

# A P P L I E D   S T A T I S T I C A L   A N A L Y S I S   I

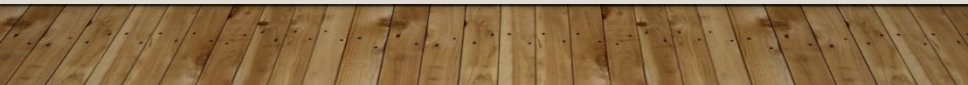
## Hypothesis testing, experiments, difference in means

Trajche Panov, PhD

[panovt@tcd.ie](mailto:panovt@tcd.ie)

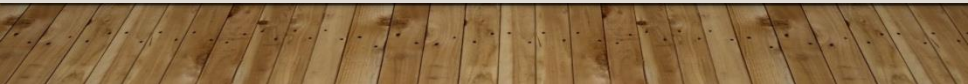
Department of Political Science  
Trinity College Dublin

September 17, 2024



# Today's Agenda

- (1) Lecture recap
- (2) Tutorial exercises



# Null-hypothesis significance testing

*What are the five steps of null-hypothesis significance testing?*

# Null-hypothesis significance testing

TABLE 6.1: The Five Parts of a Statistical Significance Test

1. **Assumptions**  
Type of data, randomization, population distribution, sample size condition
2. **Hypotheses**  
Null hypothesis,  $H_0$  (parameter value for “no effect”)  
Alternative hypothesis,  $H_a$  (alternative parameter values)
3. **Test statistic**  
Compares point estimate to  $H_0$  parameter value
4. **P-value**  
Weight of evidence against  $H_0$ ; smaller  $P$  is stronger evidence
5. **Conclusion**  
Report  $P$ -value  
Formal decision (optional; see Section 6.4)

(Agresti and Finlay 2009, 147)

# Null-hypothesis significance testing

## *Step 1: Assumptions*

- Type of data: Continuous or categorical data
- Sampling method: Data randomly obtained (e.g., random sample)
- Population distribution: Variable assumed to follow certain distribution (e.g., normal distribution)
- Sample size: Validity improves with increasing sample size

(Agresti and Finlay [2009](#), 144)

# Null-hypothesis significance testing

## Step 2: Hypotheses

- Hypothesis: “a statement about a population. It takes the form of a prediction that a parameter takes a particular numerical value or falls in a certain range of values” (Agresti and Finlay [2009](#), 143).
- Null ( $H_0$ ) and alternative hypothesis ( $H_a$ ): “The null hypothesis is a statement that the parameter takes a particular value. The alternative hypothesis states that the parameter falls in some alternative range of values. Usually the value in the null hypothesis corresponds, in a certain sense, to no effect. The values in the alternative hypothesis then represent an effect of some type” (Agresti and Finlay [2009](#), 144).
- Can be one-sided ( $<$ ,  $>$ ,  $\geq$ ,  $\leq$ ) or two-sided ( $=$ ,  $\neq$ )

# Null-hypothesis significance testing

## *Step 3: Test statistic*

- Test statistic: “The test statistic summarizes how far that estimate falls from the parameter value in  $H_0$ . Often this is expressed by the number of standard errors between the estimate and the  $H_0$  value” (Agresti and Finlay [2009](#), 145).
- Depending on probability distribution, test statistic is also called z-statistic or t-statistic.

# Null-hypothesis significance testing

## Step 4: *P*-value

- P-value: “The P-value is the probability that the test statistic equals the observed value or a value even more extreme in the direction predicted by  $H_a$ . It is calculated by presuming that  $H_0$  is true. The smaller the P-value, the stronger the evidence is against  $H_0$ ” (Agresti and Finlay [2009](#), 145).



# Null-hypothesis significance testing

## Step 5: Conclusion

		Null hypothesis ( $H_0$ ) is	
		True	False
Decision about Null hypothesis ( $H_0$ )	Don't reject	Correct inference (true negative) (probability = $1-\alpha$ )	Type II Error (false negative) (probability = $\beta$ )
	Reject	Type I Error (false positive) (probability = $\alpha$ )	Correct inference (true positive) (probability = $1-\beta$ )

→ most concerned about Type I Error, probability of obtaining a false positive result,  $\alpha$  (error probability), p-value.

# Null-hypothesis significance testing

## Step 5: Conclusion

- Validate whether the obtained test statistic is unlikely to occur, under the assumption that the null hypothesis is true. → p-value, if probability is low, we reject  $H_0$ .
- More precisely, select  $\alpha$ -level, which indicates acceptable probability of Type I error (usually 0.05, 0.01). If p-value  $< \alpha$ , we reject  $H_0$ . → proof by contradiction (disprove the null)
- Careful when making conclusion, there is still a probability that we falsely reject  $H_0$ , even if p-value  $< \alpha$ . → We use a certain language

# Two hypothesis tests for today

1. Null hypothesis test for a mean
2. Null hypothesis test for a difference in means

# Null hypothesis test for a mean (t-test)

**TABLE 6.3:** The Five Parts of Significance Tests for Population Means

1. **Assumptions**  
 Quantitative variable  
 Randomization  
 Normal population (robust, especially for two-sided  $H_a$ , large  $n$ )
2. **Hypotheses**  
 $H_0: \mu = \mu_0$   
 $H_a: \mu \neq \mu_0$  (or  $H_a: \mu > \mu_0$  or  $H_a: \mu < \mu_0$ )
3. **Test statistic**  

$$t = \frac{\bar{y} - \mu_0}{se} \text{ where } se = \frac{s}{\sqrt{n}}$$
4. **P-value**  
 In  $t$  curve, use  
 $P$  = Two-tail probability for  $H_a: \mu \neq \mu_0$   
 $P$  = Probability to right of observed  $t$ -value for  $H_a: \mu > \mu_0$   
 $P$  = Probability to left of observed  $t$ -value for  $H_a: \mu < \mu_0$
5. **Conclusion**  
 Report  $P$ -value. Smaller  $P$  provides stronger evidence against  $H_0$  and supporting  $H_a$ . Can reject  $H_0$  if  $P \leq \alpha$ -level.

# Null hypothesis test for a mean (t-test)

TABLE 6.2: Responses of Subjects on a Scale of Political Ideology

Response	Race		
	Black	White	Other
1. Extremely liberal	10	36	1
2. Liberal	21	109	13
3. Slightly liberal	22	124	13
4. Moderate, middle of road	74	421	27
5. Slightly conservative	21	179	9
6. Conservative	27	176	7
7. Extremely conservative	11	28	2
	$n = 186$	$n = 1073$	$n = 72$

$H_0$ : mean political ideology is moderate  $\rightarrow \mu = 4.0$

$H_a$ : mean political ideology falls in liberal or conservative direction  
 $\rightarrow \mu \neq 4.0, (\mu < 4 \text{ or } \mu > 4).$

(Agresti and Finlay 2009, 149)

## Null hypothesis test for a mean (t-test)

With  $\bar{y} = 4.075$ ,  $s = 1.512$ , we calculate  $se = \frac{\sqrt{s}}{n} = \frac{1.512}{186} = 0.111$

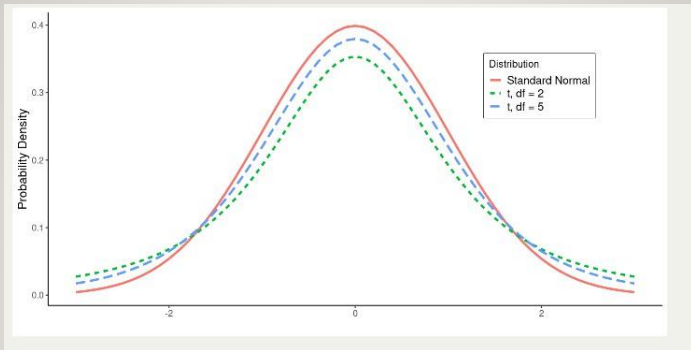
and then  $t = \frac{\bar{y} - \mu_0}{se} = \frac{4.075 - 4.0}{0.111} = 0.68$

How to interpret this value? How likely are we to observe data in sample (this test statistics), under the assumption that  $H_0$  is true?

→ Probability distribution

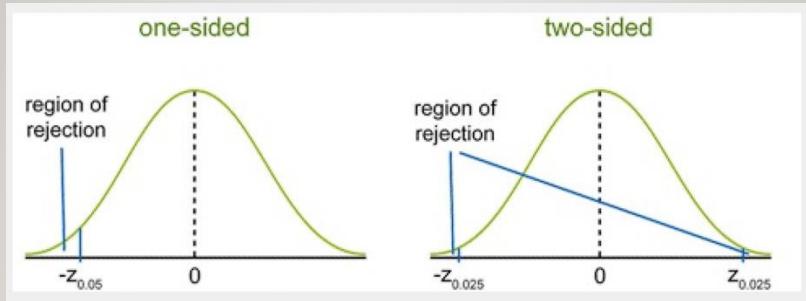
(Agresti and Finlay [2009](#), 150)

## Null hypothesis test for a mean (t-test)



What is the conclusion? P-value=0.50, and is not  $< 0.05$ , thus we cannot reject  $H_0 \rightarrow$  It is plausible that the population mean is 4.0, and therefore moderate.

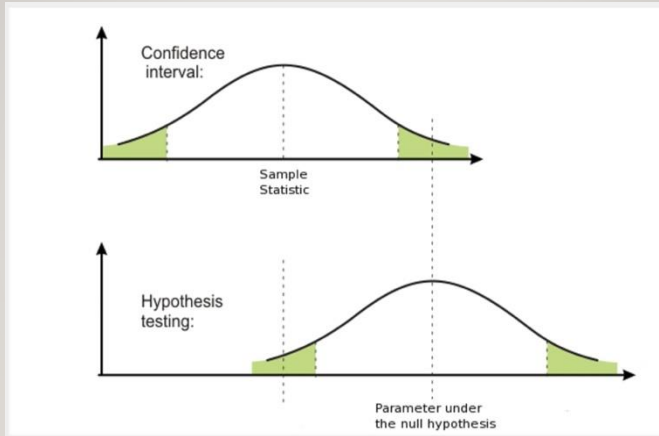
# One-sided versus two-sided test



“In most research articles, significance tests use two-sided P-values” (Agresti and Finlay [2009](#), 153). Why? But can also be two-sided, always depends on research question.



# Confidence intervals and NHST



Does confidence intervals contain parameter under null hypothesis?

# Null hypothesis test for a difference in means (t-test)

*What is a t-test for the difference in means?*

- Null and alternative hypothesis: (Step 2) The means of two groups are identical,  $\bar{y}_1 = \bar{y}_2$  or  $\bar{y}_1 - \bar{y}_2 = 0$  ( $H_0$ ), the means of two groups are different,  $\bar{y}_1 \neq \bar{y}_2$  ( $H_a$ ).
- Test statistics: (Step 3) "measures the number of standard errors between the estimate and the  $H_0$  value" (Agresti and Finlay [2009](#), 192).

$$t = \frac{\text{Estimate of parameter} - \text{Null hypothesis value of parameter}}{\text{Standard error of estimate}}$$

$$t = \frac{(\bar{y} - \bar{y}) - 0}{se}, H_0 \text{ assumes } \bar{y}_2 - \bar{y}_1 = 0, se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

# Null hypothesis test for a difference in means (t-test)

*What is a t-test for the difference in means?*

**TABLE 7.1:** Cooking and Washing Up Minutes, per Day, for a National Survey of Men and Women Working Full Time in Great Britain

Sex	Sample Size	Cooking and Washing Up Minutes	
		Mean	Standard Deviation
Men	1219	23	32
Women	733	37	16

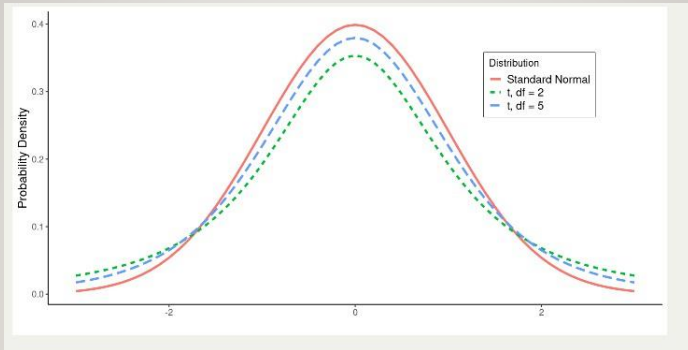
$$t = \frac{(\bar{y}_2 - \bar{y}_1) - 0}{se} = \frac{(37 - 23)}{1.09} = 12.8$$

How to interpret this value? How likely are we to observe data in sample (this test statistics), under the assumption that  $H_0$  is true?

→ Probability distribution

(Agresti and Finlay [2009](#), 192)

# Null hypothesis test for a difference in means (t-test)



What is the conclusion? P-value < 0.05, We can reject  $H_0$  with an error probability (p-value) of essentially 0%. → Suggests that population means differ, and sample means show that women have a higher mean household work time than men

# Causal effect

*What is a causal effect? What is the fundamental problem of causal inference?*



# Causal effect

*What is a causal effect?*

- Outcome/dependent variable ( $Y$ ), with potential outcomes  $Y_{T=0,i}/Y_i^0$  and  $Y_{T=1,i}/Y_i^1$
- Treatment/independent variable ( $T$ ), with control ( $T = 0$ ) and treatment condition ( $T = 1$ )
- Causal effect: “change in some feature of the world that would result from a change to some other feature of the world”,  
 $Y_{T=1,i} - Y_{T=0,i} = Y_i^1 - Y_i^0$
- Counterfactual comparison: “outcome would be different in a counterfactual world in which the action was different” → what would be the state of  $Y$ , had  $X$  not occurred?

(Bueno de Mesquita and Fowler [2021](#), 159)

# Causal effect

*What is the fundamental problem of causal inference?*

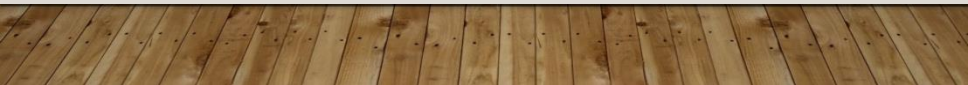
- “we can only observe, at most, one of the two quantities— $Y_{1i}$  or  $Y_{0i}$ —for any individual at a particular point in time”

(Bueno de Mesquita and Fowler [2021](#), 164).

→ causal effect is unobservable

# SATE, Difference in means

*What is the sample average treatment effect (SATE)? And what can we actually observe?*





## SATE, Difference in means

- $SATE = \frac{1}{n} \sum_{i=1}^n \{Y_i(1) - Y_i(0)\}$ , “the average of individual-level treatment effects in the sample”.
- SATE is unobservable due to fundamental problem of causal inference → we only observe sample difference in means

(Imai [2017](#), 49)

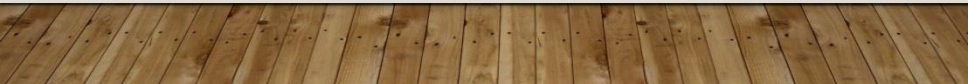
## SATE, Difference in means

- Sample difference in means =  $Mean(y_i^1) - Mean(y_i^0)$ ,  
“difference in average outcome comparing the subgroup of people who in fact received treatment to the subgroup of people who in fact did not receive treatment”.
- Sample difference in means is biased estimate of the true SATE  
→ **Correlation does not imply causation**

(Bueno de Mesquita and Fowler [2021](#), 187)

# Bias

*What are the sources of bias? And how can we overcome bias?*



# Bias

*What are the sources of bias?*

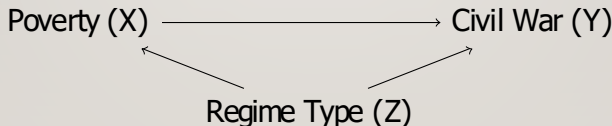
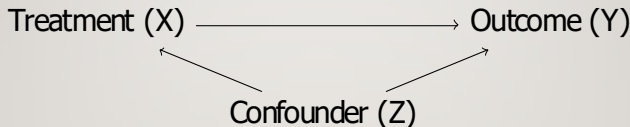
- Baseline differences: “[d]ifference in the average potential outcome between two groups (e.g., the treated and untreated groups), even when those two groups have the same treatment status” → Confounders may cause baseline differences, which may cause bias (*omitted variable bias*)

(Bueno de Mesquita and Fowler [2021](#), 187)

## Bias

*What are the sources of bias?*

1. Confounder: "A feature of the world that (1) has an effect on treatment status and (2) has an effect on the potential outcome over and above the effect it has through its effect on treatment status".



(Bueno de Mesquita and Fowler, 2021, 187)

# Bias

*What are the sources of bias?*

2. Other sources of bias: Reverse causality, unobserved unit heterogeneity (special type of omitted variable bias), post-treatment bias

# Bias

*And how can we overcome bias?*

1. Control for confounders:

- \* But “there will still be unobservable confounders that we can’t control for, reverse causation, or variables that are part confounder and part mechanism” (Bueno de Mesquita and Fowler [2021](#), 215).

2. Randomized experiments: Randomly assigning units to treatment and control group eliminates baseline differences

- \* But “the ideal experiment that we’d like to run is often impractical, infeasible, or unethical” (Bueno de Mesquita and Fowler [2021](#), 239).

3. Causal inference methods (e.g., Difference-in-Difference Design, Instrumental Variables)

# Software check

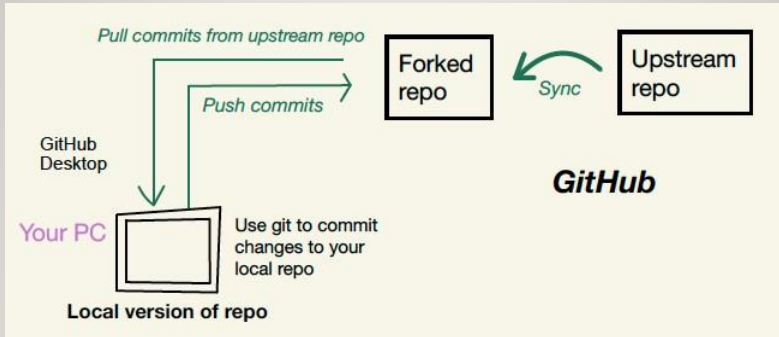
*How to update your local repository? How to git pull?*

We will learn how to git push next week.








# Software check



1. Synchronize fork
2. Fetch origin

# REFERENCES

---

-  Agresti, Alan, and Barbara Finlay. 2009. *Statistical methods for the social sciences*. Essex: Pearson Prentice Hall.
-  Bueno de Mesquita, Ethan, and Anthony Fowler. 2021. *Thinking clearly with data: A guide to quantitative reasoning and analysis*. Princeton: Princeton University Press.
-  Imai, Kosuke. 2017. *Quantitative social science: An introduction*. Princeton: Princeton University Press.

THANK YOU  
FOR YOUR ATTENTION!

---

panovt@tcd.ie