CSE 599K

Empirical Research Methods

Winter 2025

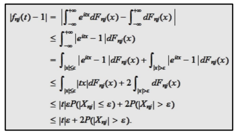
Statistical methods: overview

Today

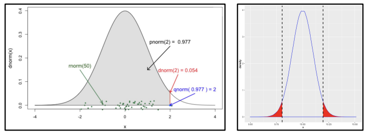
- Uniform vs. stratified sampling
- Statistical vs. practical significance
- Parametric vs non-parametric statistics

3 ways to understand and apply statistics

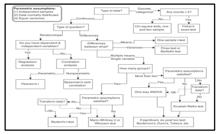
Math/Proofs



Simulations/Visualizations



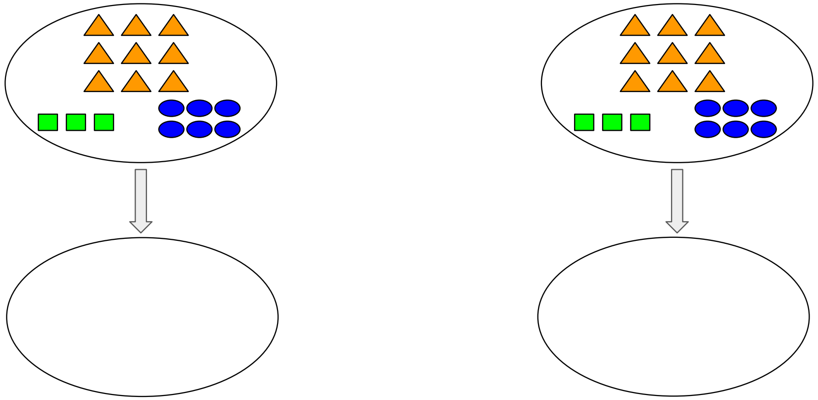
Decision diagrams



Uniform random vs. stratified random

Sampling: uniform random vs. stratified random

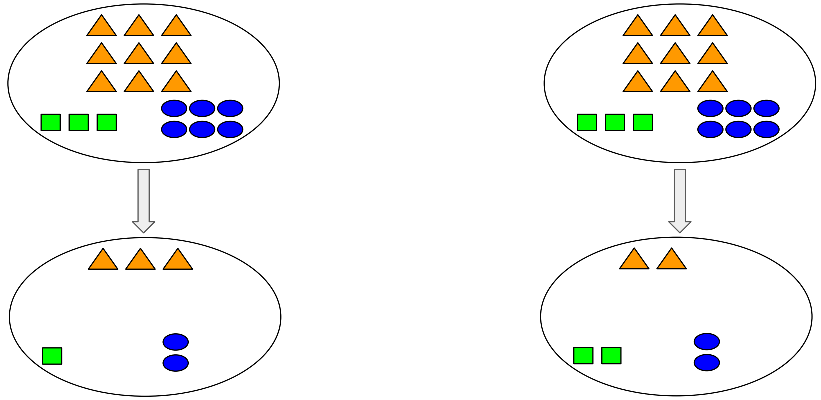
Uniform random | **Stratified random**



Sample six items: what are the expected outcomes?

Sampling: uniform random vs. stratified random

Uniform random | **Stratified random**



When would you use which sampling approach?

Statistical vs. practical significance

Statistical significance

Hypothetical study on system performance

- Compare normalized **throughput** of **two systems**.
- **Statistical test** for the **difference in mean throughput**.

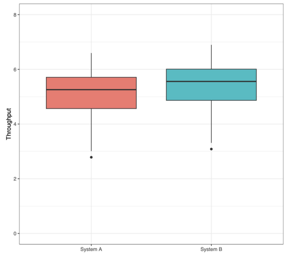
Statistical significance

Hypothetical study on system performance

- Compare normalized throughput of two systems.
- Statistical test for the difference in mean throughput.

Scenario 1: $p = 0.2137$

Scenario 2: $p < 0.05$ (~0.01)



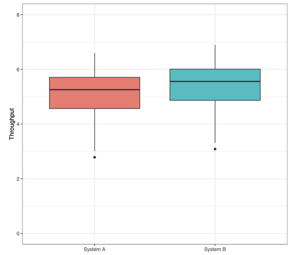
What plot do you expect for Scenario 2?

Statistical significance

Hypothetical study on system performance

- Compare normalized throughput of two systems.
- Statistical test for the difference in mean throughput.

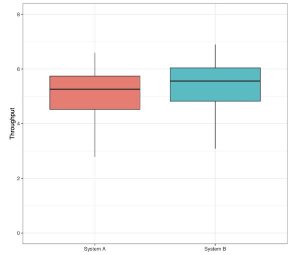
Scenario 1: $p = 0.2137$



N = 30

The p value is conflated with **sample size!**

Scenario 2: $p < 0.05$ (~0.01)



N = 120

A little quiz

1. What is the difference between statistical and practical significance?
2. What is the interpretation of the p value?
3. What is an effect size?

Small-group brainstorming

- Explain the answer to a group member.
- Come up with open questions.

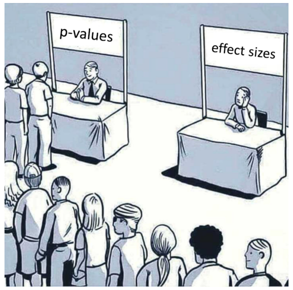
Statistical vs. practical significance

Statistical significance

- Is the difference due to chance?
- *p value*

Practical significance

- Does the difference matter in practice?
- *Effect size*



Effect size measures: examples

Correlation coefficients

- Pearson's r
- Kendall's tau (rank based)
- Spearman's rho (rank based)

"Raw" differences in central tendency

- Difference in means
- Difference in medians

Effect size measures: distinction

Distinction

- Parametric vs. non-parametric
  - Parametric: Pearson's r, Cohen's d
  - Non-parametric: Spearman's rho, A_{12}
- Standardized vs. non-standardized
  - Non-standardized: Difference in means Δ_M
  - Standardized: Δ_M divided by the (pooled) standard deviation
- Variable types
  - Continuous: Cohen's d, A_{12}
  - Ordinal: $A_{12,}$ Cliff's delta, Somers' D
  - Dichotomous: Odds ratio

Interpreting effect sizes

Example (Cohen's d):

- < 0.2: negligible
- >= 0.2: small
- >= 0.5: medium
- >= 0.8: large

Interpreting effect sizes: it's your job!

Example (Cohen's d):

- < 0.2: negligible
- >= 0.2: small
- >= 0.5: medium
- >= 0.8: large

(Standardized) effect sizes are a good starting point, but:

- Is an effect practically significant? Depends on context and domain!
- Raw differences may be easier to interpret (in context).

Generic effect sizes don't provide specific answers!

Contextualizing effect sizes

A statistically significant (large) effect may not be practically relevant:

- System response time: 20ms vs. 10ms
- Analysis runtime: 8h vs. 6h
- Top-5 vs. top-10 ranking
- Magnitude vs. location shift (superiority)

Parametric vs. non-parametric statistics

Parametric vs. non-parametric statistics

Parametric statistics

- Assumptions about the underlying distribution.
  Examples for common assumptions:
  - Normal distribution.
  - Equal variance.
- Parametric because of the reliance on distribution parameters.
- Example: Student's t-test, Welch's t-test.

Non-parametric statistics

- Fewer assumptions about the underlying distribution.
- Rank-based -> more robust to outliers.
- Example: Mann Whitney u test (Wilcoxon rank sum test).

Two common statistical tests

Student's/Welch's t test

- Assumes normality
- Hypothesis is related to equality of mean(s).

Mann Whitney u test

- Agnostic to the underlying distribution
- Hypothesis is related to location shift.

A little quiz

1. Why not always use non-parametric statistics (fewer assumptions)?
2. Is the following statement true?
   "If a parametric test is not significant, then a non-parametric test cannot be significant either due to less statistical power."
3. What conclusions can you draw from the Cohen's d vs. A_{12} effect sizes?

My new awesome system

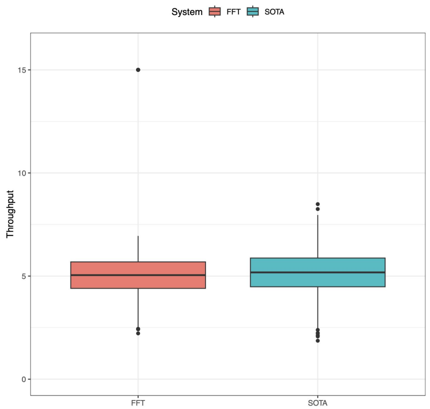
Evaluate system performance

- System: A new system (**A**) for fast file transfers: **FFT**.
- Goal: Compare the throughput against the state of the art (**B**): **SOTA**.

Results:

- ***Conclusion:*** FFT significantly outperforms SOTA:
  On average, its throughput of 5.29 files/ms -- a 2.3% increase over SOTA (5.17 files/ms).
- ***Statistical significance:*** The Mann Whitney U test showed that the difference is significant at the 0.05 significance level (p=0.0071).
- ***Practical significance***: While a relative increase of 2.3% may seem modest, we argue that this is a big achievement, given how optimized the state of the art is.

My new awesome system



Does this change your perception of the results?
What went wrong?

Statistical analysis: best practices

General advice:

- Be explicit about hypotheses and measures of interest (mean, median, location shift, proportions, etc.).
- Select appropriate statistical tests for a given hypothesis.
- Use data visualization to complement statistical tests.
- Be explicit about the effect size of interest.
- Contextualize effect size (requires domain knowledge).