

---

# Understanding and serving users

Surveys, interviews and protocols  
Intro to data analysis

# Asking users questions:

---

- *Questionnaires are probably the most frequently used method of summative evaluation of user interfaces. However, questionnaires provide a subjective evaluation of interfaces which is often greatly influenced by the type of questions asked and the way in which the questions are phrased. (Chignell, 1990)*

# Surveys and rating scales

---

- Often used to measure satisfaction:
  - And they often have acronyms..
    - QUIS
    - SUMI
    - WAMMI
- Or can be used to measure
  - intention to use (TAM)
  - cognitive effort (TLX)
- Or you can make your own up

# Four types of error in surveys:

---

- Coverage error
  - All relevant members of the sample are not equally likely to be asked to respond
- Sampling error
  - Test sample is a subset of population- Unavoidable unless you test everyone
- Measurement error
  - Inaccuracy or incomparability of response
- Non-response error
  - Failure to respond by large and/or unique parts of the sample

# Four sources of measurement error

---

- Method of eliciting answers:
  - Face to face (interview), postal, web, etc.
    - 'method effect'
- Question wording
  - Open to interpretation
- Interviewer
  - May lead or bias respondent
- Respondent
  - May not be fully honest

# Typical wording problems

---

- Questions are leading  
e.g., “Do you agree that interface was quite satisfying to use”
- Questions are vague  
e.g., “Overall, I rate the process as satisfying”
- Users do not understand terms in survey  
e.g. “Rate your affect on the following scale.....”
- Users lie or react to social pressure  
e.g. ratings for prestige items are distorted

# Close or open-ended

---

- Close-ended:
  - “This site has sufficient content for my needs” strongly agree.....strongly disagree
- Open-ended:
  - “What content would you like to see added?”  
.....

Open ended items often more difficult to answer and lead to great variation in response

# Category order effects

---

- When respondent is given a list of alternatives to select or rank their answer:
  - Written responses bias the early choices
  - Spoken responses bias the latter choices  
Dillman and Tarnai (1991)
- Try randomizing the order of choices across users

# Attitudes are subtle

---

- Asked the same question twice, users' responses may vary
- Minimize this with Scaling techniques
  - Ask the same question in different ways, compare results and derive a composite answer from all.

# When you ask changes the answer.....(Teague et al 2001)

---

- Asked users to rate ease and enjoyment:
  - During task (every 20-120 secs)
  - After each task is completed
  - After all tasks were completed
- Post-test ratings more positive
- Interrupted respondents changed verbal protocol

'The post-task's group's think aloud protocol was a running dialogue of a very descriptive nature ("I'm clicking here now looking for an index. I can't find the price of this product"). The concurrent group gave less a description of what they were doing but gave more of a verbal critique of the website and explained why they gave the ratings that they did'

Teague et al (2001) Concurrent versus post-task Usability test ratings, *Proc. of CHI'01*, ACM Press.

# Encouraging thoughtful answers

---

- Rather than ask: “Does the resource provide enough information?”
- Use a series of questions to focus the response e.g.,
  - “Consider the following situation...”
  - “Did the information system provide sufficient information for you to complete that task to your satisfaction?”

# Writing good questions

---

- Avoid vague wordings. Instead offer suitable context and be specific
- Avoid jargon, acronyms and abbreviations
- Avoid over-precision that requires perfect memory from respondent
- Do not suggest the ‘proper’ answer
- See Salant and Dillman, *How to conduct your own survey*, Wiley.1994 for more info

# Forms of survey scale:

---

- **Simple rating scale**

“I view my job as:

Very easy |...|.....|....|....|....|....|....|....|....|....| Very difficult”

- **Likert scale (Likert’s Summated Ratings)**

“This interface is very difficult to understand:

Strongly Agree, Agree, no opinion, Disagree, Strongly Disagree”

- **Semantic differentials**

“I view my work as: challenging|...|.....|....|....|....|....| overwhelming”

- **Thurstone equal appearing interval scales**

“Check the following statements if you agree with them...

This interface is very easy to use [ ]

# Technical Issues

---

- Reliability
  - Does the test reproduce the same results in the same context
- Validity
  - Does the test actually measure “satisfaction” or something else?

# How to devise a simple survey

---

- Define what you want to measure
- Generate large set of test questions
- Relate questions to your goal
- Determine scale
- Test items on real users
- Remove poor items
- Ensure test results yield answer to question

# Intro to data analysis

---

- Wherever you examine and study users, you will obtain or face data
- Analysis of data can reduce complexity, shed insights or be used to confuse....

# List of objectives

---

- Learn to interpret statistical terminology and arguments
- Appreciate the role of probability in understanding human behavior
- Be able to establish confidence intervals for your own data
- Be able to apply statistical reasoning to your own and others' test designs.
- *And as a bonus:*
  - Become comfortable with test selection and interpretation

# Statistics as language

---

- Foreign languages look strange

$$t = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{\frac{\left[ \left( \sum X_A^2 - \frac{(\sum X_A)^2}{n_1} \right) + \left( \sum X_B^2 - \frac{(\sum X_B)^2}{n_2} \right) \right] \times \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}{n_1 + n_2 - 2}}$$

- Once you learn to decode them, you read them as simply as English
- Yet people often blame themselves for not immediately understanding stats

# Role of Statistics in User studies

---

- Enable us to treat our observations systematically
- Determine likelihood of events occurring by chance or by design
- Limit bias in our data collection
- Share the communication of findings to others in an unambiguous fashion

# Test methods aim to be:

---

- **Reliable**
  - Provide reproducible results
- **Valid**
  - Measure what they claim to measure
- **Representative**
  - Sample appropriate subjects and variables
- **Ethical**
  - Involve no harm to participants

# The 7 dangers of user testing

---

- Improperly phrased research questions
- Variables not properly controlled
- Inappropriate sampling
- Not enough subjects
- Improper test administration
- Improper analysis
- Improper interpretation

# Terminology:

## Level of measurement

---

- Refers to the relationship between and among the numerical attributes we assign to our variables
  - Does a high number mean 'more than' or is it an arbitrary reference or label?
  - Is the difference between each scale value equivalent?
  - Is 4 really twice 2 or just some value larger than 2?

# Measuring –

## assigning numbers to our observations

---

- **Nominal**
  - E.g., Gender (Male / Female) Eye colour (blue /brown / grey/ green).
- **Ordinal**
  - Putting things in order, one bigger than the other. E.g., test grades, rating scales.
- **Interval**
  - Size can be ordered, plus there are equal intervals between adjacent units. Zero point is arbitrary. E.g., IQ (difference between 90 and 100 is considered equal to 100-110.)  
Temperature in °F or °C
- **Ratio**
  - Differences and ratios meaningful as scale has an absolute zero point independent of unit of measurement. E.g., Height (metres/ inches), weight, (pounds/kilos), time etc.

- **Nominal**

- Gender

- Mac/PC/Unix user

- Background or profession

- **Ordinal**

- Ranks of preference

- Many rating scales

- Clusters of experience (high to low)

User-related  
variables

- **Interval**

- (some) Survey scales

- **Ratio**

- Age

- Time on task

- Error scores

- Task completion scores

# Treating levels of measurement

---

- Level of measurement determines what we can meaningfully do with the data in terms of mathematical operations
  - Can we add/subtract them?
  - Can we multiply and divide them?
- TIP: Aim for the highest level of measurement possible when collecting data

# Interval and Ratio level variables

---

- **Continuous**
  - Infinite number of values between adjacent units (e.g., height, weight)
  - Can draw line graphs
- **Discrete/Categorical**
  - No possible values between adjacent units (e.g., number of children in family)
  - Shouldn't draw line graphs, use histograms instead

# Summary

---

- Testing is a means of obtaining reliable knowledge about a resource or system
- Statistical tests are an integral part of the evaluation process and should be considered as part of any user or usability test design
  - We do this by considering our variables and deciding what we want to know about them
- The type of statistical test used will depend in part on the level of measurement achieved

# Exercise 1

What level of measurement is suggested by the following?

---

- Number of cars owned by faculty / staff
- Make of cars owned by faculty staff
- Student grades marked from A (best) to E (worst)
- Time taken to complete an IQ test
- Scores on the IQ test

Which of the following represent a continuous variable and which a discrete variable

- Number of males in a sample
- Age of males in sample
- Height of females in sample
- Number of software packages used

# Terminology: “Variable”

---

- Variables are attributes of interest that we manipulate, observe and measure
- They provide us with an estimate or value for the property we seek to understand
  - Efficiency of performance may be assessed by time, thus “seconds per task” or “minutes spent reading” might be our variable

# Types of Variables

---

- Independent variable (IV)
  - This is what the tester manipulates
- Dependent variable (DV)
  - These are the measures we take from users, considered to be dependent on the variables tester manipulates
- Control variables (CV)
  - These are the other sources of variability in the dependent measures but whose effects we try to limit or 'control'

# In testing usability:

---

- Evaluators manipulate interface designs (*the independent variables*) and observe their effects on user performance (*the dependent variables*)
- Dependent variables tells us if our Independent variables (interfaces, training etc.) have an effect or are correlated, etc.
- We need to ‘*control*’ other extraneous variables that might also affect the users’ performance.

# Typical dependent variables

---

- Scores on a satisfaction test
- Task completion scores
- Number of links followed in navigating a site
- Verbal comments of users
- Number of times user accesses a facility
- Ratings from inspection tests
- Speed of task completion
- Expressed preferences for application A or B
  - The list is potentially endless.....

# Example - testing a web site

---

- Task - find session time and location for a class meeting
- Independent variable:
  - The Web site
    - We might test two (or more) versions - thus we would have 2 (or more) levels of Independent Variable
  - In typical usability tests, the IV is an interface, or component of an interface, such as features, layout, documentation etc.
    - Or it might be a variable among users e.g. training, experience, age etc.

# Testing the Web site (2)

---

- Dependent variables:
  - Effectiveness
    - Does user locate the answer?  
Yes/No, % score = ratio scale
  - Efficiency
    - How long does it take the user?  
Scores in seconds = ratio scale
    - How many screens does the user navigate through?  
# of screens = ratio scale
  - Satisfaction
    - Does user comment positively/negatively on interface?  
Interview comments (positive/negative)=nominal scale,  
Survey= ordinal (interval) scale

# Testing the Web site (3)

---

- Control variables
  - Other sources of variance that may affect the Dependent variables (performance) but are not part of our Independent variable (the interface)
    - E.g., gender, age, web experience, etc.
  - If we test the design without considering these variables, we may confound our findings
  - But if we consider each one of these, our test becomes much larger!

# Handling control variables

---

- If the control variables seem very important, they should be re-considered as independent variables and incorporated into the study
- If not central to the test's goal, then randomize the selection of the sample

# Independent v. Repeated measures

---

- Where each user tries one system only, you have 'independent measures'
- Where users try both (or more) systems, you have 'repeated measures'
  - Repeated measures require fewer users
  - But run risk of learning or practice effects biasing results

# Order effects

---

- For repeated measures:
  - Here, counter-balance exposure to treatments thereby reducing bias to B with impressions of A

<u>User</u>	<u>Interface A</u>	<u>Interface B</u>
Pete	1st	2nd
Joe	2nd	1st
Ann	1st	2nd
Mary	2nd	1st

# How many users is enough?

---

- For strong statistical reliability, we like to see 30 or more, but many tests are robust with smaller numbers
- In most user studies, 6 per condition is common
- And it really depends on your question.....

# Recap

---

- All levels of measurement are not equal
- The variables we chose to observe dictate what we may find out
- We manipulate ‘independent’ and observe ‘dependent’ variables, while ‘controlling’ extraneous sources of variability
- Once captured, we ‘treat’ data statistically

# Basic types of statistical treatment

---

- **Descriptive** statistics which summarize the characteristics of a *sample* of data
- **Inferential** statistics which attempt to say something about a *population* on the basis of a *sample* of data - infer to *all* on the basis of *some*

# Two kinds of descriptive statistic:

---

- Measures of central tendency
  - mean
  - median
  - mode
- Measures of dispersion (variation)
  - range
  - interquartile range
  - variance/standard deviation

# Symbol check

---

$\Sigma$

- Sigma: Means the 'sum of'

$\sum_{i=1}^n x_i$

- Sigma (1 to n) x of i: means add all values of  $i$  from 1 to n in a data set
- $X_i$  = the  $i^{\text{th}}$  data point

# Mean (arithmetic mean, or average)

---

Sum of all observations divided by the number of observations

In notation:

$$\frac{\sum_{i=1}^n x_i}{n}$$

Mean uses every item of data but is sensitive to extreme 'outliers'

Suppose we ask 10 people to perform a word processing task and measure the time in minutes which it takes them to complete it. The raw data are as follows:

---

2 4 3 2 5 2 3 2 4 2

To find the mean, we sum the observations and divide by the number of observations:

$$\frac{\sum_{i=1}^n x_i}{n} = \frac{2 + 4 + 3 + 2 + 5 + 2 + 3 + 2 + 4 + 2}{10} = 29 / 10 = 2.9$$

# Mean (cont.)

---

- Now suppose the person who took 5 minutes had *real* problems and took 60 minutes.
- The mean now becomes:

$$\frac{2 + 4 + 3 + 2 + 60 + 2 + 3 + 2 + 4 + 2}{10} = 84 / 10 = 8.4$$

Mean is sensitive to *every* data point in the set

# Median (the 50th percentile point)

---

To determine the median, put the observations in size order:

2 2 2 2 2 3 3 4 4 5



Half-way point

Hence, median = 2.5

Same is true even if last observation is 60 rather than 5

Thus even if mean shifts from 2.9 to 8.4 (as seen), median remains 2.5.

# When to use the Median?

---

Median is appropriate when

- Data are ordinal
- Data contain extreme observations which will distort the mean as a measure of central tendency  
e.g., testers' salaries:

\$45k, \$50k, \$55k, \$57k, \$60k, \$65k, \$150k

Median = \$57k

Mean = \$69k

# Mode -the most frequent observation

---

2 2 2 2 2 3 3 4 4 5

<u>N°</u>	<u>Frequency</u>
2	5
3	2
4	2
5	1

Mode = 2

The mode is the only permissible measure of central tendency for nominal data

# Measures of dispersion

---

- Range - the difference between least and largest scores:

<u>Sample</u>	<u>Scores</u>	<u>Mean</u>
I	6,4,7,5	5.5
II	7,10,1,4	5.5

Range for :

Sample I:	$7 - 4 = 3$
Sample II:	$10 - 1 = 9$

# Interquartile deviation

---

- Range is sensitive to outliers:

2 2 2 2 2 3 3 4 4 5 6 60

$$\text{Range} = 60 - 2 = 58$$

25th percentile (Q1)



75th percentile (Q3)



2 2 2 2 2 3 3 4 4 5 6 60

- Interquartile range =  $Q3 - Q1 = 4.5 - 2 = 2.5$

# When to use Interquartile deviation

---

- Useful where extremes render range un-descriptive
- Less sensitive to outliers, tells us more about the *internal spread* of data

# How do we handle outliers?

---

- Usually worth doing two analyses:
  - One with the outlier
  - One without the outlier
- Outlier might be caused by test design, some odd user behavior *or* might be a true reflection of some usability issue - so you need to investigate this further.....