Stat C131A: Statistical Methods for Data Science

Lecture 8: Hypothesis Testing I

Sep 23 2025

CLT: In practice

- The CLT states: the distribution of sample means approaches a normal distribution as sample size increases
  - Applies to independent, identically distributed (i.i.d.) samples with finite variance.
  - The limiting distribution is Normal(μ, σ^2/n), where μ and σ^2 are population mean and variance

- The **rate of convergence depends on the population distribution**:
  - Populations that are already symmetric (e.g., uniform, normal) → sample means look normal even for small n.
  - Populations with skew or heavy tails (exponential etc.) **need much larger n**
  - **Finite variance** is critical

- CLT does **not guarantee exact normality** for **small samples**
  - With $n < 30$, departures from normality may still be visible
  - Outliers have more influence in small samples, distorting the mean's distribution

- CLT concerns the distribution of **sample statistics**, **not populations**
  - The population itself may remain skewed, bimodal, or heavy-tailed
  - What becomes approximately normal is the sampling distribution of the mean

What is: a **statistical hypothesis**? What is: **hypothesis testing** ?

- A statistical hypothesis is a claim (statement) about a population parameter
  - Say, the mean, difference in means, etc.,
- Hypothesis testing is a clinical process of quantitatively ascertaining the validity of such a claim

Hypothesis types

- We deal with fundamentally **two types of hypotheses** in hypothesis testing
  - The **NULL Hypothesis**
  - The **ALTERNATE Hypothesis**

The **NULL** Hypothesis

- Is the postulate that your data is saying NOTHING, towards your claim
- Examples
  - There is **no evidence** that
    - The fitness app increases number of steps
    - Extra tutoring improves test scores
    - Dark chocolate reduces blood pressure
    - This experimental drug (chemical) trigger gene expression
    - ….
- Represents **status quo**
  - Aka **it's rejection** leads to us to **proving the ALTERNATE hypothesis** !

The Alternate Hypothesis

- Centered on what we DO want to prove !
  - The fitness app does indeed increases number of steps
  - Extra tutoring does indeed improve test scores
  - ….

The Hypothesis Testing Framework

Null Hypothesis	TRUE	FALSE
Reject	**Type I Error (α)**	**No Error**
Accept	**No Error**	**Type II Error (β)**

Type I Error (α)

- Happens when the null hypothesis is actually true, but we reject it
- **False positive** → we detect an effect/difference when there isn't one.
- Example: Concluding that the fitness app does increase walking steps, when it does not
- **Probability** of Type I error is α, which we set as the **significance level**
  - You can set it to what you like but the "default" is 0.05

Type II Error (β)

- Happens when the null hypothesis is false, but we fail to reject it
- False negative → we miss detecting a real effect/difference.
- Example: **Failing to conclude** that the fitness app does increases walking steps (when it actually does)
- **Probability** of Type II error is β

Significance (α)

- The **threshold for rejecting** H_0
- Common values: 0.05, 0.01.
- Setting a smaller α (e.g., 0.01) reduces Type I errors
  - But makes Type II errors more likely
    - WHY ?

Power $(1 - \beta)$

- Power is the **probability of correctly rejecting** H_0 when it is false
- Higher power = better chance of detecting a real effect
- Factors that increase power:
  - Larger sample size
  - Bigger effect size: say the difference in average steps between app users and non-app users is large
- Higher α (looser significance threshold)

Relationship Between α and β

- α and β are not directly complementary
- But there is a trade-off:
  - Smaller $\alpha \rightarrow$ more conservative test (fewer false positives), but higher chance of false negatives
  - Larger $\alpha \rightarrow$ easier to detect effects (higher power), but more risk of false positives
- In practice: researchers balance α and β by choosing a up front (e.g., 0.05) and designing sample size to achieve desired power (e.g., 0.8 → β = 0.2)

Population vs Sample

- Population parameter: true value (unknown).
- Sample statistic: computed from data; used to estimate the parameter
- Different random samples give different statistics.
- We assess how unusual our sample is if the hypothesis were true

Two ways to test

- **Permutation tests**
  - Based on:
    - Data-driven null distribution
    - Randomization
- **Parametric tests**
  - Based on:
    - Sample statistics
    - An assumed distributional form (Normal, etc.)
    - Population parameters
  - We formulate hypotheses based on population parameters

Permutation Test

- We do not assume Normality or any specific population distribution
- Steps
  - Pool all observations from both groups together.
  - Shuffle labels: Randomly reassign data points to “Group A” or “Group B,” as if the null hypothesis were true.
  - Recalculate the test statistic (e.g., difference in means) for each shuffle.
  - Build a reference distribution: Repeat shuffling many times (e.g., 10,000) to see what differences we’d expect by chance.
  - Compare the observed statistic to this reference distribution

Parametric Test

- **Assume** a population distribution (often Normal, or approximated by CLT).
- Frame hypotheses about population parameters (e.g., mean proportion)
- Compute a sample statistic (e.g., sample mean, difference in means).
- **Standardize** it into a **test statistic**
- Compare the test statistic against the theoretical distribution (Normal, etc.)
- Determine **the probability of observing a statistic as extreme** (or more) if the null hypothesis were true.

The **p-value**: the probability of observing that extreme

- What are the chances that this is what you would see "typically" / "usually"
- We start with the observed test statistic: x_{obs}
- The p-value asks, *if the nulll hypothesis is true then how likely is it to see a test statistic as large (or larger) than* x_{obs}
- **p-value**$(x_{obs}) = P_{H_0}(X \geq x_{obs})$
- Here
  - X = random variable for the test statistic under Ho
  - $P_{H_0}\left(X\right)$ = probability calculated assuming the null hypothesis is true
  - $X \geq x_{obs}$ = the "tail area", how extreme the observed result is (or more)

Permutation test: one- vs two-sided

- **One-sided/tail** p: proportion with stat >= observed (or <=)
- **Two-sided/tails** p: proportion with |stat| >= |observed|

Permutation test: Example

- **Question:** Do app users walk more steps per day than non-users?

- **Hypotheses:**
  - H_0 : the two groups have the same population distribution of steps;
  - H_1 : they differ (two-sided)

- Under H_0, labels are exchangeable: who is called "App" vs "Non-App" is arbitrary.

- Shuffle labels many times, recompute the difference in means each time, and see how extreme our observed difference is relative to this null distribution.

App usage: Permutation test

Daily Steps Bin	App Users (n=30)	Non-Users (n=30)
< 4,000	1	5
4,000–5,999	4	9
6,000–7,999	8	9
8,000–9,999	9	5
≥ 10,000	8	2

- Approx. means using bin midpoints: App ≈ 8,266 steps; Non-App ≈ 6,333 steps
- Observed difference (App − Non-App): Δ_obs ≈ 1,933 steps

Permutation test

- Test statistic: **difference in group means** (App − Non-App).
- **Procedure:**
  - Pool all 60 step counts.
  - Randomly shuffle the "App/Non-App" labels, split back into two groups of 30.
  - Compute Δ_perm for this shuffle.
  - Repeat many times (e.g., 10,000) to form the null distribution of Δ under H0.
- **Two-sided p-value** = fraction of shuffles with |Δ_perm| ≥ |Δ_obs|.
- Example outcome:
  - Out of 10,000 shuffles
  - 40 had |Δ_perm| ≥ 1,933 → p ≈ 0.004
- Interpretation: such a large mean difference would be very unlikely if H_0 were true → evidence that app users walk more on average.

Permutation tests: strengths & limitations

- Few assumptions; robust to skew/outliers.
- Exact for finite samples if all relabelings are used
- Requires exchangeability under H_0
- Can be slow if very large n

Hypothesis formulation: Example 1

- Question: Is the **average wait time** at Cheeseboard Collective Pizza > 15 minutes ?
  - μ = Population mean
  - Null Hypothesis: $H_0 : \mu = 15$
  - Alternate Hypothesis: $H_1 :$?

Hypothesis formulation: Example 1

- Is the **average wait time** at Cheeseboard Collective Pizza > 15 minutes ?

  - μ = Population mean
  - Null Hypothesis: $H_0 : \mu = 15$
  - Alternate Hypothesis: $H_1 : \mu > 15$

2025 startup rankings

	University	*Founder count*	*Company count*
1	**UC Berkeley**	**1,804**	**1,650**
2	**Stanford**	**1,519**	**1,380**
3	**Harvard**	**1,355**	**1,237**
4	**University of Pennsylvania**	**1,206**	**1,113**
5	**MIT**	**1,131**	**1,019**

Source: PitchBook (https://pitchbook.com/news/articles/pitchbook-university-rankings)

Hypothesis formulation: Example 2

- Question : Is the per quarter count of new startups emerging from Berkeley significantly higher than that of Stanford's ?

  - μ_B = The average number of new startups per quarter from Berkeley
  - μ_S = The average number of new startups per quarter from Stanford
  - Null Hypothesis: H_0 : ?
  - Alternate Hypothesis: H_1 : ?

Hypothesis formulation: Example 2

- Question : Is the per quarter count of new startups emerging from Berkeley significantly higher than that of Stanford's ?

  - μ_B = The average number of new startups per quarter from Berkeley
  - μ_S = The average number of new startups per quarter from Stanford
  - Null Hypothesis: $H_0 : \mu_B = \mu_S$
  - Alternate Hypothesis: H_1 : ?

Hypothesis formulation: Example 2

- Question : Is the per quarter count of new startups emerging from Berkeley significantly higher than that of Stanford's ?

  - μ_B = The average number of new startups per quarter from Berkeley
  - μ_S = The average number of new startups per quarter from Stanford
  - Null Hypothesis: $H_0 : \mu_B = \mu_S$
  - Alternate Hypothesis: $H_1 : \mu_B > \mu_S$