Applied Statistical Analysis I

Hypothesis testing, experiments, difference in means

Elena Karagianni, PhD Candidate
karagiae@tcd.ie

October 1, 2025

Department of Political Science, Trinity College Dublin

Hypothesis Testing

Hypothesis Testing

- What is a hypothesis?

Hypothesis Testing

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

Hypothesis Testing

- A **hypothesis** is a statement about a population. It is a prediction that a parameter takes a particular numerical value or falls in a certain range of values. Agresti and Finlay, 2009, p.143

Hypothesis Testing

- Usually, we want to learn about population parameters based on sample statistics $\rightarrow$ statistical inference.
- For example: is a population parameter different from zero? If you find a value different from zero in your sample, can you be sure it's not just random variation?
- Formulate competing hypotheses:
  H_0: The population parameter equals zero (null hypothesis).
  H_A: The population parameter differs from zero (alternative hypothesis).

The 5 Steps of Hypothesis 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)

Figure 1: Source: Agresti and Finlay, 2009, p.147

Hypothesis Testing: Step 1

Assumptions Agresti and Finlay, 2009, p.144

- Type of data: continuous, categorical, etc.
- Sampling method: randomly obtained (e.g., random sample)
- Population distribution: assume a certain distribution (e.g., normal)
- Sample size: validity improves with larger samples

Hypothesis Testing: Step 2

State the Hypotheses

- Example: Someone claims the mean income of a country is 1400.
- We take a random sample of size $n = 1000$ and obtain a mean of 1350 with a standard deviation of 750.

Hypothesis Testing: Step 2

State the Hypotheses

- Example: Someone claims the mean income of a country is 1400.
- We take a random sample of size $n = 1000$ and obtain a mean of 1350 with a standard deviation of 750.
- Competing hypotheses:
  $H_0 : \mu = 1400 \quad \text{vs.} \quad H_A : \mu \neq 1400$

Hypothesis Testing: Step 2

State the Hypotheses

- Example: Someone claims the mean income of a country is 1400.
- We take a random sample of size $n = 1000$ and obtain a mean of 1350 with a standard deviation of 750.
- Competing hypotheses:
  $H_0 : \mu = 1400 \quad \text{vs.} \quad H_A : \mu \neq 1400$
- In other words: H_0 = the population mean **is equal** to the sample mean and H_A = the population mean **is not equal** to the sample mean.

Hypothesis Testing: Step 2

State the Hypotheses

- Example: Someone claims the mean income of a country is 1400.
- We take a random sample of size $n = 1000$ and obtain a mean of 1350 with a standard deviation of 750.
- Competing hypotheses:
  $H_0 : \mu = 1400 \quad \text{vs.} \quad H_A : \mu \neq 1400$
- In other words: H_0 = the population mean **is equal** to the sample mean and H_A = the population mean **is not equal** to the sample mean.
- The hypothesis can be one-sided (<, >, ≥, ≤) or two-sided (=, $\neq$).

Hypothesis Testing: Step 3

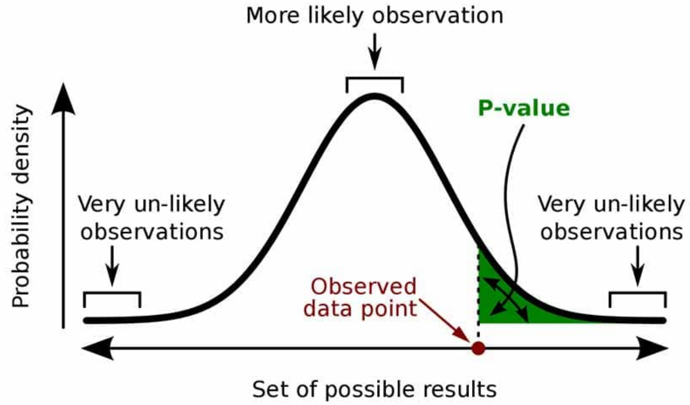
Test Statistic

- "The test statistic summarizes how far the 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, p.145
- Depending on the distribution of the test statistic, it is often referred to as a z-statistic (when the population standard deviation is known and n is large) or a t-statistic (when the n is small and/or the population variance is unknown).

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).

Step 4: p-value



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

Hypothesis Testing: Step 5

Conclusion

- Validate whether the obtained test statistic is unlikely to occure, under the assumption that the null hypothesis (H_0) is true $\rightarrow$ p-value, if probability is low, we reject H_0.
- Select an α-level, which indicates acceptable probability of Type 1 error (usually 0.05, 0.01).
- If p-value < α, we reject H_0 $\rightarrow$ proof by contradiction.
- Be careful when you are drawing a conclusion - there is still a probability that we falsely reject the null, even if p-value < α.

		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 = β)
	Reject	Type I Error (false positive) (probability = α)	Correct inference (true positive) (probability = $1-\beta$)

We are most concerned about Type I error: the probability of obtaining a false positive result.

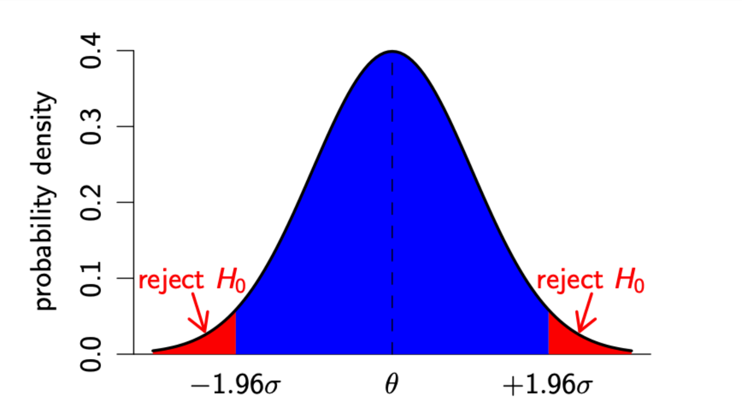
Calculating a Test Statistic

- Let's take our previous example: We take a random sample of size n = 1000 and obtain a mean of 1350 with a standard deviation of 750.
- At a level of 95% confidence, the critical z-values on a standard normal distribution are ±1.96
- How much is the sample mean (i.e., 1350) away from the hypothesized population mean (i.e., 1400) in repeated samples? Calculate the z-score of your realized sample:

$$z = \frac{\bar{x} - \mu_{pop}}{\sigma_{pop}/\sqrt{n}} = \frac{1350 - 1400}{\widehat{750}/\sqrt{1000}} = -2.1$$

- This is smaller than the critical value and you would reject H_0.

Calculating a Test Statistic



Decision Based on Confidence Intervals

- A $100(1-\alpha)\%$ **confidence interval (CI)** for a population mean is an interval that, in repeated random samples, would contain the true mean about $100(1-\alpha)\%$ of the time.
- For a two-sided test with significance level $\alpha = 0.05$, the CI is a **95% CI**.
- Our example:

$$\bar{x} \pm z_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}} = 1350 \pm 1.96 \cdot \frac{750}{\sqrt{1000}} = [1303, 1396].$$

- The hypothesized mean 1400 is not in this 95% CI $\Rightarrow$ reject H_0 at $\alpha = 0.05$.
- **Key idea:** For a two-sided test at level α,

$$\mu_0 \notin 100(1-\alpha)\% \text{ CI} \quad \Longleftrightarrow \quad p < \alpha.$$

Example

Significance test for a mean (t-test)

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

	Race		
Response	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' $\rightarrow \mu = 4$
H_A: Mean political ideology falls in liberal or conservative direction $\rightarrow \mu \neq 4$ ($\mu < 4$ or $\mu > 4$)

Significance test for a mean (t-test)

- With $\bar{x}$ = 4.075, s = 1.512, we can calculate the SE:

$$SE = \frac{s}{\sqrt{n}} = \frac{1.512}{\sqrt{186}} = 0.111$$

- And then the test statistic:

$$t = \frac{\bar{x} - \mu_0}{SE} = \frac{4.075 - 4}{0.111} = 0.68$$

- Degrees of freedom: $df = n - 1 = 185$.
- How to interpret this value?

Significance test for a mean (t-test)

- For a **two-sided** $\alpha = 0.05$ and df = 185, the critical value is : $t_{0.975,185} \approx 1.97$
- Our $t = 0.68$ is far inside $[-1.97, 1.97] \Rightarrow$ fail to reject H_0.
- p-value: $P(|T| > 0.68) \approx 0.50$ (from the t distribution).

Significance test for a mean (t-test)

- For a **two-sided** $\alpha = 0.05$ and df = 185, the critical value is : $t_{0.975,185} \approx 1.97$
- Our $t = 0.68$ is far inside $[-1.97, 1.97] \Rightarrow$ fail to reject H_0.
- p-value: $P(|T| > 0.68) \approx 0.50$ (from the t distribution).

Checking the CIs too:

$$4.075 \pm 1.97 \cdot 0.111 = [3.86,\ 4.29]$$

Since 4 is inside the 95% CI, result agrees: fail to reject H_0.

Significance test for a mean (t-test)

What is the conclusion?

p-value $\approx 0.50 > 0.05$, thus **we cannot reject the null (H_0)**. It is plausible that the population mean is 4, and therefore moderate.

Introduction to causality

Causal effect

What is a causal effect?

What is the fundamental problem of causal inference?

Key idea

A **causal effect** is the change in an outcome Y that would occur if we switched a unit from one condition to another.

Variables

- **Treatment** T**:**
  - $T = 1$ = treated
  - $T = 0$ = control
- **Outcome** Y**:** what we measure

Potential outcomes

$$Y_i(1) = \text{outcome if treated}$$

$$Y_i(0) = \text{outcome if control}$$

Individual causal effect

$$\tau_i = Y_i(1) - Y_i(0)$$

The difference between what *did* happen and what *would have happened* in the counterfactual world.

Causal effect

The fundamental problem with causal inference: “we can only observe, at most, one of the two quantities, for any individual at a particular point in time”
$\rightarrow$ **The causal effect is unobservable.**

(Bueno de Mesquita & Fowler, 2021, p.164)

SATE

What is the sample average treatment effect (SATE)?

What can we actually observe?

SATE

$$\frac{1}{n}\sum_{i=1}^{n}\{Y_i(1) - Y_i(0)\}$$

- It is the average of individual-level treatment effects in the sample.
- SATE is unobservable due to the fundamental problem of causal inference $\rightarrow$ we only observe **sample difference in means**.

SATE: Difference in Means

Naïve estimator (difference in means)

$$\hat{\Delta} = \bar{Y}_{T=1} - \bar{Y}_{T=0}$$

Difference in average outcome between treated and control groups.

Key point

- $\hat{\Delta}$ is an **unbiased estimate of SATE** only if treatment is as-if random.
- Otherwise it may be **biased by selection** (treated and control differ in unobserved ways).
- Correlation does not imply causation.

Bias

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

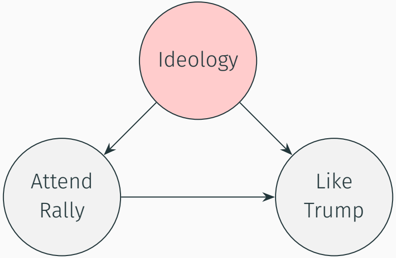
Bias

- **Baseline differences**: "...difference 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).

Bias

Confounders

- It has an effect on the treatment status.
- It also affects the potential outcome *beyond* its effect through treatment.



Other sources of bias

Bias

Other sources of bias

- Reverse causality
- Unobserved unit heterogeneity (special type of OVB)
- Post-treatment bias

Bias

How can we overcome bias?

Bias

How can we overcome bias?

1. **Control for confounders**:
   "there will still be unobservable confounders that we can't control for, reverse causation, or variables that are part confounder and part mechanism"

Bias

How can we overcome bias?

1. **Control for confounders**:
   "there will still be unobservable confounders that we can't control for, reverse causation, or variables that are part confounder and part mechanism"
2. **Randomized experiments**:
   Randomly assigning units to treatment and control groups eliminates baseline differences.
   "But, the ideal experiment that we'd like to run is often impractical, infeasible, or unethical".

Bias

How can we overcome bias?

1. **Control for confounders**:
   "there will still be unobservable confounders that we can't control for, reverse causation, or variables that are part confounder and part mechanism"
2. **Randomized experiments**:
   Randomly assigning units to treatment and control groups eliminates baseline differences.
   "But, the ideal experiment that we'd like to run is often impractical, infeasible, or unethical".
3. **Causal inference methods**: Difference-in-Differences design, Instrumental Variables etc.

(BdM and Fowler, 2021, 215-239)