CSE 593

Final Exam Review

Farnaz Jahanbakhsh

Logistics

- Poster presentation in Tishman Hall Dec 5th at 3pm
- Assignment 5 group & Final Document and Poster due Dec 4th at 5pm
- Quiz 6 assigned Thursday, due Friday at 5pm
- Course evaluation assignment due Thursday (free 0.5 extra credit!)
  - Can potentially change your letter grade
- Poll on Piazza about snacks for the poster presentation
- Don't forget about the final exam!

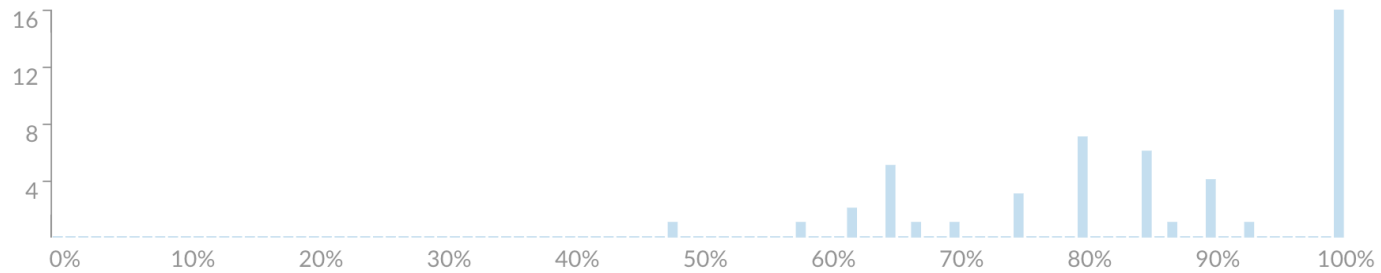# Quiz 5 review

μ Average Score	↗ High Score	↘ Low Score	σ Standard Deviation	🕐 Average Time
84%	100%	48%	14.23	04:10:04

Effect size

- Magnitude of a difference/relationship
- Independent of sample size
- Measure of effect size depends on the statistical test
- For certain effect size measures, rules of thumb thresholds exist
- E.g., for Cohen's d:

  0.2 small
  0.5 medium
  0.7 large

- These are **not** hard and fast rules!

- Significance level (alpha): probability of rejecting the null hypothesis when it is in fact true
- P value: probability of observing the data/difference from the experiment given that the null hypothesis is true
- Suppose alpha = 0.05. You run a test and p = 0.1
  - Probability of observing this difference when in reality there is no difference is 10%
  - But we said we only accept a 5% chance of this happening
  - Can't conclude the results are statistically significant

Q1

Attempts: 49 out of 49

A researcher wanted to compare two designs and set the significance level (alpha) to 0.05. After conducting the study, they found a Cohen's d effect size of 0.5 (a medium effect size) and a p-value of 0.1. What can the researcher conclude about the study? Select all that apply.

The medium effect size suggests there may be a meaningful difference between the designs, even though it was not statistically significant.	46 respondents	94 %	✓	90% answered correctly
A medium effect size with a p-value above 0.05 suggests the study may lack sufficient power to detect a significant difference.	46 respondents	94 %	✓	
The result is not statistically significant at the 0.05 level, so the researcher cannot conclude there is a difference between the designs.	48 respondents	98 %	✓	
The p-value of 0.1 is close enough to the alpha level that the researcher should consider the result statistically significant.		0 %		
The researcher can conclude that there is no difference between the designs since the result is not statistically significant.		0 %		

Determining sample size

- Sample size, alpha, power, and effect size have a relationship
- Power: probability of rejecting the null hypothesis when it is in fact false ($1 - \beta$)
- Effect size: the minimum effect size you aim to detect with sufficient statistical power
- To determine effect size:
    - Use rules of thumb (if they exist and make sense for your study)
    - Run a pilot study
    - Find a similar study and use their effect size for guidance

Q2

Attempts: 49 out of 49

+0.36

Discrimination Index ⑦

In an HCI study, a researcher tests whether a new interface design significantly improves task completion time compared to an existing design. The researcher plots the data and observes that the confidence intervals around the means for the two designs do not overlap. What conclusion can the researcher draw?

92% answered correctly

Nonoverlapping confidence intervals prove that there is a statistically significant difference in task completion time between the two designs.	4 respondents	8 %	▌
Nonoverlapping confidence intervals mean that the new design is significantly better in terms of task completion time.		0 %	▏
Nonoverlapping confidence intervals suggest there may be a statistically significant difference between the two designs, but a formal statistical test is still needed to confirm this.	45 respondents	92 %	████████ ✓
Nonoverlapping confidence intervals indicate that the task completion times for the two designs are equal.		0 %	▏

Q3

-0

Discrimination Index ⓘ

In a usability study, researchers compare two versions of a mobile app interface to see if one leads to faster task completion times. The study yields a statistically significant difference in task completion times between the two versions. What additional information should the researchers consider before recommending the new interface?

100% answered correctly

The p-value is sufficient, as statistical significance confirms that the new interface is objectively better.		0 %	▮
The effect size, as it indicates whether the observed difference is large enough to be meaningful in a real-world setting	49 respondents	**100 %**	✓
The mean difference, as a higher mean difference always implies a practical improvement.		0 %	▮
The total number of participants, as more participants would guarantee that the result is important in real life.		0 %	▮

10

Effect size (again)

- As sample size grows, small differences become statistically significant

- Why? Because tests compare the observed difference to the variability in the data

- With larger samples, variability shrinks

- Test *sensitivity* increases in larger sample sizes

Q4

+0.36

Discrimination Index ⑦

A researcher is investigating how screen brightness and notification frequency affect users' ability to focus on a computer-based task. Each participant completes a concentration task under different combinations of screen brightness (low, medium, high) and notification frequency (no notifications, infrequent, frequent). Participants experience all combinations of these conditions, and their concentration levels are measured after each task. What type of experimental design is being used?

Within-subjects, repeated measures	48 respondents	**98** %		✓
Between-subjects, full factorial	1 respondent	2 %		
Between-subjects, partial factorial		0 %		
Mixed design, partial factorial		0 %		

98% answered correctly

Q5

+0.68

Discrimination Index ⑦

Consider the following scenario.

Suppose a designer wants to show that their new keyboard (called maxi-QWERTY) that they have designed has a lower text entry error rate than two existing baselines:. In their study, each participant tries all three keyboard types: 1) mini-QWERTY keyboard where the participant can only enter text by pressing on individual keyboard keys, and 2) mini-QWERTY keyboard where the participant can only swipe to enter text.

What statistical analysis test should the designer pick to evaluate their design? Select **the best** answer. (**Hint:** consult Yatani's table; link was provided in the lecture.)

Just two paired t-test, each to compare the intervention with one of the baselines.	1 respondent	2 %	▮
Friedman test, followed by pairwise comparison using Willcoxon tests.	33 respondents	**67** %	▬▬▬▬ ✓
Repeated-measures ANOVA, followed by pairwise comparison using paired t-tests.	15 respondents	31 %	▬▬
The designer should not use a statistical analysis test because the best study design for evaluating their keyboard will not have enough participants to reach statistically significant results.		0 %	▮

67% answered correctly

13

Interval and ratio: equal intervals between values
Intervals don't have a true 0 point. Ratios do (0 means none).

	Interval/Ratio (Normality assumed)	Interval/Ratio (Normality not assumed), Ordinal	Dichotomy (Binomial)
Compare two unpaired groups	Unpaired t test	Mann-Whitney test	Fisher's test
Compare two paired groups	Paired t test	Wilcoxon test	McNemar's test
Compare more than two unmatched groups	ANOVA	Kruskal-Wallis test	Chi-square test
Compare more than two matched groups	Repeated-measures ANOVA	Friedman test	Cochran's Q test
Find relationship between two variables	Pearson correlation	Spearman correlation	Cramer's V
Predict a value with one independent variable	Linear/Non-linear regression	Non-parametric regression	Logistic regression
Predict a value with multiple independent variables or binomial variables	Multiple linear/non-linear regression		Multiple logistic regression

For sufficiently large sample sizes, we can assume the sampling distribution of the test statistic (e.g., mean difference in a t test) is normal (because of the Central Limit Theorem)

	Interval/Ratio (Normality assumed)	Interval/Ratio (Normality not assumed), Ordinal	Dichotomy (Binomial)
Compare two unpaired groups	Unpaired t test	Mann-Whitney test	Fisher's test
Compare two paired groups	Paired t test	Wilcoxon test	McNemar's test
Compare more than two unmatched groups	ANOVA	Kruskal-Wallis test	Chi-square test
Compare more than two matched groups	Repeated-measures ANOVA	Friedman test	Cochran's Q test
Find relationship between two variables	Pearson correlation	Spearman correlation	Cramer's V
Predict a value with one independent variable	Linear/Non-linear regression	Non-parametric regression	Logistic regression
Predict a value with multiple independent variables or binomial variables	Multiple linear/non-linear regression		Multiple logistic regression

	Interval/Ratio (Normality assumed)	Interval/Ratio (Normality not assumed), Ordinal	Dichotomy (Binomial)
Compare two unpaired groups	Unpaired t test	Mann-Whitney test	Fisher's test
Compare two paired groups	Paired t test	Wilcoxon test	McNemar's test
Compare more than two unmatched groups	ANOVA	Kruskal-Wallis test	Chi-square test
Compare more than two matched groups	Repeated-measures ANOVA	Friedman test	Cochran's Q test
Find relationship between two variables	Pearson correlation	Spearman correlation	Cramer's V
Predict a value with one independent variable	Linear/Non-linear regression	Non-parametric regression	Logistic regression
Predict a value with multiple independent variables or binomial variables	Multiple linear/non-linear regression		Multiple logistic regression

Can't rely on the Central Limit Theorem when:
- Sample size is small
- Data is heavily skewed/long-tailed and/or bounded

Dichotomy: dependent variable takes one of 2 values — usually success or failure

	Interval/Ratio (Normality assumed)	Interval/Ratio (Normality not assumed), Ordinal	Dichotomy (Binomial)
Compare two unpaired groups	Unpaired t test	Mann-Whitney test	Fisher's test
Compare two paired groups	Paired t test	Wilcoxon test	McNemar's test
Compare more than two unmatched groups	ANOVA	Kruskal-Wallis test	Chi-square test
Compare more than two matched groups	Repeated-measures ANOVA	Friedman test	Cochran's Q test
Find relationship between two variables	Pearson correlation	Spearman correlation	Cramer's V
Predict a value with one independent variable	Linear/Non-linear regression	Non-parametric regression	Logistic regression
Predict a value with multiple independent variables or binomial variables	Multiple linear/non-linear regression		Multiple logistic regression

Paired/matched: within-subject
Unpaired/unmatched: between-subject

	Interval/Ratio (Normality assumed)	Interval/Ratio (Normality not assumed), Ordinal	Dichotomy (Binomial)
Compare two unpaired groups	Unpaired t test	Mann-Whitney test	Fisher's test
Compare two paired groups	Paired t test	Wilcoxon test	McNemar's test
Compare more than two unmatched groups	ANOVA	Kruskal-Wallis test	Chi-square test
Compare more than two matched groups	Repeated-measures ANOVA	Friedman test	Cochran's Q test
Find relationship between two variables	Pearson correlation	Spearman correlation	Cramer's V
Predict a value with one independent variable	Linear/Non-linear regression	Non-parametric regression	Logistic regression
Predict a value with multiple independent variables or binomial variables	Multiple linear/non-linear regression		Multiple logistic regression

	Interval/Ratio (Normality assumed)	Interval/Ratio (Normality not assumed), Ordinal	Dichotomy (Binomial)
Compare two unpaired groups	Unpaired t test	Mann-Whitney test	Fisher's test
Compare two paired groups	Paired t test	Wilcoxon test	McNemar's test
Compare more than two unmatched groups	ANOVA	Kruskal-Wallis test	Chi-square test
Compare more than two matched groups	Repeated-measures ANOVA	Friedman test	Cochran's Q test
Find relationship between two variables	Pearson correlation	Spearman correlation	Cramer's V
Predict a value with one independent variable	Linear/Non-linear regression	Non-parametric regression	Logistic regression
Predict a value with multiple independent variables or binomial variables	Multiple linear/non-linear regression		Multiple logistic regression

- Pearson and Spearman correlations both measure relationships between 2 variables.
- Different in their assumptions and what they measure
- Pearson measures linear relationships
- Spearman measures monotonic relationships

Q5

+0.68

Discrimination Index ?

Consider the following scenario.

Suppose a designer wants to show that their new keyboard (called maxi-QWERTY) that they have designed has a lower text entry error rate than two existing baselines:. In their study, each participant tries all three keyboard types: 1) mini-QWERTY keyboard where the participant can only enter text by pressing on individual keyboard keys, and 2) mini-QWERTY keyboard where the participant can only swipe to enter text.

What statistical analysis test should the designer pick to evaluate their design? Select **the best** answer. (**Hint:** consult Yatani's table; link was provided in the lecture.)

Just two paired t-test, each to compare the intervention with one of the baselines.	1 respondent	2 %	▮	
Friedman test, followed by pairwise comparison using Willcoxon tests.	33 respondents	67 %	▬▬▬▬▬	✓
Repeated-measures ANOVA, followed by pairwise comparison using paired t-tests.	15 respondents	31 %	▬▬	
The designer should not use a statistical analysis test because the best study design for evaluating their keyboard will not have enough participants to reach statistically significant results.		0 %	▮	

67% answered correctly

20

Q6

An online learning platform aims to foster a culture of constructive feedback in its user forums. Which of the following strategies best reinforces descriptive norms to encourage this behavior? Select **all** that apply.

Hiding comments that may contain unconstructive criticism until reviewed by moderators.	28 respondents	57 %	✓	55% answered correctly
Highlighting and displaying constructive comments prominently on the homepage, so users see examples of positive engagement.	49 respondents	100 %	✓	
Posting a reminder at the top of each discussion, instructing users to provide constructive feedback.	16 respondents	33 %		
Adding a warning message that users who post unconstructive comments may be banned.	3 respondents	6 %		

21

Questions, comments, and/or concerns?

Farnaz Jahanbakhsh
farnaz@umich.edu
https://people.csail.mit.edu/farnazj