CSE 503

Software Engineering

Winter 2021

Empirical Research

February 26, 2021

Recap: Kinds of conceptual variables

Dependent variable

- Outcome variable -- the measured response.

Independent variable

- Experimental variable -- systematically manipulated/controlled.

Covariate

- Experimental variable -- measurable but not controllable.

Recap: Study designs

Between subjects design

- Independent variable(s) take on exactly one value for each subject.

Within subjects design

- Independent variable(s) take on multiple/all possible values for each subject.
- Repeated measures design.

Mixed design

- A mixed design of between-subjects variables and within-subjects variables.

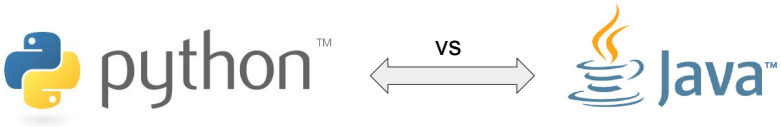
Today

- Experiment validity
- Sampling
- P-value and statistical significance
- Parametric vs. non-parametric statistics
- Effect size and practical significance
- Censored data

External, internal, and construct validity

External validity

- Does the experiment generalize (to larger population, other subjects, etc.)?
- How representative is the sample?



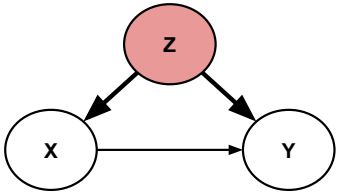
External, internal, and construct validity

External validity

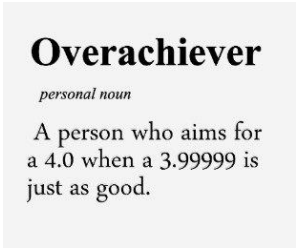
- Does the experiment generalize (to larger population, other subjects, etc.)?
- How representative is the sample?

Internal validity

- Does the experiment isolate the variable(s) of interest?
- Does the experiment control for confounders and unwanted effects?
- Be aware of **carry-over effects** (within-subjects designs)!
  - For example: order of tasks (subjects get accustomed to or tiered of a task).



External, internal, and construct validity



Construct validity

- Is the experiment adequately operationalized?
- Does the experiment use adequate proxy measures?
- Be aware of **interactions (being tested vs. treatment) and bias!**
  - For example: subjects may perform better/worse under test conditions.

External, internal, and construct validity

External validity

- Does the experiment generalize (to larger population, other subjects, etc.)?
- How representative is the sample?

Internal validity

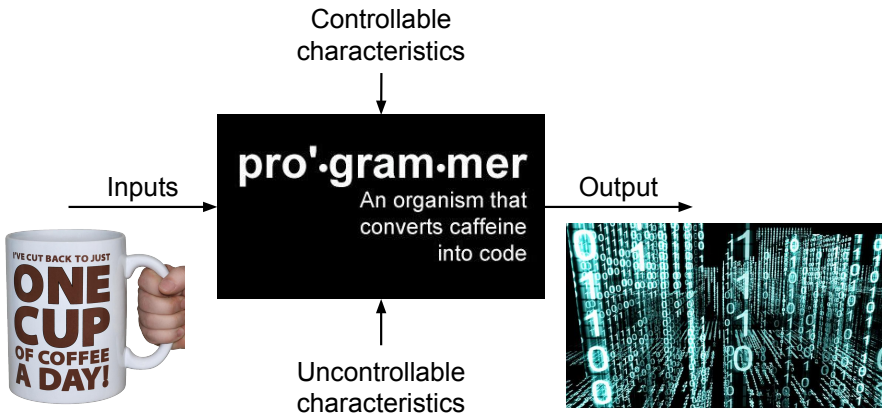
- Does the experiment isolate the variable(s) of interest?
- Does the experiment control for confounders and unwanted effects?

Construct validity

- Is the experiment adequately operationalized?
- Does the experiment use adequate proxy measures?

We also need to consider (statistical) conclusion validity.

Validity of an example experiment



Validity of an example experiment

pro'·gram·mer
An organism that converts caffeine into code

High-level research question:

Does coffee consumption improve programmer productivity and code quality?

Operationalization 1:

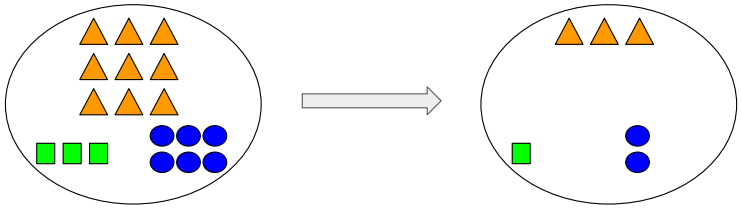
- 20 participants code for 20 weeks: on project 1 on Mondays with coffee; on project 2 on Fridays without coffee.
- Code quality: number of defects encountered in each project.
- Productivity: number of lines of code written.
- Coffee consumption: dollars spent on coffee (Monday receipts).

Operationalization 2:

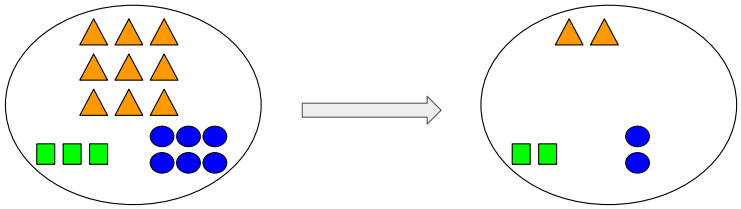
- 20 participants, randomly assigned to two groups of 10: each participant gets the same coding assignment.
- Code quality: number of defects encountered in the assignment.
- Productivity: number of lines of code written.
- Coffee consumption: Participants in group 1 get a free 64oz coffee.

Sampling: random vs. stratified random

Random

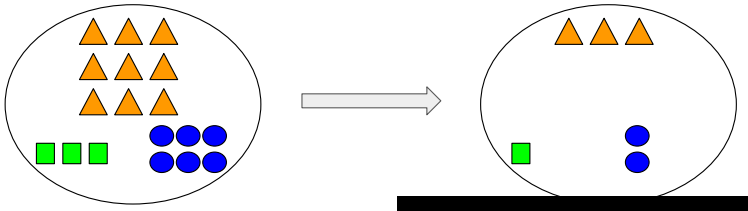


Stratified random

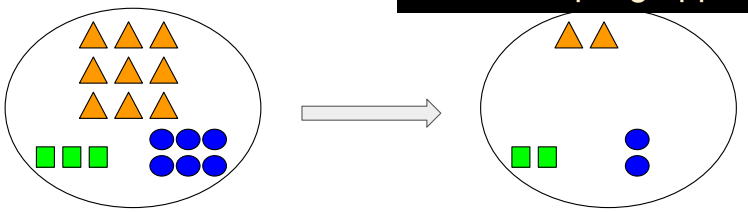


Sampling: random vs. stratified

Random



Stratified random



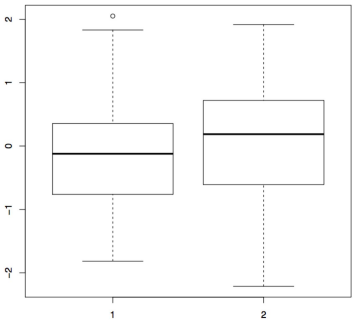
When would you use which sampling approach?

Statistical significance

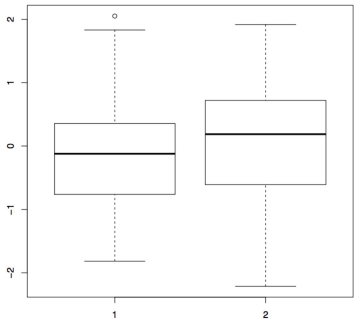
Hypothetical study on system performance

- Compare normalized runtime performance of two systems.
- Null hypothesis: No difference in mean runtime.

Scenario 1: p = 0.166



Scenario 2: p < 0.05 (~0.005)

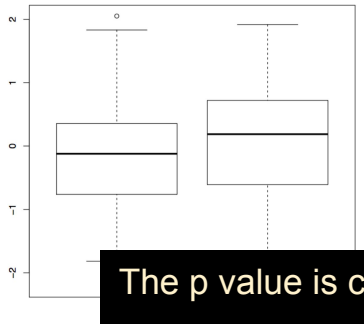


Statistical significance

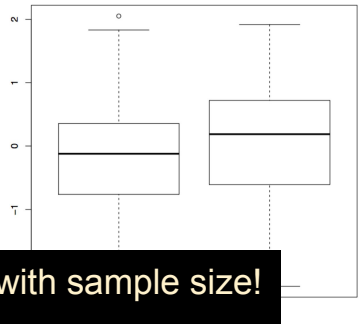
Hypothetical study on system performance

- Compare normalized runtime performance of two systems.
- Null hypothesis: No difference in mean runtime.

Scenario 1: p = 0.166
(n=50)



Scenario 2: p < 0.05 (~0.005)
(n=200)



The p value is conflated with sample size!

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.

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 overall (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: 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 a non-negligible 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!

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?

Contextualizing effect sizes

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

- System response time: 10ms vs. 20ms
- Analysis runtime: 45min vs. 1h
- Top-10 vs. overall precision
- Magnitude vs. location shift (superiority)

My new awesome system

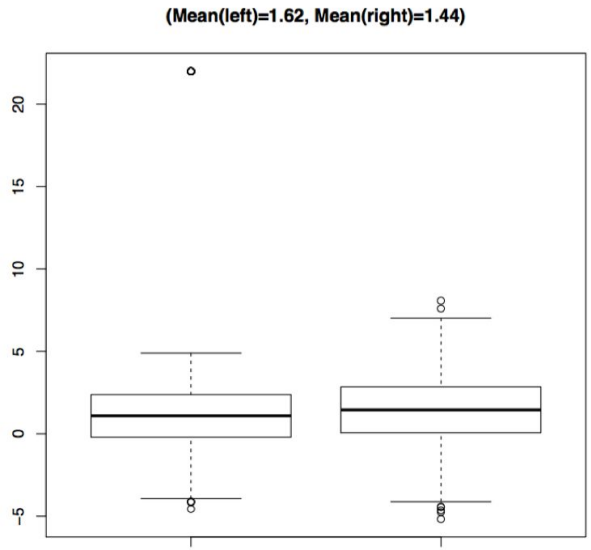
Evaluate system performance

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

Results:

- ***Conclusion****:* FFT significantly outperformed the state of the art: On average, it transferred 1.62 files per second -- a 12.5% increase over the state of the art, which only transferred 1.44 files per second.
- ***Statistical significance****:*The Mann Whitney U test showed that the difference is significant at the 0.05 significance level ($p<0.002$).
- ***Practical significance***: While a relative increase of 12.5% may seem modest, we argue that this is a big achievement, given how optimized state-of-the-art systems for fast file transfers are.

My new awesome system



Does this change your perception of the results?

Discussion

Which of the following are particularly relevant for your research area?

- Experiment validity
- Sampling
- Censored data
- Statistical vs. practical significance
- Choice of statistical tests and effect sizes