Statistics and Data science

Theory: Causality and RCT

Q. Gallea[1]

[1]Enterprise 4 Society

Overview

1. Why causality ?
2. The fundamental problem of causal inference
3. Problems
4. Solutions
5. The gold-standard (RCT, A/B testing)

Why causality ?

- Every decision we take relies on causal relationships :
  - **Individuals :**
  - If I go vegan, I'll reduce my ecological footprint.
  - If I drink this tequila shot, I'll dance better.

  - **Companies :**
  - Home-office reduces productivity.
  - Spamming users with YouTube Premium Ads will increase the number of subscribers.

  - **Policy makers :**
  - Replacing nuclear power plants with renewables will help to reach the Paris Agreement.
  - Lockdowns will reduce the spread of the covid-19.

- Failing to properly assess causality might lead to costly mistakes.

The fundamental problem of causal inference

- https://www.youtube.com/watch?v=0zvrGiPkVcs&t=58s

The fundamental problem of causal inference

- https://www.youtube.com/watch?v=0zvrGiPkVcs&t=58s
- "How do we know what would have happened without the aid ? We have no idea. We don't know what the **counterfactual** is. There is only one Africa."
- It's impossible to observe the outcome with and without treatment for the same entity at the same point in time.

The fundamental problem of causal inference

Rubin Causal Model

- Treatment : $D_i = 1$ if treated, 0 otherwise
- Potential outcome :
  - Y_{i0} : Individual i outcome without treatment
  - Y_{i1} : Individual i outcome with treatment

Expected

$\Rightarrow$ $Y_i = Y_{i0}$ if $D_i = 0$ or $Y_i = Y_{i1}$ if $D_i = 1$

$\Rightarrow$ $Y_i = Y_{i0} + (Y_{i1} - Y_{i0})D_i$

The fundamental problem of causal inference

Rubin Causal Model

- Treatment : $D_i = 1$ if treated, 0 otherwise
- Potential outcome :
  - Y_{i0} : Individual i outcome without treatment
  - Y_{i1} : Individual i outcome with treatment

  $\Rightarrow$ $Y_i = Y_{i0}$ if $D_i = 0$ or $Y_i = Y_{i1}$ if $D_i = 1$

  $\Rightarrow$ $Y_i = Y_{i0} + (Y_{i1} - Y_{i0})D_i$
- Causal effect : $Y_{i1} - Y_{i0}$

The fundamental problem of causal inference

Rubin Causal Model

- Treatment : $D_i = 1$ if treated, 0 otherwise
- Potential outcome :
  - Y_{i0} : Individual i outcome without treatment
  - Y_{i1} : Individual i outcome with treatment

  $\Rightarrow$ $Y_i = Y_{i0}$ if $D_i = 0$ or $Y_i = Y_{i1}$ if $D_i = 1$

  $\Rightarrow$ $Y_i = Y_{i0} + (Y_{i1} - Y_{i0})D_i$
- Causal effect : $Y_{i1} - Y_{i0}$ impossible to observe

The fundamental problem of causal inference

Rubin Causal Model

- Instead of $Y_{i1} - Y_{i0}$ we can measure $E(Y_i|D_i = 1) - E(Y_i|D_i = 0)$

Average/Expected outcome of the people treated minus the average/ expected outcome of people not treated

The fundamental problem of causal inference

Rubin Causal Model

- Instead of $Y_{i1} - Y_{i0}$ we can measure $E(Y_i|D_i = 1) - E(Y_i|D_i = 0)$

$$E(Y_i|D_i = 1) - E(Y_i|D_i = 0) =$$

This is a way to add 0 in our equation

$$E(Y_{i1}|D_i = 1) - E(Y_{i0}|D_i = 1) + [E(Y_{i0}|D_i = 1) - E(Y_{i0}|D_i = 0)]$$

- This "naive" comparison includes :
  - Average Treatment of the Treated (ATT)
  - "Selection" Bias

The fundamental problem of causal inference

Rubin Causal Model

- There is a bias if :
- $E(Y_{i0}|D_i = 1) \neq E(Y_{i0}|D_i = 0)$
- The average outcome for the treated and untreated would be different without treatment

Endogeneity issues

- **Reverse causality :**
  - The outcome is simultaneously influencing the explanatory variable (e.g. weapons and wars)
- **Omitted variable bias :**
  - Another variable correlated with the outcome and the explanatory variable is not observed/captured (e.g. swimming and sunscreen)
- **Measurement errors :**
  - Systematic miss-measurement related to some variable of interest (e.g COVID-19)

Solutions

1. [today] The gold-standard : RCT, A/B testing
2. [week 2] Regression discontinuity design (RDD)
3. [week 3] Difference-in-Difference (DiD)
4. [week 4] Synthetic controls

The gold-standard : RCT, A/B testing

- A Randomized Control Trial is a controlled experiment (in opposition to a natural experiment) where you randomly allocate the treatment between groups.

⇒ By **randomly allocating the treatment** to the different groups, you can **solve the selection bias**.

RCT

- **How does RCT solve the fundamental causal inference problem ?**

- Recall that :

$$E(Y_i|D_i=1)-E(Y_i|D_i=0)=$$
$$E(Y_{i1}|D_i=1)-E(Y_{i0}|D_i=1)+[E(Y_{i0}|D_i=1)-E(Y_{i0}|D_i=0)]$$

- Randomization of the treatment allocation implies : $E(Y_{i0}|D_i=1)=E(Y_{i0}|D_i=0)$

$\Rightarrow$ $E(Y_i|D_i=1)-E(Y_i|D_i=0)=E(Y_{i1}|D_i=1)-E(Y_{i0}|D_i=1)$

- **The difference between the two groups is causal.**

RCT

- $E(Y_{i0}|D_i = 1) = E(Y_{i0}|D_i = 0)$
- is called an **Identifying assumption**

- Most of the time in causal inference everything relies on the **identifying assumption(s)** and often you can't formally test for this !
- ⚠This is the key ! This is where you have to look to challenge causality (and where the fun begins). ⚠

RCT

Limitations

1. **Not always possible** (e.g. gender) or ethical (e.g. weapons) to manipulate the treatment.

RCT

Limitations

1. **Not always possible** (e.g. gender) or ethical (e.g. weapons) to manipulate the treatment.

2. **External validity vs. internal validity :**
   - Example with seating position and learning
   - https://www.sciencedirect.com/science/article/pii/S0959475217305716
   - https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0236131

RCT

Limitations

1. **Not always possible** (e.g. gender) or ethical (e.g. weapons) to manipulate the treatment.

2. **External validity vs. internal validity :**
   - Example with seating position and learning
   - https://www.sciencedirect.com/science/article/pii/S0959475217305716
   - https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0236131

3. **Simple random allocation works with large samples or if the groups are relatively homogenous :**
   - Large sample : by the law of large numbers, on average the groups will be similar.
   - Homogeneity : In a lab experiment, you have inbred strains of rats (almost identical genetically).

   $\Rightarrow$ Cluster/Stratified sampling method.

RCT

Limitations

4. **Blinding is not always possible !** If possible, the researchers and the subjects do not know who receives the treatment. At least, we should always do our best to prevent the subjects from knowing if they are in the treatment or control group.

RCT

Limitations

4. **Blinding is not always possible !** If possible, the researchers and the subjects do not know who receives the treatment. At least, we should always do our best to prevent the subjects from knowing if they are in the treatment or control group.

5. **Other potential bias :**
   - check : `https://catalogofbias.org/`
     (e.g volunteer bias, attrition bias, survivorship bias, etc.)
   - There is a gap between theory and reality (real example) : Three rules to deal with imperfect experimental designs (`https://medium.com/p/830c92f60590`)

RCT

Limitations

- **Example :** Assessing the effect of mindfullness training on mental health in a firm with a field experiment.
- Which firms ? How many persons ? How would you set up the groups etc ?

What is a sample size calculation ?

- In the industry or research, time/money is often an important constraint :

1. Collecting more data often implies more investment (time/money)
2. But, more data implies more information, and higher confidence in the results
   - $se = sd/\sqrt{n}$

$\Rightarrow$ Sample size calculations allow us to compute the smallest required sample size, such that we can test a specific statistical hypothesis with enough confidence (given some level of Type I and II errors).

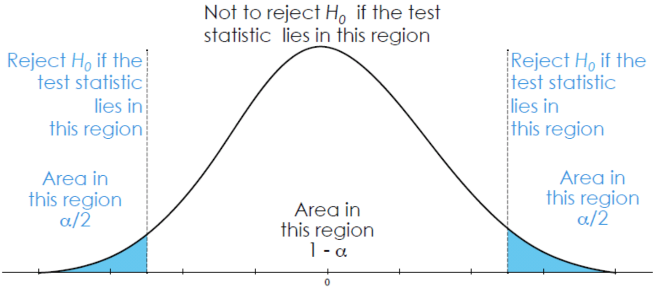
Two types of statistical errors

		True state of nature	
		Effect present	Effect absent
Conclusion of statistical analysis	Effect present (reject H_0)	Correct	Type I error (α)
	No effect (accept H_0)	Type II error (β)	Correct

Type I error : α

- In statistics you "never" get an answer with a 100% certitude
- α is the probability of wrongly rejecting the null hypothesis
- Defined by the researcher (most frequently 5%)
- α will define the rejection rule of H_0
  - if p-value$\leq \alpha \Rightarrow$ reject H_0
  - if p-value$> \alpha \Rightarrow$ not reject H_0

Example


Not to reject H_0 if the test statistic lies in this region
Reject H_0 if the test statistic lies in this region
Reject H_0 if the test statistic lies in this region
Area in this region $\alpha/2$
Area in this region $1 - \alpha$
Area in this region $\alpha/2$
0

Weakness of the p-value

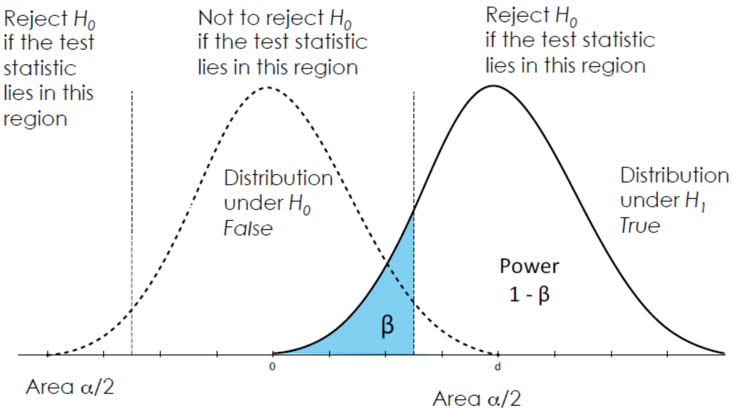
- **Statement on p-values (ASA Statement (2016))**
  - Indicates how incompatible the data are with a specified statistical model
  - Scientific conclusions should not be based only on the p-value
  - Proper inference requires full reporting and transparency
  - p-value states nothing about the magnitude

- My view on this :
  - The statistical significance should be used as a necessary condition to interpret the magnitude, nothing more.
  - (The worst idea ever is to get rid of the stars in tables. Rather show the p-values without stars.)
  - Read : *Why and how to use forest plots efficiently ?* `https://medium.com/towards-data-science/unhappy-with-statistical-significance-p-value-here-is-a-simple`

Type II error : β

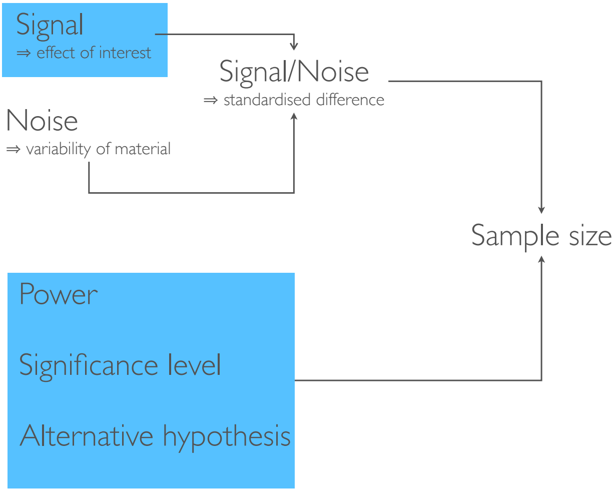
- β is the probability of not rejecting H_0 when H_1 is true
- **Power** $= 1 - \beta$, probability of correctly rejecting H_0

Example



Reject H_0 if the test statistic lies in this region
Not to reject H_0 if the test statistic lies in this region
Reject H_0 if the test statistic lies in this region
Distribution under H_0 False
Distribution under H_1 True
Power
1 - β
β
0
d
Area α/2
Area α/2

Sample size is influenced by five variables



Power and sample size

- Relationship between six variables
  1. The effect size of practical interest
  2. The standard deviation
  3. The significance level
  4. The desired power of the experiment
  5. The sample size
  6. The alternative hypothesis (i.e. one or two-sided test)

1. Effect size of practical interest (δ)

- Effect size $\nearrow \Rightarrow$ Power $\nearrow$
- The decision on how large an effect would need to be considered of scientific interest / practical relevance
  - Would a 10% change in energy consumption upon treatment be of practical relevance and should the experiment be designed to detect it ?
  - If 50% of the control group is expected to show some effect, what proportion in the treated group would be of interest to detect ?

2. Standard deviation (σ)

- Standard deviation $\nearrow \Rightarrow$ Power $\searrow$
- Needs to be estimated
    - Similar experiments, literature review, pilot study
- If not available
    - Standard Cohen's D (corresponds to the effect size divided by the standard error) :
        - Small effect $= 0.2$
        - Medium Effect $= 0.5$
        - Large Effect $= 0.8$

3. The significance level (α)

- $\alpha \nearrow \Rightarrow$ Power $\nearrow$
- However, other things being equal, specifying a low chance of a false-positive result will increase the chance of a false-negative result
- By convention, fixed to 5%

4. Power of the experiment γ

- Power $\nearrow \Rightarrow$ sample size $\nearrow$
- $\gamma = 1 - \beta$ ($\beta =$ Allowed probability of wrongly not rejecting the null hypothesis H_0) The aim should be to have powerful experiments that have a high chance of detecting an effect if it exists (i.e. low β error)
- **Usually set between 80% and 90%**

⚠ To be considered in the interpretation of "negative results" : One can only conclude that differences in effects between two treatments were certainly absent if the study would have had enough POWER to detect them. FALSE ! post-hoc test doesn't work !

5. Sample size (n)

- sample size $\nearrow \Rightarrow$ power$\nearrow$
- In many cases it is the sample size that is to be determined, and all the other variables are specified
- If the sample size is fixed $\Rightarrow$ the power can be estimated $\Rightarrow$ experiment may not be worth doing

6. The alternative hypothesis (H_1)

- The usual null hypothesis $\Rightarrow$ no differences among treatment means $\Rightarrow H_1 =$ there is an effect $\Rightarrow$ two-sided significance test
- If the alternative is that means differ in a particular direction $\Rightarrow$ one-sided test

Sample size - Principle

- Specify the smallest true difference between the treatments that would be of practical relevance
- **Choose the smallest sample size allowing to test your hypothesis efficiently (power)**

What do we need to compute the sample size

- Type of the outcome variable (continuous, categorical, proportion etc.) and number of comparisons
  - ⇒ Choice of statistical test
- Direction of the test : one or two-sided
- Standard deviation (or variance)
- Effect size

⇒ Use Python or G-power (let's see both)