CS&SS 321 - Data Science and Statistics for Social Sciences

Module III - Introduction to causal inference and linear models

Ramses Llobet

Module III

- This module introduces and reviews the topic of causation in science.
  - *randomization*.
  - *applied causal inference*.
  - *causal modeling* (module IV).
- It also introduces the **linear regression model** and the method of **least squares** (LS).

The statistics war of the late XXth century

The statistics war of the XXIth century

- Causal inferences requires a model outside of the statistical model.

Causes in, causes out

- Why do experiments work? When do they work?
- What if treatment is imperfect assigned?
- Should you *control* for anything? Everything?

Answers depend upon **causal assumptions** ($\rightarrow$).

- An **assumption** is a premise or supposition that is accepted *without direct evidence*, often forming the basis for reasoning or an argument.

Causes in, causes out

- Causal assumptions requires **causal knowledge** of social systems.
- For example, where X represents **rain** and Y represents **puddles**.
  - What **causal assumption** ($\rightarrow$) you find more reasonable?

(i) $X \leftarrow Y$

(ii) $X \rightarrow Y$

Causal design

- **Step 1**: sketch a (scientific) casual model: $X \to Y$.
  - *Causes in*: assumptions reflect **background knowledge** (*theory and literature review*).
- **Step 2**: use the model to design **data collection** and **statistical procedures**.
- **Step 3**: use statistical analyses to **hypothesis test** and report results.
  - *Causes out*: test assumptions' implications about the **causal mechanism**.

Causal design: intervention

- In causal inference, an **intervention** is a deliberate and controlled manipulation of one or more variables in a system to assess their **causal impact** on the outcome of interest.
  - *Example*: Pouring a bucket of water on the floor creates a puddle; does rain follow?
- We formalize this via the **potential outcomes** framework.

Causation in science

Treatment indicator: $T_i \in \{0, 1\}$, where i refers respondents.

- **(1) example:**
  - $T_i = 0$ indicates no membership in a union.
  - $T_i = 1$ indicates membership in a union.
- **(2) example:**
  - $T_i = 0$ indicates no daughters.
  - $T_i = 1$ indicates having daughters.

Outcome: Y_i

- **(1) example:** redistribution attitudes (*gincdif*).
- **(2) example:** pro-feminist attitudes (*progressive.vote*).

Causation in science

- Consider the treatments' (T) **causal mechanisms** ($\rightarrow$) that drives the **outcome** (Y).
  - **Why** does labor **union membership** increase support for redistribution?
  - **Why** does having a **daughter** increase pro-feminist attitudes?

Potential outcomes $Y_i(0)$, $Y_i(1)$, where:

- **(1) example:**
  - $Y_i(0)$ represents redistribution attitudes *without* membership.
  - $Y_i(1)$ represents redistribution attitudes *with* membership.
- **(2) example:**
  - $Y_i(0)$ represents pro-feminist attitudes *without* daughters.
  - $Y_i(1)$ represents pro-feminist attitudes *with* daughters.

Causation in science

The **fundamental problem of causality**, we cannot observe two outcomes at the same time:

$$\text{individual treatment effect} = Y_{\text{Ramses}}(1) - Y_{\text{Ramses}}(0) \qquad (1)$$

Instead, we **estimate** effects by taking the differences in means between **treatment**, $\bar{Y}(1)$, and **control**, $\bar{Y}(0)$, groups.

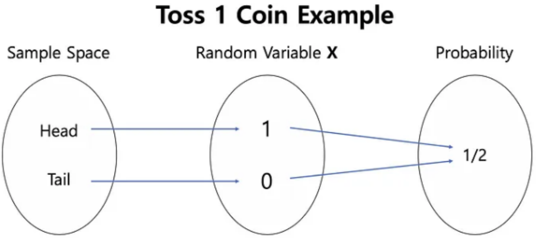
$$\text{average treatment effect} = \bar{Y}(1) - \bar{Y}(0) \qquad (2)$$

However, we can identify **ATE** if, and only if, the treatment D has been **randomly assigned** to each respondent `i`. Formally,

$$T_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \qquad (3)$$

Causation in science

- Think about random assignment as flipping a coin.
  - In **expectation** (as $n \to \infty$), a fair coin has a probability of 0.5 to show tails (0) or heads (1).
  - By definition, a random event has a probability of 0.5.



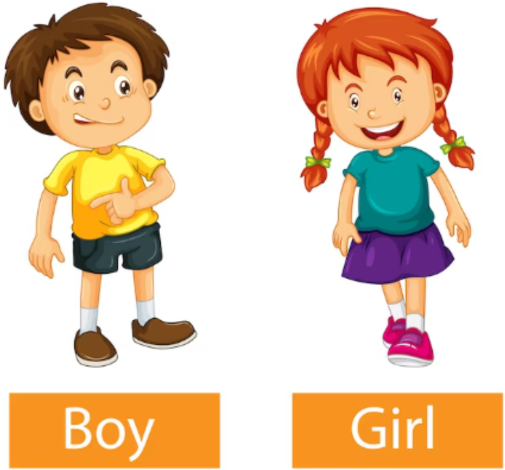
- **What if**, in expectation, a coin has a probability of 0.7 ?

Causation in science

- Is labor union membership a random occurrence?

Causation in science

- Is having a girl (instead of a boy) a random occurrence?

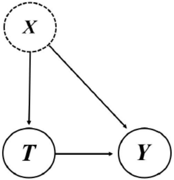


Causation in science

- **Selection bias**: Self-selection and unbalanced factors introduce bias in our statistical estimations.
    - *Self-selection*: Left-wing individuals are more likely to become labor union activists.
    - *Unbalanced factors*: Labor union members may systematically differ from non-union members in terms of factors such as occupation and income.

Causation in science

- In observational studies, unconditional treatment effects are unlikely due to the influence of **confounding** factors, both **observed** and **unobserved**.



- However, sometimes we can assume **conditional random effects**.

$$T_i \perp\!\!\!\perp \left(Y_i\left(0\right), Y_i\left(1\right)\right) | X_i. \tag{4}$$

Causation in science

- Let's work a short coding example.
- Open the file `unions_sweden.Rmd`, we will do only the **first** section.
- We will finish the remaining section next week.

From previous model: Data Generating Process

- Two very useful pieces of information from a DGP are its **mean** and **standard deviation**.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{N} X_i \quad ; \quad S = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

where

- $\bar{X}$ represents the **sample mean**.
- N is the number of **observations** in the sample.
- X_i represents **values** from a variable in the sample.
- S represents the **sample standard deviation**.

Standard devitation and variance

- The **standard deviation** and **variance** are both measures of the spread of a distribution.
  - To estimate the variance (S^2), we simply take the **square** of the standard deviation (S).

$$S^{2} = \left(\sqrt{\frac{1}{N}\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)^{2} \qquad S^{2} = \frac{1}{N}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

- S^2 is the **sample** variance.
- Q: Why choose the standard deviation over the variance to report **summary statistics**?

Mean and variance

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{n} X_i \quad ; \quad S^2 = \frac{1}{N}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

- The **sample mean** ($\bar{X}$) describes the location (*the center*) of the data (*distribution*).
- The **sample variance** (S^2) measures the variability in the data (*distribution*).
  - The variance describes the **average deviation** in a distribution.

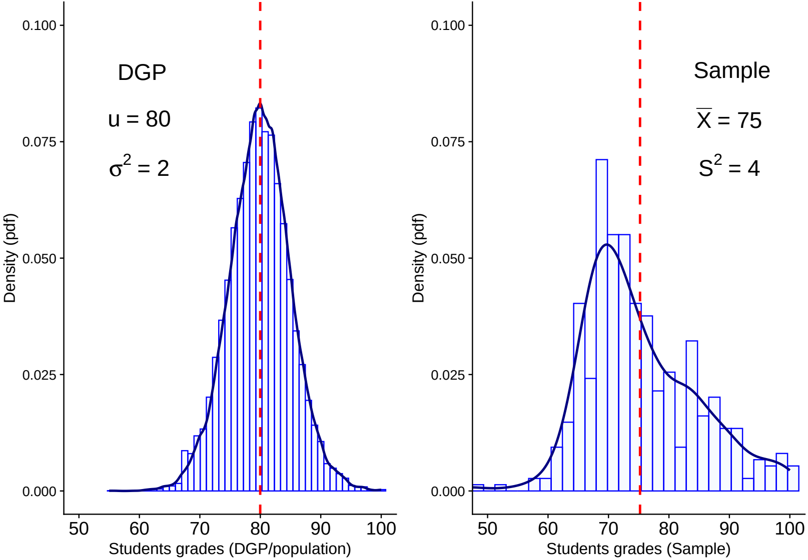
DGP vs. sample

We distinguish between the **Data Generating Process** (DGP) and the data **sample**.

- DGP or *population* is a **theoretical** concept describing how observed/sampled data is generated.
  - It follows a **distribution**, typically depicted as the *TRUE* (**!?**).
  - Its parameters, mean (μ) and variance (σ^2), are **fixed**.
- The sample is an **empirical** construct, representing realizations/occurrences of a data process.
  - Sample data maps into **distributions** of *random variables*.
  - Its parameters, mean ($\bar{X}$) and variance (S^2), are **random**.

Note: we use the sample to infer (**approach**) the underlying *TRUE* of a DGP.

DGP vs. sample



Unconditional distributions

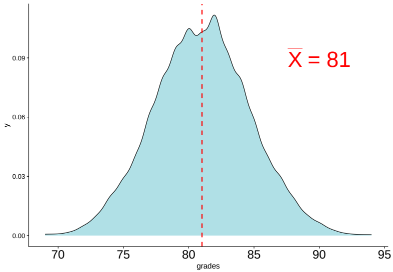
- The **expectation** $E[.]$ of a random variable X, denoted as $E[X]$, is a useful measure of central tendency of the DGP.
  - The expectation is also called the **expected value** or **mean**.
  - In the case of the normal distribution, the expectation is the first **central moment** and is denoted as μ.
- In general, a natural estimator of the expectation is the **sample mean**.

$$\mu = E[X] = \bar{X} = \frac{1}{n}\sum_{i=1}^{N} X_i$$

Unconditional distributions

- We have a sample of UW students' grades.
- What may be a good candidate to estimate the mean of this population?

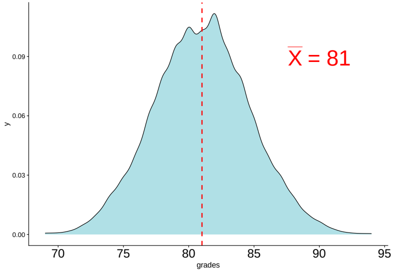
$$E[grades] = ?$$



Unconditional distributions

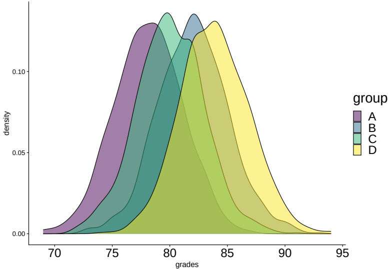
- We have a sample of UW students' grades.
- What may be a good candidate to estimate the mean of this population?

$$E[grades] = 81$$



Conditional distributions

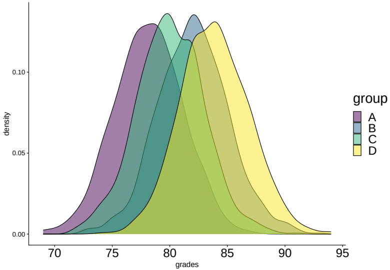
- However, is the distribution the same for all the departments?
- We can compare the grade distribution for these different sub-populations.
    - Group A
    - Group B
    - Group C
    - Group D



Conditional distributions

- We can condition the random variable grades on on a fixed value (x) of the variable group.
- We call this the **conditional mean** (or **conditional expectation**).

$$E[grades \mid group = x]$$

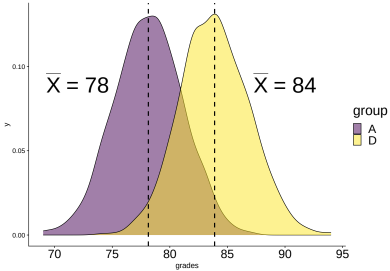


Conditional distributions

- For example, take the conditional mean of groups A and D.

$$E[grades \mid group = A] = 78$$

$$E[grades \mid group = D] = 84$$



Conditional distributions

- The values 78 and 84 are the mean grades in the subpopulations of groups A and D.

$$E[grades \mid group = D] - E[grades \mid group = A] = 84 - 78 = 6$$

- However, it is important to note that we **cannot** attribute **causality** or interpretation to these differences.
- Conditioning helps **describe variation**, but by themselves are neither a **model** nor an **explanation**.

Best predictor

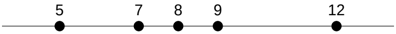
- In statistics, our focus is on predicting **quantities of interest**.
- Prediction is the **closest best guess** for a quantity, considering all sample realizations.
- The **prediction error** is the distance of each data point from our best guess.
- Hence, the best predictor, θ, minimizes the prediction error (e) as much as possible.

$$\text{MSE}: \; E[\,(Y_i - \theta)^2\,]$$

- A measure of prediction error magnitude is the expectation of its square, also called **Mean Squared Error**.
  - *Note*: The notation θ is arbitrary and, in this context, denotes the best predictor.

Prediction error: first guess

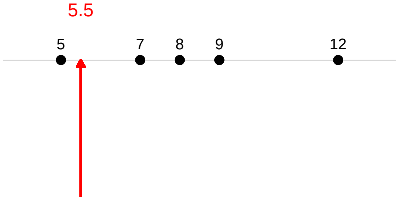
What is your **best guess** (θ) that **minimizes** the prediction error (MSE)?



N_i	Y_i	θ	$Y_i - \theta$	$error$
1	5			
2	7			
3	8			
4	9			
5	12			

$$MSE = E[\,(Y_i - \theta)^2\,]$$

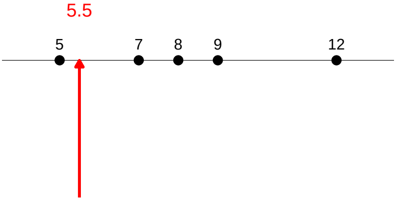
Prediction error: first guess



N_i	Y_i	θ	$Y_i - \theta$	*error*
1	5	5.5		
2	7	5.5		
3	8	5.5		
4	9	5.5		
5	12	5.5		

$$MSE = E[\,(Y_i - 5.5)^2\,]$$

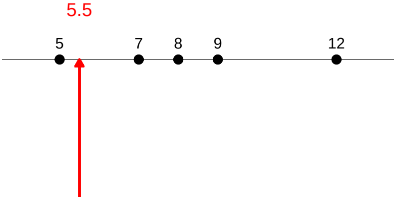
Prediction error: first guess



N_i	Y_i	θ	$Y_i - \theta$	*error*
1	5	5.5	5-5.5	
2	7	5.5	7-5.5	
3	8	5.5	8-5.5	
4	9	5.5	9-5.5	
5	12	5.5	12-5.5	

$$MSE = E[\,(Y_i - 5.5)^2\,]$$

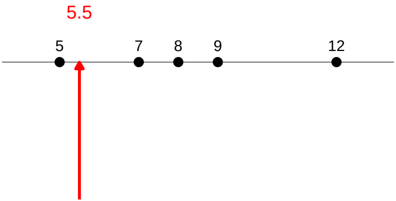
Prediction error: first guess



N_i	Y_i	θ	$Y_i - \theta$	*error*
1	5	5.5	5-5.5	-0.5
2	7	5.5	7-5.5	1.5
3	8	5.5	8-5.5	2.5
4	9	5.5	9-5.5	3.5
5	12	5.5	12-5.5	6.5

$$MSE_1 = \frac{1}{5}(-0.5 + 1.5 + 2.5 + 3.5 + 6.5)^2$$

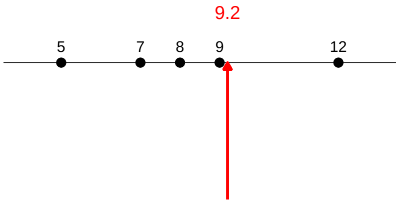
Prediction error: first guess



N_i	Y_i	θ	$Y_i - \theta$	*error*
1	5	5.5	5-5.5	-0.5
2	7	5.5	7-5.5	1.5
3	8	5.5	8-5.5	2.5
4	9	5.5	9-5.5	3.5
5	12	5.5	12-5.5	6.5

$$
\begin{aligned}
MSE_1 &= \frac{1}{5}(-0.5 + 1.5 + 2.5 + 3.5 + 6.5)^2 \\
&= \frac{(13.5)^2}{5} = \frac{182.25}{5} = \mathbf{36.45}
\end{aligned}
$$

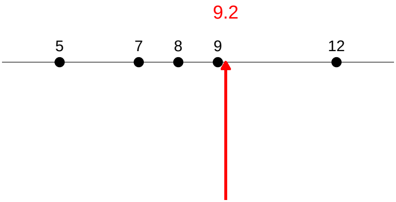
Prediction error: second guess



N_i	Y_i	θ	$Y_i - \theta$	*error*
1	5	9.2	5-9.2	
2	7	9.2	7-9.2	
3	8	9.2	8-9.2	
4	9	9.2	9-9.2	
5	12	9.2	12-9.2	

$$MSE_2 = E[\,(Y_i - 9.2)^2\,]$$

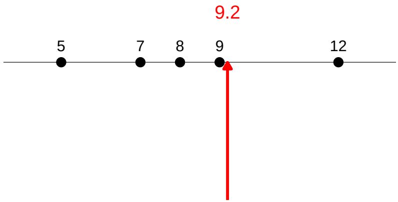
Prediction error: second guess



N_i	Y_i	θ	$Y_i - \theta$	result
1	5	9.2	5-9.2	-4.2
2	7	9.2	7-9.2	-2.2
3	8	9.2	8-9.2	-1.2
4	9	9.2	9-9.2	-0.2
5	12	9.2	12-9.2	2.8

$$MSE_2 = \frac{1}{5}(-4.2 + -2.2 + -1.2 + -0.2 + 2.8)^2$$

Prediction error: second guess



N_i	Y_i	θ	$Y_i - \theta$	result
1	5	9.2	5-9.2	-4.2
2	7	9.2	7-9.2	-2.2
3	8	9.2	8-9.2	-1.2
4	9	9.2	9-9.2	-0.2
5	12	9.2	12-9.2	2.8

$$MSE_2 = \frac{1}{5}(-4.2 - 2.2 - 1.2 - 0.2 + 2.8)^2$$
$$= \frac{(-5)^2}{5} = \frac{25}{5} = \mathbf{5}$$

Best predictor and prediction error

- Two best guesses are provided: $\theta_1 = 5.5$ and $\theta_2 = 9.2$.
- From these best guesses, two measures of prediction error are retrieved: $MSE_1 = 36.45$ and $MSE_2 = 5$.
- The best predictor minimizes prediction error given the data.
  - Which was the **best predictor**, θ_1 or θ_2?
  - It's evident that $MSE_1 > MSE_2$.
  - Therefore, 9.2 better predicts this DGP than 5.5.

Linear model: intercept only

A special case of the regression model is when there are no regressors

$$Y = \mu + e$$

In the **intercept only model**, we find out that the best predictor is μ!

Hence, the best predictor of an unconditional distribution is its **mean**. We can show this by computing the MSE:

$$\text{MSE}: \; E[\,(Y-\theta)^2\,] = \; E[\,(Y-\mu)^2\,]$$

Best predictor and conditional means

- Let's work a short coding example.
- Open the file `pred_sim.Rmd`, and complete all the exercises.