



Human-Computer Interaction IS4300

1



16 – Swing Layout Managers *due now*

- **You have two choices for requirements:**
 - 1) try to duplicate the functionality of an existing applet; or,
 - 2) create your own (ideally project-related) applet with the following minimum requirements:
 - A JFrame and a modal JDialog.
 - A JTabbedPane and JScrollPane.
 - Nested JPanels including the following layout managers: GridLayout, FlowLayout, BorderLayout
 - Some interaction widgets (JButton, etc.) on every JPanel and tab.
 - Reasonable behavior when the JFrame is resized.
- You may not use GridBagLayout or absolute layout anywhere in the project.

2

T5b – Paper Prototyping *due next class*

- Recruit 3-5 users who are as close as possible to your target demographic.
- Be sure to record demographic information (age, gender, education, occupation, etc.) for your report.
- **Testing Users** When you run your prototype on a user, you should do the following things:
 - Obtain verbal consent for participation.
 - Brief the user.
 - Present one task.
 - Watch the user do the task. Take notes of your observations.
 - Repeat with the other tasks.
 - Interview users, take any measures you think are important.

3

Problem



Usability.gov

Short Usability Test Report for [Site]

Date of Report: [Month Day Year]
 Date of Test: [Month Day Year]
 Location of Test: [City, State]
 Prepared for: [Name]
 Phone Number: [000-000-0000]
 Email: [name@address.gov]

Prepared by: [Name]
 Phone Number: [000-000-0000]
 Email: [name@address.gov]

Executive summary

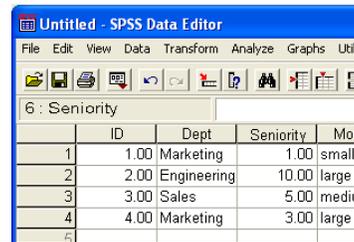
NOTE: This section describes the main goal and rationale of the study. Briefly describe the scenarios that participants completed, how the sessions were conducted, and how many participants took part in the study. This section should also discuss overall trends, such as whether or not participants were able to complete all the tasks. Data should be reported as both a number of completed scenarios as well as a percentage. Is there a reason why tasks were completed or not? Be sure to give an overall impression (theme) about what the reader will encounter in the report.

Example: Paper Prototyping usability test.

7

Coding data

- Transcribe all interviews
- Code (data entry) all measures
 - Questionnaires
 - Metrics (times, errors)
 - Check for errors, missing data
 - Unstacked format typical
 - Excel ok for very simple analyses, recommend SPSS for more complex (and R if you are a hacker)
- Do asap, by people in the room



6 : Seniority

	ID	Dept	Seniority	Mor
1	1.00	Marketing	1.00	small
2	2.00	Engineering	10.00	large
3	3.00	Sales	5.00	mediu
4	4.00	Marketing	3.00	large
5				

Summarizing Data

- Qualitative
 - Problem analysis
 - Text analysis
- Quantitative
 - Descriptive statistics
 - Inferential statistics



Summarizing Qualitative Data

Use with Interview & Think Aloud data

10



Analytic Induction (Znaniecki)

Nonexperimental, Qualitative analogue to scientific method

1. Phenomenon tentatively defined
2. Hypothesis is developed
3. A single instance is considered to determine if hypothesis is confirmed
4. If hypothesis fails, then phenomenon or hypothesis is redefined
5. Additional cases are examined and, if the new hypothesis is repeatedly confirmed, some degree of certainty results
6. Each negative case requires that the hypothesis be reformulated until there are no exceptions

11

Example: Opinions of Clippy



12

Problem (“usability defect”) analysis Example: Optometrist website

- U1: Could not find SEARCH function. Failed to complete.
- U2: Spent long time finding contents of cart. Completed.
- U3: Spent long time finding SEARCH function. Completed.
- U4: No problems.
- U5: Could not find SEARCH function. Failed to complete.
- U6: Did not like colors on checkout page.

13

Summarizing Quantitative Data

Kinds of Measures

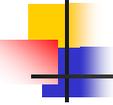
Primary source:
Bordens & Abbott, *Research
Design and Methods*

14

Scales of Measurement

- *Nominal Scale*
 - Lowest scale of measurement involving variables whose values differ by category (e.g., male/female)
 - Values of variables have different names, but no ordering of values is implied
- *Ordinal Scale*
 - Higher scale of measurement than nominal scale
 - Different values of a variable can be ranked according to quantity (e.g., high, moderate, or low self-esteem)

15



Scales of Measurement

- *Interval Scale*
 - Scale of measurement on which the spacing between values is known
 - No true zero point
 - E.g. Fahrenheit
- *Ratio Scale*
 - Similar to interval scale, but with a true zero point (e.g., number of lever presses, height)

16



What kind is it?

- Age
- Gender
- Job Category (Engineer, Manager...)
- Efficiency (time to complete)
- School Year (Freshman...)
- Temperature (Celsius)
- Think aloud quotes / themes
- Monitor Size
- Weather (Rain, Snow, ...)
- Debrief quotes / themes
- Productivity (wpd)
- Owns Pet (or not)

17

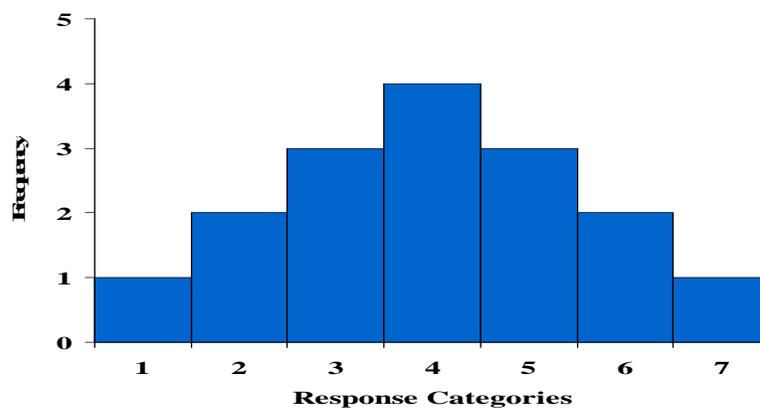
Descriptive Statistics

Practically speaking

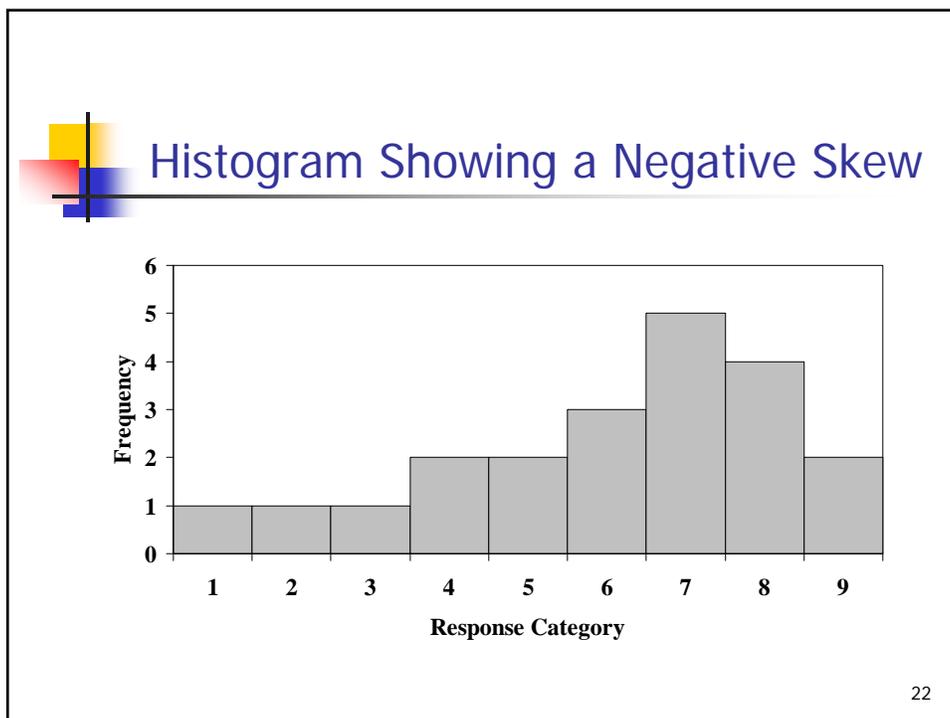
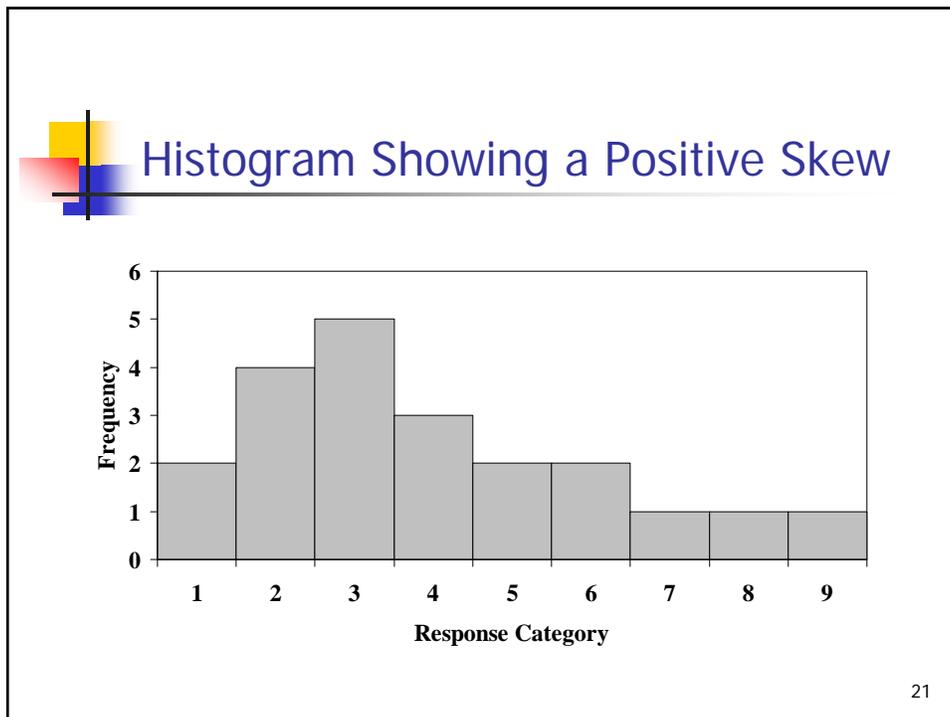
- You will decide on statistical methods depending on whether your measures are
 - Nominal, or
 - Numeric (Interval, Ratio)
- Ordinal values can sometimes be treated as numeric, sometimes as non-numeric.
- **Assume scale measures are ordinal**
- Text (structured qualitative data) should be subjected to other analyses first.

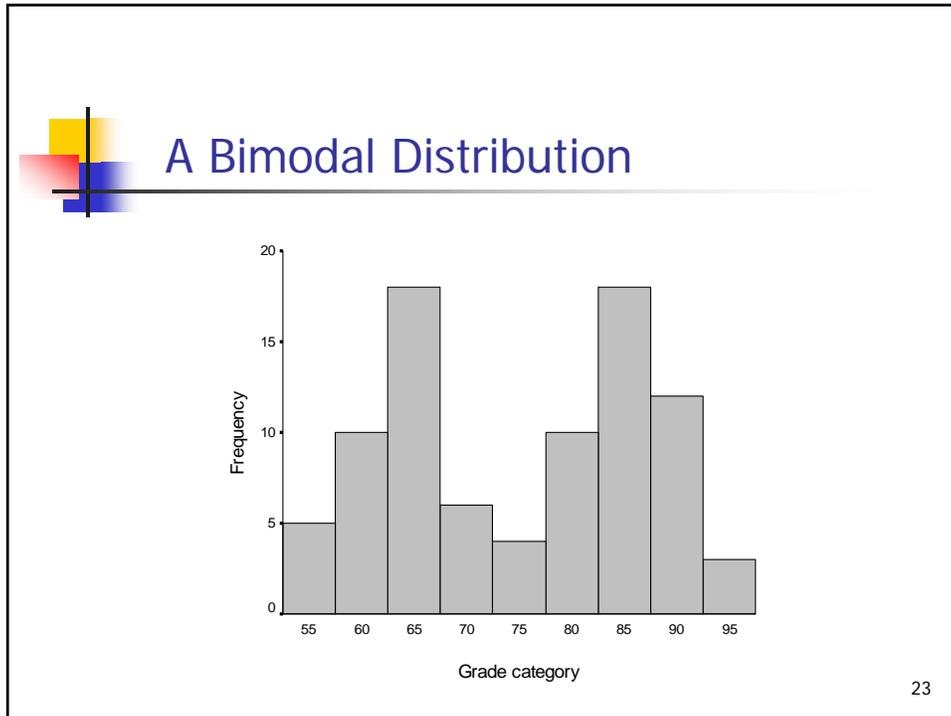
19

Histogram Showing a Normal Distribution



20





- ## Measures of Center
- Mean
 - Median
 - Mode
- 24



Measures of Center: Characteristics and Applications

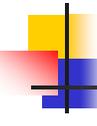
- *Mode*

- Most frequent score in a distribution
- Simplest measure of center
- Scores other than the most frequent not considered
- Limited application and value

- *Median*

- Central score in an ordered distribution
- More information taken into account than with the mode
- Relatively insensitive to outliers
- Prefer when data is skewed
- Used primarily when the mean cannot be used

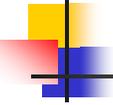
25



Decision rule

- If nominal, use mode
- Else if interval or ratio and approximately normal and no outliers, use mean
- Else, use median

26



Measures of Spread

- Std Deviation
- Inter-quartile range
- Range

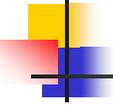
27



Measures of Spread: Characteristics

- *Range*
 - Subtract the lowest from the highest score in a distribution of scores
 - Simplest and least informative measure of spread
 - Scores between extremes are not taken into account
 - Very sensitive to extreme scores
- *Interquartile Range*
 - Less sensitive than the range to extreme scores
 - Used when you want a simple, rough estimate of spread

28



Measures of Spread: Characteristics

- *Variance*
 - Average squared deviation of scores from the mean
- *Standard Deviation*
 - Square root of the variance
 - Most widely used measure of spread

29

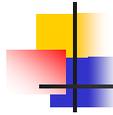


Decision rule

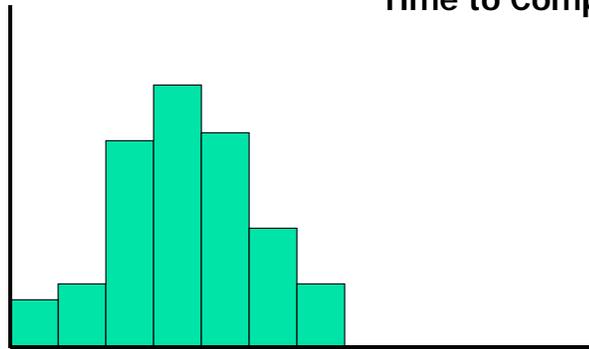
- If nominal, stop (no statistic)
- If interval or ratio and approximately normal and no outliers, use stddev
- Else use inter-quartile range

30

Which measures of center and spread?

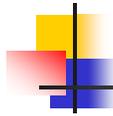


Time to Complete

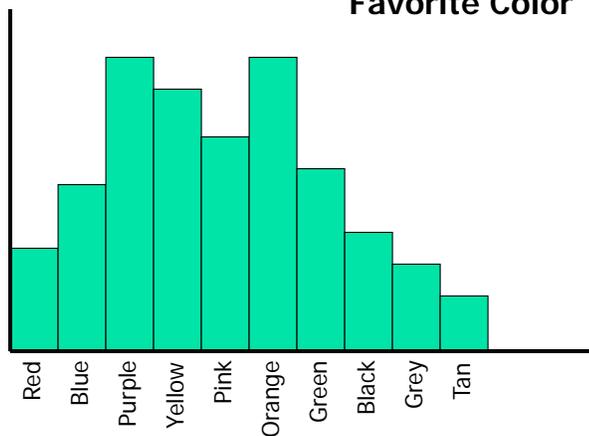


31

Which measures of center and spread?

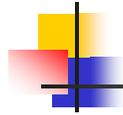


Favorite Color

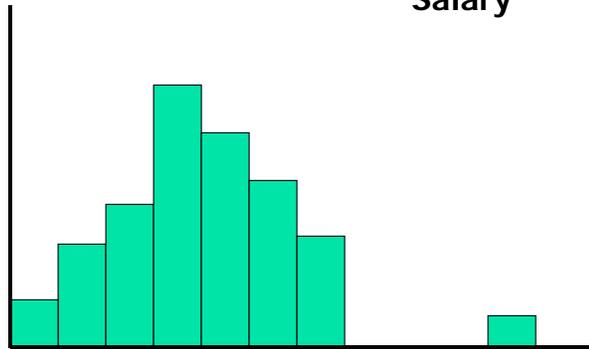


32

Which measures of center and spread?

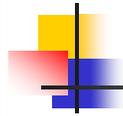


Salary

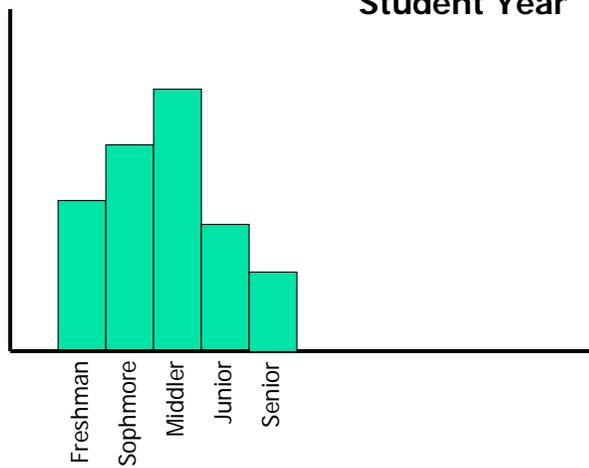


33

Which measures of center and spread?

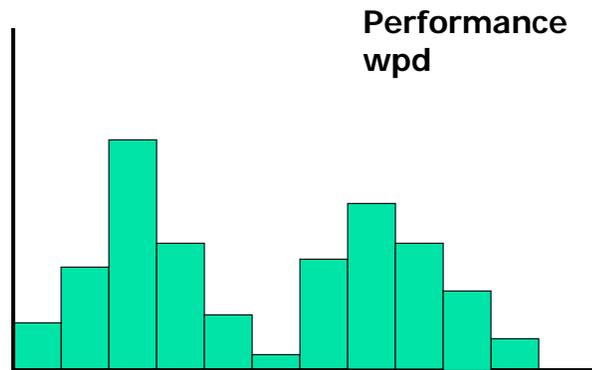


Student Year



34

Which measures of center and spread?



35

Recent controversy over analysis of scale measures

- Historically, have been treated as interval if they appear normal (i.e., with mean, stdev, and t-test)
- Some statisticians say NEVER. They are ordinal measures – must use median, no meaningful range measures, and non-parametric inferential statistics (e.g., Mann-Whitney)
- See
 - “Stats: We're Doing It Wrong” on ACM.ORG

36

Which measures of center and spread?

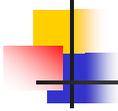


37

Inferential Analyses for Quantitative "Metric" studies

Users performed the set of standardized tasks in a significantly shorter time using interface FOO compared to interface BAR,
 $t(27)=3.4, p<.05$

38



Samples & Populations

- Population = everyone you care about
 - E.g., all of your primary stakeholders, all of your customers, all gamers in the US, etc
- Sample = everyone in your study

- Usually $|Sample| \ll |Population|$
- Inferential statistics let us make claims about the Population based on data from one or more Samples.
- If you could experiment on everyone in the population you would not need inferential statistics.

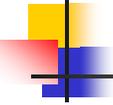
39



Typical case

- You are trying to demonstrate there is a difference between two metrics
 - E.g., performance with interface FOO vs. performance with interface BAR

40



Inferential Analyses

- Correlational
- Demonstrative
- Experimental
 - Between-subjects
 - Single factor, two-level
 - Single factor, N-level (for $N > 2$)
 - Two factor, N-level (for $N \geq 2$)
 - Within-subjects
 - Single factor, two-level

41



Types of Experimental Designs

Between-Subjects Design

- - Different groups of subjects are randomly assigned to the levels of your independent variable
 - Data are averaged for analysis
 - Use t-test for independent means
- Simplest: “single factor, two-level, between subjects” designs.

42

Types of Experimental Designs

Within-Subjects Design

- A single group of subjects is exposed to all levels of the independent variable
- Data are averaged for analysis
- aka "repeated measures design", "crossover design"
- Use t-test for dependent means aka "paired samples t-test"

- Simplest: "single factor, two-level, within subjects" designs.

- Note: If interval or ratio measures and approximately normal, use "t-test for dependent means" aka "paired samples t-test" to analyze.

43

Within-Subjects Designs

Benefits

- Can ask users to directly compare interfaces.
 - "Which did you like better?"
- More Power! *Why?*
 - Controls for all inter-subject variability
 - Randomized between-subjects design just balances the effects between groups

45

Within-Subjects Designs

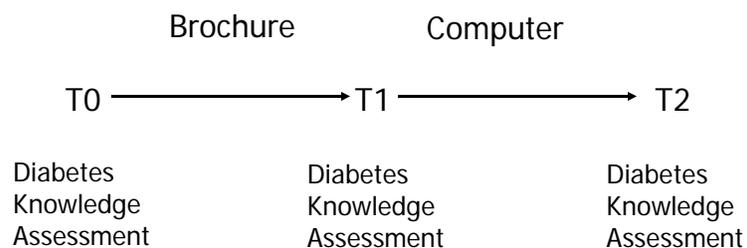
Disadvantages

- More demanding on subjects, especially in complex designs
- Subject attrition is a problem
- *Carryover effects*: Exposure to a previous treatment affects performance in a subsequent treatment

46

Carryover Example

- Embodied Conversational Agents to Promote Health Literacy for Older Adults



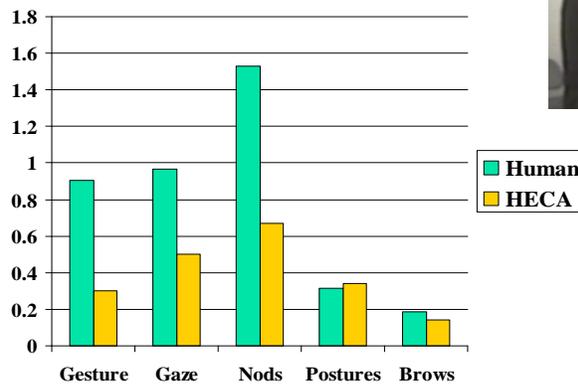
47

Some Sources of Carryover

- *Learning*
 - Learning a task in the first treatment may affect performance in the second
- *Fatigue*
 - Fatigue from earlier treatments may affect performance in later treatments
- *Habituation*
 - Repeated exposure to a stimulus may lead to unresponsiveness to that stimulus
- *Sensitization*
 - Exposure to a stimulus may make a subject respond more strongly to another
- *Contrast*
 - Subjects may compare treatments, which may affect behavior
- *Adaptation*
 - If a subject undergoes adaptation (e.g., dark adaptation), then earlier results may differ from later ones

48

Example Study: Handheld ECAs



Example – Best Design?

- You've just developed the "Matchmaker" – a handheld device that beeps when you are in the vicinity of a compatible person who is also carrying a Matchmaker.
- You evaluate the number of users who are married after six months of use compared to a non-intervention control group.



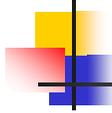
50

Example – Best Design?

- You've just developed "Reado Speedo" that reads print books using OCR and speaks them to you at twice your normal reading rate. You want to evaluate your product against the old fashioned way on reading rate, comprehension and satisfaction.



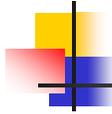
51



Example – Best Design?

- You've developed a new web-based help system for your email client. You want to compare your system to the old printed manual.

52

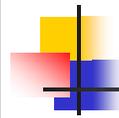


Inferential Analyses

- Correlational
- Demonstrative
- Experimental
 - Between-subjects
 - Single factor, two-level
 - Single factor, N-level (for $N > 2$)
 - Two factor, N-level (for $N \geq 2$)
 - Within-subjects
 - Single factor, two-level

53

Type of Errors in Inferential Statistics



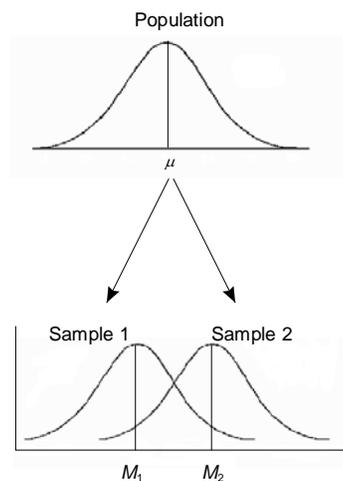
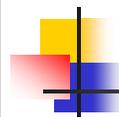
Research Hypothesis: There is a difference
(e.g., FOO better than BAR)

		"The Truth"	
		No diff	Diff
Conclude diff	Conclude diff	Type I Error	Correct Decision
	Conclude no diff	Correct Decision	Type II Error

'p' = Probability of Type I Error
The likelihood the difference observed is not real.

54

Relationship Between Population and Samples When a Treatment Had No Effect



'p' = Likelihood of this happening.

55

t-test for independent means

- Two samples, interval or ratio
- No other information about comparison distribution
- Assumptions:
 - Sample randomly selected from population.
 - The sampling distribution of means is normal
 - Variances of the two populations (whether they are the same or different) are the same.

56

Excel T.TEST, returns 'p'

Syntax

```
T.TEST(array1,array2,tails,type)
```

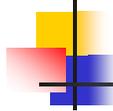
The T.TEST function syntax has the following arguments:

- **Array1** Required. The first data set.
- **Array2** Required. The second data set.
- **Tails** Required. Specifies the number of distribution tails. If tails = 1, T.TEST uses the one-tailed distribution. If tails = 2, T.TEST uses the two-tailed distribution.
- **Type** Required. The kind of t-Test to perform.

Parameters

If type equals	This test is performed
1	Paired
2	Two-sample equal variance (homoscedastic)
3	Two-sample unequal variance (heteroscedastic)

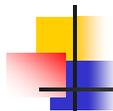
57



t-test

- If assumptions are followed, T.TEST returns 'p'
 - Likelihood of differences observed being due to chance, or error
 - Probability of Type I error
- If $p < \text{threshold}$ (conventionally 0.05), we say there is a significant difference
- If $p \geq \text{threshold}$, we conclude nothing (experiment was inconclusive)

58



Reporting results

- Significant results, scientific articles
 $t(df) = tscore, p < sig$
e.g., $t(38) = 4.72, p < .05$
- Non-significant results
e.g., $t(38) = 4.72, n.s.$
- Informal usability reports:
 - t-test for independent means indicated that performance with FOO was significantly better than performance with BAR, $p < .05$
 - t-test for independent means for performance with FOO vs. BAR was not significant.

59

Example t-test

User	UI	Time
1	FOO	5.1
2	FOO	3.5
3	FOO	4.2
4	FOO	1.7
5	FOO	4.9
6	FOO	6.4
7	BAR	2.1
8	BAR	4.1
9	BAR	1.1
10	BAR	2.8
11	BAR	3.2
12	BAR	1.4

61

Another note about "Power"

- For small, informal, qualitative, debugging usability tests
 - 5 users gets 80% of "usability defects"
- For quantitative usability experiments
 - Should do a "Power Analysis"
 - See online "Power Analysis Calculator"
 - Parameters: anticipated α , β (or power=1- β), effect size, tails
 - A "pilot study" with 12 users/condition will let you estimate effect size

62

Writing up your usability test results... use the template!

64



Short Usability Test Report for [Site]

Date of Report: [Month Day, Year]
 Date of Test: [Month Day, Year]
 Location of Test: [City, State]

Prepared for: [Name]
 Phone Number: [XXX-XXX-XXXX]
 Email: [name@address.gov]

Prepared by: [Name]
 Phone Number: [XXX-XXX-XXXX]
 Email: [name@address.gov]

Executive summary

NOTE: This section describes the main goal and rationale of the study. Briefly describe the scenarios that participants completed, how the sessions were conducted, and how many participants took part in the study. This section should also discuss overall trends, such as whether or not participants were able to complete all the tasks. Data should be reported as both a number of completed scenarios as well as a percentage. Is there a reason why tasks were completed or not? Be sure to give an overall impression (theme) about what the reader will encounter in the report.

65



HUMAN-COMPUTER INTERACTION **THIRD EDITION**

DIX
FINLAY
ABOWD
BEALE

chapter 11

user support



ALAN DIX, JANET FINLAY,
GREGORY D. ABOWD, RUSSELL BEALE

HUMAN-COMPUTER INTERACTION
THIRD EDITION

user support

- For non-trivial applications, users will need help.
- Design of help system is as important as the rest of the system.



ALAN DIX, JANET FINLAY,
GREGORY D. ABOWD, RUSSELL BEALE
**HUMAN-COMPUTER
INTERACTION**
3RD EDITION

user support

- Issues
 - different types of support at different times
 - implementation and presentation both important
 - all need careful design
- Types of user support
 - quick reference, task specific help, full explanation, tutorial
- Provided by help and documentation
 - help - problem-oriented and specific
 - documentation - system-oriented and general
 - same design principles apply to both



ALAN DIX, JANET FINLAY,
GREGORY D. ABOWD, RUSSELL BEALE
**HUMAN-COMPUTER
INTERACTION**
3RD EDITION

Requirements

- Availability
 - continuous access concurrent to main application
- Accuracy and completeness
 - help matches and covers actual system behaviour
- Consistency
 - between different parts of the help system and paper documentation
- Robustness
 - correct error handling and unpredictable behaviour
- Flexibility
 - allows user to interact in a way appropriate to experience and task
- Unobtrusiveness
 - does not prevent the user continuing with work

ALAN DIX, JANET FINLAY,
GREGORY D. ABOWD, RUSSELL BEALE
HUMAN-COMPUTER
INTERACTION
10th EDITION

Adaptive Help Systems

- Use knowledge of the context, individual user, task, domain and instruction to provide help adapted to user's needs.
- Problems
 - knowledge requirements considerable
 - who has control of the interaction?
 - what should be adapted?
 - what is the scope of the adaptation?

ALAN DIX, JANET FINLAY,
GREGORY D. ABOWD, RUSSELL BEALE
HUMAN-COMPUTER
INTERACTION
10th EDITION

Adaptive Help Systems



Problems?



Problems

- User modelling
- Domain and task modelling
- Initiative

- All research problems still
- Very very difficult to infer user intent

- Adaptable interfaces / help - OK
- Adaptive interfaces / help – very problematic



Designing user support

- User support is not an `add on'
 - should be designed integrally with the system.

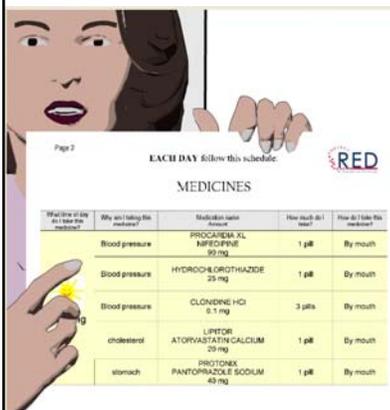
- Concentrate on content and context of help
- User testing!

Gaze in Document Explanation



- Collaborated with Akiko Aizawa's "Gaze/NLP Lab" – Tobii TX300
- Research Question: Can gaze information be used to aide in automated explanation of documents?

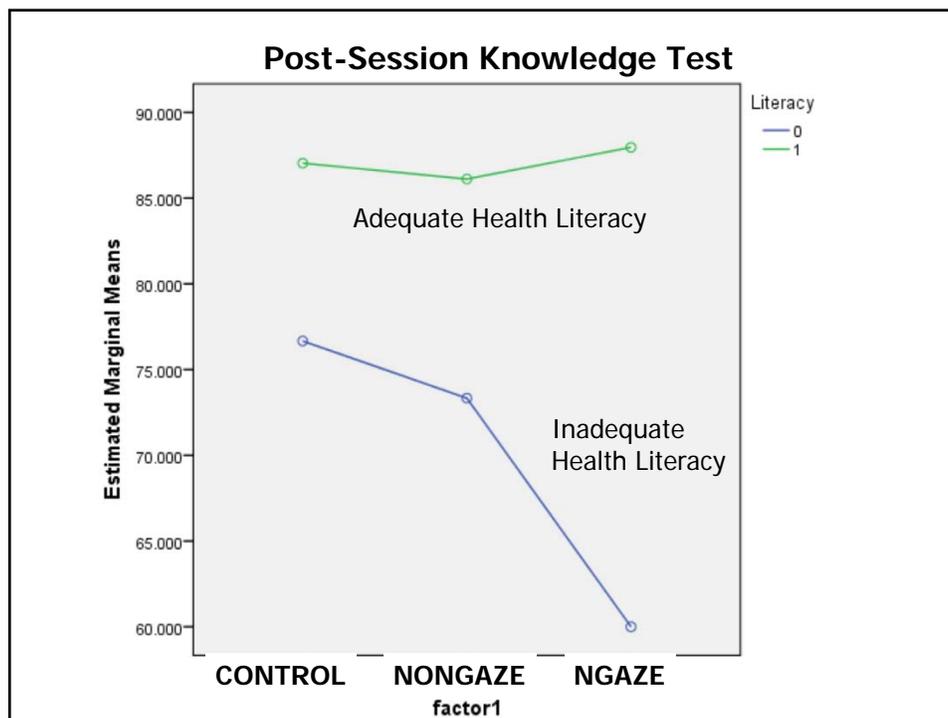
Gaze DocEx

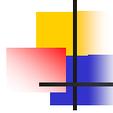


- Research Threads
 - Real-time assistance
 - Gazing at agent as request for help
 - Tracking of last point read as index for help
 - Gazing away from document as "lost engagement" cue
 - Assessment of Health Literacy based on gaze
 - Fixation time a function of word frequency

Study

- **Design:** 3-treatment, counterbalanced, within-subjects
- **Treatments:**
 - **CONTROL** S just reads document on own (no agent).
 - **NON-GAZE** Agent is present and can be asked for help (definitions of complex phrases) by explicitly asking via button pushes.
 - **GAZE** Agent is present and offers help if either 1) S fixates on a phrase too long; or 2) S gazes at agent (help offered on last phrase gazed at); or 3) clicks the help button (help offered on last phrase gazed at).
- **Subjects**
 - N=14 good data sets.
 - Age 22-67 (mean 50.0), 8 (57%) female, 5 (35.7%) low health literacy (REALM).





To Do

- Read 3 papers on game UI design
 - Guest lecture Monday – James Yin
- Finish P5 – paper prototyping