PS 211: Introduction to Experimental Design

Fall 2025 · Section C1

Lecture 18: Review, Non-Parametric Tests, P-Hacking, & Open Science

Updates & Reminders

- Homework 4 is due **today.**
  - Remember, only your top three homework grades count.
  - Answers will be posted tomorrow night.
- Exam 4 is next Tuesday, **Dec. 9**.
  - Review sheet will be posted today or tomorrow.
  - Focuses on Lectures 14 - 18.
  - **NO MAKE-UPS**
- Final office hours are this week!
- Course evaluations are open.

Review: Correlation and Regression

Significance Testing: Correlation & Regression

- Correlation and regression both assess linear relationships between variables.
- These relationships can be tested for **statistical significance.**
- *Null hypothesis:* There is no linear association between X and Y (correlation = 0; slope = 0).
- *Alternative hypothesis:* There is a linear association between X and Y (correlation ≠ 0; slope ≠ 0).
- For both, the test statistic depends on sample size (n) and effect size (r or slope).
- Larger samples → more power to detect smaller effects.
- Just like with all our inferential tests, **the p-value indicates probability of observing data if null is true.**

Relation between correlation & regression

- If there is a linear relationship between X and Y, **both** correlation and regression will capture it.
- If there is no linear relationship, both will indicate no association.
- The regression slope can be very, very small even if correlation is strong, depending on the scales of X and Y.
- Correlation quantifies strength of association.
- Regression quantifies change in Y per unit change in X.

Relation between ANOVA & regression

- ANOVA can be viewed as a special case of regression where the predictor(s) is/are categorical.
- Remember, ANOVA asks: Is the variance between group means larger than the variance within groups?
  - In regression terms, this is equivalent to asking: Does knowing the group membership (categorical predictor) help predict the outcome variable?
- Both ANOVA and regression use the same underlying principles of determining how much variance in the outcome can be explained by the predictors versus how much remains unexplained (error).
- Both can handle multiple predictors (factors in ANOVA; variables in regression), interactions, and mixed designs (combinations of within- and between-subjects factors/variables).

Review of Inferential Statistics: The Core Idea

Why do we need inferential statistics?

- We typically observe **samples**, not populations.
  - Remember, samples are just subsets of the larger population we care about.
- We want to make **inferences** about populations.
  - For example, when we're testing a treatment for depression, we want to know if it works for all people with depression, not just the people who happen to be in our study.

Variability in Samples

- Samples have **variability** due to many factors.
  - Sampling error: random variability between samples
  - Measurement error: imprecision in our measurements
  - Individual differences: true variability among people
  - **True population variability: differences in the population that we care about!**

Review of Inferential Statistics: The Core Idea (Continued)

- We use inferential statistics to evaluate whether the patterns we see in our sample data are likely to reflect true effects in the population, or if they could have arisen just by chance due to variability.
- We can think of all statistical tests as comparing **signal** to **noise**:
  - Signal: the meaningful effect we are trying to detect (e.g., difference between groups, association between variables)
  - Noise: the random variability that obscures the signal (e.g., individual differences, measurement error)
- The key question: **Is the signal strong enough relative to the noise to conclude that there is a real effect in the population?**
- Our test statistics quantify this signal-to-noise ratio.

Signal-to-Noise in Practice

Examples of how we quantify signal

- Group mean differences (ANOVA, t-tests)
- Strength of association (correlation, regression coefficients)

Examples of how we quantify noise

- Standard deviation
- Standard error
- Within-group variance
- Residual error in regression

Review: Standard Deviation and Standard Error

Standard Deviation (SD)

- One measure of the **spread of individual scores** (variability) in a sample/population
- High SD: lots of variability among individuals
- Low SD: individuals are similar to each other
- When there is more variability in the data, it is harder to detect true effects (signal) because the noise is larger.

Example: High variability in anxiety levels can make it more challenging to identify whether a treatment is effective, as the noise (individual differences) may obscure the signal (treatment effect).

Standard Error (SE)

- Equal to the SD of the sampling distribution of the mean.
- Describes **how much sample means vary** across repeated samples. Quantifies uncertainty in our estimate of the population mean.
- Answers the question: *If we took many samples from the same population, how much would their means differ from each other?*
- Formula:
  SE = SD / √n

Why Sample Size Matters

- Remember, test statistics depend on both effect size (signal) and variability (noise).
- **Sample size (n)** directly affects the standard error (SE):
  SE = SD / √n
- Larger samples reduce the SD of the sampling distribution (SE), making our estimates more precise.
- This reduces noise in our signal-to-noise ratio, and makes it easier to detect true effects.

Tests We've Covered

Z-test

Does our sample mean come from a population with a known mean?

- Population SD (σ) **must** be known (rare in practice).
- Uses the standard normal distribution.
- Example: testing if a new teaching method changes test scores compared to student scores from prior years with known population mean and SD.

Tests We've Covered

Single-sample *t*-test

Does our sample mean come from a population with a known mean?

- Population SD (σ) is unknown and must be estimated with sample SD (s).
- Uses *t* distribution (wider tails than *z* distribution).
- Example: testing if average sleep duration in a sample differs from the recommended 8 hours.

Tests We've Covered

Paired-samples *t*-test

Are the means of two related measurements different?

- Compare two measurements from the **same participants**.
- Uses **difference scores** (e.g., post-test - pre-test).
- Tests whether the mean difference is significantly different from zero.
- Uses *t* distribution.
- Example: testing if a training program improves fitness scores by comparing pre- and post-training scores.

Tests We've Covered

Independent-samples t-test

Are the means of two independent groups different?

- Compare means of **two separate groups**
- Tests whether the difference in means is significantly different from zero.
- Uses pooled estimate of variance from both groups.
- Uses *t* distribution.
- Example: testing if a new drug leads to lower blood pressure compared to a placebo.

Tests We've Covered

One-way Between-Subjects ANOVA

Are the means of 3 or more independent groups different?

- Compares **3+ independent groups**
- *F* statistic compares:
  - **Between-group variance** (signal)
  - **Within-group variance** (noise)
- To determine *which* means differ, post-hoc tests (Tukey, Bonferroni) are required. Example: testing if different teaching methods lead to different average test scores across multiple classes.

Tests We've Covered

One-way Repeated-Measures ANOVA

Are the means of 3 or more related measurements different?

- Compares **3+ conditions** in same participants.
- Accounts for within-subject variability (reduces noise).
- *F* statistic compares:
  - **Between-condition variance** (signal)
  - **Within-condition variance** (noise) after accounting for subject variability
- Example: testing if reaction times differ across three types of stimuli presented to the same participants.

Tests We've Covered

Factorial ANOVA

Are the means of groups defined by multiple independent variables different?

- Compares means across levels of **2+ independent variables (factors)**
- Tests:
  - Main effects
  - Interactions
- Interactions allow effects of one IV to depend on levels of another
- Example: testing effects of drug type and dosage on symptom reduction.

Tests We've Covered

Correlation

Are two continuous variables linearly related?

- Measures linear association between two continuous variables (X and Y)
- Quantified by **Pearson's *r***, which indicates:
  - Strength (magnitude) of association
  - Direction (positive/negative) of association
- Can test significance of correlation by converting *r* to a *t*-statistic.
  - Significance indicates whether the observed correlation is likely due to chance.
  - Depends on effect size (*r*) and sample size (n).

Tests We've Covered

Simple Linear Regression

Can we predict Y from X?

- Predicts continuous outcome variable (Y) from a single continuous predictor (X)
- Finds best-fit line: $Y = bX + a$ by minimizing squared errors between observed and predicted Y values.
- Produces: slope (b), intercept (a), and residuals (errors).
- Can examine standardized (β) or unstandardized (b) coefficients.
- Can assess model fit (R^2) and significance of slope.
- Example: predicting shark attacks based on water temperature.

Tests We've Covered

Multiple Regression

Can we predict Y from multiple Xs?

- Predicts continuous outcome variable (Y) from multiple continuous predictors (X1, X2, ...) simultaneously.
- Allows testing unique contributions of each predictor.
- Can include interaction terms to assess moderation effects.
- Produces coefficients for each predictor, overall model fit (R^2), and significance tests.
- There are methods that allow us to include both continuous and categorical predictors (dummy coding) *as well as* both between- and within-subjects variables (mixed-effects models).
- Example: predicting job performance from experience, education, and personality traits.

Assumptions of Parametric Tests

The tests we have discussed so far assume that:

- Data are randomly sampled
- Observations are independent
- Distribution of scores (or residuals) in the population is **approximately normal**
- Outcome variable is numeric
    - Interval or ratio scale
    - Not ordinal or categorical
- When assumptions hold: parametric tests are powerful.
- When badly violated: results can be biased.

Non-Parametric Tests: Overview

- Non-parametric tests do **not** assume normality or specific distributions.
- They are more **robust** to violations of assumptions.
- Often used with:
  - Small sample sizes
  - Ordinal or categorical data
  - Extreme outliers

Common Non-Parametric Tests

(do not memorize; just be aware of options)

Rank-based alternatives

- Mann–Whitney U (independent groups)
- Wilcoxon signed-rank (paired)
- Kruskal–Wallis (one-way ANOVA alt.)
- Friedman test (repeated-measures ANOVA alt.)
- Spearman's ρ (correlation alt.)

For categorical / binary outcomes

- Chi-square goodness-of-fit
- Chi-square test of independence
- Fisher's exact test
- McNemar's test (paired binary)

Example: Chi-Square (χ^2) Tests

- Used for **categorical** data.
- Equivalent to ANOVA / *t*-tests for categorical outcomes.
- Chi-square evaluates whether **observed** frequencies differ from **expected** frequencies.

Two main types:

- Goodness-of-fit (one categorical / nominal variable)
- Test of independence (two categorical / nominal variables)

Chi-Square Goodness-of-Fit Test

- Used when examining whether **one categorical variable** matches expected distribution.
- Compares observed frequencies in each category to expected frequencies based on a theoretical distribution (e.g., uniform distribution).
- Null hypothesis: observed frequencies match expected frequencies.
- If large discrepancies: reject null hypothesis.
- Null distribution: chi-square distribution with (k - 1) degrees of freedom (k = number of categories).
- Example: Testing whether the distribution of favorite Top 5 Spotify artists differs from uniform distribution.

R function: chisq.test()

Chi-Square Test of Independence

- Used when examining whether **two categorical variables** are associated.
- Null hypothesis: no association between variables.
- Alternative hypothesis: association exists.
- Example: Testing whether voting preference (Democrat, Republican, Independent) is associated with age group (18-29, 30-44, 45-60, 60+).

Limitations of Non-Parametric Tests

Disadvantages

- Lower power
- Cannot model complex designs easily
- Harder to interpret effect sizes
- Discard metric information

Advantages

- Fewer assumptions
- More robust to outliers
- Useful for ordinal or categorical variables
- Still allow valid inference in messy real-world data

Statistical Tests: Recap

- We have covered a variety of parametric inferential tests.
  - *z*-tests, *t*-tests, ANOVA, correlation, regression
- We have also discussed non-parametric alternatives.
- But it is also important to remember that selecting the appropriate statistical test is only part of conducting meaningful, ethical, and useful research.

Beyond Statistics: Research Ethics

- Ethical research practices are essential for credible, trustworthy science.
- Misuse of statistics can lead to misleading or false conclusions.
- Researchers have a responsibility to conduct and report research honestly and transparently.
- There are entire courses on research ethics. Today, we're just going to briefly cover some key issues related to data analysis and reporting.

P-Hacking & Researcher Degrees of Freedom

- **P-hacking**: manipulating data or analyses to achieve statistically significant results ($p < 0.05$).
- Examples:
  - **Cherry-picking**: selecting only certain data points or analyses that support a desired conclusion.
  - **Data dredging**: conducting multiple analyses without pre-specifying them, increasing the chance of finding false positives.
  - **Post-hoc hypothesizing**: creating hypotheses after seeing the data, rather than before.
- **Researcher degrees of freedom**: choices researchers make during data collection, analysis, and reporting that can influence results.
  - Examples: deciding when to stop data collection, which variables to include, how to handle outliers, etc.

P-Hacking & Researcher Degrees of Freedom: Example

- Imagine a researcher testing a new drug's effect on anxiety.
- They collect data from 20 participants and find $p = .08$ (not significant).
- They decide to:
  - Add 10 more participants → $p = .04$ (significant).
  - Exclude 2 outliers → $p = .03$ (more significant).
  - Test multiple outcome measures → find one with $p = .02$ (significant).
- They report only the final significant result, ignoring the initial non-significant findings.

Why is P-Hacking a Problem?

- Increases false positive rates (Type I errors).
- Undermines the credibility of research findings.
- Can lead to incorrect conclusions and wasted resources.
- Erodes public trust in science.
- Contributes to the "replication crisis" in psychology and other fields.

The replication crisis in psychology

- Many high-profile findings fail to replicate in subsequent studies, often due to questionable research practices like p-hacking.
- Some findings that have been difficult to replicate include:
  - Power posing effects (e.g., standing in a "powerful" pose increases confidence and risk-taking)
  - Priming effects (e.g., reading words related to old age makes people walk slower)
  - Neonatal imitation (e.g., newborns imitating facial expressions)
- These and other replication failures have led to increased scrutiny of research practices and calls for greater transparency and rigor in psychological science.

P-Hacking & Researcher Degrees of Freedom: Solutions

Some solutions:

- Pre-register studies to specify analyses in advance.
- Use robust statistical methods to reduce false positives.
  - Example: Adjust for multiple comparisons.
- Report effect sizes and confidence intervals, not just p-values.
- Report all analyses, including failed ones.
- Share data and code for transparency.

Open Science: What & Why

Open science includes:

- Pre-registration: Determine hypotheses, methods, and analyses before data collection.
- Registered reports: Peer review before data collection.
- Sharing data: Make data publicly available.
- Sharing code: Provide code for analyses.

Benefits:

- Transparency
- Reproducibility
- Builds cumulative knowledge
- Reduces questionable research practices

That’s all for today!

See you Thursday for Exam 4 review!