Reliability and Validity

- If the goal was to hit the “Bullseye” with each dart...
  - Think of reliability ~ consistency, validity ~ accuracy
  - Reliability = reproducibility factor (consistency of a measure)
  - Validity = whether you measure what you think you measure

Questions

- Preceding questions, while unquestionably relevant, are not testable
- Try to re-cast as testable questions (...even though the new question may appear less important)

Controlled Experiment

- Why do we conduct controlled experiments?
  - Simply...
    - To answer (or raise!) questions about a new or existing UI design or interaction technique!
- Questions include...
  - Is it viable?
  - Is it as good as or better than current practice?
  - Which of several design alternatives is best?
  - What are its performance limits and capabilities?
  - What are its strengths and weaknesses?
  - Does it work well for novices, for experts?
  - How much practice is required to become proficient?

Testable Questions

- You have invented a new text entry technique for cell phones. In your view, it’s pretty good. In fact, you think it’s better than the most widely used current technique, multi-tap. You decide to undertake some empirical research to evaluate your invention and to compare it with multi-tap? What are your questions?
A Tradeoff

<table>
<thead>
<tr>
<th>Accuracy of Answer</th>
<th>Narrow</th>
<th>Broad</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- If error rates are kept under 2%, is the new technique faster than multi-tap within one hour of use?
- Is the new technique better than multi-tap?

Validities

- Internal validity: the extent to which the effects observed are due to the test conditions
  - Differences in the means are due to inherent properties of the test conditions
  - Variances are due to participant differences
  - Other potential sources of variance are controlled
  - Note: Uncontrolled sources of variance are bad news and compromise internal validity
- External validity: the extent to which results are generalizable to other people and other situations
  - Re people, the participants are representative of the broader intended population of users
  - Re situations, test environment and experimental procedures are representative of real world situations where the UI/technique will be used

Test Environment Example

- Scenario...
  - You wish to compare two input devices for remote pointing (e.g., at a projection screen)
- External validity is improved if the test environment mimics expected usage
- Test environment should probably...
  - Use a projection screen (not a CRT)
  - Position participants at a significant distance from screen (rather than close up)
  - Have participants stand (rather than sit)
  - Include an audience!
- But... is internal validity compromised?

Experimental Procedure Example

- Scenario...
  - You wish to compare two text entry techniques for mobile devices
- External validity is improved if the experimental procedure mimics expected usage
- Test procedure should probably require participants to...
  - Enter representative samples of text (e.g., phrases containing letters, numbers, punctuation, etc.)
  - Edit and correct mistakes as they would normally
  - Text while walking
- But... is internal validity compromised?

The Tradeoff

- Tension between internal & external validity
- The more the test environment and experimental procedures mimic real-world situations, the more the experiment is susceptible to uncontrolled sources of variation, e.g. pondering, distractions
- Internal and external validity are increased by...
  - Posing multiple narrow (testable) questions that cover the range of outcomes influencing the broader (untestable) questions
  - E.g., a technique that is faster, is more accurate, takes fewer steps, is easy to learn, and is easy to remember, is generally better
- There is usually a positive correlation between the testable and untestable questions

Answering Empirical Questions

- If you asked participants which one they preferred, they will answer.
- We want to know if the measured performance is different between test conditions, so...
  - We conduct a user study and measure the performance on each test condition over a group of participants
- Three questions:
  1. Is there a difference? Obvious – some difference is likely
  2. Is the difference large or small? Descriptive statistics can help
  3. Is the difference significant or is it due to chance? Inferential statistics can help (ANOVA)
Answering Questions still
► Design the experiment to collect the data to test the hypotheses to evaluate the interface to refine the design
► Three elements: quantitative, experimental, with end users.
► Benchmark tasks - gather quantitative data
  ▪ Specific, clearly stated task for users to carry out
  ▪ Example: "Find the message from Mary and reply with a response of 'Tuesday morning at 11'.”
  ▪ Users perform these under a variety of conditions and you measure performance (time, errors, etc).
► Representative tasks - add breadth, can help understand process

Variables
► “independent” = the things you compare (e.g. using mouse or joystick)
► “dependent” = the things you observe and measure (e.g. time taken to navigate)
► “control” = the things you don’t want to interfere (e.g. age, order of trying each alternative)
  ▪ Don’t allow it to vary: e.g., all males
  ▪ Allow it to vary randomly: e.g., randomly assign participants to different groups
  ▪ Counterbalance - systematically vary it: e.g., equal number of males, females in each group
► “nuisance” = the things you forgot to control!

Hypotheses
► What you predict will happen
► More specifically, the way you predict the dependent variable (i.e., accuracy) will depend on the independent variable(s)
► “Null” hypothesis ($H_0$)
  ▪ Stating that there will be no effect, e.g., “There will be no difference in performance between the two groups”
► “Significance level” ($p$):
  ▪ The probability that your null hypothesis was wrong, simply by chance
  ▪ The cutoff or threshold level of $p$ (“alpha” level) is often set at 0.05, or 5% of the time you’ll get the result you saw, just by chance

Example
► Do people read faster using iPad or Kindle?
  ▪ Independent – reader type (iPad vs. Kindle)
  ▪ Dependent - time to complete reading (minutes)
  ▪ Controlled variables - same number of males and females in each group
  ▪ Hypothesis: Time to read will be shorter using iPad
  ▪ $H_0$: Time$_{iPad}$ = Time$_{Kindle}$
► Within/between design issues
  ▪ Within Subjects: Every participant provides a score for all levels or conditions
  ▪ Between Subjects: Each participant provides results for only one condition

Trade-offs
► Within subject:
  ▪ More efficient: Each subject gives you more data
  ▪ More statistical “power”: Each person is their own control
  ▪ Require fewer participants
  ▪ May mean more complicated design to avoid “order effects”, e.g. participants may learn from first condition or fatigued after the first condition
► Between subject:
  ▪ Simpler design & analysis
  ▪ Easier to recruit participants (only one session)
  ▪ Less efficient
  
<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>iPad</td>
<td>Kindle</td>
<td>iPad</td>
</tr>
<tr>
<td>P1</td>
<td>12 s</td>
<td>P1</td>
</tr>
<tr>
<td>P2</td>
<td>19 s</td>
<td>P3</td>
</tr>
<tr>
<td>P3</td>
<td>13 s</td>
<td>P4</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>P5</td>
</tr>
<tr>
<td>P6</td>
<td>21 s</td>
<td>P6</td>
</tr>
</tbody>
</table>

What about subjects?
► How many?
  ▪ Book advice: at least 10
  ▪ Other advice: 6 subjects per experimental condition
  ▪ Real advice: depends on statistics
► How do you know you had enough?
  ▪ First look at each participant’s data
  ▪ Were there outliers, people who fell asleep, anyone who tried to mess up the study, etc.?
  ▪ Then look at aggregate results and descriptive statistics

Scatter plot

Time in s
Scales of Measurement

- Major task in measurement: systematically apply numbers to variables.
- Nominal (naming/category scale):
  - Differences between categories – qualitative.
  - Represent categories where there is no basis for ordering the categories, e.g., male vs. female, Ford vs. Toyota.
- Ordinal (order):
  - Involve categories that can be ordered along a pre-established dimension.
  - No way of knowing how different the categories are from one another, e.g., white, green, blue, brown belts.
- Ratio (numbers):
  - Distance between adjacent numbers are equal.
  - Most ratio scales are counts of things (e.g., temperature).
  - There is reference to zero point.

Effective range of the scale:

- Every measure has an effective range for the population under study.
- Attenuation effect: if effective range is inadequate (distorts data & threatens the validity of the study).
  - Ceiling effect – restricted higher range
  - Floor effect – restricted lower range

Levels of Measurement

<table>
<thead>
<tr>
<th>Levels of Measurement</th>
<th>Nominal</th>
<th>Ordinal</th>
<th>Interval</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examples</td>
<td>Gender, name of places, Socioeconomic class ranks</td>
<td>Scores, personality &amp; attitude scales</td>
<td>Weight, height, length, time, # of responses</td>
<td></td>
</tr>
<tr>
<td>Properties</td>
<td>Identity</td>
<td>Identity Magnitude</td>
<td>Identity Magnitude Equal intervals</td>
<td>Identity Magnitude Equal intervals True zero point</td>
</tr>
<tr>
<td>Mathematical operations</td>
<td>Rank order</td>
<td>+, -</td>
<td>+, -, x, ÷</td>
<td></td>
</tr>
<tr>
<td>Type of data</td>
<td>Nominal</td>
<td>Ordered</td>
<td>Score</td>
<td>Score</td>
</tr>
<tr>
<td>Typical statistics used</td>
<td>Chi-square</td>
<td>Mann Whitney</td>
<td>t-test, ANOVA</td>
<td>t-test, ANOVA</td>
</tr>
</tbody>
</table>

Parametric and Nonparametric Tests

- Parametric tests estimate at least one parameter (in t-test it is population mean)
  - Usually for normal distributions and when the dependent variable is interval/ratio
  - Less likely to have type II error
  - Prone to violation to normality of data
- Nonparametric tests do not test hypothesis about specific population parameters
  - Distribution-free tests
  - Although appropriate for all levels of measurement most frequently applied for nominal or ordinal measures
  - Easier to compute and have less restrictive assumptions

Strategy of Experimentation

- “Best-guess” experiments
  - Used a lot
  - More successful than you might suspect, but there are disadvantages...
- One-factor-at-a-time (OFAT) experiments
  - Sometimes associated with the “scientific” or “engineering” method
  - Devastated by interaction, also very inefficient
- Statistically designed experiments
  - Based on Fisher’s factorial concept
  - Full factorial, fractional factorial, latin square, etc

Full Factorial Design

- All possible combinations of factor levels are tested
- Start w. two-level design: experiments which include all decision variables at only two levels (usually coded as - and +)
- With this you get the main effects and interactions between pairs and among all 3 variables
- Example: the time to get there in ms (y) from all combinations of three decision variables:
  - T = target distance at 60 pixels, 180 pixels
  - C = CD gain at 20%, 40%
  - K = input device A (mouse) or B (joystick)
**Full factorial design with 2 levels per factor**

<table>
<thead>
<tr>
<th>Trial</th>
<th>T</th>
<th>C</th>
<th>K</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>60</td>
<td>20</td>
<td>A</td>
<td>y-</td>
</tr>
<tr>
<td>2</td>
<td>180</td>
<td>20</td>
<td>A</td>
<td>y-</td>
</tr>
<tr>
<td>3</td>
<td>60</td>
<td>40</td>
<td>A</td>
<td>y+</td>
</tr>
<tr>
<td>4</td>
<td>180</td>
<td>40</td>
<td>A</td>
<td>y+</td>
</tr>
<tr>
<td>5</td>
<td>60</td>
<td>20</td>
<td>B</td>
<td>y+</td>
</tr>
<tr>
<td>6</td>
<td>180</td>
<td>20</td>
<td>B</td>
<td>y+</td>
</tr>
<tr>
<td>7</td>
<td>60</td>
<td>40</td>
<td>B</td>
<td>y+</td>
</tr>
<tr>
<td>8</td>
<td>180</td>
<td>40</td>
<td>B</td>
<td>y+</td>
</tr>
</tbody>
</table>

**Fractional Factorial Design**

- Full factorial design is time and resource intensive (think if each has more than 2 levels)
- Fractional factorial experiment, meaning simply an experiment involving a subset of the experimental conditions.
- Latin square: get main effects but no interactions (so don’t do it if you suspect interaction)
  - Condition: every factors must have the same number of levels
  - Every level of every factor appears with every level of every other factor exactly once (think Sudoku)

**Data Analysis: Descriptive**

- Mean (=average), median, mode: measures of location. Ex. Data: 3 8 3 3 1
  - A variable that is way different than others is called an “outlier” (=8).
  - Mean: \((3+8+3+3+1)/5 = 3.6\) → affected by outliers
  - Median = the middle point of sequentially arranged data points \((1,3,3,3,8) = 3\) → Robust to outliers
  - Mode: the most frequently showing data = 3 → Robust to outliers
  - The most appropriate measure depends on distribution shape (symmetric vs. skewed, unimodal vs. multimodal)
- Spread/dispersion: The degree to which scores are clumped around the mean.
  - Standard deviation and variance

**2.4.3 Data distribution shapes**

- If data are symmetric, the mean, median, and mode will be approximately the same.
- If data are multimodal, report the mean, median and/or mode for each subgroup.
- If data are skewed, report the median.

**Two groups data – case 1**

<table>
<thead>
<tr>
<th>Participant</th>
<th>Method</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>5.3</td>
<td>5.7</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>3.6</td>
<td>4.6</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>5.2</td>
<td>5.1</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>3.3</td>
<td>4.5</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>4.6</td>
<td>6.0</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>4.1</td>
<td>7.0</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>4.0</td>
<td>6.0</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>5.0</td>
<td>4.6</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>5.2</td>
<td>5.5</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>5.1</td>
<td>5.6</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td></td>
<td>4.5</td>
<td>5.5</td>
</tr>
<tr>
<td><strong>SD</strong></td>
<td></td>
<td>0.73</td>
<td>0.78</td>
</tr>
</tbody>
</table>

**Two groups data – case 2**

<table>
<thead>
<tr>
<th>Participant</th>
<th>Method</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>2.4</td>
<td>6.9</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>2.7</td>
<td>7.2</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>3.4</td>
<td>2.6</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>6.1</td>
<td>1.8</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>6.4</td>
<td>7.8</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>5.4</td>
<td>9.2</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>7.9</td>
<td>4.4</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>1.2</td>
<td>6.6</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>3.0</td>
<td>4.8</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>6.6</td>
<td>3.1</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td></td>
<td>4.5</td>
<td>5.5</td>
</tr>
<tr>
<td><strong>SD</strong></td>
<td></td>
<td>2.23</td>
<td>2.45</td>
</tr>
</tbody>
</table>
Experimental Results
► How does one know if an experiment’s results mean anything or confirm any beliefs?
► Example: 40 people participated, 28 preferred interface 1, 12 preferred interface 2
► What do you conclude?
► Inferential (diagnostic) statistics
  ▪ Tests to determine if what you see in the data (e.g., differences in the means) are reliable (replicable), and if they are likely caused by the independent variables, and not due to random effects
  ▪ e.g. t-test to compare two means
  ▪ e.g. ANOVA (Analysis of Variance) to compare several means
  ▪ e.g. test “significance level” of a correlation between two variables

ANOVA (Analysis of Variance)
► While t-test is for comparing 2 means, ANOVA is for >2
► Why not do multiple t-tests? If you want to test H0: m1 = m2 = m3
► Why not test:
  ▪ m1 = m2
  ▪ m1 = m3
  ▪ m2 = m3
  \[
  \begin{align*}
  &\text{For each test 95% probability to correctly fail to reject (accept?) null, when null is really true} \\
  &m_1 = m_2, m_1 = m_3, m_2 = m_3
  \end{align*}
  \]
  \[
  0.95^3 = \text{probability of correctly failing to reject all 3} = 0.86
  \]
► ANOVA: calculate ratios of different portions of variance of total dataset to determine if group means differ significantly from each other (Excel – Data analysis – ANOVA single f.)
► Calculate ‘F’ ratio, named after R.A. Fisher
► Same rule as t-test, observe p to see significance

Correlation
► A total of 4000 cans are opened in Texas every minute. 10 babies are conceived in Texas every minute. Therefore, each time you open a can in Texas, you stand a 1 in 400 chance of becoming pregnant.
► R = correlation coefficient (under Data analysis on Excel, check p to see if the correlation is significant)

Simple Linear Regression
► linear relationship between a predictor variable, plotted on the x-axis, and a response variable, plotted on the y-axis
► \[ Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \]
  \[
  \text{independent variable (X)} \quad \text{dependent variable (Y)} \quad \epsilon
  \]
  \[
  \text{intercept} \quad \text{slope} \quad \text{residuals}
  \]
► Remember R = correlation coefficient
► Regression on Excel:
  ▪ Scatter plot
  ▪ Format Trendline
  ▪ Type: linear
  ▪ Option: Display R-squared and equation on chart

Drawing Conclusions
► Errors in analysis do occur
► Main types:
  ▪ Type I/False positive - You conclude there is a difference, when in fact there isn’t
  ▪ Type II/False negative - You conclude there is no difference when there is
► Make your conclusions based on the descriptive stats, but back them up with inferential stats
  ▪ e.g., “The expert group performed faster than the novice group \( t(1,34) = 4.6, p > .01.\)”
► Translate the stats into words that regular people can understand
  ▪ e.g., “Thus, those who have computer experience will be able to perform better, right from the beginning…”

Feeding Back Into Design
► What were the conclusions you reached?
► How can you improve on the design?
► What are quantitative benefits of the redesign?
  ▪ e.g. 2 minutes saved per transaction, which means 24% increase in production, or $45,000,000 per year in increased profit
► What are qualitative, less tangible benefit(s)?
  ▪ e.g. workers will be less bored, less tired, and therefore more interested → better customer service
Homework

▶ Icon design – compare learnability of concrete vs. abstract icons
- Hypothesis?
- Variables?
- Tasks?
- Method?
- Data analysis?