Statistical Inference

Matti Sarvimäki

Mini-Course on Causal Inference
Lecture 2

Hypothesis testing and statistical significance

- The question: How likely it is that the difference between treatment and control groups could be due to chance?
    - i.e. test the null hypothesis that the treatment had no effect

Hypothesis testing and statistical significance

- The question: How likely it is that the difference between treatment and control groups could be due to chance?
  - i.e. test the null hypothesis that the treatment had no effect
- Learning objectives. You understand the following concepts:
  1. point estimates
  2. standard errors
  3. p-values
  4. statistical significance
  5. t-statistics
  6. critical values
  7. confidence intervals

  and how to use them to **interpret basic empirical results**.

Example: Gender and policy decisions

Econometrica, Vol. 72, No. 5 (September, 2004), 1409–1443

WOMEN AS POLICY MAKERS: EVIDENCE FROM A RANDOMIZED POLICY EXPERIMENT IN INDIA

BY RAGHABENDRA CHATTOPADHYAY AND ESTHER DUFLO[1]

This paper uses political reservations for women in India to study the impact of women's leadership on policy decisions. Since the mid-1990's, one third of Village Council head positions in India have been randomly reserved for a woman: In these councils only women could be elected to the position of head. Village Councils are responsible for the provision of many local public goods in rural areas. Using a dataset we collected on 265 Village Councils in West Bengal and Rajasthan, we compare the type of public goods provided in reserved and unreserved Village Councils. We show that the reservation of a council seat affects the types of public goods provided. Specifically, leaders invest more in infrastructure that is directly relevant to the needs of their own genders.

KEYWORDS: Gender, decentralization, affirmative action, political economy.

Example: Gender and policy decisions

- Here is an extract from their Table V:

	West Bengal		
	Mean, Reserved GP	Mean, Unreserved GP	Difference
Dependent Variables	(1)	(2)	(3)
A. *Village Level*			
Number of Drinking Water Facilities	23.83	14.74	9.09
Newly Built or Repaired	(5.00)	(1.44)	(4.02)

  - Data: **161 GPs out of which 54 were reserved** for women leaders
    - first row of columns (1) and (2) reports averages
    - first row of column (3) report difference in averages
    - second row, col (3) reports the standard error (SE)
- This lecture: How to correctly interpret point estimates and SEs

Point estimate and statistical significance

- In the example above, we had the following sample averages

$$\bar{y}^1 = Avg[y|D = 1] = 23.8$$
$$\bar{y}^0 = Avg[y|D = 0] = 14.7$$

  where $D = 1$ denotes the GP being reserved for female leader
- $\bar{y}^1 - \bar{y}^0 = 9.1$ is the **point estimate**
  - *the most likely* impact is that, on average, 9.1 more drinking facilities are build per village when a GP is led by a woman
  - research design / identification: GPs were randomly assigned into treatment and control groups and thus selection bias is unlikely

Point estimate and statistical significance

- However, the point estimate may differ from zero because:
  1. female leaders are more likely to invest in drinking water
  2. the 54 treatment GPs just happen to invest more in drinking water (for reasons that have nothing to do with the gender of their leader)

Point estimate and statistical significance

- However, the point estimate may differ from zero because:
  1. female leaders are more likely to invest in drinking water
  2. the 54 treatment GPs just happen to invest more in drinking water (for reasons that have nothing to do with the gender of their leader)
- Question: How likely are we to get a point estimate of at least 9.1 just due to random variation across GPs?
  - the convention is to call an estimate "statistically significant" if the likelihood of a chance finding is below 5%

Simulating a test distribution

- An intuitive way to think about randomly occurring differences between groups is to create a distribution of "placebo" treatments

- Split the GPs into two random groups and calculate their averages
  - you can get the data here
  - ... and my simulation code here

Simulating a test distribution

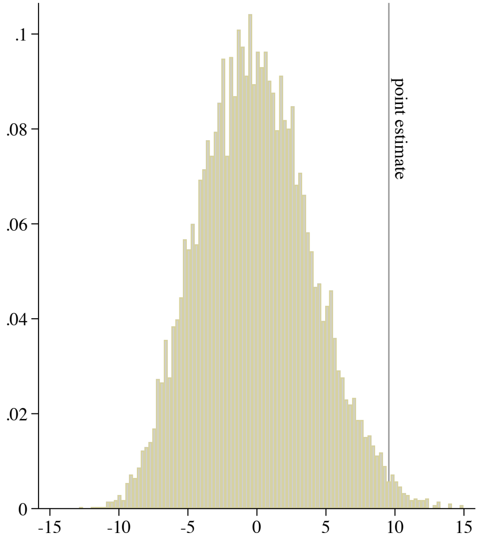
- An intuitive way to think about randomly occurring differences between groups is to create a distribution of "placebo" treatments

- Split the GPs into two random groups and calculate their averages
  - you can get the data here
  - ... and my simulation code here
- Note that $\mathbb{E}[y|D_{pl} = 1] = \mathbb{E}[y|D_{pl} = 0]$
  - the "placebo" assignments D_{pl} are made-up and thus have no impact
  - but: as the table shows, with just 54 GPs in the "treatment" group, the differences can sometimes be large

"Treatment"	"Control"	Diff
15.80	19.66	-3.86
14.63	20.22	-5.59
17.10	19.03	-1.92
17.85	18.67	-0.81
13.22	20.90	-7.68
15.23	19.93	-4.70
16.91	19.12	-2.21
16.21	19.46	-3.24
21.69	16.81	4.88
19.98	17.64	2.34

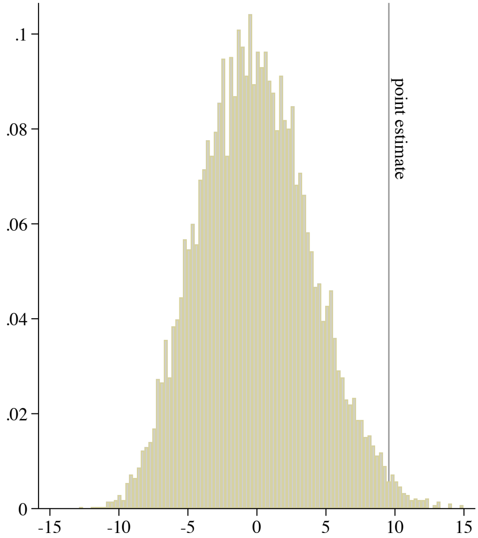
10 "placebo" simulations

Simulating a test distribution



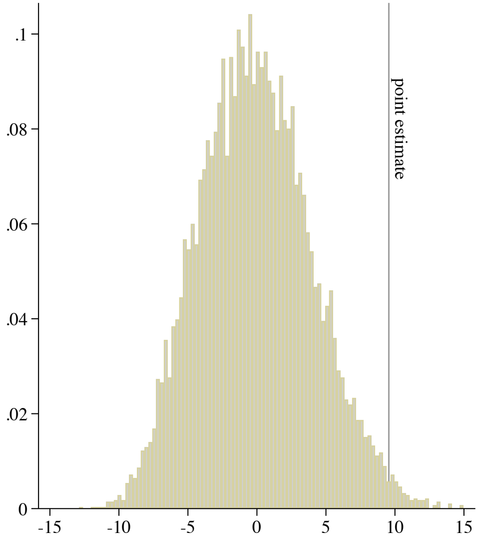
- Simulation with 10,000 rounds
  - average: -0.099
  - standard deviation: 4.03

Simulating a test distribution



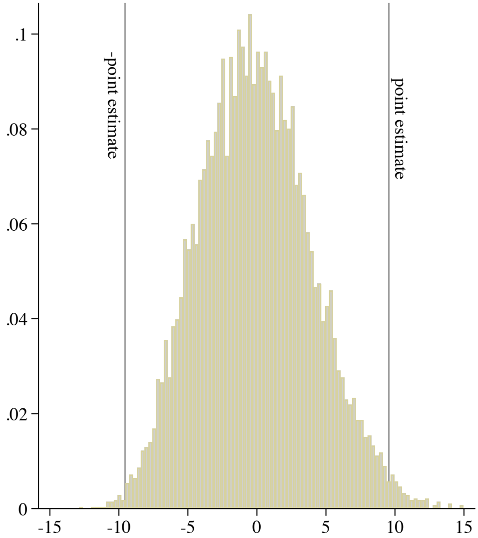
- Simulation with 10,000 rounds
  - average: -0.099
  - standard deviation: 4.03
- As you see from the histogram, sometimes random splits of the sample yield differences that are larger than the point estimate
  - the largest difference is 14.97

Simulating a test distribution

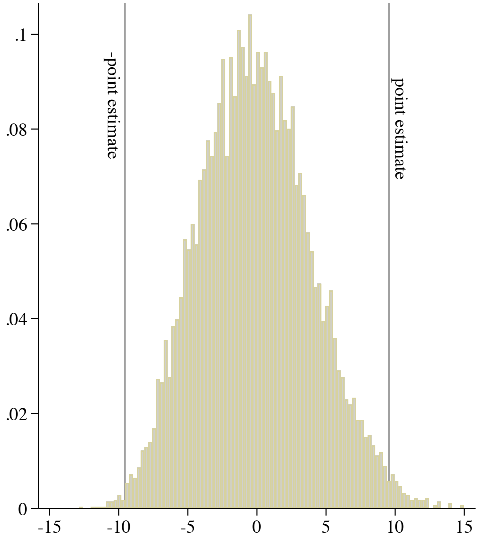


- Simulation with 10,000 rounds
  - average: -0.099
  - standard deviation: 4.03
- As you see from the histogram, sometimes random splits of the sample yield differences that are larger than the point estimate
  - the largest difference is 14.97
- However, this is quite rare:
  - difference > point estimate in 1.1% of the simulation rounds

P-value

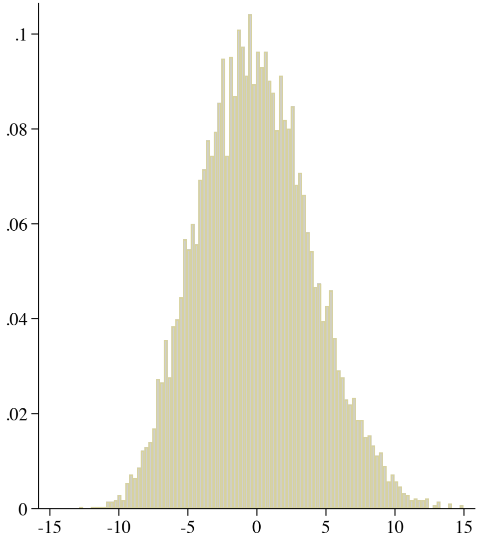


- **p-value**: the probability of obtaining a result at least as extreme as the result actually observed under the null hypothesis
  - here, the null hypothesis is zero treatment effect, i.e. $H_0 : \mathbb{E}[y|D = 1] = \mathbb{E}[y|D = 0]$



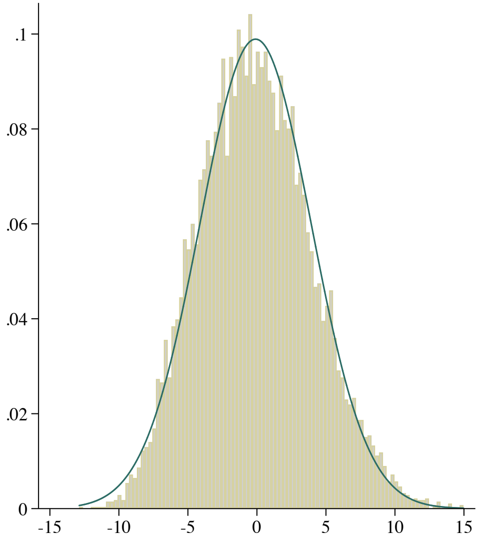
- **p-value**: the probability of obtaining a result at least as extreme as the result actually observed under the null hypothesis
    - here, the null hypothesis is zero treatment effect, i.e. $H_0 : \mathbb{E}[y|D = 1] = \mathbb{E}[y|D = 0]$
- "2-sided" test: what is the likelihood that we'd find such a large deviation (in absolute value) from zero by chance?
    - here, the answer is 1.4%
    - by convention, estimates are called "statistically significant" (we reject the null hypothesis) if their p-value is less than 5%
    - this is not necessarily a good convention

Central limit theorem



- Above, we used a simulated **test distribution** to calculate p-values

Central limit theorem



- Above, we used a simulated **test distribution** to calculate p-values
  - the simulated distribution looks a lot like a Normal distribution
- Indeed, one of the most striking results in statistics is the *Central Limit Theorem*
  - the sampling distribution of the sample mean of a large number of independent random variabes is approximately Normal

→ We can approximate the test distribution instead of simulating it

  - saves a lot of computing time

Standard errors

- Standard error (SE) summarizes the variability in the treatment effect estimate.

Standard errors

- Standard error (SE) summarizes the variability in the treatment effect estimate. In our example:

$$\hat{SE} = S(Y_i)\sqrt{\frac{1}{n_1} + \frac{1}{n_0}}$$
$$= 18.4\sqrt{\frac{1}{54} + \frac{1}{107}}$$
$$= 4.02$$

where $S(y_i)$ is the sample standard deviation of y in the pooled sample, and n_1 and n_0 are the number of observations in the treatment and control groups

Standard errors

- Standard error (SE) summarizes the variability in the treatment effect estimate. In our example:

$$\hat{SE} = S(Y_i)\sqrt{\frac{1}{n_1} + \frac{1}{n_0}}$$
$$= 18.4\sqrt{\frac{1}{54} + \frac{1}{107}}$$
$$= 4.02$$

where $S(y_i)$ is the sample standard deviation of y in the pooled sample, and n_1 and n_0 are the number of observations in the treatment and control groups

  - close to the standard deviation of 4.03 in our simulated test distribution
  - it is also the number reported in parentheses of Table V

Standard errors

- Standard error (SE) summarizes the variability in the treatment effect estimate. In our example:

$$\hat{SE} = S(Y_i)\sqrt{\frac{1}{n_1} + \frac{1}{n_0}} = 18.4\sqrt{\frac{1}{54} + \frac{1}{107}} = 4.02$$

  where $S(y_i)$ is the sample standard deviation of y in the pooled sample, and n_1 and n_0 are the number of observations in the treatment and control groups
  - close to the standard deviation of 4.03 in our simulated test distribution
  - it is also the number reported in parentheses of Table V
- Estimates more precise when:
  1. the outcome variable has less variation [lower $S(y_i)$]
  2. the experiment is larger [higher n_1 and/or n_0]

t-statistic and significance testing

- **t-statistic = point estimate / SE**. In our example:

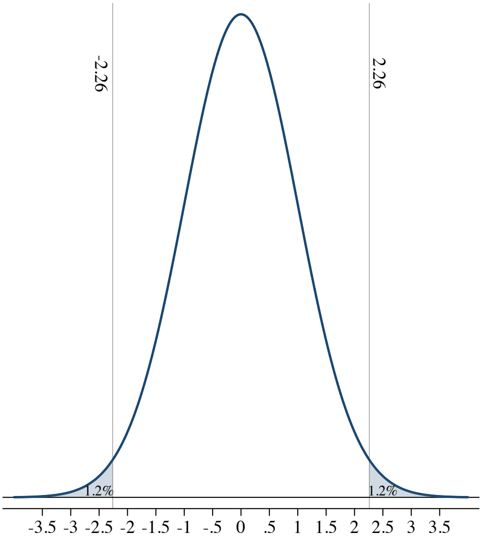
$$t = \frac{9.1}{4.02} = 2.26$$

t-statistic and significance testing

- **t-statistic = point estimate / SE**. In our example:

$$t = \frac{9.1}{4.02} = 2.26$$

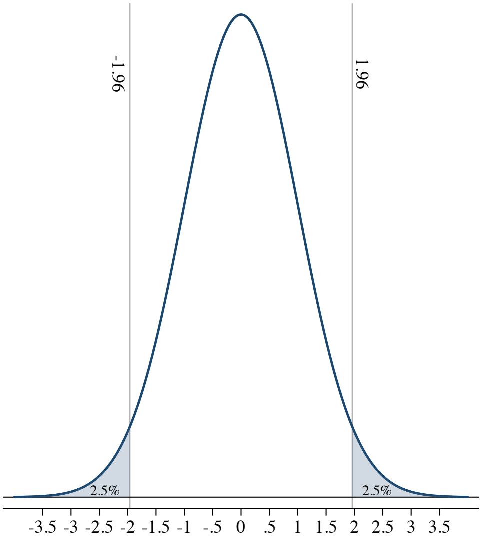
- How exceptional would it be to draw 2.26 or more from a standard Normal distribution?
  - fact: $t \sim \mathcal{N}(0, 1)$ (approximately)
    $\rightarrow$ the likelihood of drawing -2.26 (or less) is 1.19%
    $\rightarrow$ the (two-sided) p-value is $2 \times 0.0119 = 0.0238$



Critical values and a rule-of-thumb

- Critical value is a point in the test distribution corresponding to a specific p-value
  - in large samples, a t-statistic of 1.96 corresponds to a p-value of 0.05 in a 2-sided test

→ A common rule-of-thumb is to call a result "statistically significant" if the point estimate is at least twice as large as its standard error



Confidence intervals

- Often the relevant question is how large/small effects we can rule out
  - instead of testing whether we can reject the null hypothesis of no effect at some confidence level (as in the previous slides)

Confidence intervals

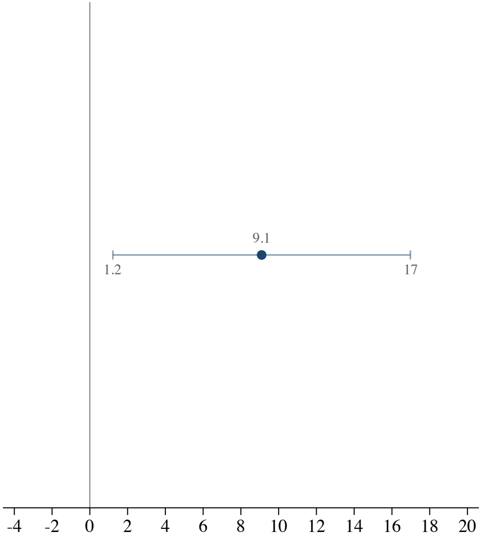
- Often the relevant question is how large/small effects we can rule out
  - instead of testing whether we can reject the null hypothesis of no effect at some confidence level (as in the previous slides)
- We answer this using **confidence intervals**. For example, the 95% confidence interval is

$$[\hat{\beta} - 1.96 \times \hat{SE}, \hat{\beta} + 1.96 \times \hat{SE}]$$

  where $\hat{\beta}$ is the point estimate and $\hat{SE}$ the estimated standard error
- In our example, we had $\hat{\beta} = 9.1$, $\hat{SE} = 4.02 \rightarrow$ the 95% CI is

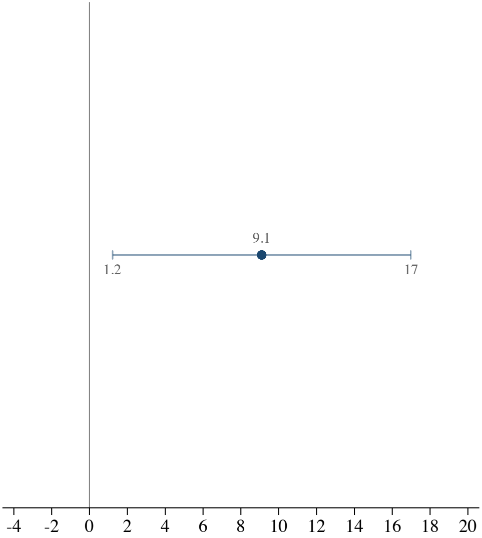
$$[9.1 - 1.96 \times 4.02, 9.1 + 1.96 \times 4.02] \Leftrightarrow [1.2, 17.0]$$

Confidence intervals

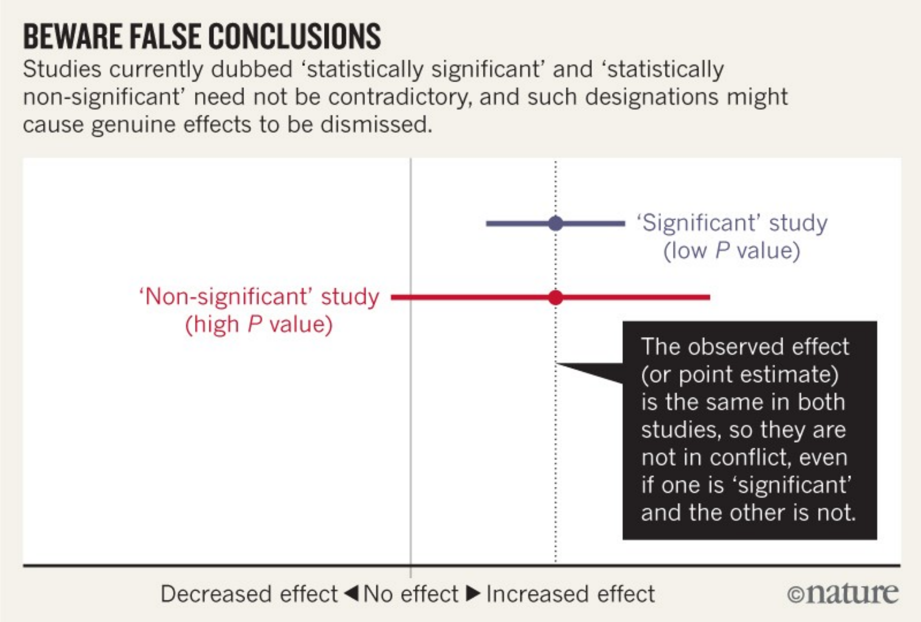


- CIs are often presented graphically
  - e.g. the point estimate and 95% CI for our running example would look like this

Confidence intervals



- CIs are often presented graphically
  - e.g. the point estimate and 95% CI for our running example would look like this
- Helps to clarify that an estimate can be statistically insignificant because
  1. estimate is small and precisely estimated $\rightarrow$ we can rule out economically significant effects
  2. estimate is imprecisely estimated $\rightarrow$ we cannot rule out much



Source: Amrhein et al. (2019)

Summary

- **Standard error** is the standard deviation of a statistic
  - tells how *precise* our point estimate is
  - estimates become more precise (smaller SE) as the sample size increases or variation in the outcome variable decreases
- **P-value** is the probability of obtaining a result at least as extreme as the result actually observed if the null hypothesis is true
  - convention to call results "statistically significant" if $p < .05$
  - corresponds to $|\text{point estimate}| \geq 2 \times$ standard error
- **Confidence interval** includes values most compatible with the data
  - the point estimate is *the* most compatible value