Measuring the User Experience

Collecting, Analyzing, and Presenting Usability Metrics

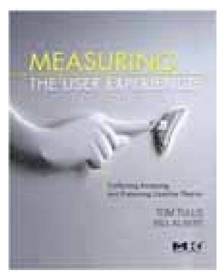
Chapter 2
Background

Tom Tullis and Bill Albert

Morgan Kaufmann, 2008
ISBN 978-0123735584

Introduction

- Purpose
  - Provide basic information about understanding data and designing usability study
  - Practical step-by-step guide to analyzing data without large number of formulas or complicated statistics
  - Use Microsoft Excel for calculations



Designing a Usability Study

- Well thought out test
  - Save you time and effort
  - Answer your research question clearly
- Questions
  - What type of participants do I need?
  - How many participant do I need?
  - Am I going to compare the data from a single group of participants or from several different groups?
  - Do I need to counterbalance (adjust for) the order of tasks?

3

Designing a Usability Study

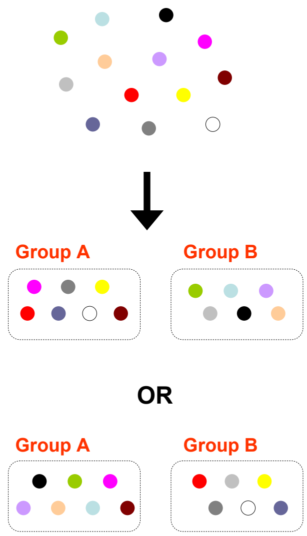
Selection Participants

- Selecting Participants
  - Decision based on cost, availability, appropriateness, and study goals
- How well should the participants match the target audience?
  - If application for doctors, then try to get practicing physicians as participants
  - If you only have participants who are close approximations, then be aware of the limitation of the collected data
  - Many of statistics assume the sample data reflects the larger population

4

Designing a Usability Study
Selection Participants

- Are you going to divide the data by different types of participants?
  - Self-reported expertise in some domain (novice, intermediate, expert)
  - Frequency of use (e.g. number of website interactions per month)
  - Amount of experience with something relevant (days, months, years)
  - Demographics (gender, age, location, etc.)
  - Activities (use of a particular function or features)



5

Designing a Usability Study
Selection Participants

- What is your sample strategy?
  - How will you be able to generalize your findings to larger populations

- Sampling Techniques
  - Random Sampling
    - Everyone in population has roughly equal probability of being selected
  - Systematic Sampling
    - Participants selected based on predefined criteria
    - Every 10th person on list, every 100th person through a turnstile
  - Stratified Sampling
    - Create subsamples of the entire population and ensure sample size for each subgroup achieved
    - Ensure sample reflects the population – 50% male and 50% female, 20% over age 65
  - Samples of convenience
    - Include anyone willing to participate
    - Important to know how well this sample reflects the general population and be aware of any special biases that is reflected in the feedback

6

Designing a Usability Study
Sample Size

- How many participants are enough?
  - No equation/rule that says if you have x data will/won't be valid
  - Based on goals of usability study and tolerance for a margin or error
- Usability Goal
  - Identify major design issues as part of an iterative process
    - 3 or 4 users can provide useful data
    - Won't identify all (or most) of the usability issues, but can find significant ones
  - Product close to release, find remaining usability issues
    - Need more participants

Designing a Usability Study
Sample Size

- How much error are you willing to accept?
  - Last time we talked about changing confidence interval to meet the desired confidence level
  - Table illustrates that given an 80% success rate with a 95% confidence level, how the interval changes with user size
    - As sample size increases, the upper and lower bounds move closer together

Number Successful	Number of Participants	Lower 95% Confidence	Upper 95% Confidence
4	5	36 %	98 %
8	10	48 %	95 %
16	20	58 %	95 %
24	30	62 %	91 %
40	50	67 %	89 %
80	100	71 %	86 %

8 out of 10 participants successfully completed a task

Will 80% of larger population also be able to complete the task?

No -- with a 95% confidence level, somewhere between 48% and 95% of the people in the larger population succeed

From Lecture 9

Fraction of Data	# of Standard Deviations from Mean
50.0%	0.674
68.3%	1.000
90.0%	1.645
95.0%	1.960
95.4%	2.000
98.0%	2.326
99.0%	2.576
99.7%	3.000

Designing a Usability Study
Within-Subjects or Between-Subjects Study

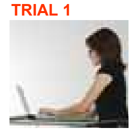
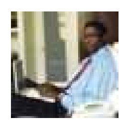
- Are you going to be comparing different data for each participant or data from each participant to other participants?


9

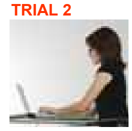
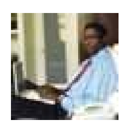
Designing a Usability Study
Within-Subjects or Between-Subjects Study

- Within-subjects (repeated-measures)
  - Comparing different data for each participant across several trials
  - Commonly used when you want to evaluate how easily participant can learn to use a product
- Benefits
  - Small sample sizes
  - Participant being compared to self so difference in data isn't because of differences in participants
- Weakness
  - "Carryover effects" – performance in one condition impacts performance in another condition
  - Example, practice may improve performance where fatigue may decrease performance
  - Need to counterbalance for this effect as you design the study

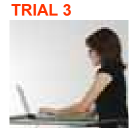
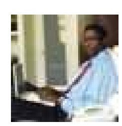
TRIAL 1

TRIAL 2

TRIAL 3

How many errors does User A have in Trial 1, 2, and 3?

How quickly can user B complete a task in Trial 1, 2, and 3?

10

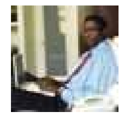
Designing a Usability Study

Within-Subjects or Between-Subjects Study

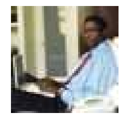
- Between-subjects
  - Compare data from each participant to other participants
  - Used to compare results from different participants
    - Satisfaction between novice and experts
    - Task completion time for younger vs. older users
    - Users randomly assigned to different groups, given different prototypes
- Benefit
  - No “carryover effects” – impacts both groups
- Weakness
  - Need a larger sample size, more variance across participants

TASK 1


TASK 2

TASK 3

Did User A complete the task faster than User B?

Was User B more satisfied with the inter

11

Designing a Usability Study

Counterbalancing

- Order in which participants perform tasks can impact results
  - Performance increases over time
  - Is Task 5 just easier than Task 1?
  - Was there learning between Task 1 and Task 5?
- Counterbalancing
  - Changing the order in which different tasks are performed
  - Randomly shuffle tasks, or create different orders so everyone is different
- When not to counterbalance?
  - Tasks are totally unrelated to each other, performing one task will not help with the next
  - Natural order of tasks is present, changing order would not make sense

Participant	First Task	Second Task	Third Task	Fourth Task
P1	T1	T2	T3	T4
P2	T3	T1	T4	T2
P3	T2	T4	T1	T3
P4	T4	T3	T2	T1

12

Designing a Usability Study
Independent and Dependent Variables

- Need to have a clear idea of what you plan to manipulate and what you plan to measure
- Independent Variables
  - Aspect of a study that you manipulate
  - Chosen based on research question
- Dependent Variables (outcome or response variable)
  - Describes what happened as a result of the study
  - Something you measure as the result, or as dependent on, how you manipulate the independent variables

Research Question : Differences in performance between males and females

Independent variable : Gender

Dependent variable : Task completion time

Research Question : Differences in satisfaction between novice and expert users

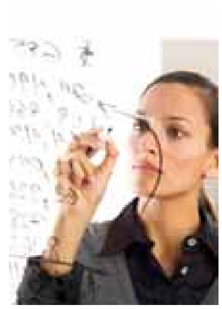
Independent variable : Experience level

Dependent variable : Satisfaction

13

Types of Data

- Data exists in many forms
  - To analyze data need to understand types of data
  - What you can and can't do with different types of data
- Four general types of data
  - Nominal Data
  - Ordinal Data
  - Interval Data
  - Ration Data



14

Types of Data

- Nominal Data
  - Unordered groups or categories
  - Without order, cannot say one is better than another
- May provide characteristics of users, independent variables that allow you to segment data
  - Windows versus Mac users
  - Geographical location
  - Males versus females
- What about dependent variables?
  - Number of users who clicked on A vs. B
  - Task success
- Usage
  - Counts and frequencies

Types of Data

- Ordinal Data
  - Ordered groups and categories
  - Data is ordered in a certain way but intervals between measurements are not meaningful
- Ordinal data comes from self-reported data on questionnaires
  - Website rated as excellent, good, fair, or poor
  - Severity rating of problem encountered as high, medium, or low
- Usage
  - Looking at frequencies
  - Calculating average is meaningless (distance between high and medium may not be the same as medium and low)

American Film Institute (AFI) Top 100 Movies

1. CITIZEN KANE
2. CASABLANCA
3. THE GODFATHER
4. GONE WITH THE WIND
5. LAWRENCE OF ARABIA
6. THE WIZARD OF OZ
7. THE GRADUATE
8. ON THE WATERFRONT
9. SCHINDLER'S LIST
10. SINGIN' IN THE RAIN

• • •

20. ONE FLEW OVER THE CUCKOO'S NEST

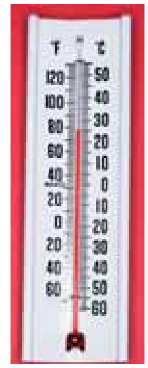
Is "Singin' in the Rain" twice as good as "One Flew Over the Cuckoo's Nest"?

Types of Data

- Interval Data
  - Continous data where differences between the measurements are meaningful
  - Zero point on the scale is arbitrary

- System Usability Scale (SUS)
  - Example of interval data
  - Based on self-reported data from a series of questions about overall usability
  - Scores range from 0 to 100
    - Higher score indicates better usability
    - Distance between points meaningful because it indicates increase/decrease in percieved usability

- Usage
  - Able to calculate descriptive statistics such as average, standard devication, etc.
  - Inferal statistics can be used to generalize a population



60°C vs 50 °C is meaningful

Does 0°C mean no heat?

0°C meaningful – freezing point of water but it's arbitrary

17

Types of Data

- Severity rating example
  - Is the data purely ordinal or is it interval?

○ **Poor** ○ **Fair** ○ **Good** ○ **Excellent**

Explicit labels on items makes data ordinal

Poor ○ ○ ○ ○ ○ **Excellent**

Leaving interval labels off and using anchors (labels on the end) makes data more "interval-like"

Poor ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ **Excellent**

With 10 points along the scale, make it more obvious that the distances between data points along a scale are equal

Interval data

18

Types of Data

- Ratio Data
  - Same as interval data with the addition of absolute zero
  - Zero has inherit meaning
- Example
  - Difference between a person of 35 and a person 38 is the same as the difference between people who are 12 and 15
  - Time to completion, you can say that one participant is twice as fast as another
- Usage
  - Most analysis that you do work with ratio and interval data
  - Geometric mean is an exception, need ratio data

19

Metrics and Data

- Choosing right statistics is important
  - Using wrong test, end up with incorrect conclusion
  - Invalidate your entire usability test
- Type of data examined will dictate different test
  - Commonly used statistical tests based on data type provided below

Data Type	Common Metrics	Statistical Procedure
Nominal (categories)	Task success (binary), errors (binary), top-2-box scores	Frequencies, crosstabs, Chi-squared
Ordinal (ranks)	Severity ratings, rankings (designs)	Frequencies, crosstabs, chi-square, Wilcoxon rank sum tests, Spearman rank correlation
Interval	Likert scale data, SUS scores	All descriptive statistics, *t*-tests, ANOVAs, correlation, regression analysis
Ratio	Completion time, time (visual attention), average task success (aggregated)	All descriptive statistics (including geometric means), *t*-tests, ANOVAs, correlation, regression analysis

20

Descriptive Statistics

- Descriptive Statistics
  - Describes the data without saying anything about the larger population
- Inferential statistics
  - Able to draw conclusions about larger population beyond the sample
- Analyze data using Excel
  - Tools > Data Analysis
  - "Descriptive Statistics"

	A	B	C	D	E
1	Participant	Task Time			
2	P1	34			
3	P2	33		Mean	35.08333333
4	P3	28		Standard Error	3.246112671
5	P4	44		Median	33.5
6	P5	46		Mode	22
7	P6	21		Standard Deviation	11.24486415
8	P7	22		Sample Variance	126.4469697
9	P8	53		Kurtosis	-1.321525965
10	P9	22		Skewness	0.251441718
11	P10	29		Range	32
12	P11	39		Minimum	21
13	P12	50		Maximum	53
14				Sum	421
15				Count	12
16				Confidence Level (95.0%)	7.144645813
17					

Raw task time for 12 users

Descriptive statistics provided by excel

21

Descriptive Statistics

- Measures of Central Tendency
  - Middle, or central part of data
  - Mean
    - Average time for user to complete task
    - Just over 35 seconds
  - Median
    - Midway point in the distribution
    - Half of the users were faster than 33.5 seconds
    - Half of the users were slower than 33.5 seconds
    - Usage – salaries, executive salaries skew mean so much that average salary appears higher than the majority really are
  - Mode
    - Most commonly occurring value
    - Two participants finished the task in 22 sec
    - More useful in subjective rating scales

D	E
Mean	35.08333333
Standard Error	3.246112671
Median	33.5
Mode	22
Standard Deviation	11.24486415
Sample Variance	126.4469697
Kurtosis	-1.321525965
Skewness	0.251441718
Range	32
Minimum	21
Maximum	53
Sum	421
Count	12
Confidence Level (95.0%)	7.144645813

22

Descriptive Statistics

- Measures of Variability
  - Shows how the data are spread or dispersed across the range of all the data
  - Variability important in confidence
    - Greater the variability, the less dependable the data is to understanding general population
  - Range
    - Distance between the minimum and maximum data points
    - Helps to identify outliers
  - Variance
    - How spread out the data is relative to the average or mean
    - Equation provided on right
  - Standard Deviation
    - Slightly easier to interpret, unit of SD is same as data
    - Square root of variance

D	E
Mean	35.08333333
Standard Error	3.246112671
Median	33.5
Mode	22
Standard Deviation	11.24486415
Sample Variance	126.4469697
Kurtosis	-1.321525965
Skewness	0.251441718
Range	32
Minimum	21
Maximum	53
Sum	421
Count	12
Confidence Level (95.0%)	7.144645813

individual data point

mean of all values

$$S^2 = \frac{\Sigma(X-M)^2}{N-1}$$

Sample size

23

Descriptive Statistics

- Confidence Intervals
  - Range that estimates the true population value for a statistic

D	E
Mean	35.08333333
Standard Error	3.246112671
Median	33.5
Mode	22
Standard Deviation	11.24486415
Sample Variance	126.4469697
Kurtosis	-1.321525965
Skewness	0.251441718
Range	32
Minimum	21
Maximum	53
Sum	421
Count	12
Confidence Level (95.0%)	7.144645813

You want to be 95% certain about the mean for the entire population is

Data shows that population mean is 35 ± 7 sec or 28 to 42 seconds

24

Comparing Means

- Compare different means (interval and ratio data only)
  - Determine if one design has higher satisfaction than another design
  - Error rates are higher for one group compared to another
- Again we can use excel do comparison
  - Look at various ways of comparing means

Comparing Means

- Independent Samples
  - Comparing means across different set of participants
- Examples
  - Compare data for men and women
  - Compare satisfaction rates for novice and expert users
- Analyze using Excel
  - Tools > "*t*-Test: Two Samples Assuming Equal Variance"

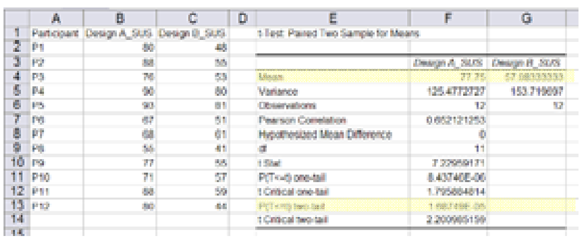
	A	B	C	D	E	F
1	Expert time	Novice time		t-Test: Two-Sample Assuming Equal Variances		
2	34	45				
3	33	48			Expert time	Novice time
4	28	53		Mean	35.08333333	49.33333333
5	44	66		Variance	128.4469697	229.6969697
6	46	67		Observations	12	12
7	21	35		Pooled Variance	178.0719697	
8	22	39		Hypothesized Mean Difference	0	
9	53	21		df	22	
10	22	34		t Stat	-2.615728765	
11	29	55		P(T<=t) one-tail	0.007892632	
12	39	59		t Critical one-tail	1.717144335	
13	50	70		P(T<=t) two-tail	0.015785265	
14				t Critical two-tail	2.073873058	

Experts are faster (35 sec) compared to novices (49 sec)

p-value is the probability that random sampling would lead to a difference between sample means as large (or larger) than you observed

Comparing Means

- Paired-Samples
  - Comparing means within the same set of participants
- Examples
  - Same set of participants perform tasks using Prototype A and then Prototype B
  - Measure variables such as self-reported ease of use and time
- Analyze using Excel
  - Tools > "*t*-Test: Paired Two Samples for Means"



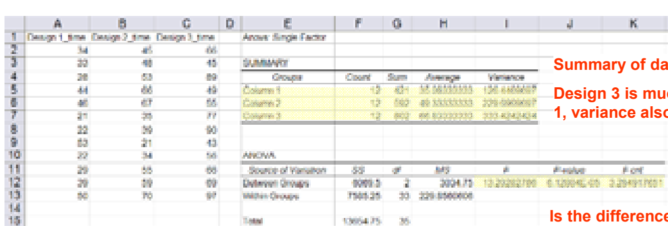
	A	B	C	D	E	F	G
1	Participant	Design A_SUS	Design B_SUS		t-Test: Paired Two Sample for Means		
2	P1	80	48				
3	P2	88	55			Design A_SUS	Design B_SUS
4	P3	76	53		Mean	77.75	57.08333333
5	P4	90	80		Variance	125.4772727	153.719697
6	P5	93	81		Observations	12	12
7	P6	67	51		Pearson Correlation	0.652121253	
8	P7	68	61		Hypothesized Mean Difference	0	
9	P8	56	41		df	11	
10	P9	77	55		t Stat	7.22959171	
11	P10	71	57		P(T<=t) one-tail	8.43740E-06	
12	P11	88	59		t Critical one-tail	1.795884814	
13	P12	80	44		P(T<=t) two-tail	1.68748E-05	
14					t Critical two-tail	2.200985159	

Important to look at mean

p-value indicates there is a significant difference between the two designs

Comparing Means

- Comparing More than Two Samples
  - Analysis of variance (ANOVA) determines if there is a significant difference across more than two groups
- Analyze using Excel
  - Tools > "Data Analysis"
  - "ANOVA: Single Factor"



	A	B	C	D	E	F	G	H	I	J	K
1	Design 1_time	Design 2_time	Design 3_time		Anova: Single Factor						
2	34	45	66								
3	33	48	45		SUMMARY						
4	28	53	89		Groups	Count	Sum	Average	Variance		
5	44	66	49		Column 1	12	421	35.08333333	136.4469697		
6	46	67	55		Column 2	12	592	49.33333333	229.9090909		
7	21	35	77		Column 3	12	802	66.83333333	333.4242424		
8	22	39	90								
9	53	21	43								
10	22	34	56		ANOVA						
11	29	55	66		Source of Variation	SS	df	MS	F	P-value	F crit
12	39	59	69		Between Groups	6069.5	2	3034.75	13.20282788	6.12004E-05	3.284917651
13	50	70	97		Within Groups	7585.25	33	229.8560606			
14											
15					Total	13654.75	35				

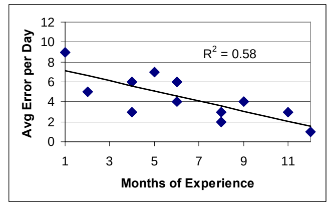
Summary of data

Design 3 is much slower than Design 1, variance also higher

Is the difference significant?
F crit says need 3.28 to achieve significant, our F value is 13.2
Means there is a significant effect due to the "design" variable

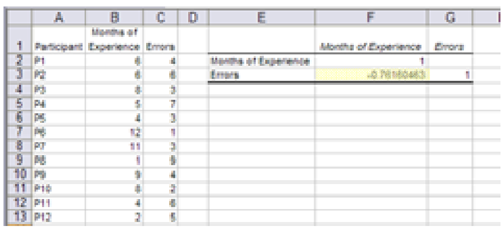
Relationships between Variables

- Observe that what participants say and what they do don't always correspond
  - Participants struggle to complete a few tasks
  - Give good ratings when asked how easy or difficult task is
- How to investigate relationship between variables (or lack thereof)?
- Analyze using Excel
  - Create scatter plot, add trend line and R^2 value
  - Tools > "Data Analysis"
  - "Correlation"



Negative relationship – as one variable increase (months) the other variable decreases (errors)

Trend line helps to better visualize relationship



	A	B	C	D	E	F	G
1	Participant	Months of Experience	Errors			Months of Experience	Errors
2	P1	6	4		Months of Experience	1	
3	P2	6	6		Errors	-0.76160463	1
4	P3	8	3				
5	P4	5	7				
6	P5	4	3				
7	P6	12	1				
8	P7	11	3				
9	P8	1	9				
10	P9	9	4				
11	P10	8	2				
12	P11	4	6				
13	P12	2	5				

R value is correlation coefficient

Measures the strength of the relationship between two variables (-1 to 1)

Stronger the relationship the closer to -1 or 1

29

Nonparametric Tests

- How to analyze nominal and ordinal data?
  - Is there a significance between men and women for success and failure on a particular task?
  - Is there a difference in the way novice, intermediate, and expert users ranked different websites?
- Need nonparametric test
  - Make different assumptions than in previous tests (e.g. normal distribution and equal variances)
- Chi-square Test
  - Used to compare categorical (nominal) data

30

Nonparametric Tests

- Chi-square Test (continued)
  - Want to see if there is a significant difference between three groups of users
    - Novices, intermediates, and experts
  - 60 participants, 20 per group
  - Measure task success or failure for a single task
    - Novice – 6 out of 20 successful
    - Intermediate – 12 out of 20 successful
    - Expert – 18 out of 20 successful

	A	B	C
1	**Group**	**Observed**	**Expected**
2	Novice	6	11
3	Intermediate	9	11
4	Expert	18	11
5	**Total**	**33**	**33**
6			
7		Chi-test	0.028856

Is there a statistically significant difference between groups?

Are the difference between the observed and expected due to chance

Likelihood that this distribution is due to chance – 2.9%

Number is less than 0.05 (95% confidence) we can say there is a difference in success rates between the groups

31

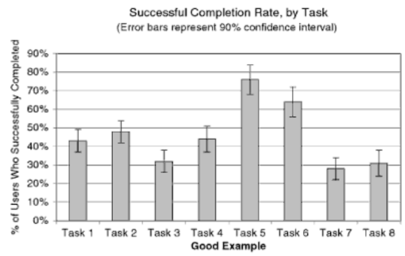
Presenting Your Data Graphically - General Tips

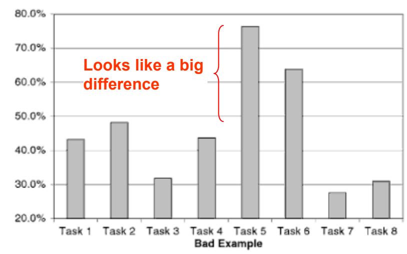
- Label the axes and unit
  - Does the scale of 0 to 100 percent represent task completion time? May not be obvious to audience
  - What are the time units?
  - If axis already makes it clear (e.g. “Task1”, “Task2”, etc.) don’t need to add another label “Tasks”
- Don’t imply more precision in your data then they deserve
- Don’t use color alone to convey information
  - Color is a good principle for designing information display
  - What is the report is in black and white? Someone is color blind?
- Display labels horizontally whenever possible
  - Exception – title for vertical axis is normally horizontal
- Show confidence intervals whenever possible
  - Mainly applies to bar graphs and line graphs that are presenting means of individual participant data
- Don’t overload your graphs
  - Just because you can create a single graph that shows all of the ratings for 20 different tasks doesn’t mean you should
- Be careful with 3D graphs
  - Does it really help? In some cases it’s harder to see the values being plotted

32

Presenting Your Data Graphically

- Column or Bar Graphs
- Appropriate to present values of continuous data for discrete items or categories
  - If both continuous, use line graph
- Axis for continuous variable should start a zero
  - Length of bar indicates values
  - Don't artificially manipulate lengths
- Don't let maximum value for continuous variable go higher than it is theoretically able to
  - 110% percent

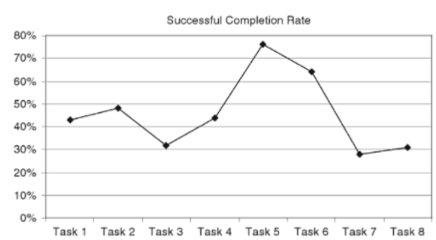




33

Presenting Your Data Graphically

- Line versus bar graph
  - In using a line graph ask if the places along the line between data points make sense
  - If not, use a bar graph

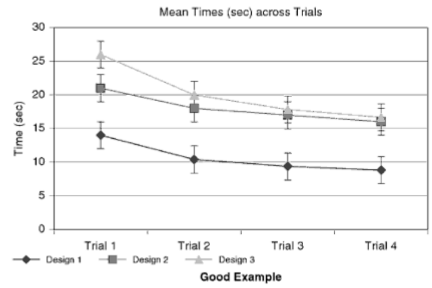


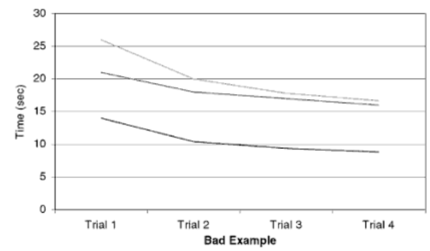
Previous figure as a line graph

34

Presenting Your Data Graphically

- Line Graphs
  - Show trends in continuous variable, often over time
- Appropriate when you want to present values of one continuous variable as a function of another
- Show data points
  - Actual points matter, not the lines
- Use lines that have sufficient weights (points too)
- Include a legend if more than one line used
- Vertical axis normally starts at zero, but not as important with a line graph compared to bar graph

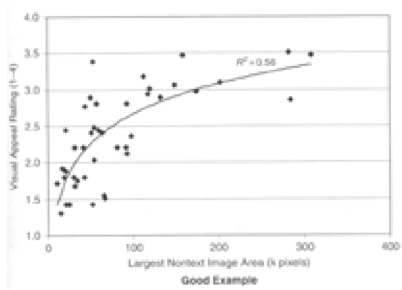


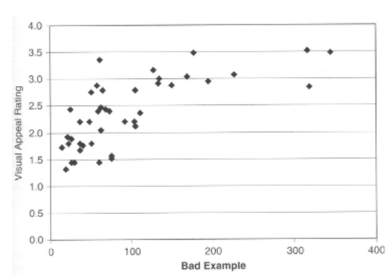


35

Presenting Your Data Graphically

- Scatterplots
  - Shows pairs of values
- Must have paired values you want to plot
- Normally both variables are continuous
- Use appropriates scales
  - Vertical axis can't be lower than 1.0, start the scale at that point
- Purpose is normally to show a relationship between the two variables
  - Helpful to add trend line
  - May also include R^2 to indicate goodness of fit

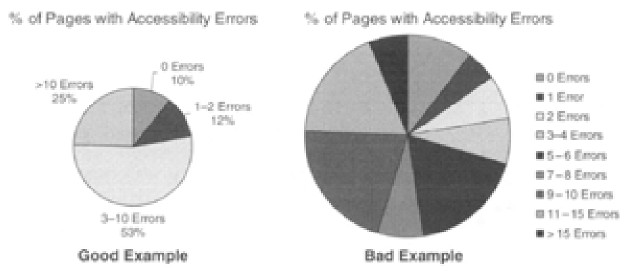




36

Presenting Your Data Graphically

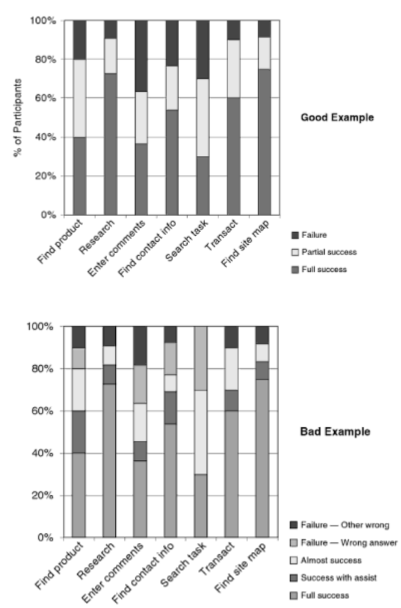
- Pie Charts
  - Illustrate parts or percentages of a whole
- Appropriate when the parts add up to 100 percent
  - To account for all cases you can add an “other” category
- Minimize number of segments
  - Difficult to make sense
- Include percentage and label for each segment



37

Presenting Your Data Graphically

- Stacked Bar Graphs
  - Basically multiple pie charts shown in bar form
- Appropriate when parts for each item in a series add up to 100 percent
- Items in series are normally categorical
- Minimize number of segments in each bar
  - More than three can be difficult to interpret
- Make use of color coding conventions that audience is likely to be familiar with
  - In US – green is good, yellow is marginal, and red is bad



38

Summary

- Important to know your data
- Consider questions about how to select participants for your study, how to order tasks, what participants perform what tasks, and how many participants you need to get reliable data
- Knowing your data is critical when analyzing your results. The specific type of data you have will dictate what statistics you can (and can't) perform
- When presenting your data graphically, use the appropriate types of graphs
  - Bar graphs for categorical data and line graphs for continuous data
  - Pie charts or stack bar graphs when the data sum to 100 percent.

39

Summary

- Nominal data are categorical
  - Nominal data are usually expressed as frequencies or percentages
  - Chi-square test can be used when you want to learn whether the frequency distribution is random or there is some underlying significance to the distribution pattern
- Ordinal data are rank orders
  - Ordinal data are also analyzed suing sequences, and the distribution patterns can be analyzed with a chi-square test
- Interval data are continuous data where the interval between each point are meaningful with without a natural zero
  - Can be described by means, standard deviations, and confidence intervals
  - Means can be compare to each other for the same set of users (paired samples t-test) or across different users (independent sample t-test)
  - ANOVA can be used to compare more than two sets of data
  - Relationships between variable can be examined through correlations.
- Ration data are the same as interval data but with a natural zero
  - Same statistics that apply to interval data also apply to ratio data

40