


# Quantitative Methods (Day 5)



CMSC 33231 - Diana Franklin



# Goals for today

---

Look at the process for performing quantitative research

What types of research questions does quantitative analysis answer?

What were the data sources?

How did they set up the study to collect appropriate data?

How did they analyze it?

How do we read quantitative results?

# Papers read

---

Language design (syntax)

<https://dl.acm.org/doi/pdf/10.1145/2089155.2089159>

Peer Feedback

<https://dl.acm.org/doi/10.1145/2445196.2445250>

Supporting diverse learners with TIPP&SEE

<https://dl.acm.org/doi/10.1145/3408877.3432366>

# ANOVA

---

An ANOVA test is a type of statistical test used to determine if there is a statistically significant difference between two or more categorical groups by testing for differences of means using variance.

The assumptions of the ANOVA test are the same as the general assumptions for any parametric test:

1. An ANOVA can only be conducted if there is **no relationship between the subjects** in each sample. This means that subjects in the first group cannot also be in the second group (e.g. independent samples/between-groups).
2. The different groups/levels must have **equal sample sizes**.
3. An ANOVA can only be conducted if the dependent variable is **normally distributed**, so that the middle scores are most frequent and extreme scores are least frequent.
4. Population variances must be equal (i.e. homoscedastic). **Homogeneity of variance** means that the deviation of scores (measured by the range or standard deviation for example) is similar between populations.

# One-Way ANOVA

---

A one-way ANOVA (analysis of variance) has one categorical independent variable (also known as a factor) and a normally distributed continuous (i.e., interval or ratio level) dependent variable.

The independent variable divides cases into two or more mutually exclusive levels, categories, or groups.

Useful for testing an intervention - did it work? How well did it work?

# Two-Way ANOVA

---

A two-way ANOVA (analysis of variance) has two or more categorical independent variables (also known as a factor), and a normally distributed continuous (i.e., interval or ratio level) dependent variable.

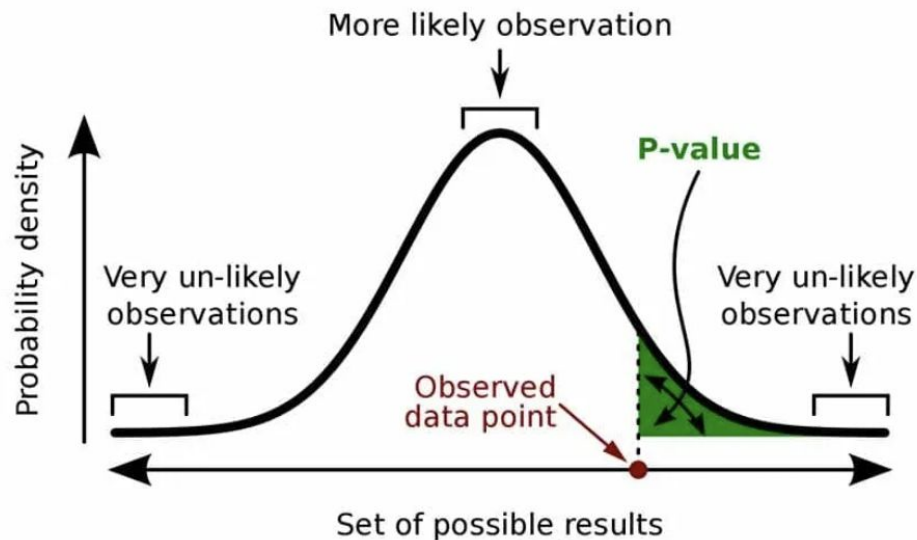
The independent variables divide cases into two or more mutually exclusive levels, categories, or groups. A two-way ANOVA is also called a factorial ANOVA.

Useful for finding out what factor is dominant in causing a certain result.

An example of a factorial ANOVAs include testing the effects of social contact (high, medium, low), job status (employed, self-employed, unemployed, retired), and family history (no family history, some family history) on the incidence of depression in a population.

# Reading quantitative results

<https://www.simplypsychology.org/p-value.html>



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

# Reading quantitative results: p values

---

A p-value, or probability value, is a number describing how likely it is that your data would have occurred by random chance (i.e. that the null hypothesis is true).

- P value is between 0 and 1
- The smaller the p value, the more likely it is significant
- A *p*-value less than 0.05 (typically  $\leq 0.05$ ) is statistically significant.

When reporting *p* values, report exact *p* values (e.g.,  $p = .031$ ) to two or three decimal places. However, report *p* values less than .001 as  $p < .001$ . The tradition of reporting *p* values in the form  $p < .10$ ,  $p < .05$ ,  $p < .01$ , and so forth, was appropriate in a time when only limited tables of critical values were available.

To understand the strength of the difference between two groups (control vs. experimental) a researcher needs to calculate the **effect size**.



# Reading quantitative results: effect size

Effect size is a quantitative measure of the magnitude of the experimental effect. The larger the effect size the stronger the relationship between two variables

Cohen's d:

$$\text{Effect Size} = \frac{[\text{Mean of experimental group}] - [\text{Mean of control group}]}{\text{Standard Deviation}}$$

| Relative size | Effect size | % of control group below the mean of experimental group |
|---------------|-------------|---|
|               | 0.0         | 50%   |
| Small         | 0.2         | 58%   |
| Medium        | 0.5         | 69%   |
| Large         | 0.8         | 79%   |
|               | 1.4         | 92%   |

# Language Design

---

Interesting design considerations for Quorum

- Screen readers should read it well

- Syntax should be easy for users

What studies did they use to design it?

- Studies on blind students for screen reading ease

- Studies (surveys) on sighted students for keywords

Stefik is groundbreaking in his use of empirical user studies for language design

# Comparison study: Study design

---

What activities did the students do?

Programming exercises -  $\frac{1}{2}$  with examples,  $\frac{1}{2}$  without

How were the results evaluated?

Rubric - divided the problems up and marked correct / incorrect - 0 vs 1

How did they make sure evaluations were consistent?

Everyone grades the same problems and calculate Inter-rater reliability

80% is the threshold for being allowed to grade independently

# Comparison study: Results

- A  $p$ -value less than 0.05 (typically  $\leq 0.05$ ) is statistically significant.
- 

How did they determine whether or not learning occurred?

Pre- post- for the same subjects

Compare performance on early activities vs late activities

How did they determine whether or not there was a difference between languages?

First, calculate the  $p$  value to determine anything significant occurred

Second, post-hoc Tukey, calculated pair-wise

# Peer Feedback: Constructed experiment

---

Does peer instruction improve learning?

- Randomly assign students to a control and treatment group

- Taught by the same teacher

- Consistent assessments, assignments, time on material

- At the same time

- Pre- and post- assessments to ensure learning

# Peer Feedback

---

What was their data?

Grades for many courses over 10 years

In what ways does it diverge from the perfect experiment?

Different instructors

Different years

Different materials, assignments, assessments

Students not randomly assigned

# Lots of subtests to address variables

---

Is it just different instructors?

- Same instructor, different groups

Different years (may students got better)?

- Looked at CS X over the years to look at trends

- Only partly works because only students who passed move on

- Students self select into the more advanced course

- Dropout rate over the years

- Should have done SI over the years!!!

Two-way ANOVA

- It may not have met the criteria for Two-way ANOVA

If you expect grades to go up over time, more appropriate to do a regression to answer to what degree years impacted the grades.

# Diversity and TIPPnSEE

---

Was a controlled experiment with control & treatment, teachers randomly assigned to control vs treatment

Use a multi-level model to look at students within classrooms and address that teachers are also different

- One level, see how classrooms differ

- Next level, take into account student factors within the classroom



# Differences in graphs

---

Bar graphs - show average

Box & whisker graphs - show mean, standard deviation

---

What statistical test do you use if you want to find a correlation between a continuous or ordinal data (time spent on an enrichment learning activity) and performance (final grade or assessment)?

Regression

# Let's try it

---

Partner with someone and focus on one of your topics

Develop 2-3 quantitative research questions around that topic (you'll probably need to get pretty specific)

# Let's try it

---

Choose a data collection source and design an activity to collect data on one quantitative RQ.