POLI 30 D: Political Inquiry

Professor Umberto Mignozzetti

(Based on DSS Materials)

Lecture 14 | Hypothesis Testing

Before we start

Announcements:

- Quizzes and Participation: On Canvas.
- GitHub page: https://github.com/umbertomig/POLI30Dpublic
- Piazza forum: Not sure what the link is. Ask your TA!
- Note to self: Turn on the mic!

Before we start

Recap: We learned:

- The definitions of theory, scientific theory, and hypotheses.
- Data, datasets, variables, and how to compute means.
- Causal effect, treatments, outcomes, and randomization.
- Sampling, descriptive statistics, correlation, and prediction.
- Strengths and weaknesses of observational and experimental studies.
- Probability, law of large numbers, and central limit theorem.

Great job!

- Do you have any questions about these contents?

Plan for Today

- Hypothesis Testing Intuition
  - Null Hypothesis
  - Alternative Hypothesis
  - Test Statistic
  - P-Values

- Hypothesis Testing Formal Procedure

- Example: Do Small Classes Improve Math Scores?
  - What Is the Estimated Average Treatment Effect?
  - Is the Effect Statistically Significant?

The Context

- Suppose we are estimating the average causal effect of a treatment on an outcome.
- In this context, X is the treatment variable, and Y is the outcome variable.
- What do we need to calculate to estimate the average causal effect?
  - The **difference-in-means** estimator
- We want to compute it by fitting a linear regression: $\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X$.
- Which coefficient is equivalent to the difference-in-means estimator?

The Context

- What we can estimate is $\widehat{\beta}_1$, which is the average causal effect *at the sample level*
- What we care about is β_1, which is the average causal effect *at the population level*
- As we discussed in the last class, samples often differ from population:
  - Noise produced by sampling makes $\widehat{\beta}_1 \neq \beta_1$

The Context

- The question we want to answer is:
    - Looking at the *sample*, do we have enough evidence to conclude that the *population* ATE differs from zero?
    - In other words, can we say that β_1 is unlikely to be zero?
- By the way, why do you think we focus on zero?
- To answer this question, we need to do something called a **hypothesis testing**

Hypothesis Testing

Hypothesis Testing

- We assume the contrary of what we want to prove and test if this leads to a contradiction.
- Suppose a person is on trial for murder. To be fair to the person, we assume that she is innocent.
  - Then, we look at the evidence:
    - **Person 1**: No good alibi, DNA, or footage.
    - **Person 2**: No good alibi, has blood in the crime scene with matched DNA, and footage showing the person leaving minutes after the crime.
- Which person is more likely to be innocent?

Hypothesis Testing

- By the way, **both** could. be *innocent*. Or **both** could be *guilty*. For a given person:

	H_0 is true Truly not guilty	H_1 is true Truly guilty
Do not reject the null hypothesis **Acquittal**	Right decision	Wrong decision Type II Error
Reject null hypothesis **Conviction**	Wrong decision Type I Error	Right decision

- This makes hypothesis testing hard: We use what we see to infer about something we did not see.

Hypothesis Testing

- **Person 1**: No good alibi, DNA, or footage.

	H_0 is true Truly not guilty	H_1 is true Truly guilty
Do not reject the null hypothesis **Acquittal**	Right decision	Wrong decision Type II Error
Reject null hypothesis **Conviction**	Wrong decision Type I Error	Right decision

Hypothesis Testing

- **Person 2**: No good alibi, has blood in the crime scene with matched DNA, and footage showing the person leaving minutes after the crime.

	H_0 is true **Truly not guilty**	**H_1 is true** **Truly guilty**
Do not reject the null hypothesis **Acquittal**	Right decision	Wrong decision Type II Error
Reject null hypothesis **Conviction**	Wrong decision Type I Error	Right decision

Hypothesis Testing

We have two powerful friends: LLN (Law of Large Numbers) and the CLT (Central Limit Theorem).

- We always assume **no** relationship between variables.
- This is called **null hypothesis**. It states that β_1 is zero:
$$H_0 \colon \beta_1 = 0$$
- Our **alternative hypothesis** state that β_1 is different than zero:
$$H_1 \colon \beta_1 \neq 0$$

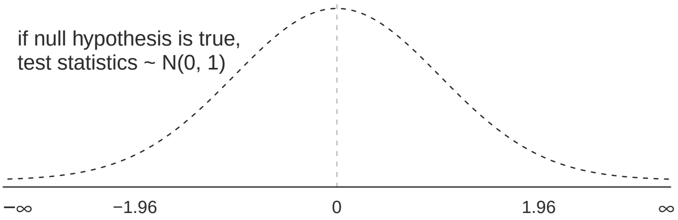
Hypothesis Testing

- Thanks to the LLN, we know that the larger the sample, the closer we are to the actual value.
- Thanks to CLT, we know that if H_0 is true, then the test statistic over multiple samples:

$$\boxed{\text{test-statistic} = \frac{\widehat{\beta}_1 - 0}{\text{standard error of } \widehat{\beta}_1} \sim N(0, 1)}$$

Hypothesis Testing

- If we were to draw multiple large samples from the same target population, $\frac{\widehat{\beta}_1 - 0}{\text{standard error of } \widehat{\beta}_1}$ would be distributed as a standard normal:

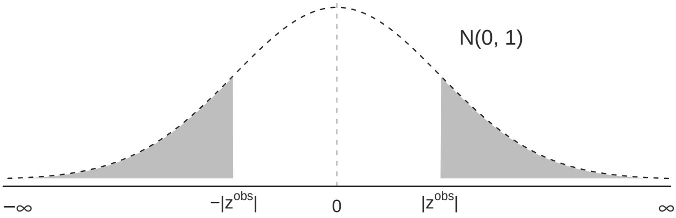


Hypothesis Testing

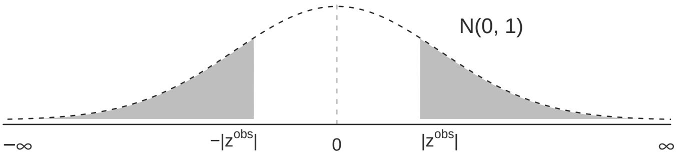
- In reality, we only draw one sample:
  - We will only observe one test statistic: z^{obs}
  - But we know the distribution of the test statistics if the null hypothesis is true.
- We can calculate the chance that we observe a test statistic as extreme or more extreme as the one we do observe if H_0 is true.
  - This is know as the **p-value**: $P(Z \leq -|z^{obs}|) + P(Z \geq |z^{obs}|)$

Hypothesis Testing

- **p-value**: $P(Z \leq -|z^{obs}|) + P(Z \geq |z^{obs}|)$



Hypothesis Testing



- **If the p-value is large**: the probability that we observe z^{obs} or more extreme is large if H_0 is true
  - z^{obs} is common relative to the distribution of test statistics under the null
- *Our evidence is consistent with H_0 being true*
- Conclusion: We fail to reject H_0. This is called **not statistically significant**

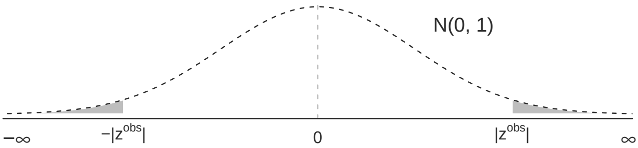
Hypothesis Testing

- **Person 1** again: No good alibi, DNA, or footage.

	H_0 is true Truly not guilty	H_1 is true Truly guilty
Do not reject the null hypothesis **Acquittal**	Right decision	Wrong decision Type II Error
Reject null hypothesis **Conviction**	Wrong decision Type I Error	Right decision

- There is little evidence to reject this person's innocence (the null hypothesis in criminal justice)!

Hypothesis Testing



- **If the p-value is small**: the probability that we observe z^{obs} or more extreme is small if H_0 is true
  - z^{obs} is extreme relative to the distribution of test statistics under the null
- *Our evidence is inconsistent with H_0 being true*
  - Either H_0 is not true, or we got unlucky by drawing a fringe sample
- **Conclusion:** We reject H_0 and conclude that the estimate is **statistically significant**

Hypothesis Testing

- **Person** 2 again: No good alibi, has blood in the crime scene with matched DNA, and footage showing the person leaving minutes after the crime.

	H_0 is true **Truly not guilty**	**H_1 is true** **Truly guilty**
Do not reject the null hypothesis **Acquittal**	Right decision	Wrong decision Type II Error
Reject null hypothesis **Conviction**	Wrong decision Type I Error	Right decision

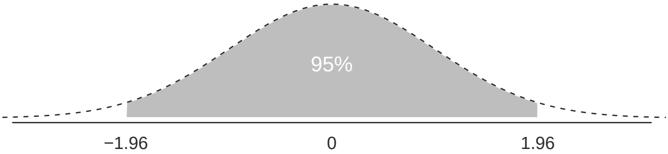
- Again, no certainty, but much more evidence that this person here may be guilty.

Hypothesis Testing

- How small does the p-value need to be to reject the null hypothesis?
  - No good answer to that. In fact, many (not so interesting) papers on the subject.
  - The smaller, the better. We will use the conventional 5% value.

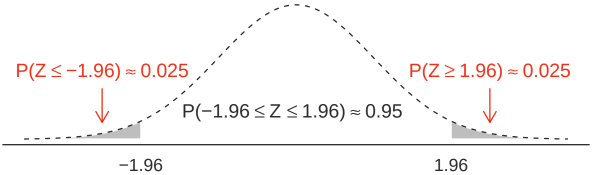
- When **p-value** $> 5\%$: We conclude that the estimate is statistically **insignificant** (*likely to **be zero** at the population level*)
- When **p-value** $\leq 5\%$: We conclude that the estimate is statistically **significant** (*likely to **not be zero** at the population level*)

Shortcut



- Recall: $P(\text{-}1.96 \leq Z \leq 1.96) \;\approx\; 0.95$
- Probability that Z takes a value less than or equal to -1.96 plus the probability that Z takes a value greater than or equal to 1.96 is approximately 5% (1-0.95=0.05)
- $P(Z \leq \text{-}1.96) + P(Z \geq 1.96) \;\approx\; 0.05$

Shortcut



- In short, given the characteristics of Z:
  - When p-value $> 5\%$, it means that $|z^{obs}| < 1.96$
  - When p-value $\leq 5\%$, it means that $|z^{obs}| \geq 1.96$
- We can draw conclusions based on either the value of $|z^{obs}|$ or the value of p-value
  - Both procedures are mathematically equivalent and lead to the same conclusion

Hypothesis Testing

Algorithm:

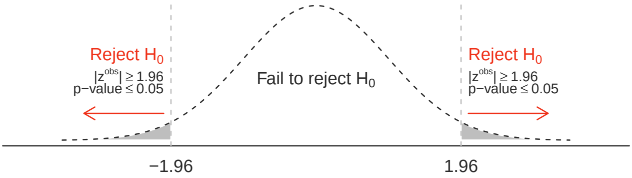
1. Specify null and alternative hypotheses

- H_0: $\beta_1 = 0$ (The true average causal effect at the population level is zero)
- H_1: $\beta_1 \neq 0$ (The true average causal effect at the population level is either positive or negative)

2. Compute the observed value of the test statistic and (perhaps also) the associated p-value

- $z^{obs} = \frac{\widehat{\beta}_1}{\text{standard error of } \widehat{\beta}_1}$
- p-value $= 2 \times P(Z \leq -|z^{obs}|)$

Hypothesis Testing



3. Conclude

- If $|z^{obs}| < 1.96$ or p-value > 0.05, we conclude that the estimate is not statistically significant
- If $|z^{obs}| \geq 1.96$ or p-value ≤ 0.05, that the estimate is statistically significant

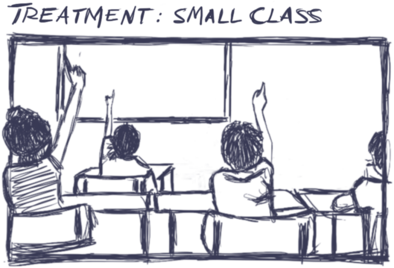
The importance of replication

- When an effect is statistically significant at the 5% level, do we know that the true estimate at the population level is not zero?
  - No, we do not
- But thanks to the LLN and the CLT, we know that if the null hypothesis is true, only in 5% of the samples drawn from the target population we will wrongly reject the null.

The importance of replication

- It is important to replicate social scientific studies:
  - Arriving at similar conclusions when analyzing different samples is reassuring.
- While the probability of falsely rejecting the null hypothesis in any one sample is 5%, the probability of falsely rejecting the null twice in a row is only 0.25%
- Let us return to the STAR dataset and estimate the average causal effect of attending a small class *on math test scores*.

Do Small Classes Improve Math Scores?



(Based on Mosteller. 1995."The Tennessee Study of Class Size in the Early School Grades," *Future of Children* 5 (2): 113–27.)

Do Small Classes Improve Math Scores?

- The data come from a randomized experiment conducted in Tennessee:
  - Students were randomly assigned to attend either a small class or a regular-size class
- To estimate the average causal effect, what estimator can we use?

0. Get Ready for the Analysis

- Load the data and create any variables needed

```
## load and look at the data
star <- read.csv("https://raw.githubusercontent.com/umbertomig/POLI30Dpublic/ma

## create treatment variable
star$small <- # stores return values as new variable
  ifelse(star$classtype=="small", # logical test
         1, # return value if true
         0) # return value if false
```

0. Get Ready for the Analysis

- Look at the data

```
## make sure the variable was created correctly
head(star) # shows first observations
##   classtype reading math graduated small
## 1     small     578  610         1     1
## 2   regular     612  612         1     0
## 3   regular     583  606         1     0
## 4     small     661  648         1     1
## 5     small     614  636         1     1
## 6   regular     610  603         0     0
```

- The treatment variable is?
- The outcome variable is?
- What is the outcome's unit of measurement?

1. What Is the Estimated Average Treatment Effect?

- Fit a linear model so that the estimated slope coefficient is equivalent to the difference-in-means estimator.
- In this case, the fitted line is: $\widehat{math} = \widehat{\beta}_0 + \widehat{\beta}_1 small$
- R code?

```
mod <- lm(math ~ small, data = star) # fits linear model
mod # shows the contents of the object
##
## Call:
## lm(formula = math ~ small, data = star)
##
## Coefficients:
## (Intercept)        small
##      628.84         5.99
```

1. What Is the Estimated Average Treatment Effect?

- $\widehat{\beta}_1 = 5.99$
- Direction, size, and unit of measurement of the effect?
  - An increase of about 6 (or 5.99) points

1. What Is the Estimated Average Treatment Effect?

CONCLUSION STATEMENT

Assuming that [the treatment and control groups are comparable] (a reasonable assumption because ...), we estimate that [the treatment] [increases/decreases] [the outcome] by [size and unit of measurement of the effect], on average.

1. What Is the Estimated Average Treatment Effect?

- *Assuming that students who attended a small class were comparable to students who attended a regular-size class (a reasonable assumption because the data come from a randomized experiment), we estimate that attending a small class increases math test scores by about 6 points, on average.*

2. Is the Effect Statistically Significant?

- Is the average treatment effect statistically distinguishable from zero at the population level?

1. Specify null and alternative hypotheses

- H_0: $\beta_1 = 0$ (attending a small class has no average causal effect on math test scores at the population level)
- H_1: $\beta_1 \neq 0$ (attending a small class either increases or decreases math test scores at the population level)

2. Is the Effect Statistically Significant?

- R computes the coefficient and the p-value by running **summary()**:

```
summary(mod)
##
## Call:
## lm(formula = math ~ small, data = star)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -119.827  -27.585   -0.827   26.163  145.163
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  628.837      1.476  426.09  < 2e-16 ***
## small          5.990      2.178    2.75  0.00604 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.74 on 1272 degrees of freedom
## Multiple R-squared:  0.005911,   Adjusted R-squared:  0.00513
## F-statistic: 7.564 on 1 and 1272 DF,  p-value: 0.006039
```

2. Is the Effect Statistically Significant?

- Is the effect statistically significant at the 5% level?

```
summary(mod)
##
## Call:
## lm(formula = math ~ small, data = star)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -119.827  -27.585   -0.827   26.163  145.163
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  628.837      1.476  426.09  < 2e-16 ***
## small          5.990      2.178    2.75  0.00604 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.74 on 1272 degrees of freedom
## Multiple R-squared:  0.005911,   Adjusted R-squared:  0.00513
## F-statistic: 7.564 on 1 and 1272 DF,  p-value: 0.006039
```

Summary

- **Today's Class:**
  - Hypothesis Testing with Estimated Regression Coefficients
  - Example: *Do Small Classes Improve Math Scores*?
    - What Is the Estimated Average Treatment Effect?
    - Is the Effect *Statistically Significant*?
- **Next class**:
  - Do's and do not's in political methodology.

Questions?

See you in the next class!