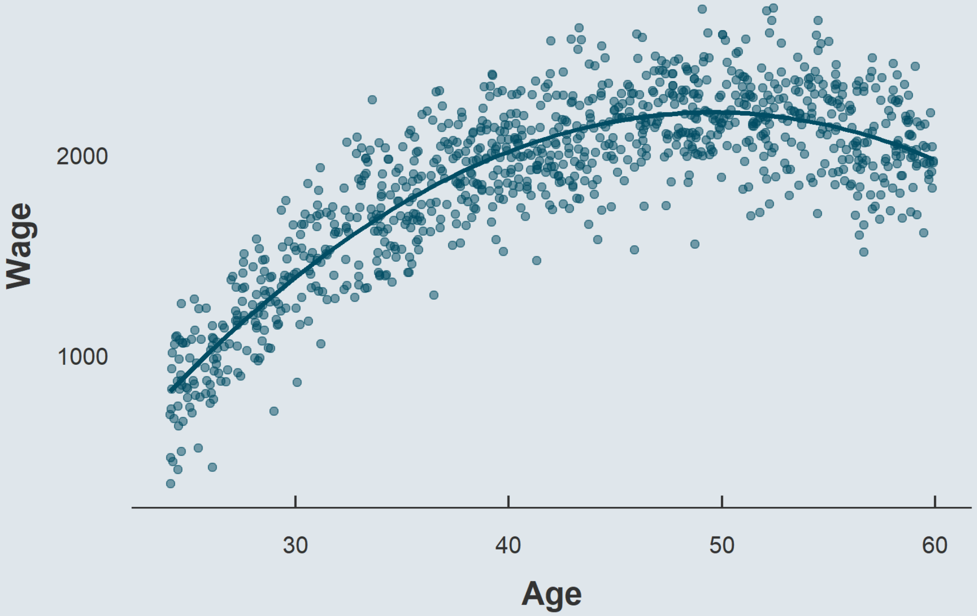
- Log vs. level

Interpretation of the regression coefficient

| | y | log(y) |
|---|---|---|
| **x** | $\hat{\beta}$ is the unit increase in $y$ due to a 1 unit increase in $x$ | $\hat{\beta} \times 100$ is the % increase in $y$ due to a 1 unit increase in $x$ |
| **log(x)** | $\hat{\beta} \div 100$ is the unit increase in $y$ due to a 1% increase in $x$ | $\hat{\beta}$ is the % increase in $y$ due to a 1% increase in $x$ |

- Polynomials

$$\text{Wage}_i = -2644 + 192 \times \text{Age}_i - 2 \times \text{Age}_i^2 + \hat{\varepsilon}_i$$



# Today: Causality and randomness

**1. Causality**

- 1.1. Omitted variable bias
- 1.2. Selection bias and counterfactual

**2. Randomness**

- 2.1. Data Generating Process & Random variables
- 2.2. Theoretical vs. empirical moments
- 2.3. $\beta$ vs. $\hat{\beta}$

**3. Causality from randomness**

- 3.1. Randomized Controlled Trials
- 3.2. Types of randomization
- 3.3. Multiple testing

**4. Wrap up!**

# Today: Causality and randomness

**1. Causality**

- 1.1. Omitted variable bias
- 1.2. Selection bias and counterfactual

# 1. Causality

## 1.1. Omitted variable bias

- Consider the following regression

$$\text{Earnings}_i = \hat{\alpha} + \hat{\beta}1\{\text{Sex}_i = \text{Male}\} + \hat{\varepsilon}_i$$

```
summary(lm(Earnings ~ Sex, asec_2020))$coefficients
```

```
##             Estimate Std. Error   t value      Pr(>|t|)
## (Intercept) 50915.03   436.8459 116.55148  0.000000e+00
## SexMale     21612.33   606.3649  35.64245 1.519977e-275
```

- Taking $\hat{\beta}$ at face value, the *'expected returns'* to being a male amount to $21,612.33 in annual earnings
- What do you think about this estimation?
  - Can we say that the estimated effect is causal?
  - What could bias our estimation?

# 1. Causality

### 1.1. Omitted variable bias

- The relationship could be impacted by many variables
    - For instance, inflated by the fact that males tend to both be better paid and work part-time less often

- The variable for hours of work here acts as a **confounding factor** because it is correlated to both $x$ and $y$
    - We need to put it as a control variable

$$\text{Earnings}_i = \hat{\alpha} + \hat{\beta}_1 1\{\text{Sex}_i = \text{Male}\} + \hat{\beta}_2\text{Hours}_i + \hat{\varepsilon}_i$$

```
summary(lm(Earnings ~ Sex + Hours, asec_2020))$coefficients
```

```
##                Estimate Std. Error   t value      Pr(>|t|)
## (Intercept) -22296.150 1138.83531 -19.57803  4.216399e-85
## SexMale      13794.385  595.79894  23.15275 4.157614e-118
## Hours         1953.829   28.23501  69.19880  0.000000e+00
```

# 1. Causality

### 1.1. Omitted variable bias

- The coefficient of the variable Sex drops from 21,612.33 in the univariate regression to 13,794.39 when including sex in the regression
    - Why?

- Part of the estimated effect of sex on earnings was due to the fact that men tend to work more hours per week
    - This effect is now captured by the coefficient associated to the variable for hours of work

- There are plenty of omitted factors that could influence our estimates
    - How would the coefficients change if we were to control for having an executive position?

→ ***We cannot consider this effect as causal!***

# 1. Causality

### 1.1. Omitted variable bias

- When interpreting coefficients, you should be unambiguous about what the estimation actually means

- Here we would say that "*controlling for the number of hours worked per week, being a male relative to a female is associated with a $13,794.39 increase in annual earnings on average, **everything else equal***".
  - Everything else equal (or *ceteris paribus*), means that the coefficient we estimate should be interpreted as the effect of $x$ on $y$ if everything else would remain constant when varying $x$

- Being a male relative to a female is also likely to be associated with an increase in the probability to have an executive position
  - But having an executive position would also increase the expected annual earnings
  - So when we say *ceteris paribus*, we say that the effect we estimate is attributable to the Sex variable *assuming that when the Sex variable changes other factors remain constant*

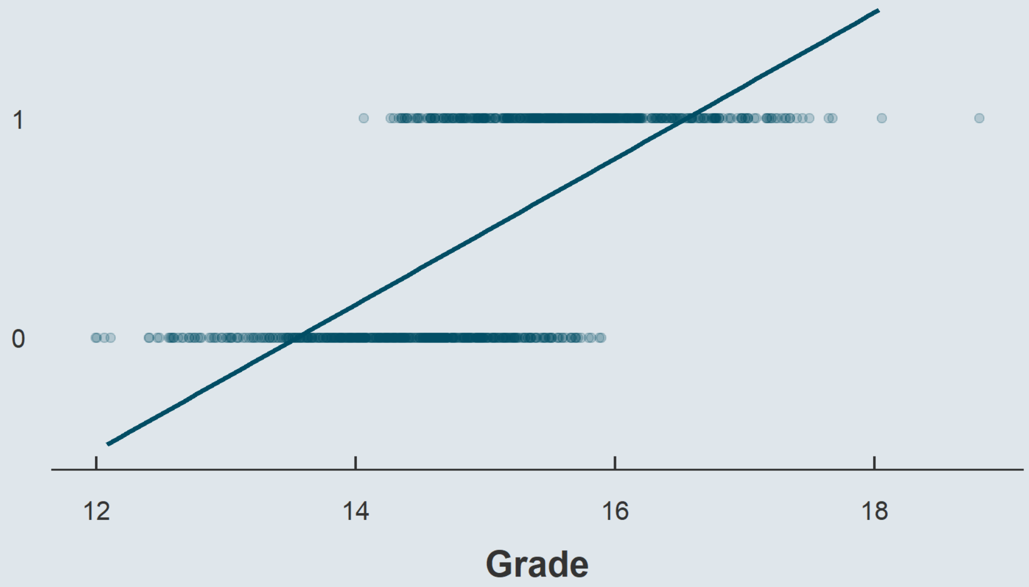
- We know this assumption is not correct ➜ we should mention it not because we believe it is the case but to be clear about what the coefficient means

# 1. Causality

### 1.2. Selection bias and counterfactual

- Omitted variable bias is a common problem
    - But estimations can be biased for many other reasons

- Remember the example about candidates to a position from last time:



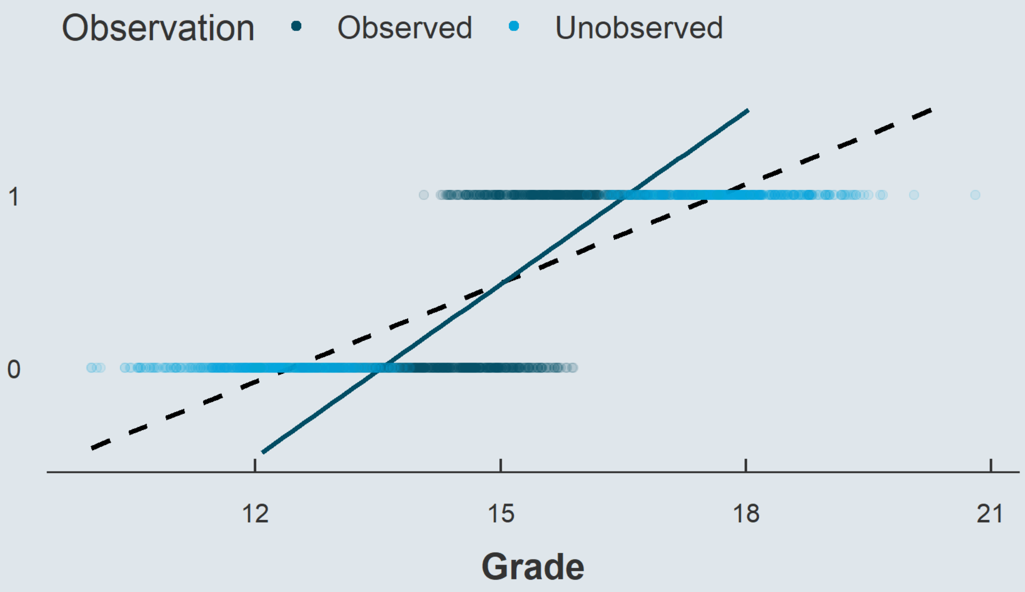
- We estimated that having a 1 unit increase in Grade (/20) would increase the probability to be accepted by about a third on expectation, *ceteris paribus*

- What do you think about this effect?
    - Look at the support of the $x$ variable

# 1. Causality

### 1.2. Selection bias and counterfactual

- The fact that grades range from 12 to 18 hints at a selection problem:
  - Individuals with very low grade won't apply to the position because they know they will be rejected
  - Individuals with very high grade won't apply to the position because they apply to better positions



- Had these individuals applied, the estimated effect would be lower

- Here the estimated coefficient is specific to a sample which is not representative of the whole population
  - Issue of **external validity**
  - The interpretation only holds in our specific setting, we cannot extrapolate

# 1. Causality

### 1.2. Selection bias and counterfactual

- Such selection problems are very common threats to causality

- What is the impact of going to a better neighborhood on your children outcomes?
  - Those who move may be different from those who stay: **self-selection issue**
  - Here it is not that the sample is not representative of the population, but that **the outcomes of those who stayed are different from the outcomes those who moved would have had, if they had stayed**

- This related to the notion of **counterfactual**
  - If those who moved were comparable to those who stayed, it would be valid to use the outcome of those who stayed as the counterfactual outcome of those who moved, i.e., the outcome they would have if they had stayed
  - But because of selection we do not have a credible counterfactual

- The notion of counterfactual is key to answer many questions
  - What is the impact of an immigrant inflow on the labor market outcomes of locals?
  - We need to know how the labor market outcomes of locals would have evolved absent the immigrant inflow but we do not observe this situation

# Overview

**1. Causality ✓**

- 1.1. Omitted variable bias
- 1.2. Selection bias and counterfactual

**2. Randomness**

- 2.1. Data Generating Process & Random variables
- 2.2. Theoretical vs. empirical moments
- 2.3. $\beta$ vs. $\hat{\beta}$

**3. Causality from randomness**

- 3.1. Randomized Controlled Trials
- 3.2. Types of randomization
- 3.3. Multiple testing

**4. Wrap up!**

# Overview

**1. Causality ✓**

- 1.1. Omitted variable bias
- 1.2. Selection bias and counterfactual

**2. Randomness**

- 2.1. Data Generating Process & Random variables
- 2.2. Theoretical vs. empirical moments
- 2.3. $\beta$ vs. $\hat{\beta}$

# 2. Randomness

### 2.1. Data Generating Process & Random variables

- So far in practice we have used many variables such as age, sex, earnings, etc.
    - But when we were working on theory we referred to these variable $x$ and $y$

- In Statistics and Econometrics these $x$ and $y$ we manipulate are called random variables
    - These variables can take values according to a data generating process (DGP)
    - The data generating process is the mechanism that causes the data to be the way we observe it

- For instance your grades can be seen as a random variable
    - Which takes given values according to an unknown data generating process
    - The DGP probably depends on how much effort you exert, on your background, on many environmental factors, ...

# 2. Randomness

### 2.1. Data Generating Process & Random variables

- Consider for instance the outcome of two dice as a random variable
  - Contrarily to the variables we usually study, we know the DGP of this one

- The DGP causes our random variable to take the following values with the following probabilities:

2 - 1/36 (⚀⚀)

3 - 2/36 (⚀⚁ - ⚁⚀)

4 - 3/36 (⚀⚂ - ⚂⚀ - ⚁⚁)

5 - 4/36 (⚁⚂ - ⚂⚁ - ⚁⚃ - ⚃⚁)

6 - 5/36 (⚂⚂ - ⚁⚃ - ⚃⚁ - ⚄⚀ - ⚀⚄)

7 - 6/36 (⚂⚃ - ⚃⚂ - ⚄⚁ - ⚁⚄ - ⚀⚅ - ⚅⚀)

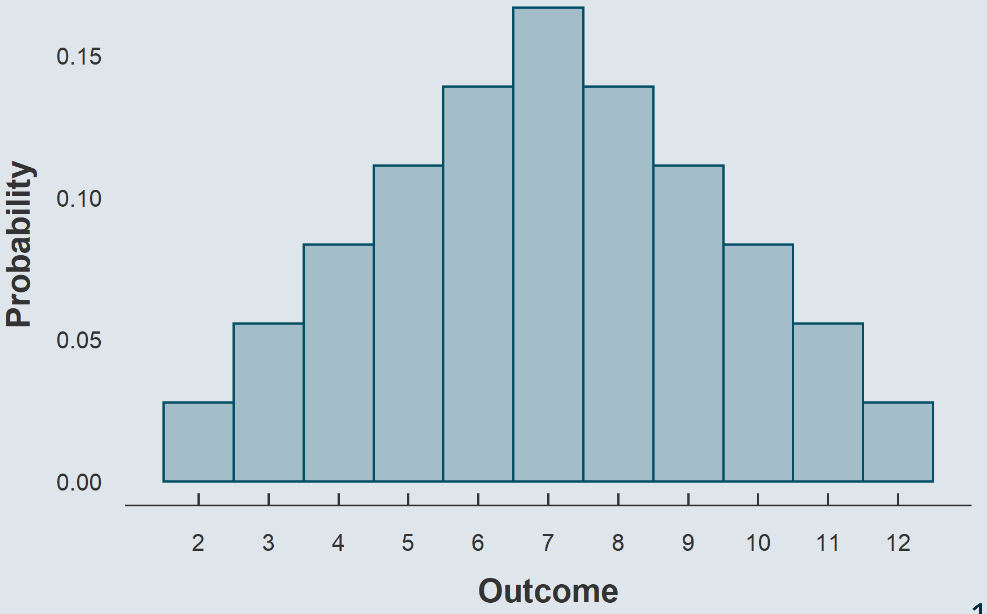
8 - 5/36 (⚃⚃ - ⚄⚂ - ⚂⚄ - ⚅⚁ - ⚁⚅)

9 - 4/36 (⚃⚄ - ⚄⚃ - ⚅⚂ - ⚂⚅)

10 - 3/36 (⚅⚃ - ⚃⚅ - ⚄⚄)

11 - 2/36 (⚅⚄ - ⚄⚅)

12 - 1/36 (⚅⚅)



# 2. Randomness

### 2.2. Theoretical vs. empirical moments

- Because we know the data generating process of our random variable, we can compute its expected value:

$$\mathrm{E}\left[X\right] = \frac{(2\times 1)+(3\times 2)+(4\times 3)+(5\times 4)+(6\times 5)+(7\times 6)}{36}+ \\ \frac{(8\times 5)+(9\times 4)+(10\times 3)+(11\times 2)+(12\times 1)}{36} = \frac{252}{36} = 7$$

- Note that we talk about **expected value** for the **theoretical moment** of the distribution, while we talk about the **mean value** for its **empirical counterpart**
  - For a given number of draws the mean won't necessarily be exactly 7
  - But if we were to do infinitely many draws, the mean would converge towards 7

***Let's try it out!***

# 2. Randomness

### 2.2. Theoretical vs. empirical moments

- We can simulate the data generating process by:
  - Storing every possible outcome in a vector
  - Randomly picking outcomes from this vector with the function sample()

```
outcomes <- c(2, 3, 3, 4, 4, 4, 5, 5, 5, 5, 6, 6, 6,
              6, 6, 7, 7, 7, 7, 7, 7, 8, 8, 8, 8, 8,
              9, 9, 9, 9, 10, 10, 10, 11, 11, 12)

sample(outcomes, 10, replace = T)
```

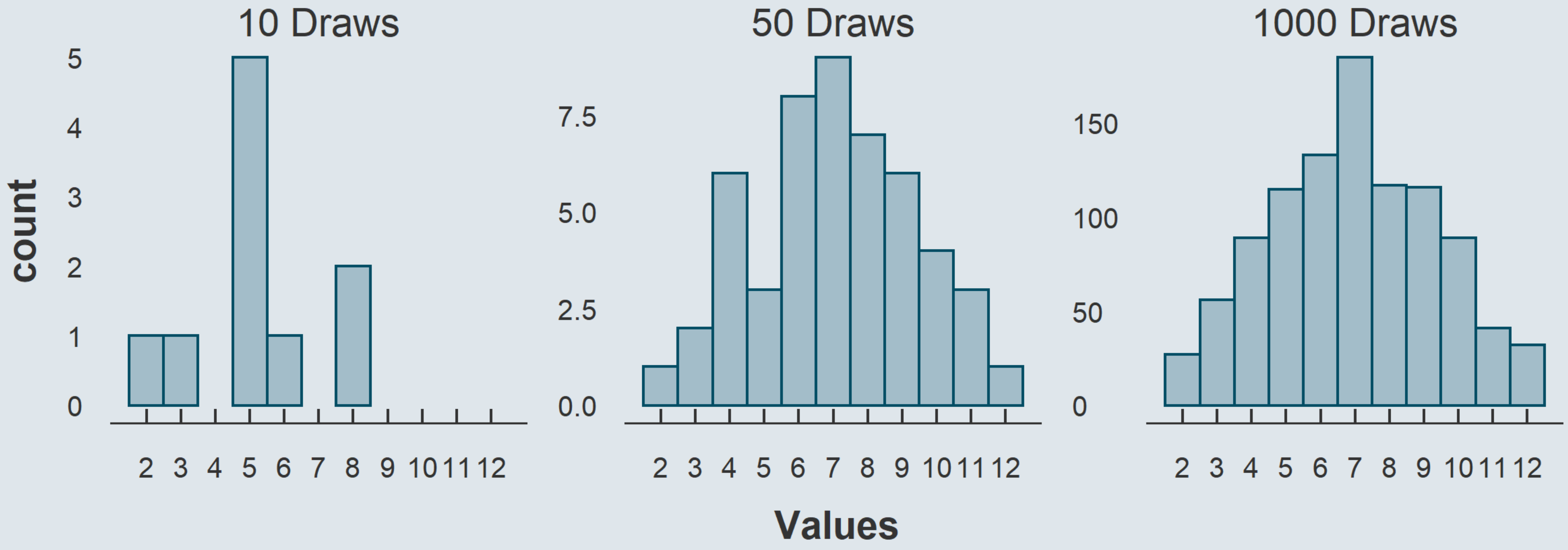
```
##  [1]  8  8  6  4  8  9  4 10  8  3
```

- The `replace` argument of the sample() function allows to indicate whether we want each randomly drawn value to be *'replaced'* before the next draw or removed from the vector

***Let's have a look at the distribution of outcomes for different numbers of draws***

# 2. Randomness

### 2.2. Theoretical vs. empirical moments



- The larger the number of draws the closer the mean to the expected value (here 5.2, 7.06, and 6.95)
  - If we were to roll two dice infinitely many times the mean would converge towards its expected value
  - This is what we call the *Law of Large Numbers* (LLN)

# 2. Randomness

### 2.2. Theoretical vs. empirical moments

- In this respect, we can view the mean as an estimation of the expected value
  - So just like we computed a standard error to get a sense of how far $\hat{\beta}$ is likely to be from the true $\beta$...
  - ... we can compute a standard error to get a sense of how far $\overline{X}$ is likely to be from $\mathrm{E}\left[X\right]$

- The formula of the standard error writes:

$$SE(\overline{X}) = \frac{SD(X)}{\sqrt{N}}$$

- And the reasoning for the 95% confidence interval is the same as with $\hat{\beta}$

$$\Pr\left[\overline{X} - t_{97.5\%} \times \mathrm{se}(\overline{X}) \leq \mathrm{E}\left[X\right] \leq \overline{X} + t_{97.5\%} \times \mathrm{se}(\overline{X})\right] = 95\%$$

  - With $t_{97.5\%}$ computed from a student t distribution with N-1 degrees of freedom

# 2. Randomness

## 2.2. Theoretical vs. empirical moments

- Just like the mean that we compute empirically is an estimate of the first moment of the distribution,
  - the variance that we compute empirically is an estimate of the second moment of the distribution

| | **Theoretical moment** | **Empirical moment** |
|---|---|---|
| **First moment:** | $\mathrm{E}(X_{\text{discrete}}) = \sum_{i=1}^{k} x_i p_i$<br>$\mathrm{E}(X_{\text{continuous}}) = \int_{\mathrm{R}} x f(x) dx$ | $\overline{X} = \frac{1}{N} \sum_{i=1}^{N} x_i$ |
| **Second moment:** | $\mathrm{Var}(X) = \mathrm{E}\left[(X - \mathrm{E}(X))^2\right] \equiv \sigma^2$ | $\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2$ |

# 2. Randomness

### 2.3. $\beta$ vs. $\hat{\beta}$

- The distinction we made between $\beta$ and $\hat{\beta}$ is of the same nature
  - Let's illustrate that with a little simulation exercise

- The function mvrnorm from the MASS package generates two normally distributed random variables
  - According to: their **covariance** and respective **expected value** and **variance**

```
data <- mvrnorm(n = #observations, mu = c(E[X1], E[X2]), Sigma = covariance matrix)
```

- The covariance matrix should be of the form

Variance-covariance matrix

| | **X1** | **X2** |
|---|---|---|
| **X1** | Var(X1) | Cov(X1, X2) |
| **X2** | Cov(X2, X1) | Var(X2) |

- For instance, with $\text{Var}(X_1) = 1$, $\text{Var}(X_2) = 2$, and $\text{Cov}(X_1, X_2) = .5$:

```
matrix(c(1, .5, .5, 2), nrow = 2)
```

```
##      [,1] [,2]
## [1,]  1.0  0.5
## [2,]  0.5  2.0
```

# Practice

**1) Use the `mvrnorm` function from the `MASS` package to generate two variables $X$ and $Y$ with:**

- $N = 1000$
- $\mathrm{E}\left[X\right] = 5$
- $\mathrm{E}\left[Y\right] = 30$
- $\mathrm{Var}(X) = 2$
- $\mathrm{Var}(Y) = 10$
- $\mathrm{Cov}(X, Y) = 4$

**2) Compute the empirical counterpart of the first and second moments of the joint distribution**

*You've got 5 minutes!*

# Solution

- Generate the joint distribution

```
library(MASS)
data <- mvrnorm(1000, c(5, 30), matrix(c(2, 4, 4, 10), 2))
x <- data[, 1]
y <- data[, 2]
```

- First empirical moment

```
c(mean(x), mean(y))
```

```
## [1]  4.995905 29.959186
```

- Second empirical moment

```
matrix(c(var(x), cov(x, y), cov(y, x), var(y)), 2)
```

```
##          [,1]     [,2]
## [1,] 1.987344 3.862527
## [2,] 3.862527 9.582982
```

# 2. Randomness

### 2.3. $\beta$ vs. $\hat{\beta}$

- Because we know the joint DGP of $X$ and $Y$, we do know the actual values of $\alpha$ and $\beta$ from the regression

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

$$\beta = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} = \frac{4}{2} = 2$$

$$\alpha = \text{E}\left[Y\right] - \beta \times \text{E}\left[X\right] = 30 - 2 \times 5 = 20$$

- And the coefficients $\hat{\alpha}$ and $\hat{\beta}$ we can compute using observed data are estimates of these true parameters

```
summary(lm(y ~ x))$coefficients
```

```
##              Estimate Std. Error   t value Pr(>|t|)
## (Intercept) 20.249336 0.16793377 120.57929        0
## x            1.943562 0.03235219  60.07514        0
```

# 2. Randomness

### 2.3. $\beta$ vs. $\hat{\beta}$

- The higher the number of observations, the closer $\hat{\beta}$ from $\beta$ on expectation:

```r
beta_hat <- function(n){
  data <- mvrnorm(n, c(5, 30), matrix(c(2, 4, 4, 10), 2))
  x <- data[, 1]
  y <- data[, 2]
  return(summary(lm(y ~ x))$coefficients[2, 1])
}
c(beta_hat(10), beta_hat(1000), beta_hat(100000))
```

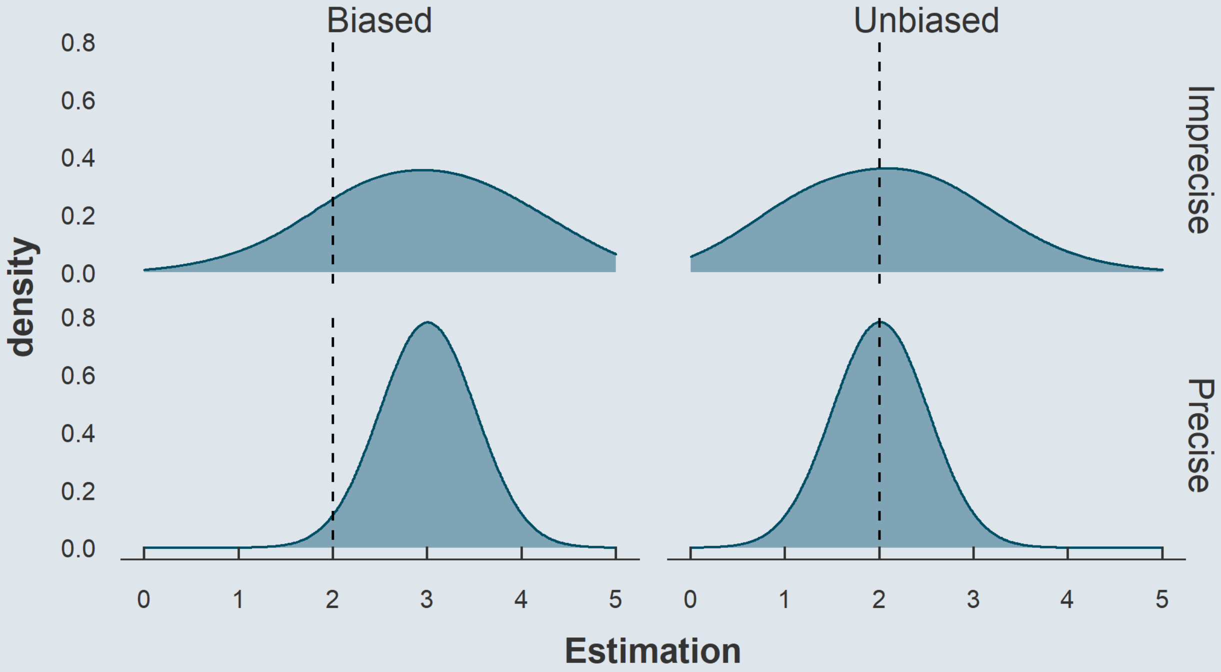
```
## [1] 2.524306 2.034427 1.997814
```

- This is what we call *consistency*
  - With this DGP the OLS estimator is consistent
  - But this is not always the case
  - You'll see the conditions for that next year

# 2. Randomness

### 2.3. $\beta$ vs. $\hat{\beta}$

- Keep in mind that consistency, unbiasedness, and precision, are very distinct concepts
    - Consider these 4 cases where we compare the distribution of estimations $\hat{\beta}$ from 1,000 randomly drawn samples to the true $\beta$



- An estimator is **unbiased** if on expectation it gives the true value we want to estimate
- An estimator is **precise** if the estimations it provides are close to each other (low variance)
- An estimator is **consistent** if the larger the sample size the higher the probability that we obtain the true value we want to estimate

# Overview

**1. Causality ✓**

- 1.1. Omitted variable bias
- 1.2. Selection bias and counterfactual

**2. Randomness ✓**

- 2.1. Data Generating Process & Random variables
- 2.2. Theoretical vs. empirical moments
- 2.3. $\beta$ vs. $\hat{\beta}$

**3. Causality from randomness**

- 3.1. Randomized Controlled Trials
- 3.2. Types of randomization
- 3.3. Multiple testing

**4. Wrap up!**

# Overview

**1. Causality ✓**

- 1.1. Omitted variable bias
- 1.2. Selection bias and counterfactual

**2. Randomness ✓**

- 2.1. Data Generating Process & Random variables
- 2.2. Theoretical vs. empirical moments
- 2.3. $\beta$ vs. $\hat{\beta}$

**3. Causality from randomness**

- 3.1. Randomized Controlled Trials
- 3.2. Types of randomization
- 3.3. Multiple testing

# 3. Causality from randomness

### 3.1. Randomized Controlled Trials

- A Randomized Controlled Trial (RCT) is a type of experiment in which the thing we want to know the impact of (called the treatment) is randomly allocated in the population
  - It is a way to obtain causality from randomness

Take for instance the `asec_2020.csv` dataset we've been working with:

```
asec_2020 %>% group_by(n()) %>%
  summarise(`Mean Earnings` = mean(Earnings),
         `% Female` = 100 * mean(Sex == "Female"),
         `% Black` = 100 * mean(Race == "Black"),
         `% Asian` = 100 * mean(Race == "Asian"),
         `% Other` = 100 * mean(Race == "Other"),
         `Mean Hours` = mean(Hours))
```

```
## # A tibble: 1 x 7
##   `n()` `Mean Earnings` `% Female` `% Black` `% Asian` `% Other` `Mean Hours`
##   <int>           <dbl>      <dbl>     <dbl>     <dbl>     <dbl>        <dbl>
## 1 64336          62132.       48.1      10.6      7.04      3.76         39.5
```

# 3. Causality from randomness

### 3.1. Randomized Controlled Trials

- Let's compare the average characteristics for two randomly selected groups:

```
asec_2020 %>%
  mutate(Group = ifelse(rnorm(n(), 0, 1) > 0, 1, 0)) %>%
  group_by(Group) %>%
  summarise(`Mean Earnings` = mean(Earnings),
            `% Female` = 100 * mean(Sex == "Female"),
            `% Black` = 100 * mean(Race == "Black"),
            `% Asian` = 100 * mean(Race == "Asian"),
            `% Other` = 100 * mean(Race == "Other"),
            `Mean Hours` = mean(Hours))
```

```
## # A tibble: 2 x 7
##   Group `Mean Earnings` `% Female` `% Black` `% Asian` `% Other` `Mean Hours`
##   <dbl>           <dbl>      <dbl>     <dbl>     <dbl>     <dbl>        <dbl>
## 1     0          61836.       48.3      10.7      6.97      3.78         39.6
## 2     1          62429.       47.9      10.6      7.11      3.75         39.5
```

# 3. Causality from randomness

### 3.1. Randomized Controlled Trials

- Their average characteristics are very close!
  - On expectation their average characteristics are the same
- And just as the two randomly selected populations are comparable in terms of their observable characteristics
  - On expectation they are also **comparable** in terms of their **unobservable characteristics!**
  - Randomization, if properly conducted, thus solves the problem of omitted variable bias

***If we assign a treatment to Group 1, Group 2 would then be a valid counterfactual to estimate a causal effect!***

- But RCTs are not immune to every problem:
  - If individuals self-select in participating to the experiment their would be a selection bias
  - Even without self-selection, if the population among which treatment is randomized is not representative there is a problem of external validity
  - For the RCT to work, individuals should comply with the treatment allocation
  - The sample must be sufficiently large for the average characteristics across groups to be close enough to their expected value
  - ...

# 3. Causality from randomness

### 3.2. Types of randomization

- To some extent their are ways to deal with these problems

- For instance if we want to ensure that a characteristic is well balanced among the two groups, we can **randomize within categories of this variable**
  - Instead of giving the treatment randomly and hoping that we will obtain the same % of females in both groups
  - We assign the treatment randomly among females and among males separately
  - This is called **randomizing by block**
  - *Note that this only works with observable characteristics!*

```
asec_2020 %>%
  group_by(Sex) %>% # Randomize treatment by sex
  mutate(Group = ifelse(rnorm(n(), 0, 1) > 0, 1, 0)) %>%
  ungroup() %>% group_by(Group) %>%
  summarise(...)
```

# 3. Causality from randomness

### 3.2. Types of randomization

- Now imagine that you want to estimate the impact of calory intake at the 10am break on pupils grades
  - You regularly give a snack to a sample of randomly selected children and a few months later you test whether there is a significant difference between their grades and that of untreated children
  - Do you expect the estimated effect to reflect the actual impact you aim to measure?

- What if some children shared their snack with untreated children?
  - These *treated children* would have *less calories* and then possibly lower grades than under full compliance
  - And their *untreated* friends would have *more calories* than expected and then possibly higher grades
  - Thus, this ***spillover effect*** would tend to fallaciously shrink the observed effect of the treatment

- One solution to that problem is to **randomize by cluster**
  - Instead of considering the treatment to be at the child level
  - Consider that the treatment is a the school level
  - A treated unit is a school where some/all children are treated
  - An untreated school is a school where no child is treated

*Beware that in terms of inference, computing standard errors the usual way while the treatment is at a broader observational level than the outcome would give fallaciously low standard errors, which would need to be corrected*

# 3. Causality from randomness

### 3.3. Multiple testing

- Another inference issue that RCTs can be subject to is multiple testing
    - If you conduct a well-designed RCT you might be tempted to exploit the causal framework to test a myriad of effects

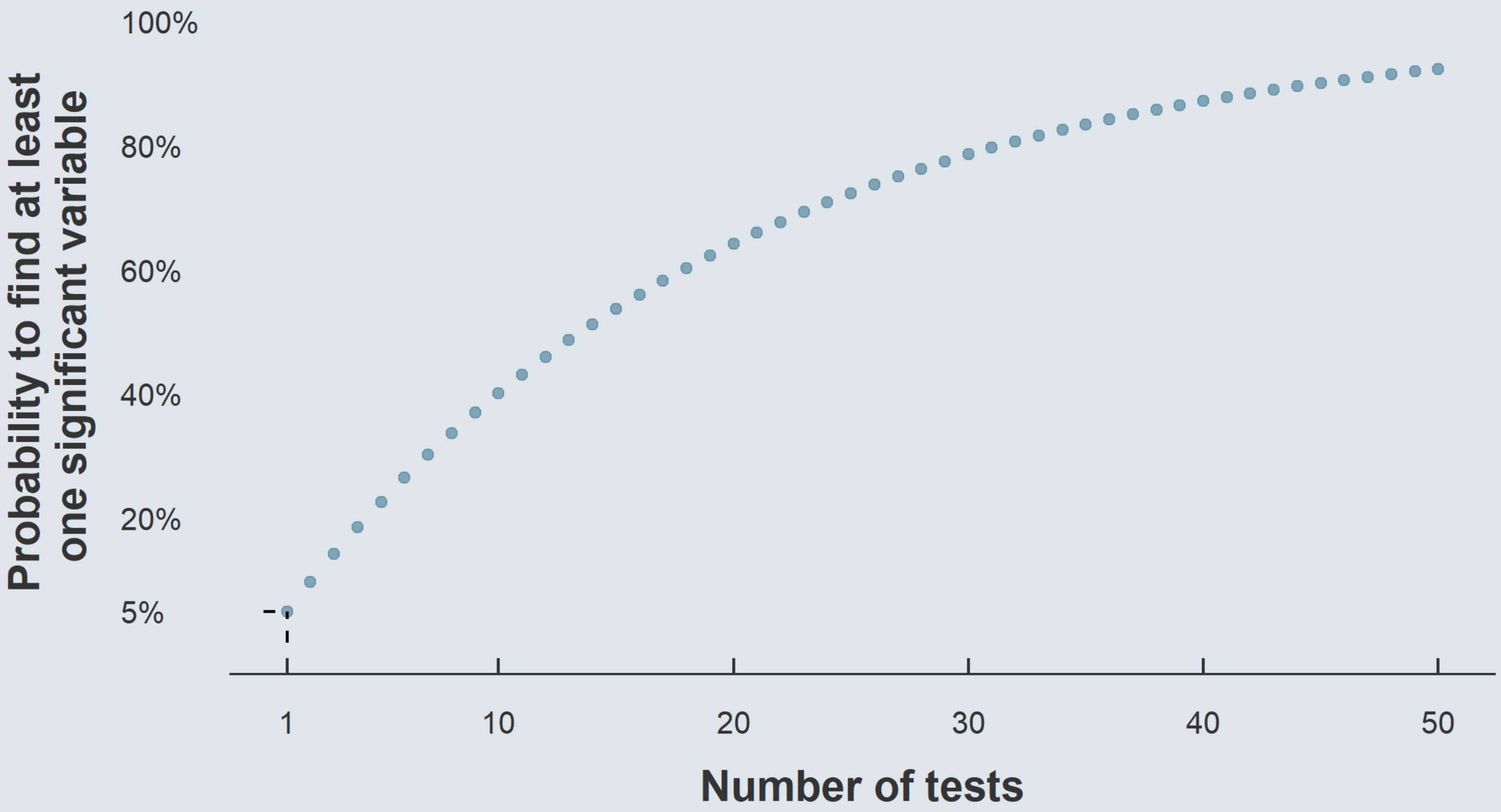
- You randomize your treatment and you compare the averages of many outcomes between treated and untreated individuals
    - You would be tempted to conclude that there is a significant effect for every variable whose corresponding p-value is lower that .05
    - But you cannot do that!

- The probability to have a p-value lower than .05 just by chance for one test is indeed 5%
    - But if you do multiple tests in a row, the probability to have a p-value lower than .05 among these multiple tests is greater than 5%
    - The greater the number of tests you perform, the higher the probability to get a significant result just by chance

**This is what we call *multiple testing***

# 3. Causality from randomness

### 3.3. Multiple testing



# 3. Causality from randomness

### 3.3. Multiple testing

- There are many ways to correct for multiple testing

- The simplest one is called the **Bonferroni** correction
  - It consists in **multiplying the p-value by the number of tests**
  - But it also leads to a large **loss of power** (the probability to find an effect when there is indeed an effect decreases a lot)

- There are more sophisticated ways to deal with the problem, which can be categorized into two approaches
  - **Family Wise Error Rate**: Control the probability that there is at least one true assumption rejected
  - **False Discovery Rate**: Control the share of true assumptions among rejected assumptions

→ *We won't cover these methods in this course but keep the multiple testing issue in mind when you encounter a long series of statistical tests*

# Overview

**1. Causality ✓**

- 1.1. Omitted variable bias
- 1.2. Selection bias and counterfactual

**2. Randomness ✓**

- 2.1. Data Generating Process & Random variables
- 2.2. Theoretical vs. empirical moments
- 2.3. $\beta$ vs. $\hat{\beta}$

**3. Causality from randomness ✓**

- 3.1. Randomized Controlled Trials
- 3.2. Types of randomization
- 3.3. Multiple testing

**4. Wrap up!**

# 4. Wrap up!

**Causality**

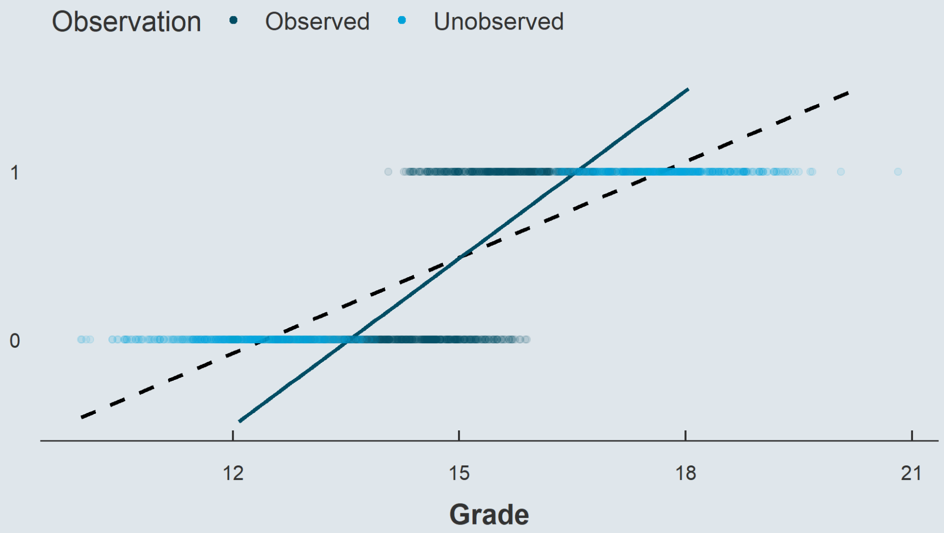
1) Omitted variable bias

- Regressing Earnings on Sex = Male without controls yields $\hat{\beta} = 21612.33$
- But controlling for weekly hours we obtain $\hat{\beta} = 13794.39$

➜ Variables that are correlated to both $x$ and $y$ should be controlled for

➜ And the coefficients must unambiguously be interpreted *ceteris paribus*

2) Selection bias



- Self-selection selection into the population studied causes problems of **external validity**
- Self-selection into the treatment variable causes problems of **counterfactual validity**

# 4. Wrap up!

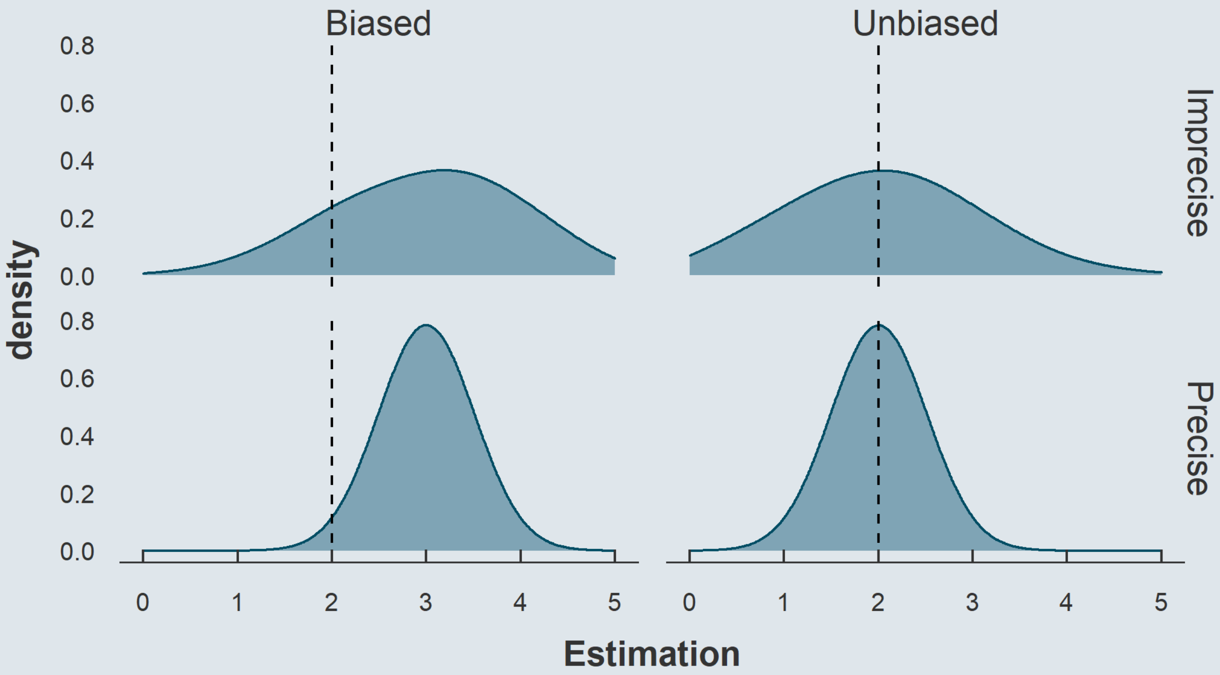
### Theoretical vs. empirical moments

| | **Theoretical moment** | **Empirical moment** |
|---|---|---|
| **First moment:** | $\mathrm{E}(X_{\text{discrete}}) = \sum_{i=1}^{k} x_i p_i$<br>$\mathrm{E}(X_{\text{continuous}}) = \int_{\mathrm{R}} x f(x) dx$ | $\overline{X} = \frac{1}{N} \sum_{i=1}^{N} x_i$ |
| **Second moment:** | $\mathrm{Var}(X) = \mathrm{E}\left[(X - \mathrm{E}(X))^2\right] \equiv \sigma^2$ | $\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2$ |

# 4. Wrap up!

## $\beta$ vs. $\hat{\beta}$

- Keep in mind that consistency, unbiasedness, and precision, are very distinct concepts
    - Consider these 4 cases where we compare the distribution of estimations $\hat{\beta}$ from 1,000 randomly drawn samples to the true $\beta$



- An estimator is **unbiased** if on expectation it gives the true value we want to estimate
- An estimator is **precise** if the estimations it provides are close to each other (low variance)
- An estimator is **consistent** if the larger the sample size the higher the probability that we obtain the true value we want to estimate

# 4. Wrap up!

**Randomized controlled trials**

- A Randomized Controlled Trial (RCT) is a type of experiment in which the thing we want to know the impact of (called the treatment) is randomly allocated in the population
  - It is a way to obtain causality from randomness as on expectation two randomly drawn population have the same average observable and unobservable characteristics, which solves the omitted variable bias

- But RCTs are not immune to every problem:
  - Self-selection issues can arise
  - The population should be representative for external validity
  - Individuals should comply with the treatment allocation
  - The sample must be sufficiently large
  - ...

- There are different types of randomization to help dealing with such problems
  - **Randomization by block for small samples:** Randomly assign the treatment within groups of individuals whose characteristic should be balanced
  - **Randomization by cluster for spillovers:** If spillovers may occur within given units, consider these units as the observational level for the treatment allocation
  - ...