

Assignment 2

	TreeMap	Phylotree	ArcTree	Total (rows)
Female	36.0	31.8	56.2	41.3
Male	41.0	56.0	55.6	50.9
Total (columns)	38.5	43.9	55.9	46.1

Non-Parametric Tests

After the next section, you will be able to identify and apply the appropriate test to use for several types of data.

Types of Data

- Nominal
 - e.g., apples, oranges
- Ordinal
 - e.g., low, medium, high
- Interval
 - e.g., temperature in °C (+, -)
- Ratio
 - e.g., temperature in Kelvin, time (\times , \div)

Does it make sense to do a t-test/ANOVA on categorical data?

Choice of Test

Level of Measurement	Sample Characteristics					Correlation
	1 Sample	2 Samples		K Samples (i.e., >2)		
		Independent	Dependent	Independent	Dependent	
Categorical/ Nominal	χ^2 or binomial	χ^2	McNemar χ^2	χ^2	Cochran's Q	
Rank/Ordinal		Mann Whitney U	Wilcoxon Signed Ranks	Kruskal Wallis H	Friedman's ANOVA	Spearman's rho
Interval/Ratio	z-test or t-test	unpaired t-test	paired t-test	one-way ANOVA (between- subjects)	one-way ANOVA (within- subjects)	Pearson's r
		Factorial ANOVA				

Adapted from:
Plonskey, 2001

Slides by Mark Hancock

Choice of Test

Level of Measurement	Sample Characteristics					Correlation
	1 Sample	2 Samples		K Samples (i.e., >2)		
		Independent	Dependent	Independent	Dependent	
Categorical/ Nominal	χ^2 or binomial	χ^2	McNemar χ^2	χ^2	Cochran's Q	
Rank/Ordinal		Mann Whitney U	Wilcoxon Signed Ranks	Kruskal Wallis H	Friedman's ANOVA	Spearman's rho
Interval/Ratio	z-test or t-test	unpaired t-test	paired t-test	one-way ANOVA (between- subjects)	one-way ANOVA (within- subjects)	Pearson's r
		Factorial ANOVA				

Adapted from:
Plonskey, 2001

Slides by Mark Hancock

Chi-Square (χ^2)

- Frequency data for each category
- Compare *observed* frequency to *expected* frequency (equal distribution)

$$\chi^2 = \sum_{i=1}^n \frac{(f_{\text{observed}_i} - f_{\text{expected}_i})^2}{f_{\text{expected}_i}}$$

Null Hypothesis

- The frequencies are equally distributed
- What does high χ^2 mean?

Example

- 40 people were asked to state their favourite fruit.
- 9 chose apples, 21 chose oranges, 4 chose peaches, and 6 chose another fruit

Choice of Test

Level of Measurement	Sample Characteristics					Correlation
	1 Sample	2 Samples		K Samples (i.e., >2)		
		Independent	Dependent	Independent	Dependent	
Categorical/ Nominal	χ^2 or binomial	χ^2	McNemar χ^2	χ^2	Cochran's Q	
Rank/Ordinal		Mann Whitney U	Wilcoxon Signed Ranks	Kruskal Wallis H	Friedman's ANOVA	Spearman's rho
Interval/Ratio	z-test or t-test	unpaired t-test	paired t-test	one-way ANOVA (between- subjects)	one-way ANOVA (within- subjects)	Pearson's r
		Factorial ANOVA				

Adapted from:
Plonskey, 2001

Slides by Mark Hancock

Choice of Test

Level of Measurement	Sample Characteristics					Correlation
	1 Sample	2 Samples		K Samples (i.e., >2)		
		Independent	Dependent	Independent	Dependent	
Categorical/ Nominal	χ^2 or binomial	χ^2	McNemar χ^2	χ^2	Cochran's Q	
Rank/Ordinal		Mann Whitney U	Wilcoxon Signed Ranks	Kruskal Wallis H	Friedman's ANOVA	Spearman's rho
Interval/Ratio	z-test or t-test	unpaired t-test	paired t-test	one-way ANOVA (between- subjects)	one-way ANOVA (within- subjects)	Pearson's r
				Factorial ANOVA		

Adapted from:
Plonskey, 2001

Slides by Mark Hancock

Mann-Whitney U

- Can compare factors with two levels *only*
- Design is *between-participants*
- Method:
 - assign each data point a rank (in the *entire* set)
 - T_A = sum of ranks in group A
 - $U_A = (\text{max possible } T_A) - (\text{observed } T_A)$
 - $U_A + U_B = n_A n_B$ (use either)

Null Hypothesis

- Ranks in each group are on average the same
- $U_A = U_B = n_A n_B / 2$
- If n large enough ($n > 5$), can use z-test (a simpler t-test) on *ranks*

Example

- 15 Mac users and 15 Windows users were asked to rate the usability of their operating system on a scale of 1 (very unusable) to 9 (very usable)

Choice of Test

Level of Measurement	Sample Characteristics					Correlation
	1 Sample	2 Samples		K Samples (i.e., >2)		
		Independent	Dependent	Independent	Dependent	
Categorical/ Nominal	χ^2 or binomial	χ^2	McNemar χ^2	χ^2	Cochran's Q	
Rank/Ordinal		Mann Whitney U	Wilcoxon Signed Ranks	Kruskal Wallis H	Friedman's ANOVA	Spearman's rho
Interval/Ratio	z-test or t-test	unpaired t-test	paired t-test	one-way ANOVA (between- subjects)	one-way ANOVA (within- subjects)	Pearson's r
		Factorial ANOVA				

Adapted from:
Plonskey, 2001

Slides by Mark Hancock

Choice of Test

Level of Measurement	Sample Characteristics					Correlation
	1 Sample	2 Samples		K Samples (i.e., >2)		
		Independent	Dependent	Independent	Dependent	
Categorical/ Nominal	χ^2 or binomial	χ^2	McNemar χ^2	χ^2	Cochran's Q	
Rank/Ordinal		Mann Whitney U	Wilcoxon Signed Ranks	Kruskal Wallis H	Friedman's ANOVA	Spearman's rho
Interval/Ratio	z-test or t-test	unpaired t-test	paired t-test	one-way ANOVA (between- subjects)	one-way ANOVA (within- subjects)	Pearson's r
		Factorial ANOVA				

Adapted from:
Plonskey, 2001

Slides by Mark Hancock

Wilcoxon Signed-Ranks

- Can compare factors with two levels *only*
- Design is *within-participants*
- Method:
 - assign *signed* rank for each participant
 - each participant has value in each level (x_i & y_i)
 - obtain rank by ordering $|x_i - y_i|$ and reapply sign
 - W = sum of signed ranks

Null Hypothesis

- Difference in ranks is on average zero
- $W = 0$
- If n large enough ($n > 9$), use z-test, otherwise use table
- Note $n < 5$ never significant at $\alpha = .05$

Example

- 20 people were asked to rate the amount of fun they have playing both soccer and ultimate on a scale of 1 (worst. game. ever.) to 9 (best. game. ever.)

Choice of Test

Level of Measurement	Sample Characteristics					Correlation
	1 Sample	2 Samples		K Samples (i.e., >2)		
		Independent	Dependent	Independent	Dependent	
Categorical/ Nominal	χ^2 or binomial	χ^2	McNemar χ^2	χ^2	Cochran's Q	
Rank/Ordinal		Mann Whitney U	Wilcoxon Signed Ranks	Kruskall Wallis H	Friedman's ANOVA	Spearman's rho
Interval/Ratio	z-test or t-test	unpaired t-test	paired t-test	one-way ANOVA (between- subjects)	one-way ANOVA (within- subjects)	Pearson's r
		Factorial ANOVA				

Adapted from:
Plonskey, 2001

Slides by Mark Hancock

Choice of Test

Level of Measurement	Sample Characteristics					Correlation
	1 Sample	2 Samples		K Samples (i.e., >2)		
		Independent	Dependent	Independent	Dependent	
Categorical/ Nominal	χ^2 or binomial	χ^2	McNemar χ^2	χ^2	Cochran's Q	
Rank/Ordinal		Mann Whitney U	Wilcoxon Signed Ranks	Kruskal Wallis H	Friedman's ANOVA	Spearman's rho
Interval/Ratio	z-test or t-test	unpaired t-test	paired t-test	one-way ANOVA (between- subjects)	one-way ANOVA (within- subjects)	Pearson's r
		Factorial ANOVA				

Adapted from:
Plonskey, 2001

Slides by Mark Hancock

Kruskal-Wallis H

- Can compare factors with $k > 2$ levels
- Design is *between-participants*
- Method:
 - assign each data point a rank (in the *entire* set)
 - T_A = sum of ranks in group A, $\mu_A = T_A / n_A$
 - $SS_{BG} = \text{sum } (n_i(\mu_i - \mu_{all})^2)$
 - $H = SS_{BG} / (n(n+1) / 12)$

Null Hypothesis

- Mean ranks in each group are the same
- H value equivalent to χ^2
 - H is ratio of sum of ranks squared to expected rank

Example

- 10 Mac users, 10 Windows users, and 10 Linux users were each given 10 simple computer tasks to complete in 2 minutes or less (each).
- The total number (out of 10) tasks that were successfully completed in the time given were tallied.

Choice of Test

Level of Measurement	Sample Characteristics					Correlation
	1 Sample	2 Samples		K Samples (i.e., >2)		
		Independent	Dependent	Independent	Dependent	
Categorical/ Nominal	χ^2 or binomial	χ^2	Macnar- mar's χ^2	χ^2	Cochran's Q	
Rank/Ordinal		Mann Whitney U	Wilcoxon Signed Ranks	Kruskal Wallis H	Friedman's ANOVA	Spearman's rho
Interval/Ratio	z-test or t-test	unpaired t-test	paired t-test	one-way ANOVA (between- subjects)	one-way ANOVA (within- subjects)	Pearson's r
		Factorial ANOVA				

Adapted from:
Plonskey, 2001

Slides by Mark Hancock

Choice of Test

Level of Measurement	Sample Characteristics					Correlation
	1 Sample	2 Samples		K Samples (i.e., >2)		
		Independent	Dependent	Independent	Dependent	
Categorical/ Nominal	χ^2 or binomial	χ^2	Macnar- mar's χ^2	χ^2	Cochran's Q	
Rank/Ordinal		Mann Whitney U	Wilcoxon Signed Ranks	Kruskal Wallis H	Friedman's ANOVA	Spearman's rho
Interval/Ratio	z-test or t-test	unpaired t-test	paired t-test	one-way ANOVA (between- subjects)	one-way ANOVA (within- subjects)	Pearson's r
		Factorial ANOVA				

Adapted from:
Plonskey, 2001

Slides by Mark Hancock

Friedman's ANOVA

- Can compare factors with $k > 2$ levels
- Design is *within-participants*
- Method:
 - assign rank of 1 to k for each *participant*
 - $SS_{BG} = n \cdot \text{sum} (\mu_i - \mu_{\text{all}})^2$
 - $\chi^2 = SS_{BG} / (k(k+1) / 12)$

Null Hypothesis

- Mean rank in each group is the same:
 - $(k + 1)/2$
- χ^2 is same as before
 - higher value \rightarrow lower p

Example

- 5 of my closest friends were asked how close they would be willing to stand next to me when I had (a) coded all day, (b) played squash for 1 hour, or (c) played ultimate for 6 hours.
- Answers included: “an arm’s length”, “a 10 foot pole away”, “I’d be willing to hug you”, etc.

Handy Chart

Level of Measurement	Sample Characteristics					Correlation
	1 Sample	2 Samples		K Samples (i.e., >2)		
		Independent	Dependent	Independent	Dependent	
Categorical/ Nominal	χ^2 or binomial	χ^2	McNemar χ^2	χ^2	Cochran's Q	
Rank/Ordinal		Mann Whitney U	Wilcoxon Signed Ranks	Kruskal Wallis H	Friedman's ANOVA	Spearman's rho
Interval/Ratio	z-test or t-test	unpaired t-test	paired t-test	one-way ANOVA (between- subjects)	one-way ANOVA (within- subjects)	Pearson's r
		Factorial ANOVA				

Adapted from:
Plonskey, 2001

Slides by Mark Hancock

Break: 15 Minutes

Interpreting Analyses

By the end of this lecture, you will be able to interpret and criticise the description of an experiment in a typical HCI paper.

Parts of a Paper (APA)

- Methods
 - Participants
 - Apparatus
 - Design
 - Procedure
- Results
- Discussion

Method: Participants

- Why is this section important?
- Important Details:
 - age, sex (demographics)
 - how did you select them?
 - how did you compensate them for their time?
- What biases might they bring to the study?
 - e.g., experience

Example

“12 volunteers (8 males and 4 females), aged between 23 and 38 participated in the experiment. 11 participants were right handed and controlled the pen with their right hand. All participants were frequent computer users and had experience with Windows. Eight participants had previously worked with digital tabletop systems.

Eleven of them had already pen- and/or touch-based interface experience (e.g. Tablet PC).”

Leithinger & Haller, Tabletop 2007

Common Mistakes

- Too little detail
- Too much detail (why might this be bad?)
- Not enough participants
- Participants don't represent population
 - e.g., computer science students

Method: Apparatus

- Describes the environment used to test the participants
- What criteria are necessary for a “good” Apparatus section?
 - possible to reproduce

Example

“The display used for all four tasks was a 1024 × 768 pixels, 73.3 cm × 55.0 cm wall display. In all cases, the display space was a perspective grid representing five walls of a virtual room with lighting as an additional depth cue (see Figure 5). Participants were able to control a 3D cursor using a tracked light pen in the 73.3 cm × 55.0 cm × 55.0 cm volume directly in front of the display. Participants were given the option to sit, but all chose to stand throughout all trials. The pen was tracked using vision algorithms and input from two cameras mounted directly above the control space. Infrared filters were used to detect the near infrared light from the pen light (non-LED).”

Keijser et al., 3DUI 2007

Common Mistakes

- Not enough detail to reproduce
 - e.g., resolution without screen size
 - e.g., is person seated or standing
- Apparatus/setup has inherent errors in measurement
 - e.g., device not sufficiently precise
 - e.g., unnecessary strain on participants

Method: Design

- Describes the way the experiment was controlled
- Important details:
 - factors, dependent variables, constants
 - how did you assign participants to conditions?

Example

“We used a 3x2x9 design with the following independent variables: display size (small, medium, large corresponding to within, equal to, and beyond visual acuity - data density was constant), visualization design (space-centric, attribute-centric), and task (3 detail, 4 overview, and 2 complex)... Display size was a between subjects variable while visualization and task were within subject variables. Task completion time, accuracy, subjective workload, and user preference were recorded.”

Yost et al., CHI 2007

Common Mistakes

- Not enough detail
 - not reproducible

Method: Procedure

- What did each participant do?
- Important details:
 - step-by-step description of task
 - what conditions must be met to end a *trial*?
 - practice
 - order of conditions, randomization, repetitions

Example

“To begin a trial, participants pressed “enter”, and the word was presented. We wished to measure the time participants took to read the word before beginning to type. They were, instructed to press “enter” as soon as they recognised the word, which would then disappear, ensuring that they were no longer reading the text. A text box was then displayed, into which they would type the word and then press “enter” to finish the trial. If they typed an incorrect word, a tone would sound and the trial word would appear again, with the trial counted as an error. If, after beginning to enter their response, the participant wished to review the word, they could press “escape” to return to viewing the word – all such trials were also counted as errors. Lastly, if the participant failed to enter the correct word after 30 seconds, the trial would end and be counted as an error.”

Grossman et al., CHI 2007

Common Mistakes

- Not enough detail
- Confusing
 - reader can't imagine what happened
- Task doesn't represent what you are trying to measure
 - e.g., measure strength of wrist by how far a person can throw a Frisbee

Results (the fun begins...)

- What do people need to know?
- Important details:
 - what test you used
 - the important values for each test (e.g., F , df , etc.)
 - the means, SDs for (each) cell

Example

“Total time on task and fixation measures were analyzed using 2 (*Task Type*) x 6 (*Target Position*) repeated measures analysis of variance (ANOVA). The click accuracy was analyzed using a chi-square analysis.”

Guan & Cutrell, CHI 2007

Example

“We found a significant main effect of *Target Position on the total time on task* ($F(5,85)=3.544, p=.006$). This indicates that people spent significantly more time on a task when the target was displayed at a lower position. We also found a main effect for *Task Type*, $F(1,17)=54.718, p<0.001$... There was no significant interaction between *Target Position and Task Type*.”

“A chi-square analysis on the number of accurate clicks showed a significant effect for target position ($\chi^2(5)=58.5, p<0.001$). The click accuracy rate dropped from 84% (average of 78% and 89%) to about 11% when the target was displayed at position 8...”

Guan & Cutrell, CHI 2007

Common Mistakes

- Incorrect analysis
 - e.g., ran pairwise t-tests, but no ANOVA
- Attempt to interpret results
 - e.g., Participants with higher education consistently preferred ultimate to soccer ($Z = 2.33$, $p = .02$), which is likely due to their higher intellect.

Discussion

- What do the results mean?
- Important details:
 - leave *all* interpretation of results to this section
 - make claims here *based on the evidence* provided in your experiment (perhaps in light of results from other work).

Example

Our results also confirmed our earlier observations about Offset Cursor, namely that it impacts task times even for large targets. For targets 24px and greater *Offset* was 1.57 times slower than *Touch* with mean median times of 938ms and 597ms respectively (Figure 13). This is somewhat surprising since estimating the offset distance should be much easier with large targets, given the increased error tolerance.

One possible reason why Offset Cursor is slower may be because users often overshoot or undershoot the target, resulting in a higher *net correction distance* – the screen distance between initial contact and final lift-off.

Vogel & Baudisch, CHI 2007

Common Mistakes

- Claims are not supported by the evidence
 - e.g., People were faster using ArcTrees than Phylotrees because aesthetics are less important than functionality.
- Sneak in a result
 - e.g., Not only did people prefer Ultimate to soccer, they also enjoyed it more (average rating 7.3 out of 10).

Break: 15 Minutes

Rubbing and Tapping for Precise and Rapid Selection on Touch-Screen Displays

Alex Olwal, Steven Feiner, Susanna Heyman

Participants: anything wrong?

“Twenty right-handed volunteers participated in the study and each received two cinema tickets as compensation. The 8 female and 12 male participants were between 19 and 34 years old (average 23.9, standard deviation 3.89). They were, or had previously been, university students. The majority of the participants were students in Media Technology at the Royal Institute of Technology...”

Common Mistakes

- Too little detail
- Too much detail (why might this be bad?)
- Not enough participants
- Participants don't represent population
 - e.g., computer science students

Participants: anything wrong?

- Do university students represent the population?
 - Note: this may be an unavoidable flaw

Apparatus: anything wrong?

“The experiment used a dual 3.0 GHz Pentium Xeon PC running Windows XP with a 15" resistive touch screen display (MultiQ MQ 158 POS). The display's 1024×768 native XGA resolution results in a pixel triad width of approximately 0.3 mm and the display was tilted approximately 15° backwards for user comfort. The experimental software was implemented with OpenGL and C++.”

Common Mistakes

- Not enough detail to reproduce
 - e.g., resolution without screen size
 - e.g., is person seated or standing
- Apparatus/setup has inherent errors in measurement
 - e.g., device not sufficiently precise
 - e.g., unnecessary strain on participants

Apparatus: anything wrong?

- Did the user sit/stand?
- Does OpenGL + C++ sufficiently describe the software?

Design: anything wrong?

A repeated-measures, within-subjects study was performed. There were five target widths (1, 2, 4, 8, and 16 pixels = 0.3, 0.6, 1.2, 2.4 and 4.8 mm) and seven techniques (Take- Off, Zoom-Pointing, Rub-Pointing, Rub-Pointing.Click, Zoom-Tapping, Zoom-Tapping.Click, and Rub-Tapping). The order in which the techniques were presented was randomized, and the order in which sizes were presented was randomized for each block of trials. An analysis found no significant effects of order on the results.

Design: anything wrong?

A repeated-measures, within-subjects study was performed. There were five target widths (1, 2, 4, 8, and 16 pixels = 0.3, 0.6, 1.2, 2.4 and 4.8 mm) and seven techniques (Take- Off, Zoom-Pointing, Rub-Pointing, Rub-Pointing.Click, Zoom-Tapping, Zoom-Tapping.Click, and Rub-Tapping). The order in which the techniques were presented was randomized, and the order in which sizes were presented was randomized for each block of trials. An analysis found no significant effects of order on the results.

Design: anything wrong?

After seeing a demonstration of a technique, a participant performed an initial block of 10 practice trials (2 trials \times 5 widths) with that technique, where each trial needed to result in a successful selection for the program to proceed, to ensure that the participant experienced successful selections for the technique...

Design: anything wrong?

After seeing a demonstration of a technique, a participant performed an initial block of 10 practice trials (2 trials × 5 widths) with that technique, where each trial needed to result in a successful selection for the program to proceed, to ensure that the participant experienced successful selections for the technique...

Design: anything wrong?

Consequently, we had:

1 trials × 5 widths	=	5 demonstration trials
2 trials × 5 widths	=	10 practice trials (must succeed)
+ 3 trials × 5 widths	=	15 practice trials
+ 14 trials × 5 widths	=	70 test trials
		<hr/>
		95 trials
		×
		<hr/>
		7 techniques
		<hr/>
		665 selections per participant

Procedure: anything wrong?

Each participant was asked to alternately select two targets placed 250 pixels apart, well away from the edges of the screen, in a reciprocal 1D pointing task, where zoom level was reset after each target selection. To maximize contrast, targets were green squares of varying size on a black background. A large grey offset rectangular outline helped the participant identify the position of the target at the beginning of each trial. The rectangular outline was hidden upon touch. Auditory feedback was provided with a low frequency beep for errors and a high-frequency beep when the participant successfully selected a target. The software logged times and hit positions, such that completion times and error rates could be derived.

Common Mistakes

- Not enough detail
- Confusing
 - reader can't imagine what happened
- Task doesn't represent what you are trying to measure
 - e.g., measure strength of wrist by how far a person can throw a Frisbee

Procedure: anything wrong?

- Could you reproduce?
 - note: there is a video
- What are they measuring?

Results: anything wrong?

A within-subjects ANOVA of mean errors show that target size ($F_{4, 76} = 34.22, p < 0.001$) and technique ($F_{6, 114} = 24.74, p < 0.001$) had a significant effect on error rate, with a significant interaction between size and technique ($F_{24, 456} = 24.39, p < 0.001$). Paired samples t-tests with a Bonferroni adjustment show that Take-Off had significantly more errors than all other techniques for 1-pixel targets. It was also significantly worse than all but Zoom-Pointing for 2-pixel targets. Finally, for 4-pixel targets, it was significantly worse than Rub-Pointing.Click.

Results: anything wrong?

A within-subjects ANOVA of mean errors show that target size ($F_{4, 76} = 34.22, p < 0.001$) and technique ($F_{6, 114} = 24.74, p < 0.001$) had a significant effect on error rate, with a significant interaction between size and technique ($F_{24, 456} = 24.39, p < 0.001$). Paired samples t-tests with a Bonferroni adjustment show that Take-Off had significantly more errors than all other techniques for 1-pixel targets. It was also significantly worse than all but Zoom-Pointing for 2-pixel targets. Finally, for 4-pixel targets, it was significantly worse than Rub-Pointing.Click.

Results: anything wrong?

A within-subjects ANOVA of mean errors show that target size ($F_{4, 76} = 34.22, p < 0.001$) and technique ($F_{6, 114} = 24.74, p < 0.001$) had a significant effect on error rate, with a significant interaction between size and technique ($F_{24, 456} = 24.39, p < 0.001$). Paired samples t-tests with a Bonferroni adjustment show that Take-Off had **significantly more errors** than all other techniques for 1-pixel targets. It was also **significantly worse than** all but Zoom-Pointing for 2-pixel targets. Finally, for 4-pixel targets, it was **significantly worse than** Rub-Pointing.Click.

Results

...we had two blocks without a single completed trial for the smallest targets. These blocks happened for two of the participants in the difficult condition of Take-Off with 1-pixel targets. We therefore chose to divide our analysis of completion times into two parts.

Results: anything wrong?

First, we conducted an ANOVA on target sizes 2–16 pixels over all techniques, which showed that technique had a significant effect on completion time ($F_{6, 114} = 84.67, p < 0.001$). Paired samples t-tests with a Bonferroni adjustment show that Take-Off was significantly slower than all our rubbing and tapping techniques for 2-, 4- and 8-pixel targets. Rub-Pointing was significantly faster than Take-Off for 16-pixel targets ($t_{19} = 5.96, p < 0.001$)...

Results: anything wrong?

First, we conducted an ANOVA on target sizes 2–16 pixels over all techniques, which showed that technique had a significant effect on completion time ($F_{6, 114} = 84.67, p < 0.001$). Paired samples t-tests with a Bonferroni adjustment show that Take-Off was significantly slower than all our rubbing and tapping techniques for 2-, 4- and 8-pixel targets. Rub-Pointing was significantly faster than Take-Off for 16-pixel targets ($t_{19} = 5.96, p < 0.001$)...

Results: anything wrong?

- What kind of ANOVA did they do?

Results: anything wrong?

First, we conducted an ANOVA on target sizes 2–16 pixels over all techniques, which showed that technique had a significant effect on completion time ($F_{6, 114} = 84.67, p < 0.001$). Paired samples t-tests with a Bonferroni adjustment show that Take-Off was significantly slower than all our rubbing and tapping techniques for 2-, 4- and 8-pixel targets. Rub-Pointing was significantly faster than Take-Off for 16-pixel targets ($t_{19} = 5.96, p < 0.001$)...

Results: anything wrong?

- What kind of post-hoc analysis did they do?

Results: anything wrong?

First, we conducted an ANOVA on target sizes 2–16 pixels over all techniques, which showed that technique had a significant effect on completion time ($F_{6, 114} = 84.67, p < 0.001$). Paired samples t-tests with a Bonferroni adjustment show that Take-Off was significantly slower than all our rubbing and tapping techniques for 2-, 4- and 8-pixel targets. Rub-Pointing was significantly faster than Take-Off for 16-pixel targets ($t_{19} = 5.96, p < 0.001$)...

Common Mistakes

- Incorrect analysis
 - e.g., ran pairwise t-tests, but no ANOVA
- Attempt to interpret results
 - e.g., Participants with higher education consistently preferred Ultimate to soccer ($Z = 2.33$, $p = .02$), which is likely due to their higher intellect.

Results: anything wrong?

- Ran one-way ANOVA when could have run two-way
- Interpretation?

Discussion: anything wrong?

Our results also supported the importance of having a distinct separation of operations, whether in a single-handed gesture (e.g., rubbing in and rubbing out) or a bimanual interaction (e.g., Zoom-Tapping). Similarity between the dominant hand click and the non-dominant hand tap for Zoom-Tapping.Click, for example, confused several users, and led to a higher error rate

Common Mistakes

- Claims are not supported by the evidence
 - e.g., People were faster using ArcTrees than Phylotrees because aesthetics are less important than functionality.
- Sneak in a result
 - e.g., Not only did people prefer Ultimate to soccer, they also enjoyed it more (average rating 7.3 out of 10).

Discussion: anything wrong?

Our results also supported the importance of having a distinct separation of operations, whether in a single-handed gesture (e.g., rubbing in and rubbing out) or a bimanual interaction (e.g., Zoom-Tapping). Similarity between the dominant hand click and the non-dominant hand tap for Zoom-Tapping.Click, for example, confused several users, and led to a higher error rate

Summary

- Parts of a Paper:
 - Methods
 - Participants
 - Apparatus
 - Design
 - Procedure
 - Results
 - Discussion
- Common mistakes

Assignment 2, Part II

Summary of Day

- Non-parametric tests
 - which tests to use and when
- Parts of a paper
 - Method, Results, Discussion
- How to interpret and criticise each section

Next Week

- Types of validity
- How to design your own study
- Homework:
 - non-parametric tests
 - paper interpretation