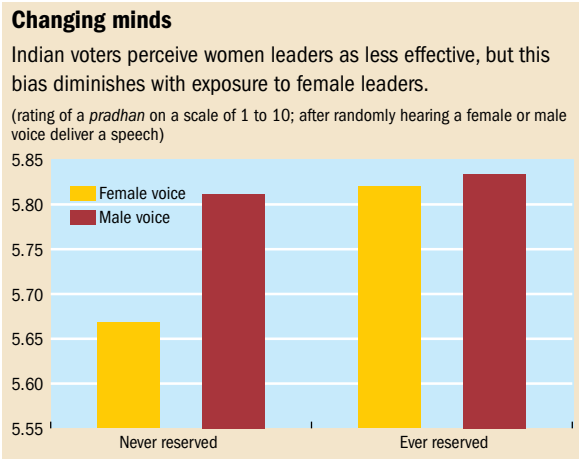
Statistical Inference

Matti Sarvimäki

Principles of Empirical Analysis
Lecture 5

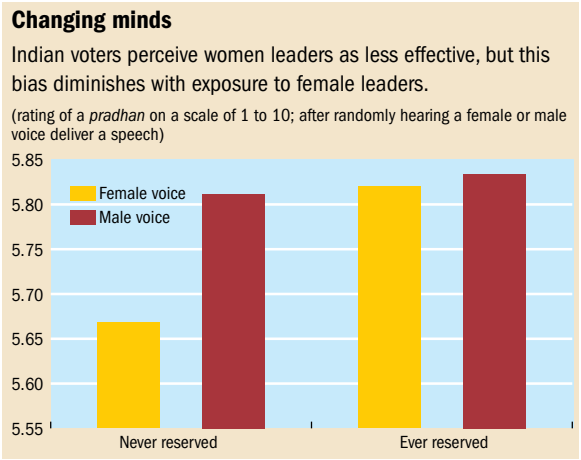
In-class discussion: Quotas in Indian local politics

- Let's start with a closer look at one of the summary figures in the summary article Women in Charge
  - what do we learn from this figure?



In-class discussion: Quotas in Indian local politics

- Let's start with a closer look at one of the summary figures in the summary article Women in Charge
  - what do we learn from this figure?
  - would you like to have any further information before making up your mind about whether women leader truly reduce bias?



Hypothesis testing and statistical significance

- Today's 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

- Today's 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**.

Another example: Gender and policy decisions

- The first study to examine India's 1993 reform was Chattopadhyay and Duflo's 2004 *Econometrica* paper on policy outcomes
  - take-away: leaders invest more in infrastructure that is directly relevant to the needs of their own genders (e.g. drinking water for women)

Another example: Gender and policy decisions

- The first study to examine India's 1993 reform was Chattopadhyay and Duflo's 2004 *Econometrica* paper on policy outcomes
  - take-away: leaders invest more in infrastructure that is directly relevant to the needs of their own genders (e.g. drinking water for women)
- For example, 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

Another example: Gender and policy decisions

- The first study to examine India's 1993 reform was Chattopadhyay and Duflo's 2004 *Econometrica* paper on policy outcomes
  - take-away: leaders invest more in infrastructure that is directly relevant to the needs of their own genders (e.g. drinking water for women)
- For example, 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) report averages
    - first row of column (3) reports difference in averages
    - second row **reports standard errors** (SE)

Another example: Gender and policy decisions

- The first study to examine India's 1993 reform was Chattopadhyay and Duflo's 2004 *Econometrica* paper on policy outcomes
  - take-away: leaders invest more in infrastructure that is directly relevant to the needs of their own genders (e.g. drinking water for women)
- For example, 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) report averages
    - first row of column (3) reports difference in averages
    - second row **reports standard errors** (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 randomly into "treatment" and "control" groups and calculate their averages
  - you can get the data here
  - ... and my simulation code here

Simulating a test distribution

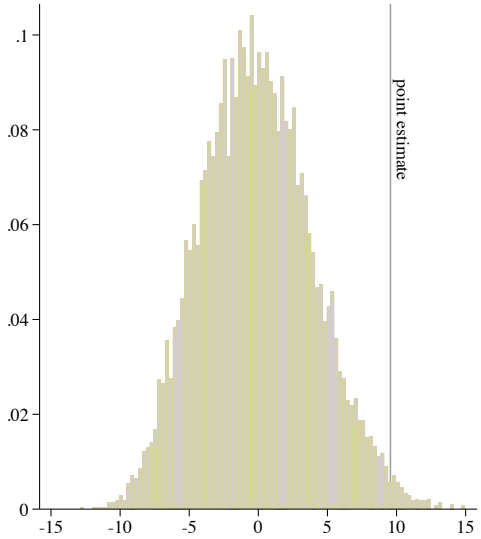
- An intuitive way to think about randomly occurring differences between groups is to create a distribution of "placebo" treatments

- Split the GPs randomly into "treatment" and "control" 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

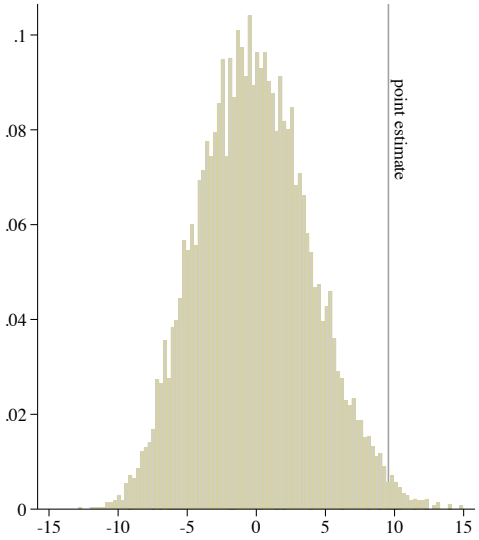
10 "placebo" simulations

Simulating a test distribution



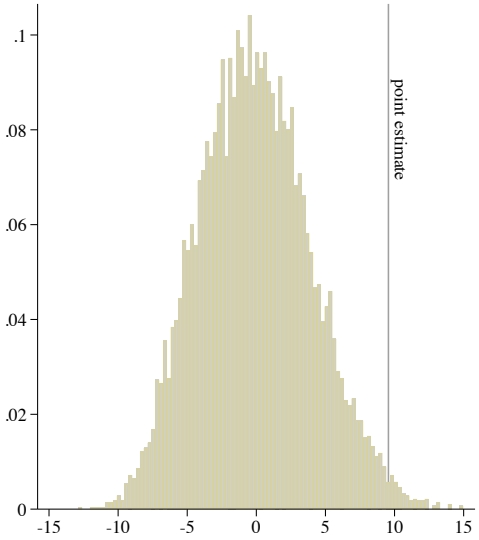
- Simulation with 10,000 rounds
  - average: -0.099
  - standard deviation: 4.03

Simulating a test distribution



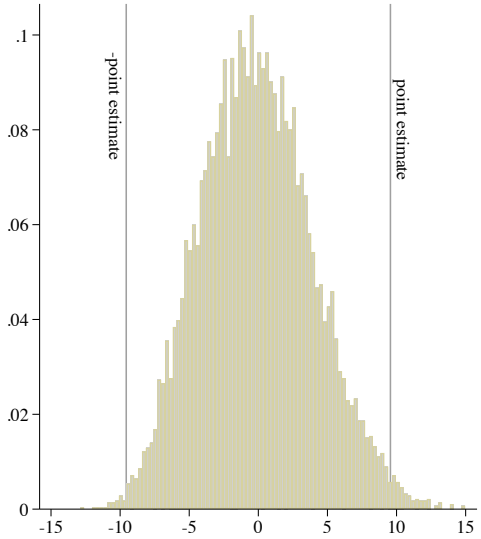
- Simulation with 10,000 rounds
  - average: -0.099
  - standard deviation: 4.03
- As you see from the histogram, sometimes random splits of the sample yield differences that are larger than the point estimate
  - the largest difference is 14.97

Simulating a test distribution



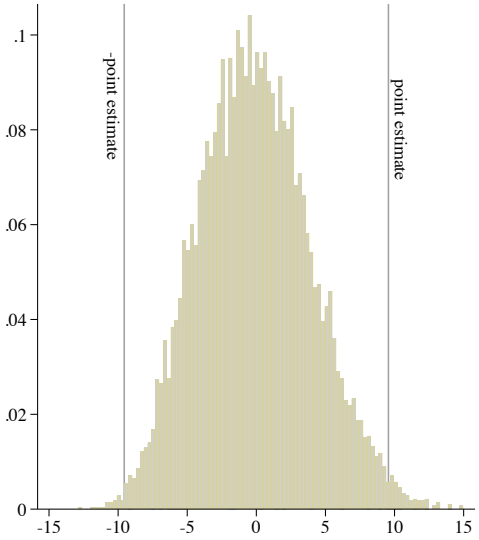
- Simulation with 10,000 rounds
  - average: -0.099
  - standard deviation: 4.03
- As you see from the histogram, sometimes random splits of the sample yield differences that are larger than the point estimate
  - the largest difference is 14.97
- However, this is quite rare:
  - difference $>$ point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

P-value



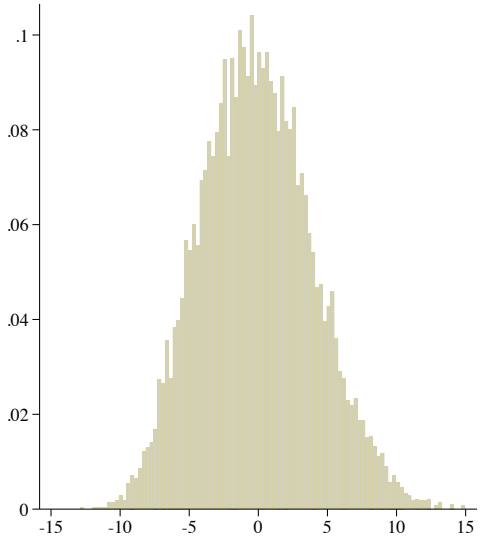
- **p-value**: the probability of obtaining a result at least as extreme as the result actually observed under the null hypothesis
  - here, the null hypothesis is zero treatment effect, i.e. $H_0 : \mathbb{E}[y|D = 1] = \mathbb{E}[y|D = 0]$

P-value



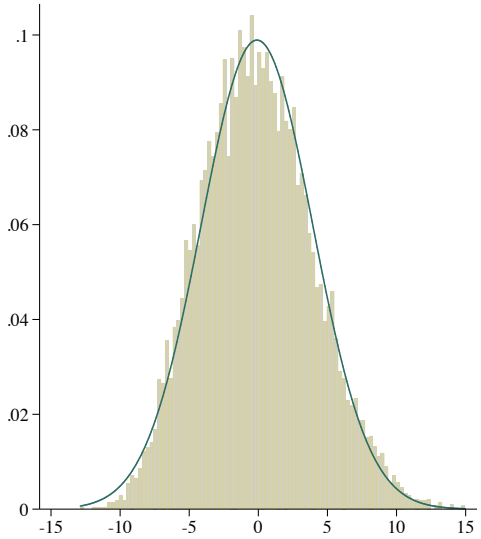
- **p-value**: the probability of obtaining a result at least as extreme as the result actually observed under the null hypothesis
  - here, the null hypothesis is zero treatment effect, i.e. $H_0 : \mathbb{E}[y|D = 1] = \mathbb{E}[y|D = 0]$
- "2-sided" test: what is the likelihood that we'd find such 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%

Central limit theorem



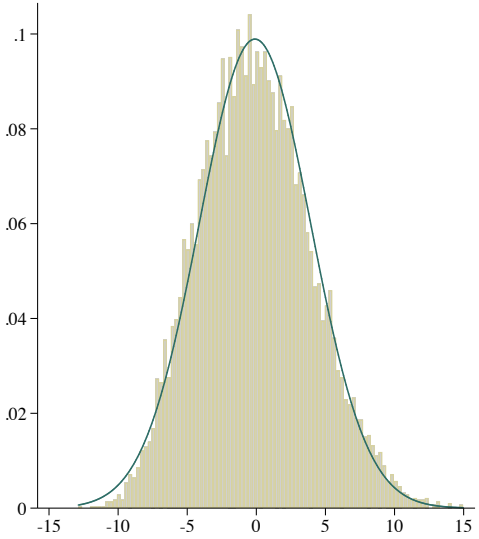
- Above, we used a simulated **test distribution** to calculate p-values

Central limit theorem



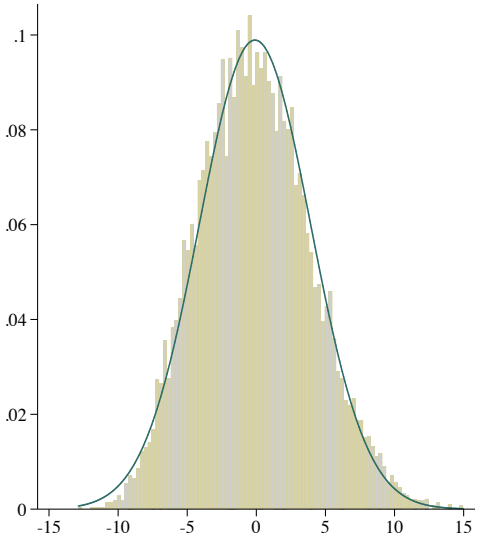
- Above, we used a simulated **test distribution** to calculate p-values
  - the simulated distribution looks a lot like a Normal distribution

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

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

$\rightarrow$ We can approximate the test distribution instead of simulating it

- saves a lot of computing time

Standard errors

- Standard error is the **standard deviation of a statistic**
  - here, the statistic of interest is the treatment effect estimate (difference between treatment and control group means)

Standard errors

- Standard error is the **standard deviation of a statistic**
  - here, the statistic of interest is the treatment effect estimate (difference between treatment and control group means)
- We can estimate the standard error for the difference in averages between two groups with

$$\hat{SE}(\bar{y}^1 - \bar{y}^0) = S(y_i)\sqrt{\frac{1}{n_1} + \frac{1}{n_0}}$$

where $S(y_i) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$ is the sample standard deviation of y, and n_1 and n_0 are the number of observations in the treatment and control groups

Standard errors

- Standard error is the **standard deviation of a statistic**
  - here, the statistic of interest is the treatment effect estimate (difference between treatment and control group means)
- We can estimate the standard error for the difference in averages between two groups with

$$\hat{SE}(\bar{y}^1 - \bar{y}^0) = S(y_i)\sqrt{\frac{1}{n_1} + \frac{1}{n_0}}$$

where $S(y_i) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$ is the sample standard deviation of y, and n_1 and n_0 are the number of observations in the treatment and control groups

  - many alternative estimators for SEs exists, each corresponds to different assumptions about the data generating process (later courses)
  - here we assume that the given estimate of the standard error is approriate and focus on its interpretation

Standard errors

- Standard error **summarizes the variability in the treatment effect estimate** due to
  1. random sampling (lecture 2)
     - hence the SEs for averages in Table V
  2. randomness in treatment/control assignment
     - who happens to end up in the treatment vs. control group

Standard errors

- Standard error **summarizes the variability in the treatment effect estimate** due to
  1. random sampling (lecture 2)
     - hence the SEs for averages in Table V
  2. randomness in treatment/control assignment
     - who happens to end up in the treatment vs. control group
  - note that even when the data includes the full population (and thus there is no random sampling), the second source of variability remains

Standard errors

- Standard error **summarizes the variability in the treatment effect estimate** due to
  1. random sampling (lecture 2)
     - hence the SEs for averages in Table V
  2. randomness in treatment/control assignment
     - who happens to end up in the treatment vs. control group
  - note that even when the data includes the full population (and thus there is no random sampling), the second source of variability remains
- Experiments yield more precise evidence when:
  1. the outcome variable has less variation [lower $S(y_i)$]
  2. the experiment is larger [higher n_1 and/or n_0]

Standard errors for female leader treatment effect

- Going back to our earlier example, the corresponding numbers are

$$\hat{SE}(\bar{Y}^1 - \bar{Y}^0) = S(Y_i)\sqrt{\frac{1}{n_1} + \frac{1}{n_0}} = 18.4\sqrt{\frac{1}{54} + \frac{1}{107}} = 4.02$$

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

Point estimate / standard error = the t-statistic

- Let's denote the statistic of interest with κ and its value under the null hypothesis with μ. Then the t-statistic is

$$t(\mu) = \frac{\kappa - \mu}{SE(\kappa)}$$

- For treatment effects, the most common null hypothesis is $H_0 : \mu = 0$
  - under this null hypothesis, the t-value for an estimate of the average treatment effect is

$$t(0) = \frac{\bar{Y}^1 - \bar{Y}^0}{\hat{SE}(\bar{Y}^1 - \bar{Y}^0)}$$

- The t-value is distributed, approximately, $t \sim \mathcal{N}(0, 1)$
  - in words: the t-value approximately follows the Normal distribution with mean zero, standard deviation one ("standard Normal distribution")

t-statistic and significance testing

- Again, let's go back to our example and calculate the t-statistic

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

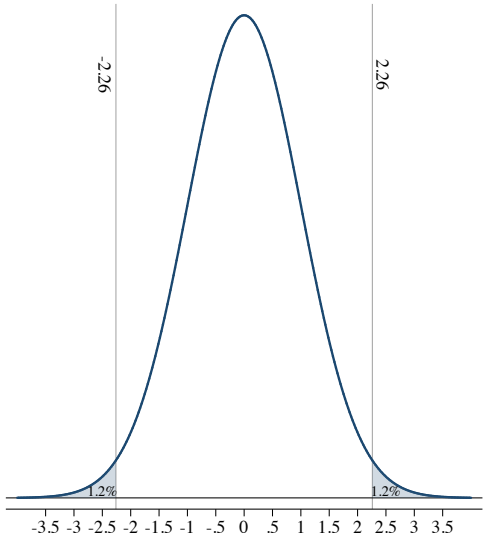
t-statistic and significance testing

- Again, let's go back to our example and calculate the t-statistic

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

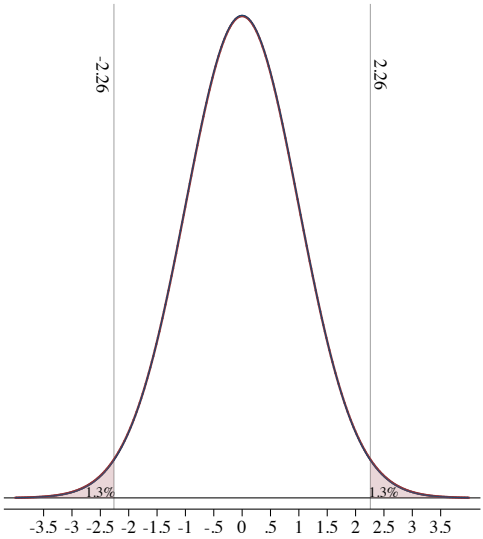
- How exceptional would it be to draw 2.26 or more from a standard Normal distribution?
  - turns out this would happen with 1.19% probability
  - the likelihood of drawing -2.26 (or less) is also 1.19%

→ the (two-sided) p-value is 0.0238



t-statistic and significance testing

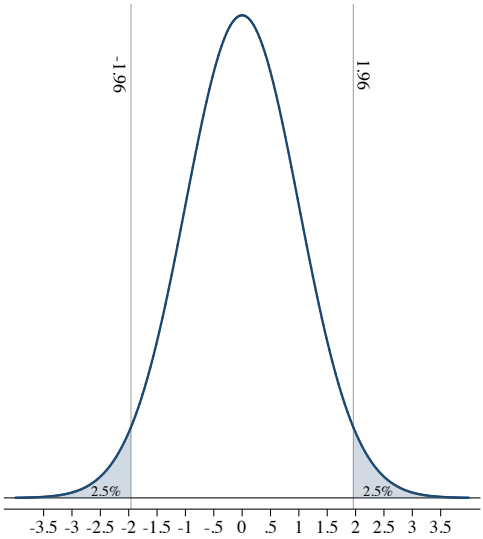
- Strictly speaking, we use Student's t-distribution for calculating p-values
  - it approaches the Normal distribution when the sample size increases
- Most applications have sufficient sample size to make this distinction irrelevant
  - here, p-value increases from 0.0238 to 0.0252



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

Confidence intervals

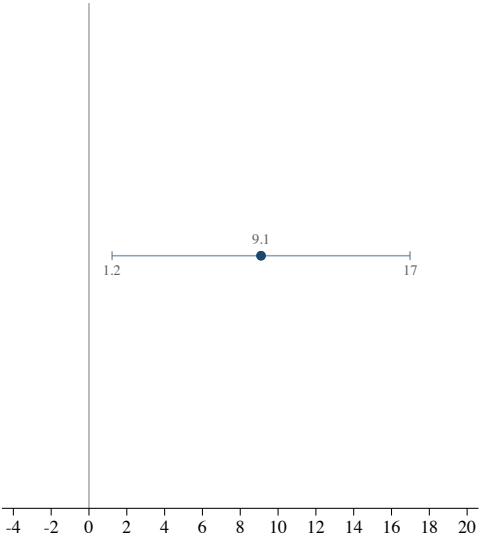
- Often the relevant question is how large/small effects we can rule out
  - instead of testing whether we can reject the null hypothesis of no effect at some confidence level (as in the previous slides)
- We answer this using **confidence intervals**. For example, the 95% confidence interval is

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

  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

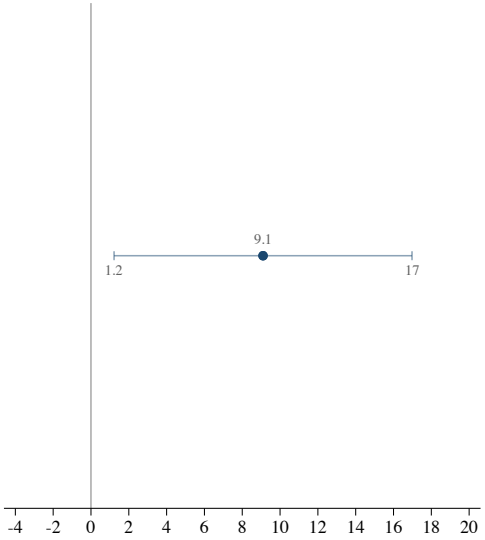
$$[9.1 - 1.96 \times 4.02, 9.1 + 1.96 \times 4.02] \Leftrightarrow [1.2, 17.0]$$

Confidence intervals



- CIs are often presented graphically
  - e.g. the point estimate and 95% CI for our running example would look like this
- This is an informative and compact way to present results

Confidence intervals



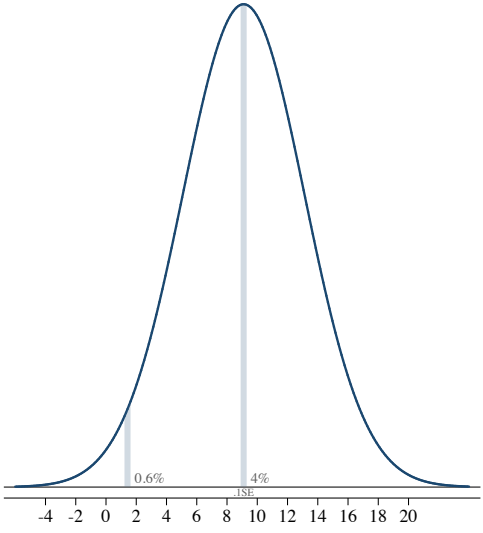
- CIs are often presented graphically
  - e.g. the point estimate and 95% CI for our running example would look like this
- This is an informative and compact way to present results
  - but: the exact interpretation of confidence intervals is a surprisingly subtle subject
  - these subtleties are left to much more advanced courses
  - here, I follow Amrhein et al. (2019); most applied economists probably have this kind of an interpretation in mind

Conditional descriptive statistics

- Confidence interval contains the values *most* compatible with the data
  - values outside the CI are *not* incompatible; they are just less compatible
- Values just outside the CI do *not* differ substantively from those just inside

Conditional descriptive statistics

- Confidence interval contains the values *most* compatible with the data
  - values outside the CI are *not* incompatible; they are just less compatible
- Values just outside the CI do *not* differ substantively from those just inside
- Not all values inside CI are equally compatible
  - point estimate is the most compatible, values near it are more compatible than those near the limits (this is the contentious part)



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 $|\text{point estimate}| \geq 2 \times \text{standard error}$
- **Confidence interval** includes values most compatible with the data
  - the point estimate is *the* most compatible value