• Last week: Introduction, types of empirical studies

• Today: Types of data, Internal and external validity (slides from Human-Computer Interaction, An Empirical Research Perspective)

• Also: scribe assignments, project details
Observe and Measure

- Foundation of empirical research
- Observation is the starting point; observations are made…
  - By the apparatus
  - By a human observer
- Manual observation
  - Log sheet, notebooks
  - Screen capture, photographs, videos, etc.
- Measurement
  - With measurement, anecdotes (*April showers bring May flowers*) turn to empirical evidence
  - “When you cannot measure, your knowledge is of a meager and unsatisfactory kind” (Kelvin)
Scales of Measurement

- Ratio
- Interval
- Ordinal
- Nominal
Nominal Data

• Nominal data (aka *categorical data*) are arbitrary codes assigned to attributes; e.g.,
  – 1 = male, 2 = female
  – 1 = mouse, 2 = touchpad, 3 = pointing stick
• The code needn’t be a number; i.e.,
  – M = male, F = female
• Obviously, the statistical mean cannot be computed on nominal data
• Usually it is the count that is important
  – “Are females or males more likely to…”
  – “Do left handers or right handers have more difficulty with…”
  – Note: The count itself is a ratio-scale measurement
• (example on next slide)
### Nominal Data – Example

- **Task:** Observe students “on the move” on university campus
- **Code and count students by…**
  - Gender (male, female)
  - Mobile phone usage (not using, using)

<table>
<thead>
<tr>
<th>Gender</th>
<th>Mobile Phone Usage</th>
<th>Total</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Not Using</td>
<td>Using</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>683</td>
<td>98</td>
<td>781</td>
</tr>
<tr>
<td>Female</td>
<td>644</td>
<td>102</td>
<td>746</td>
</tr>
<tr>
<td>Total</td>
<td>1327</td>
<td>200</td>
<td>1527</td>
</tr>
<tr>
<td>%</td>
<td>86.9%</td>
<td>13.1%</td>
<td></td>
</tr>
</tbody>
</table>

*Real Data!*
Ordinal Data

• Ordinal data associate an order or rank to an attribute
• The attribute is any characteristic or circumstance of interest; e.g.,
  – Users try three GPS systems for a period of time, then rank them: 1\textsuperscript{st}, 2\textsuperscript{nd}, 3\textsuperscript{rd} choice
• More sophisticated than nominal data
  – Comparisons of “greater than” or “less than” possible
• (example on next slide)
How many email messages do you receive each day?

1. None (I don’t use email)
2. 1-5 per day
3. 6-25 per day
4. 26-100 per day
5. More than 100 per day
Interval Data

• Equal distances between adjacent values
• But, no absolute zero
• Classic example: temperature (°F, °C)
• Statistical mean possible
  – E.g., the mean midday temperature during July
• Ratios not possible
  – Cannot say 10 °C is twice 5 °C
Interval Data – Example

- Questionnaires often solicit a level of agreement to a statement
- Responses on a Likert scale
- Likert scale characteristics:
  1. Statement soliciting level of agreement
  2. Responses are symmetric about a neutral middle value
  3. Gradations between responses are equal (more-or-less)
- Assuming “equal gradations”, the statistical mean is valid (and related statistical tests are possible)
- Likert scale example ➔ (next slide)
Please indicate your level of agreement with the following statements.

<table>
<thead>
<tr>
<th></th>
<th>Strongly disagree</th>
<th>Mildly disagree</th>
<th>Neutral</th>
<th>Mildly agree</th>
<th>Strongly agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>It is safe to talk on a mobile phone while driving.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>It is safe to read a text message on a mobile phone while driving.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>It is safe to compose a text message on a mobile phone while driving.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>
Ratio Data

• Most sophisticated of the four scales of measurement
• Preferred scale of measurement
• Absolute zero, therefore many calculations possible
• Summaries and comparisons are strengthened
• A “count” is a ratio-scale measurement
  – E.g., “time” (the number of seconds to complete a task)
• Enhance counts by adding further ratios where possible
  – Facilitates comparisons
  – Example – a 10-word phrase was entered in 30 seconds
    • Bad: $t = 30$ seconds
    • Good: Entry rate $= 10 / 0.5 = 20$ wpm
Recap

• Measurement is the foundation of empirical research
• Four types of data: Nominal, ordinal, interval, ratio

• Recall that we conduct empirical research to answer questions about an algorithm, a technique, a design, etc.
Testable Research Questions

• Scenario…
  - You have invented a new text entry technique for touchscreen mobile phones, and you think it’s pretty good. In fact, you think it is better than the Qwerty soft keyboard (QSK). You decide to undertake a program of empirical enquiry to evaluate your invention. What are your research questions?
Research Questions (2)

- Very weak
  *Is the new technique any good?*
- Weak
  *Is the new technique better than QSK?*
- Better
  *Is the new technique faster than QSK?*
- Better still
  *Is the measured entry speed (in words per minute) higher for the new technique than for QSK after one hour of use?*
Internal Validity

• Definition:
  – The extent to which the effects observed are due to the test conditions (e.g., multitap vs. new)

• Statistically, this means…
  – Differences (in the means) are due to inherent properties of the test conditions
  – Variances are due to participant differences ("pre-dispositions")
  – Other potential sources of variance are controlled or exist equally or randomly across the test conditions
External Validity

• Definition:
  – The extent to which results are generalizable to other people and other situations

• People
  – The participants are representative of the broader intended population of users

• Situations
  – The test environment and experimental procedures are representative of real world situations where the interface or 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 large display or projection screen (not a desktop monitor)
  - 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 have participants…
  – Enter personalized paragraphs of text (e.g., a paragraph about a favorite movie)
  – Edit and correct mistakes as they normally would

• But… is internal validity compromised?
There is tension between internal and external validity.

The more the test environment and experimental procedures are "relaxed" (to mimic real-world situations), the more the experiment is susceptible to uncontrolled sources of variation, such as pondering, distractions, fiddling, or secondary tasks.