THE UNIVERSITY of EDINBURGH

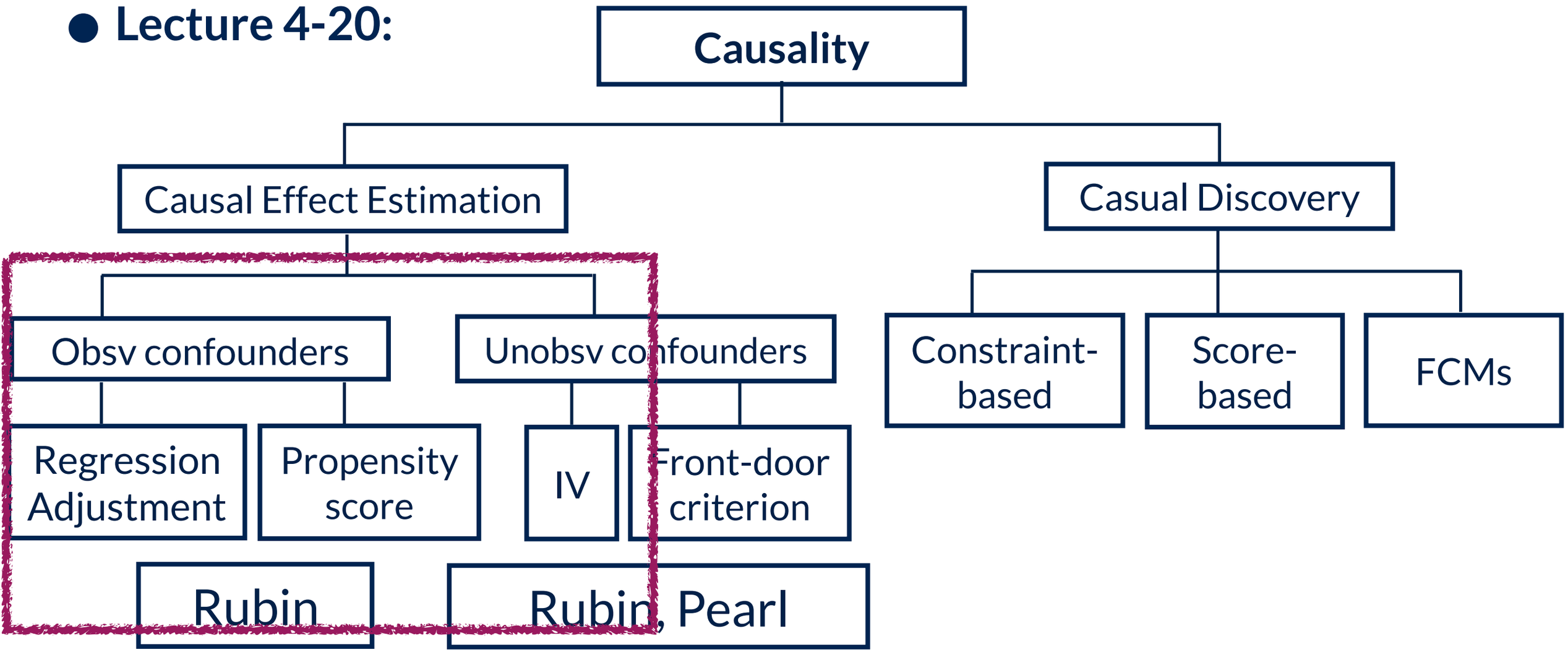
Methods for Causal Inference
Lecture 7: Sensitivity Analysis

Ava Khamseh

School of Informatics
2025-2026

So far …

- **Lecture 1**: Introduction & Motivation, why do we care about causality? Why deriving causality from observational data is non-trivial.
- **Lecture 2:** Recap of probability theory, variables, events, conditional probabilities, independence, law of total probability, Bayes' rule
- **Lecture 3**: Recap of regression, multiple regression, graphs, SCM
- **Lecture 4-20:**



Back to observed confounders

Matching: Stratification, balancing (propensity) score, IPTW, …

$$x \perp\!\!\!\perp t | b(x)$$

Estimation of propensity scores directly from the data & algorithms

$$e(x) = p(t = 1|x)$$

Sensitivity analysis: No guarantee that matching leads to balance on variables we did not match for, people who look comparable may differ.

If there is hidden bias, how severe is it:

- Does the conclusion change from statistically significant to not?
- Does it change the direction of effect?

Notice: This is separate from uncertainty due to (causal) statistical estimates, rather due to biased introduced by unobserved variables.

Sensitivity Analysis

Randomised trials are unconfounded **by design**

Observational data may have possible hidden bias/unobserved confounder that is not controlled for

No guarantee that matching leads to balance on variables we did **not** match for!

People who look comparable may differ

This violates unconfoundedness assumption

Unconfoundedness is fundamentally (directly) unverifiable

Rosenbaum Design of Observational Studies, Springer, 2010

Types of sensitivity Analysis (non-exhaustive)

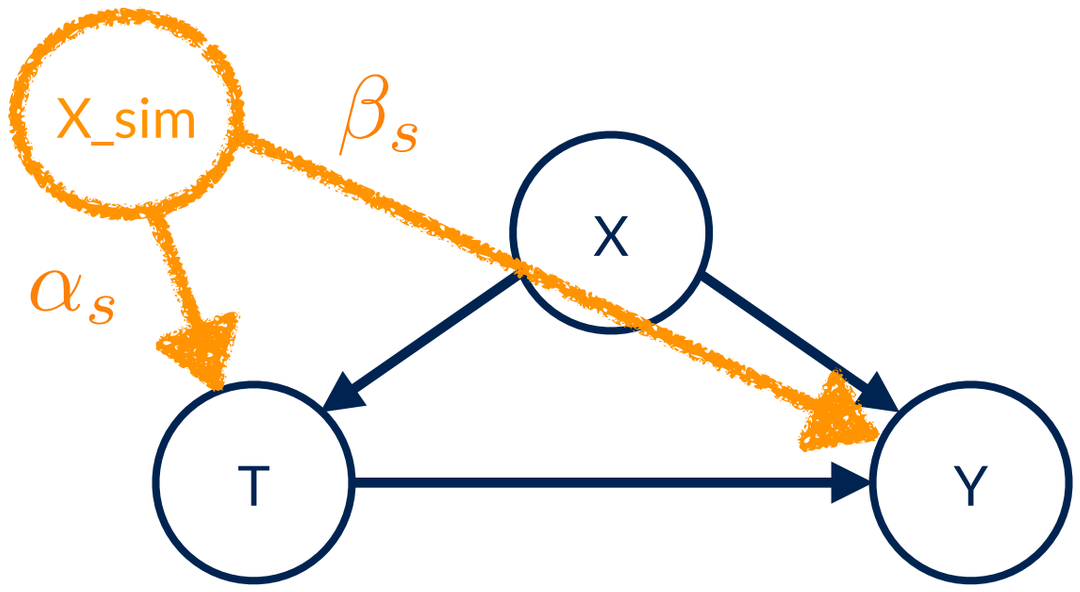
- Quick and simple sanity checks
- Super Learning other potential (‘less likely’) confounders
- Deriving bounds on the causal statistical estimates

Sensitivity Analysis: Quick sanity checks

1) **Random ‘unobserved’ common cause**: Add an independently and randomly drawn confounder affecting treatment and outcome,
re-run the analysis

Example: Specify how the simulated confounder affects treatment and outcome. This could be done via a linear model with two equal/different coefficients for a continuous treatment or a binary flip (probability that simulated confounder’s effect flips the value of treatment/outcome from 0 to 1).

If our original causal estimate was significant, this operation should not change the results ‘much’.

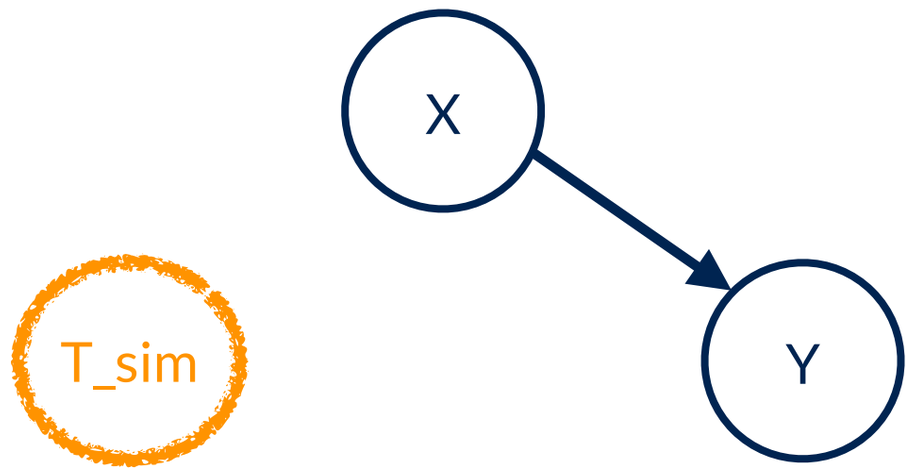


Sensitivity Analysis: Quick sanity checks

2) **Placebo treatment effect**: Replace treatment with randomly generated placebo e.g. we the same marginal distribution of # treatment and # control

The new estimate should be statistically zero.

Generate T_sim randomly, or,
Permute values of T

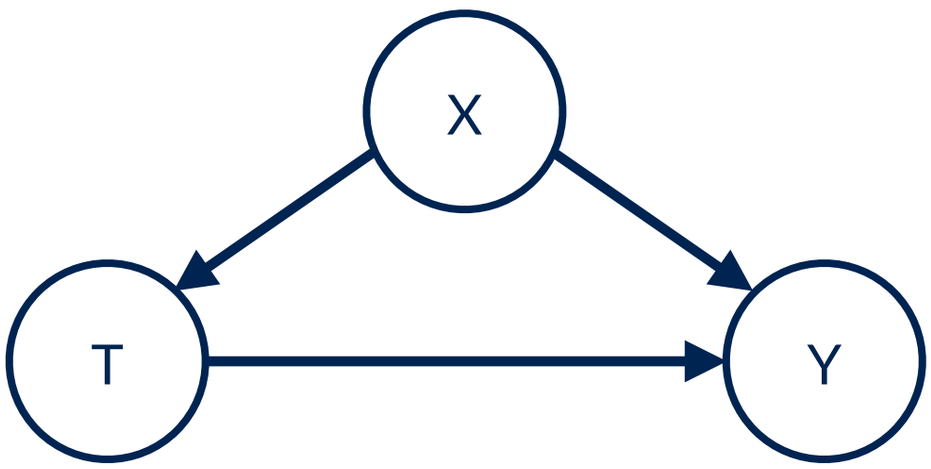


Sensitivity Analysis: Quick sanity checks

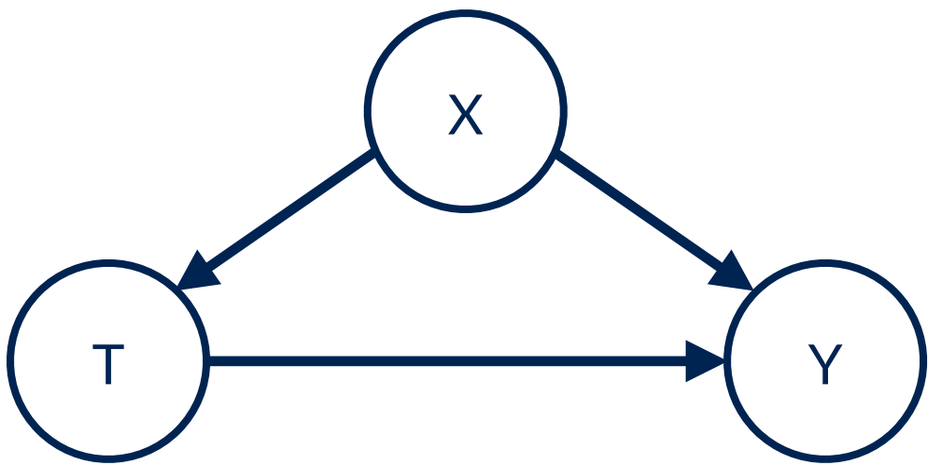
3) **Subset/validate the data:** Subsetting the data is similar to cross-validation, checking if the causal estimate remain statistically the same
(Can also use bootstrap samples of original data).

If possible validate on a different data set
(where the distribution of T, X, Y is expected to be the same)

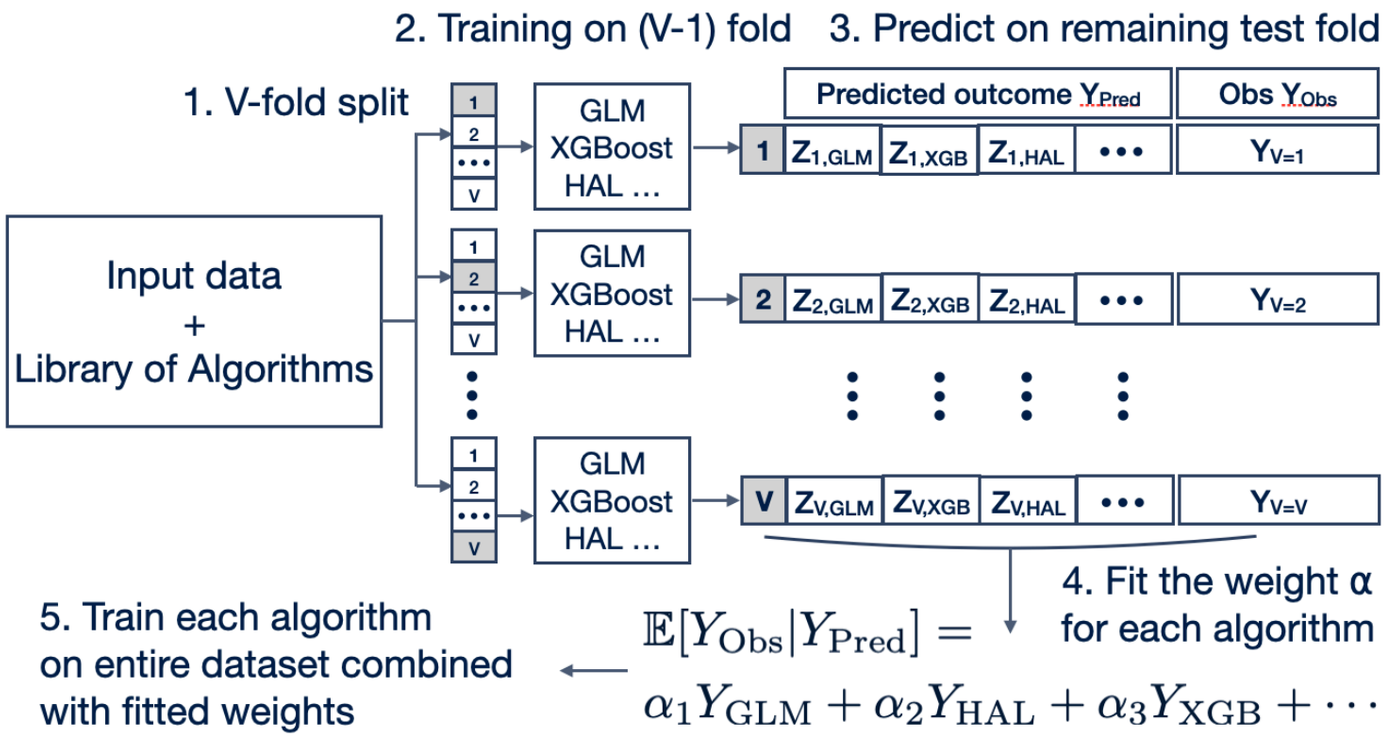
Random subset 1



Random subset 2

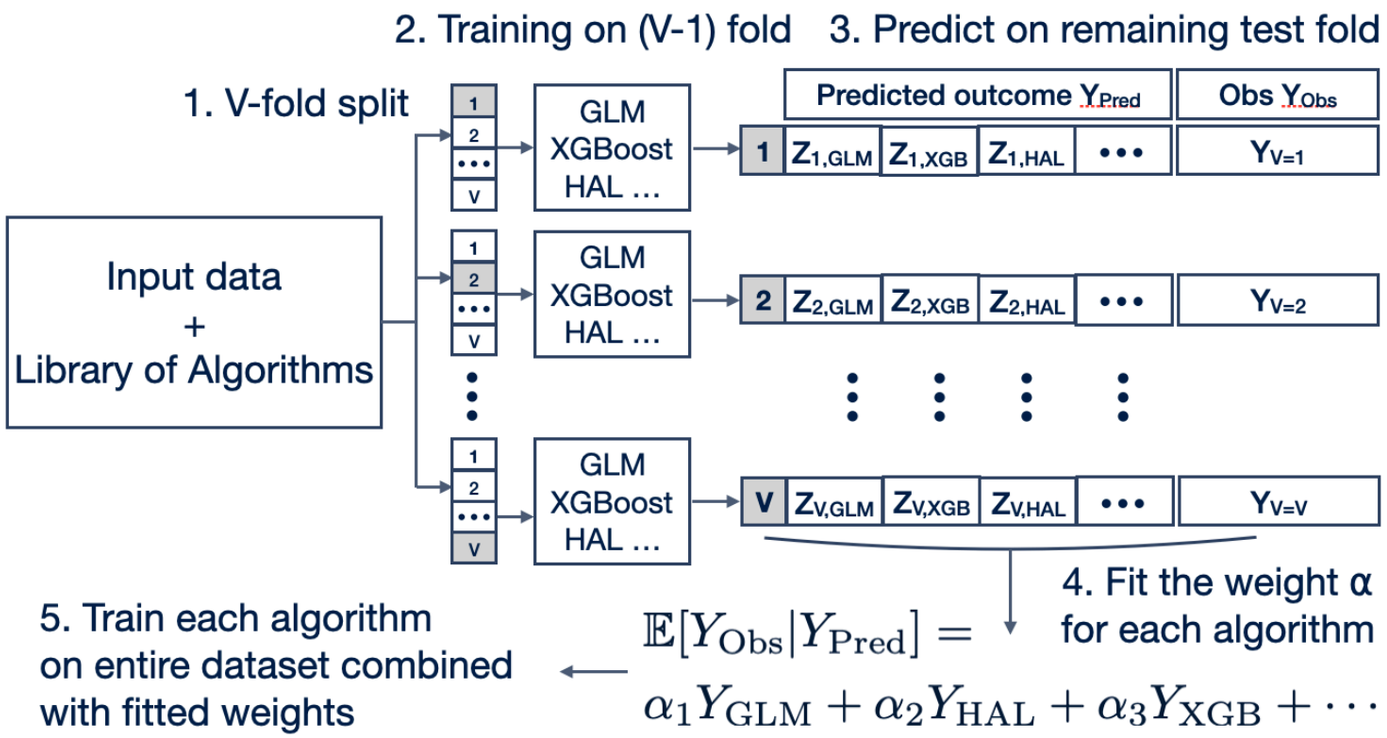


Sensitivity: Super Learning



If the subject expert suspect that a variable can be confounder, we should include it in the Super Learner, and allow the model to be chosen via V-fold cross-validation.

Sensitivity: Super Learning

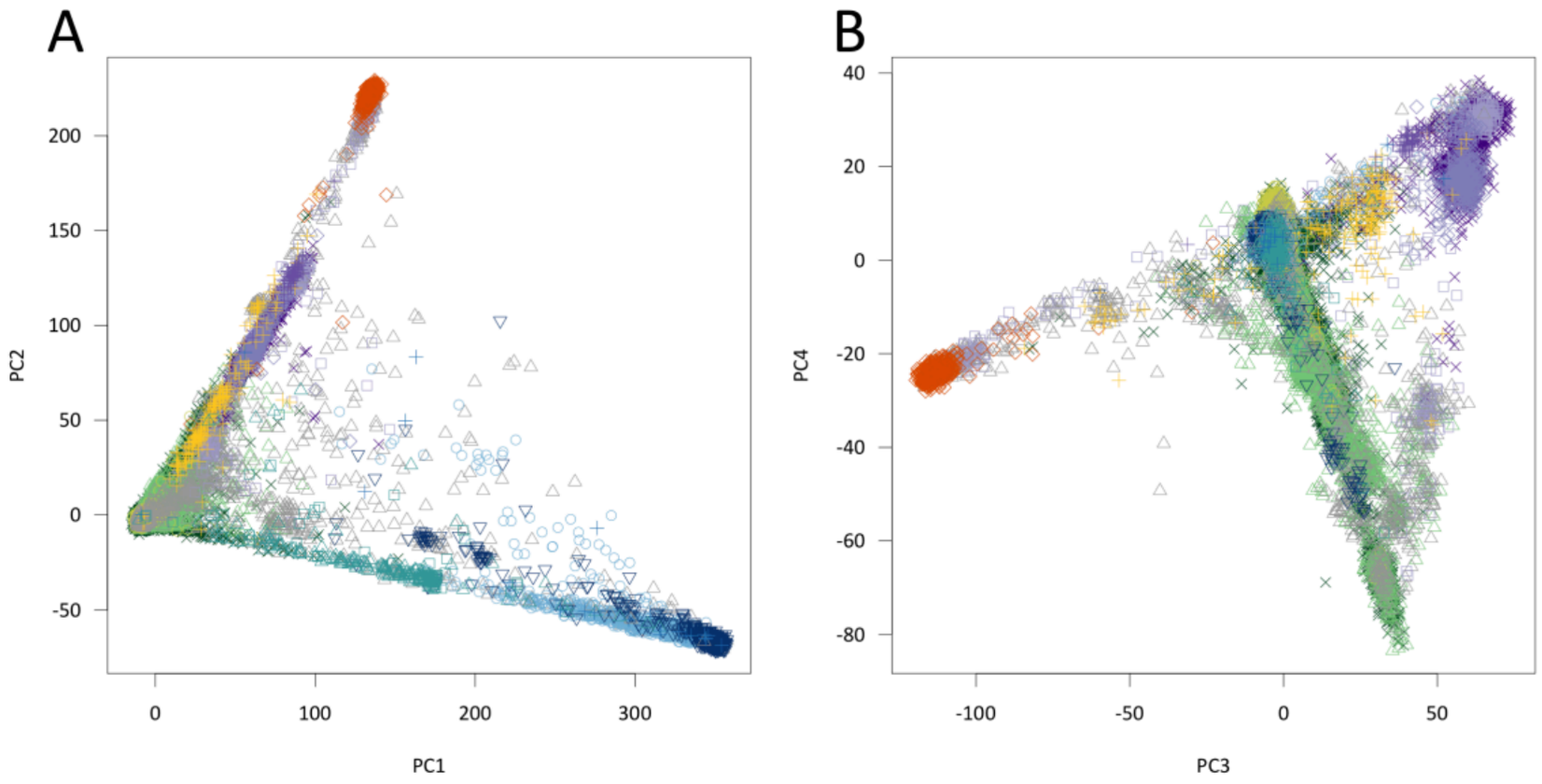


If the subject expert suspect that a variable can be confounder, we should include it in the Super Learner, and allow the model to be chosen via V-fold cross-validation.

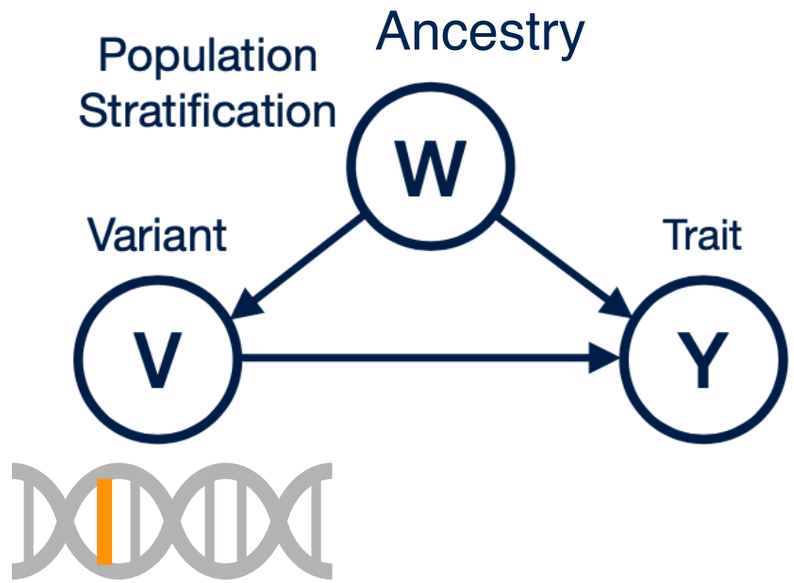
But some times there are too many potential candidate confounders … Perhaps we wish to use feature selection, then perform sensitivity tests including selected/non-selected features.

Sensitivity: Super Learning

Example: PCA plots capture variation in a population

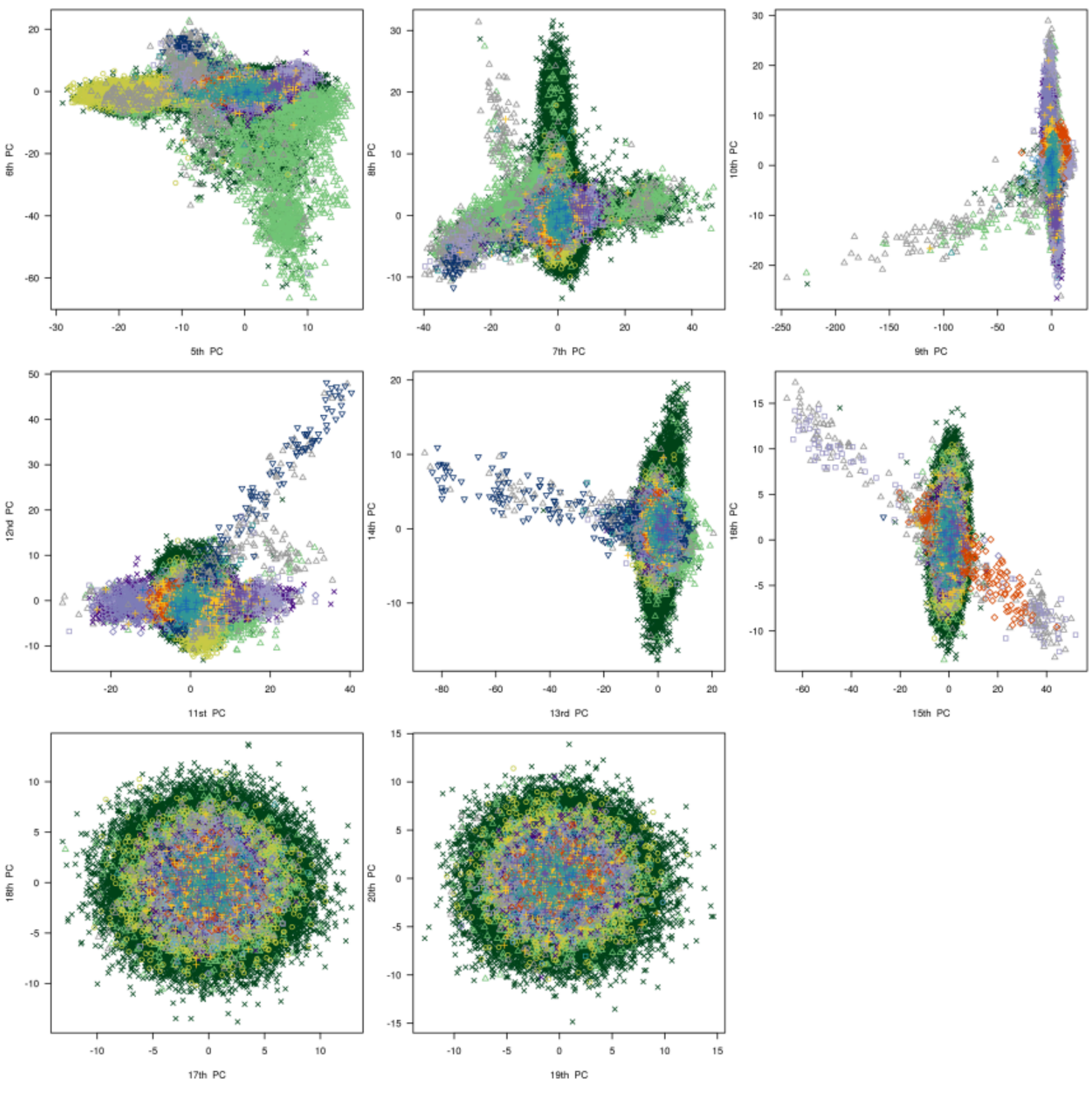


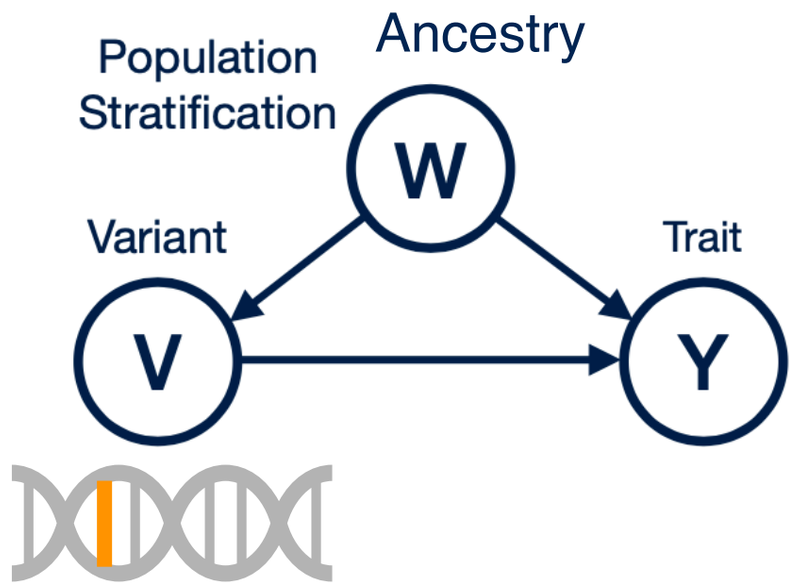
- British
- Irish
- Any other white background
- White and Asian
- Chinese
- African
- White and Black African
- Caribbean
- White and Black Caribbean
- Any other Black background
- Indian
- Pakistani
- Bangladeshi
- Any other Asian background
- Any other mixed background
- Other/Unknown



Sensitivity: Super Learning

Example: PCA plots capture variation in a population





Add higher-order PCs as confounders in the SL and test if the estimates change (they should stabilise at some order).

biobank uk

Sensitivity Analysis: Bounds

- “This difference in the unobserved covariate u, the critic continues, is the real reason outcomes differ in the treated and control groups: it is not an effect caused by the treatment, but rather a failure on the part of the investigators to measure and control imbalances in u. Although not strictly necessary, the critic is usually aided by an air of superiority: “This would never happen in my laboratory.””

Rosenbaum Design of Observational Studies, Springer, 2010

Sensitivity Analysis: Bounds

- “This difference in the unobserved covariate u, the critic continues, is the real reason outcomes differ in the treated and control groups: it is not an effect caused by the treatment, but rather a failure on the part of the investigators to measure and control imbalances in u. Although not strictly necessary, the critic is usually aided by an air of superiority: “This would never happen in my laboratory.””

- “It is important to recognize at the outset that our critic may be, but need not be, on the side of the angels. The tobacco industry and its (sometimes distinguished) consultants criticized, in precisely this way, observational studies linking smoking with lung cancer.”

Rosenbaum Design of Observational Studies, Springer, 2010

Sensitivity Analysis: Bounds

- Take individuals (i) and (j), such that their observed covariates are the same: $X^{(i)} = X^{(j)}$ hence $e^{(i)} = e^{(j)}$ no hidden bias
- Consider e.g., the odds ratio:

$$\frac{1}{\Gamma} \leq \frac{\frac{e^{(i)}_{\text{true}}}{1-e^{(i)}_{\text{true}}}}{\frac{e^{(j)}_{\text{true}}}{1-e^{(j)}_{\text{true}}}} \leq \Gamma \qquad \longrightarrow \qquad \Gamma \approx 1$$

- Otherwise if there is a hidden bias, e.g., $\Gamma = 2$, one subject is twice as likely to receive treatment than not, because of unobserved pre-treatment feature

- Γ quantifies degree of bias.

Hypothesis testing detour

Suppose we have estimated the causal effect of treatment T on outcome Y and we wish to quantify if this difference is significantly away from zero

Pictorially:

Hypothesis testing detour

Suppose we have estimated the causal effect of treatment T on outcome Y and we wish to quantify if this difference is significantly away from zero

Pictorially:

Significant
($> 3\sigma$ away from zero)

Not significant
(zero within 1σ)

Hypothesis testing detour

Suppose we have estimated the causal effect of treatment T on outcome Y and we wish to quantify if this difference is significantly away from zero

Suppose we have the **null hypothesis** H_0 that the causal effect of treatment on outcome is zero.

Hypothesis testing detour

Suppose we have estimated the causal effect of treatment T on outcome Y and we wish to quantify if this difference is significantly away from zero

Suppose we have the **null hypothesis** H_0 that the causal effect of treatment on outcome is zero.

The **alternative** hypothesis H_1 is that the causal effect of treatment on outcome is ‘significantly’ non-zero.

This significance is quantified by a **p-value**, obtained via statistical testing.

Hypothesis testing detour

Suppose we have estimated the causal effect of treatment T on outcome Y and we wish to quantify if this difference is significantly away from zero

Suppose we have the **null hypothesis** H_0 that the causal effect of treatment on outcome is zero.

The **alternative** hypothesis H_1 is that the causal effect of treatment on outcome is 'significantly' non-zero.

This significance is quantified by a **p-value**, obtained via statistical testing.

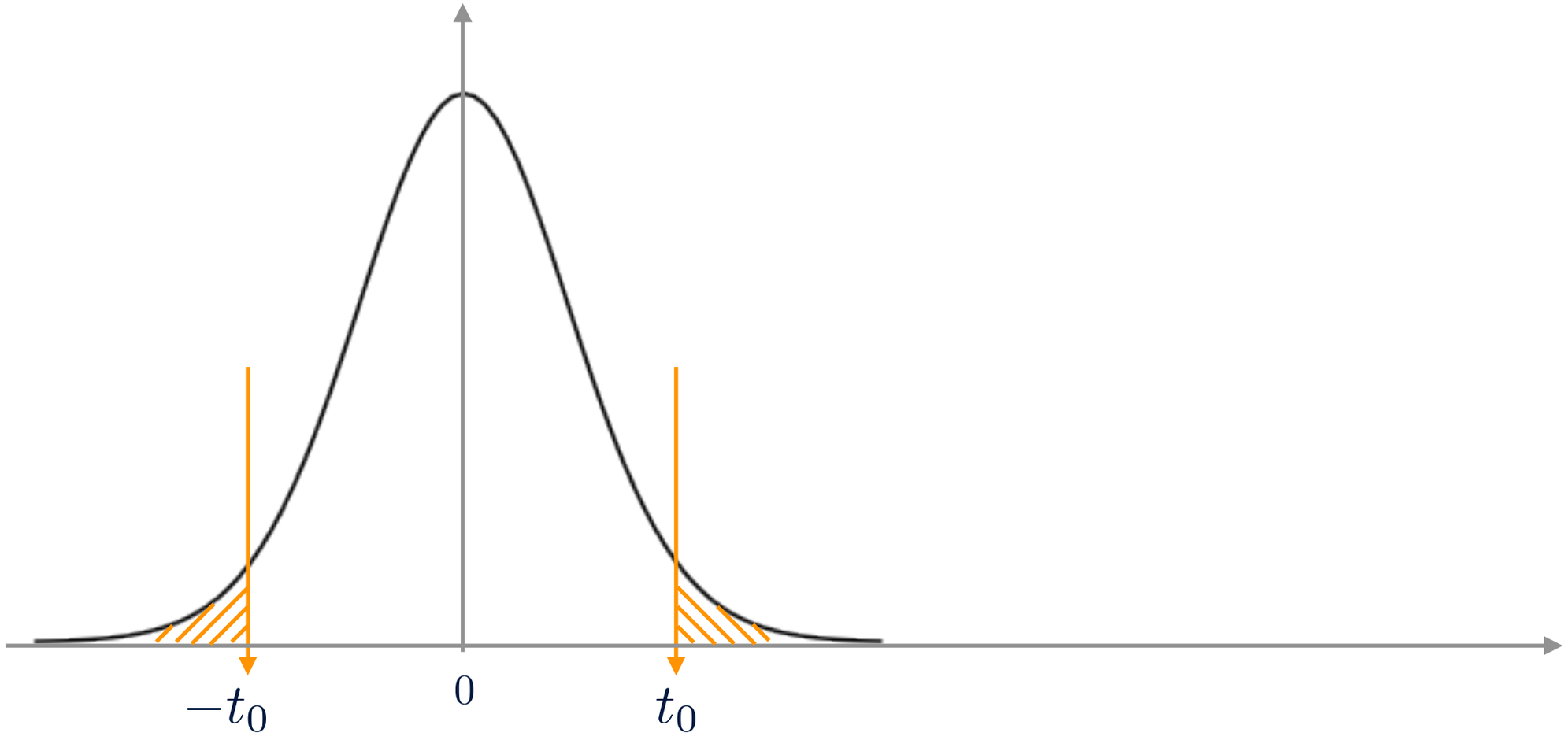
A commonly used statistic in this context is a t-test (or z-test).

$$\frac{\text{signal}}{\text{noise}} = \frac{\text{ATE}}{\sigma_{\text{ATE}}} \sim t\text{-distributed (or } z\text{-distributed)}$$

Significance: p-value

Probability of obtaining a measurement of statistics that is more extreme than the value t_0, **given the null hypothesis**.

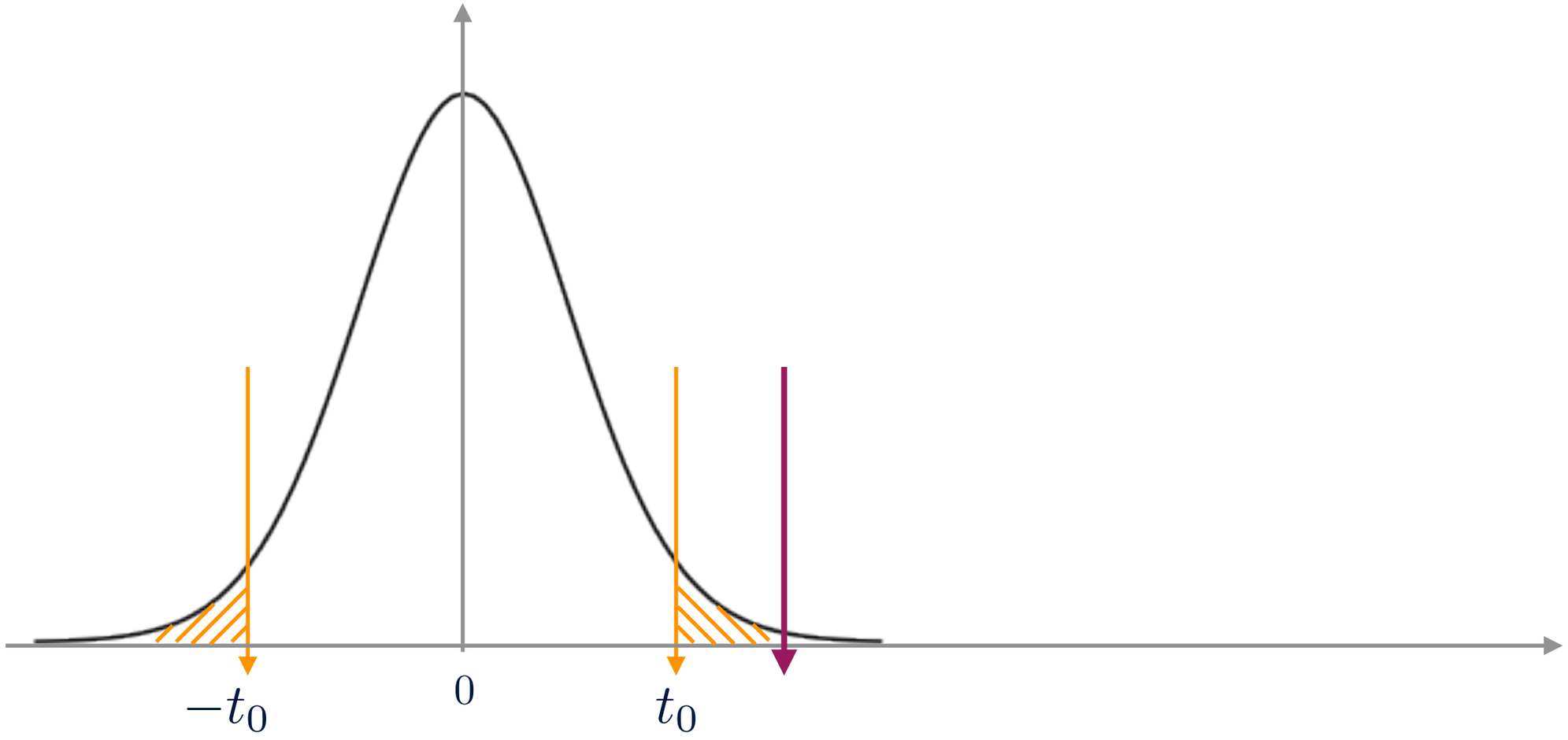
$$p\text{-value} = \Pr\left(\left| \frac{\text{signal}}{\text{noise}} \right| > t_0 \middle| H_0 \right)$$



Significance: p-value

Probability of obtaining a measurement of statistics that is more extreme than the value t_0, **given the null hypothesis**.

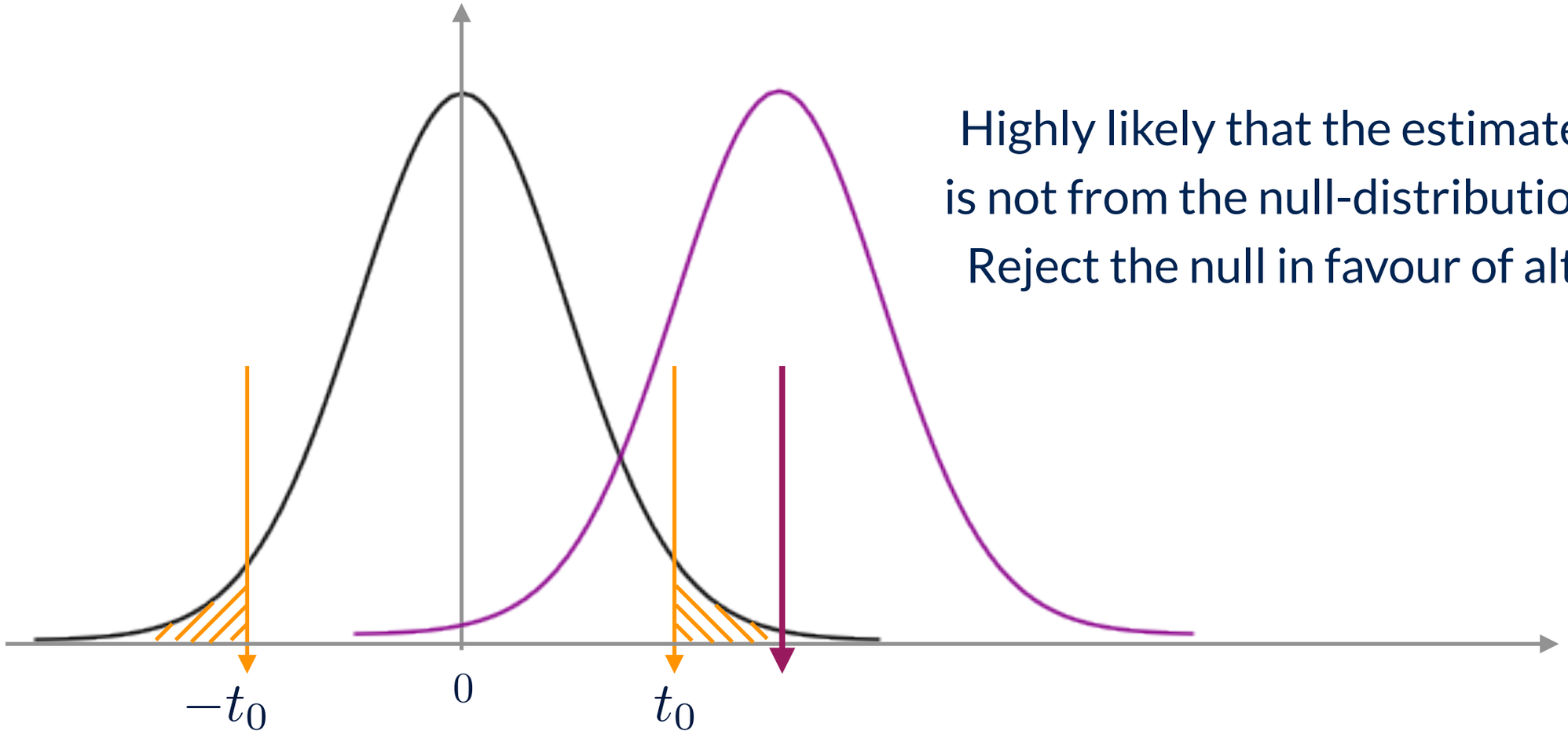
$$p\text{-value} = \Pr\left(\left|\frac{\text{signal}}{\text{noise}}\right| > t_0 \middle| H_0\right)$$



Significance: p-value

Probability of obtaining a measurement of statistics that is more extreme than the value t_0, **given the null hypothesis**.

$$p\text{-value} = \Pr\left(\left|\frac{\text{signal}}{\text{noise}}\right| > t_0 \middle| H_0\right)$$



Highly likely that the estimate is not from the null-distribution
Reject the null in favour of alt

Significance: p-value

- Correct inference:

True negative: H_0 not rejected, and the estimate is indeed from H_0

True Positive: H_0 is rejected correctly, the estimate is indeed from H_1

- Incorrect inference:

False negative: H_0 not rejected, **but** the estimate is from H_1 (type II error)

False positive: H_0 is rejected incorrectly but the estimate is from H_0 not H_1 (Type I error)

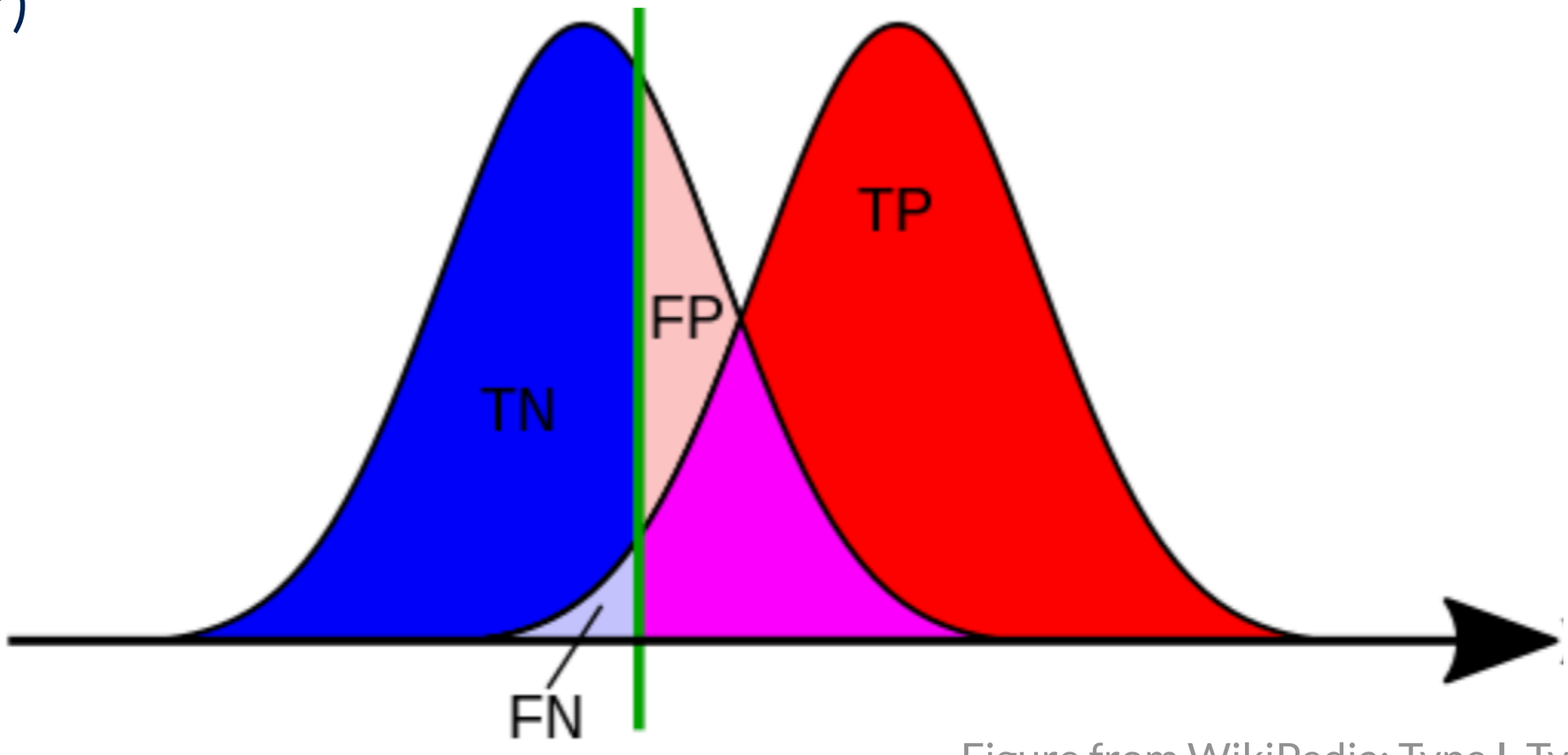


Figure from WikiPedia: Type I, Type II errors

Sensitivity Analysis bound: An example [non-examinable]

- S pairs, s = 1,...,S of two subjects, one treated, one control, **matched** for observed covariates

- Statistical test: **Wilcoxon's signed rank test** (non-parametric), W is the sum of the ranks of the positive differences between treatment and control

Control	Treatment
85	98
82	87
102	92
...	...
100	80
95	110

Rosenbaum Sensitivity Analysis in Observational Studies, 2005

Sensitivity Analysis bound: An example [non-examinable]

- S pairs, s = 1,...,S of two subjects, one treated, one control, **matched** for observed covariates

- Statistical test: **Wilcoxon's signed rank test** (non-parametric), W is the sum of the ranks of the positive differences between treatment and control

Control	Treatment	Difference
85	98	13
82	87	5
102	92	-10
...	...	...
100	80	-20
95	110	15

Rosenbaum Sensitivity Analysis in Observational Studies, 2005

Sensitivity Analysis bound: An example [non-examinable]

- S pairs, s = 1,...,S of two subjects, one treated, one control, **matched** for observed covariates

- Statistical test: **Wilcoxon's signed rank test** (non-parametric), W is the sum of the ranks of the positive differences between treatment and control

Control	Treatment	Difference	Abs Diff
85	98	13	13
82	87	5	5
102	92	-10	10
...	...	...	...
100	80	-20	20
95	110	15	15

Rosenbaum Sensitivity Analysis in Observational Studies, 2005

Sensitivity Analysis bound: An example [non-examinable]

- S pairs, s = 1,...,S of two subjects, one treated, one control, **matched** for observed covariates

- Statistical test: **Wilcoxon's signed rank test** (non-parametric), W is the sum of the ranks of the positive differences between treatment and control

Control	Treatment	Difference	Abs Diff	Rank
85	98	13	13	3
82	87	5	5	1
102	92	-10	10	2
…	…	…	…	Based on number here
100	80	-20	20	5
95	110	15	15	4

Rosenbaum Sensitivity Analysis in Observational Studies, 2005

Sensitivity Analysis bound: An example [non-examinable]

- S pairs, s = 1,...,S of two subjects, one treated, one control, **matched** for observed covariates

- Statistical test: **Wilcoxon's signed rank test** (non-parametric), W is the sum of the ranks of the positive differences between treatment and control

Control	Treatment	Difference	Abs Diff	Rank
85	98	13	13	3
82	87	5	5	1
102	92	-10	10	2
…	…	…	…	Based on number here
100	80	-20	20	5
95	110	15	15	4

Rank sum of -ves: 7
Rank sum of +ve: 8
W_{stat} = 7 (the smaller of above)

$W_{critical}$: Look-up table
Total number of individuals: N
Threshold: 0.05
If $W_{stat} < W_{critical}$ reject
i.e. drug group significantly different from control

Rosenbaum Sensitivity Analysis in Observational Studies, 2005

Sensitivity Analysis bound: An example [non-examinable]

- S pairs, s = 1,...,S of two subjects, one treated, one control, **matched** for observed covariates

- Statistical test: **Wilcoxon's signed rank test** (non-parametric), W is the sum of the ranks of the positive differences between treatment and control

- In a moderately large randomized (here matched) experiment, under the **null hypothesis of no effect**, W is approximately normally distributed

$$\mathbb{E}[W] = S(S+1)/4 \quad , \quad \mathrm{Var}[W] = S(S+1)(2S+1)/24$$

Rosenbaum Sensitivity Analysis in Observational Studies, 2005

Sensitivity Analysis bound: An example [non-examinable]

- Example: W=300, S=25 pairs in a randomised experiment
- In a randomised experiment ($\Gamma \approx 1$, well-matched):

$$\mathbb{E}[W] = 162.5 \ , \ \mathrm{Var}[W] = 1381.25 \ , \ \text{deviate } Z = (300 - 162.5)/\sqrt{1381.25} = 3.70$$

- Compared to a normal distribution: p-value = 0.0001
- In a moderately large observational study, under the null hypothesis of no effect, the distribution of W is approximately bounded between two Normal distributions (notice: $\Gamma \approx 1$)

$$\mu_{\max} = \lambda S(S+1)/2 \quad , \quad \mu_{\min} = (1-\lambda)S(S+1)/2$$

$$\sigma^2 = \lambda(1-\lambda)S(S+1)(2S+1)/6$$

$$\lambda = \Gamma/(1+\Gamma)$$

Notice $\Gamma = 1$

Rosenbaum Sensitivity Analysis in Observational Studies, 2005

Sensitivity Analysis bound: An example [non-examinable]

- Example: W=300, S=25 pairs in a randomised experiment
- For $\Gamma = 2,\ \ \lambda = \Gamma/(1+\Gamma) = 2/3$

$$\mu_{\max} = \lambda S(S+1)/2 = 216.67 \quad , \quad \mu_{\min} = (1-\lambda)S(S+1)/2 = 108.33$$

$$\sigma^2 = \lambda(1-\lambda)S(S+1)(2S+1)/6 = 1227.78$$

$$Z_1 = 5.47 \ \Rightarrow \ p = 0.00000002$$

$$Z_2 = 2.38 \ \Rightarrow \ p = 0.009$$ still significant, even with $\Gamma = 2$

- For the tobacco and lung cancer example, $\Gamma = .6$

Notice: Here there are **two** sources of uncertainty:
1) Due to the causal identifiability (e.g., unobserved variables)
2) Due to statistical estimation (inference)