BOSTON UNIVERSITY

bu.edu

PS 211: Introduction to Experimental Design

Fall 2025 · Section C1

Exam 4 Review Session

Updates & Reminders

- Today is our last regular class session!
- Exam 4 is on **Tuesday, Dec. 9** during class.
- We will shortly post:
  - Homework 4 grades
  - Discussion section grades
- We will post Exam 4 grades and course grades on Tuesday afternoon *if* everyone takes the exam on time.

Review: ANOVA

- Remember, *t* tests compare **two means**.
- ANOVA allows us to compare **three or more means** while controlling the **Type I error rate**.
- A significant ANOVA tells us:
  - At least **one group mean differs** from the others
  - But NOT **which** groups differ
- To find out *which* means differ, we run **post-hoc tests** (e.g., Tukey's HSD, Bonferroni).

Review: One-Way ANOVA

- ANOVA partitions variability into:
  - **Between-groups variability** (signal)
  - **Within-groups variability** (noise)

Review: One-Way ANOVA: The F Statistic

F statistic:

$$F = \frac{MS_{\text{Between}}}{MS_{\text{Within}}}$$

Where:

- $SS_{\text{Between}} = \sum_i n_i(M_i - M_G)^2$
- $MS_{\text{Between}} = SS_{\text{Between}}/df_{\text{Between}}$
- $MS_{\text{Within}} = SS_{\text{Within}}/df_{\text{Within}}$

If F is large:

- Between-group variance is large relative to within-group variance.
- This is evidence that groups differ.

Visualizing ANOVA

Practice: One-Way ANOVA

If your between-group variability increases while within-group variability stays constant, what happens?

A. F increases

B. F decreases

C. F stays the same

D. We cannot determine this without knowing our degrees of freedom.

Answer: A. Increasing “signal” increases F.

> We don't know how much F increases without knowing the exact values, but we can say it will increase.

Practice: Interpreting F

Suppose $F(2, 45) = 0.89$, $p = .42$. What's the correct conclusion?

A. At least one mean is different from the other two.

B. We did not find significant differences between means.

C. Post-hoc tests will reveal differences between groups.

D. The critical value was miscalculated; F cannot be less than 1.

Answer: B Because $p = .42$, we fail to reject the null hypothesis, meaning we did not find significant differences between group means.

> Note: Even though *F* tests are always one-tailed, *F* **can** be less than 1 when within-group variability exceeds between-group variability.

Post-Hoc Tests (Tukey & Bonferroni)

Key rule: When the **overall ANOVA is significant**, we can run post-hoc tests to determine which means differ.

Post-hoc tests control for **Type I error** across multiple comparisons.

Tukey's HSD

- Designed for comparing **all possible pairs** of means.
- Controls overall Type I error.
- Balances power and error control ("safety").

Bonferroni correction

- Adjusts α by: $\alpha_{\text{new}} = \alpha / \text{number of tests}$
- Very conservative → protects Type I error at cost of power.

Practice: When to Use Post-Hocs

When are post-hoc tests appropriate?

A. Whenever there are ≥ 3 groups being compared.

B. After a significant ANOVA.

C. Only when within-group variances differ.

D. When means “look different."

Answer: B Post-hoc tests should only be run after a significant overall ANOVA.

Effect Size for ANOVA: η^2

Eta-squared:

$$\eta^2 = \frac{SS_{\text{Between}}}{SS_{\text{Total}}}$$

Interpretation:

- .01 = small
- .06 = medium
- .14 = large

Represents the **proportion of variance explained** by the IV.

Practice: F and Effect Size

Which scenario gives the **largest F**?

A. Large between-group differences + small within-group variability

B. Small between-group differences + large within-group variability

C. Small between-group differences + small within-group variability

D. Identical means across groups with different sample sizes

Answer: A. Large between-group differences increase the numerator ($MS_{Between}$, the *signal*), and small within-group variability decreases the denominator (MS_{Within}, the *noise*), resulting in a larger F statistic.

Repeated-Measures ANOVA

Used when:

- Same participants measured across **multiple conditions** or time points.

Advantages:

- Higher power
- Fewer participants needed
- Controls for between-subject variability

Practice: Repeated-Measures ANOVA

Why does a repeated-measures ANOVA have more power than a between-subjects ANOVA?

A. Repeated-measures ANOVA uses larger sample sizes.

B. Repeated-measures ANOVA removes between-person variability.

C. Repeated-measures ANOVA typically is run with a larger α.

D. Repeated-measures ANOVA has a larger effect size by definition.

Answer: B Repeated-measures ANOVA removes between-person variability. The "noise" term, MS_{Within}, is smaller because the subject-specific variability is accounted for, leading to a larger *F* statistic.

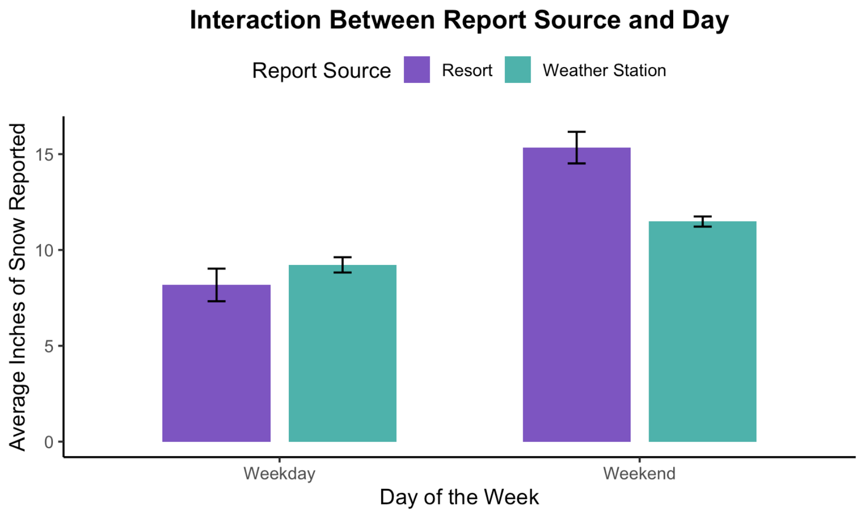
Factorial ANOVA (Two-Way)

- Factorial ANOVAs test effects of **two or more IVs** simultaneously.
- Each IV is called a **factor**.
- Factors can have **multiple levels**.

A factorial ANOVA allows testing:

1. **Main effect of A**
2. **Main effect of B**
3. **Interaction effect**
    - Does the effect of A depend on B?

Visualizing Main Effects and Interactions



Practice: Main Effects & Interactions

A researcher examines the influence of price (cheap vs. expensive) and type (red vs. white) on wine ratings. She runs a 2x2 ANOVA and finds:

- Main effect of price: $p = .001$
- Main effect of type: $p = .70$
- Interaction effect: $p = .02$

Which of the following conclusions are correct?

A. Participants rated cheap wines differently than expensive wines.

B. Participants rated red and white wines differently.

C. The effect of price on wine ratings depended on type of wine.

D. None of the above.

Answers: A and C

- A is correct because the main effect of price is significant ($p = .001$).
- C is correct because the interaction effect is significant ($p = .02$).

Correlation: Key Ideas

Correlation assesses **linear relationship** between two continuous variables.

Pearson's *r*:

- Range: –1 to +1
- Sign: direction of relationship
  - Positive → as X increases, Y increases
  - Negative → as X increases, Y decreases
- Magnitude: strength of relationship
  - Larger absolute value = stronger relationship

Remember, correlation does NOT imply causation!

Interpreting *r*

Guidelines:

- .10 small
- .30 medium
- .50 large

Coefficient of determination:
r^2 = proportion of variance in Y explained by X.

Practice: Correlation Strength

A researcher finds a correlation between driving experience (years) and accident rate (number of accidents) of r = –0.62. Which of the following conclusions are correct?

A. There is a strong negative relationship between driving experience and accident rate.

B. As driving experience increases, accident rate decreases.

C. Driving experience causes accident rates to decrease.

D. About 38% of the variance in accident rate is explained by driving experience.

E. About 62% of the variance in accident rate is NOT explained by driving experience.

Answers: A, B, D, and E

> Note: E is correct because $r^2 = (-0.62)^2 = 0.3844$, so approximately 38% of the variance is explained, meaning about 62% is not explained. The fact that it is exactly .62 is coincidental.

Practice: Third Variables

Shoe size and vocabulary correlate in children, $r = .80$. However, when controlling for age, the *partial* correlation drops to $r = .10$. What does this suggest?

A. Increases in shoe size causes children to walk earlier, which leads to faster language development and vocabulary growth.

B. There is a direct causal relationship between shoe size and vocabulary, but it is small ($r = .10$).

C. The original correlation between shoe size and vocabulary was largely due to the influence of a third variable, age.

D. Age explains all of the variance in children's vocabularies.

Answer: C The large drop in correlation when controlling for age suggests that age was a confounding variable influencing both shoe size and vocabulary.

Regression: Key Concepts

- Regression analyzes the relationship between a dependent variable (the **outcome**) and one or more independent variables (the **predictors**).
- Simple linear regression involves one predictor variable.
- Simple linear regression involves finding the **line of best fit** that minimizes the sum of squared differences between observed and predicted values:

$$\hat{Y} = bX + aX$$

Where:

- b is the slope = expected change in Y for each unit X
- a is the intercept = predicted Y when X = 0

Practice: R^2 Interpretation

If a model explains 30% of the variance in the outcome variable, what is the value of R^2?:

A. $R^2 = .03$

B. $R^2 = .30$

C. $R^2 = .30^2 = .09$

D. R^2 cannot be computed for a regression model.

Answer: B

APA Reporting: Correlation & Regression

Correlation

> There was a significant correlation between age and vocabulary, *r*(28) = .45, *p* = .01.

Regression

> A linear regression revealed that increases in hours studied were related to increases in quiz scores, *F*(1, 28) = 9.21, *p* = .005, $R^2 = .25$.

In both of these cases, how many participants were in the study?

Answer: 30 participants.

In correlation, degrees of freedom = *n* – 2 Where:

- *n* = number of participants

In regression, degrees of freedom = *n* – *k* – 1. Where:

- n = number of participants
- k = number of predictors

Chi-Square Tests

Used for **categorical variables**.

Goodness-of-fit

- Compare observed vs. expected frequencies

Test of independence

- Assess whether two categorical variables are related

Assumptions of both:

- Frequencies (not means!) should be used
- Expected counts ≥ 5 in most cells (otherwise, results may be invalid)

Practice: Chi-Square Scenarios

Which scenario requires a chi-square test?

A. Examining GPA vs. average hours slept each night.

B. Examining students' happiness scores across three different university majors.

C. Examining whether people who own cats vs. dogs differ in their favorite movie genres.

D. Examining the relationship between income and education level.

Answer: C

Which chi-square test should be used?

Answer: Chi-square test of independence

P-Hacking and Researcher Degrees of Freedom

Questionable research practices (QRPs)

- **p-hacking:**
  - Trying multiple analyses and reporting only significant results.
- **HARKing:** (hypothesizing after results known)
  - Presenting post-hoc hypotheses as if they were a priori.
- **Selective reporting:**
  - Reporting only significant conditions, measures, or studies.
- **Optional stopping:**
  - Collecting data until significant results are found.

Why are these harmful?

- Inflate Type I error:
  - Increase false positives
- Undermine replicability:
  - Results may not be consistent across studies
- Mislead literature:
  - Create a false impression of support for a theory

Open Science Practices

- Pre-registration of hypotheses & analysis plans
- Sharing materials, data, & code
- Transparent reporting

These improve **trust**, **rigor**, and **replicability** of scientific findings.

Practice: Questionable research practices

Testing 12 outcomes and reporting only the significant one is an example of:

A. Open science.

B. Exploratory data analysis.

C. *P*-hacking and selective reporting.

D. Data fraud.

Answer: C

- Testing multiple outcomes and reporting only significant ones is a form of *p*-hacking and selective reporting, which inflates the Type I error rate.

Practice: Open Science

Why is pre-registration useful?

A. It helps to prevent *p*-hacking and limits researcher degrees of freedom.

B. It allows researchers to stop data collection once they achieve significance.

C. It increases the likelihood of finding significant results.

D. It eliminates the need for peer review.

Answer: A

- Pre-registration helps prevent *p*-hacking by specifying hypotheses and analysis plans in advance, reducing researcher degrees of freedom.

Thank you!

- We made it through the semester! (Almost.)
- I hope that you learned something and didn't find this course too painful.
- This was my first time ever teaching and I learned a lot.
- I am looking forward to improving this course in the future.
- I really appreciate all of your engagement and effort, especially those of you who consistently participated during class.

Thank you also to:

- Juneau, who was an amazing TF!
- Chloe Jordan, who gave me all her course materials.

That's all for the semester!

Except: Course evaluations.