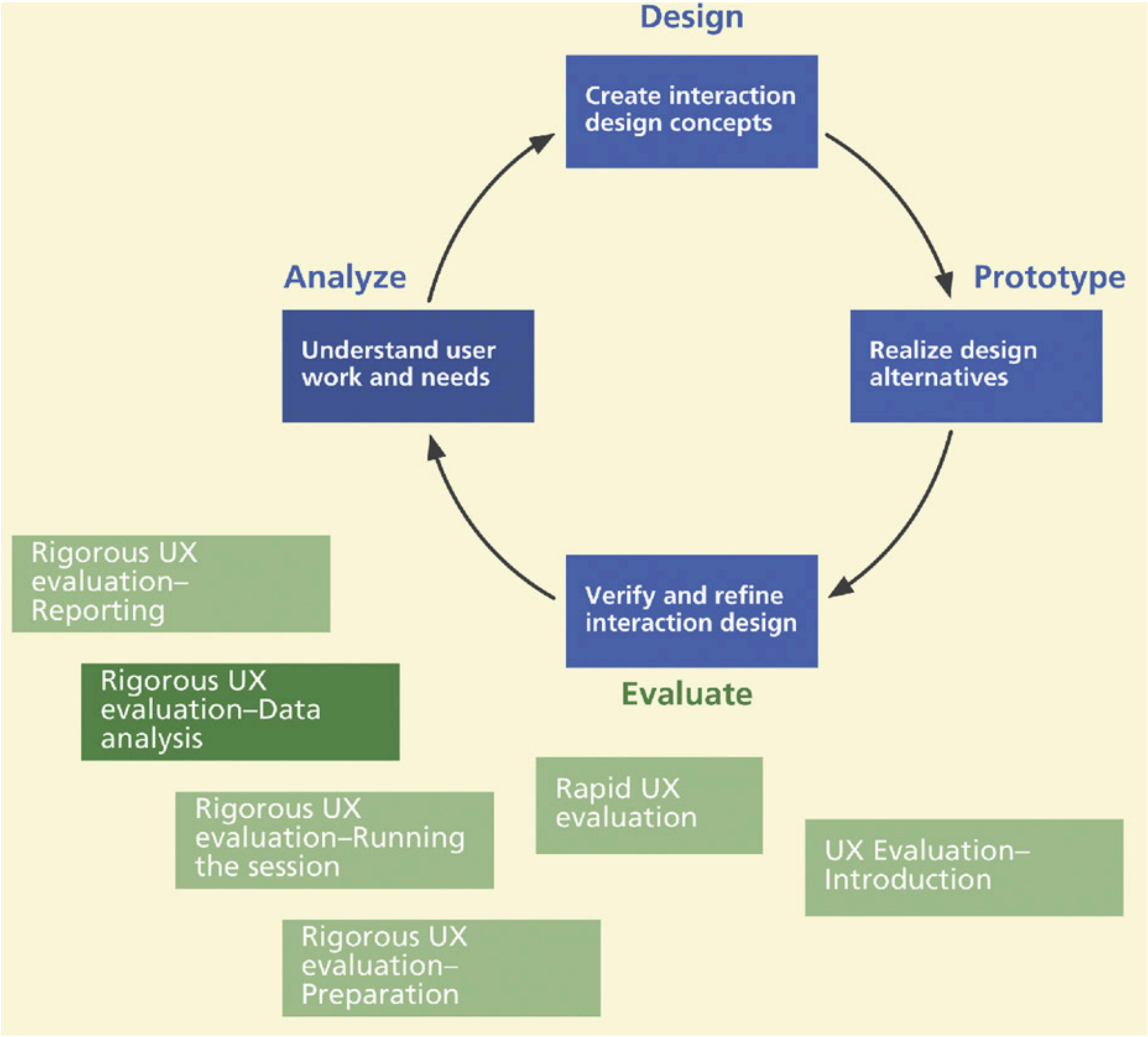
Rigorous Evaluation

Analysis and Reporting

Structure is from A Practical Guide to Usability Testing by J. Dumas, J. Redish

Software Engineering
Rochester Institute of Technology



Software Engineering
Rochester Institute of Technology

Results from Usability Tests

- Quantitative data:
  - Performance data - times, error rates, etc.
  - Subjective ratings, from post test surveys
- Qualitative data:
  - Participant comments from notes, surveys, etc.
  - Test team observations, notes, logs
  - Background data from user profiles, pretest surveys and questionnaires

Work Role: User Class	UX Goal	UX Measure	Measuring Instrument	UX Metric	Baseline Level	Target Level	Observed Results	Meet Target?
Ticket buyer: Casual new user, for occasional personal use	Walk-up ease of use	Initial user performance	BT1: Buy special event ticket	Average time on task	3 min as measured at the kiosk	2.5 min	3.5 min	No
Ticket buyer: Casual new user, for occasional personal use	Walk-up ease of use for new user	Initial user performance	BT2: Buy movie ticket	Average number of errors	<1	<1	2	No
Ticket buyer: Casual new user, for occasional personal use	Initial customer satisfaction	First impression	Questions Q1–Q10 in questionnaire XYZ	Average rating across users and across questions	7.5/10	8/10	7.5	No

Summarize and Analyze Test Data

- Qualitative data …
  - For survey multiple choice questions, count responses or average (if large groups)
  - For survey open-questions/comments, interviews, and observations …
    - Identify critical comments
    - Group into meaningful categories (+ or – for a particular task/ screen)
- Quantitative data …
  - Tabulate
  - Use statistics for analysis when appropriate

Look for Data Trends/ Surprises

- Examine the quantitative data …
  - Trends or patterns in task completion, error rates, etc.
  - Identify extremes, outliers
- Outliers - what can they tell us, ignore at your peril
  - Non-usability anomaly such as technical problem?
  - Difficulties unique to one participant?
  - Unexpected usage patterns?
- Correlate with qualitative data such as written comments – why?
- If appropriate compare old versus new program versions, different user groups

Examining the Data for Problems

- *Have you achieved the usability goals*
  – learnable, memorable, efficient, understandable, satisfying …?
- Unanticipated usability problems?
  - Usability concerns that are not addressed in the design
- Have the quantitative criteria that you have set been met or exceeded?
- Was the expected emotional impact observed?

Task and Error Analysis

- What tasks did users have the most problems with (usability goals not met)?
- Conduct error analysis
  - Categorize errors/task by type
    - Requirement or design defect (or bug)
  - % of participants performing successfully within the benchmark time
  - % of participants performing successfully regardless of time (with or without assistance)
    - If low then BIG problems

Prioritize Problems

- Criticality = Severity + Probability
- Severity
  - 4: Unusable – not able/want to use that part of product due to design/implementation
  - 3: Severe – severely limited in ability to use product (hard to workaround)
  - 2: Moderate – can use product in most cases, with moderate workaround
  - 1: Irritant – intermittent issue with easy workaround; cosmetic
- Factor in scope– local to a task (e.g., on screen) versus global to the application (e.g., main menu)

Rubin, Jeffrey, and Chisnell, Dana. Handbook of Usability Testing : How to Plan, Design, and Conduct Effective Tests (2). Hoboken, US: Wiley, 2008. ProQuest ebrary.

Prioritize Problems (cont.)

- Probability of occurrence

Frequency ranking	Estimated frequency of occurrence
4	Will occur ≥90% of the time the product is used
3	Will occur 51–89% of the time
2	Will occur 11–50% of the time
1	Will occur ≤10% of the time

- When done – sort by Criticality (priority)

Rubin, Jeffrey, and Chisnell, Dana. Handbook of Usability Testing : How to Plan, Design, and Conduct Effective Tests (2). Hoboken, US: Wiley, 2008.

Statistical Analysis

- Summarize quantitative data to help discover patterns of performance and preference, and detect usability problems
- Descriptive and inferential techniques

Descriptive Statistics

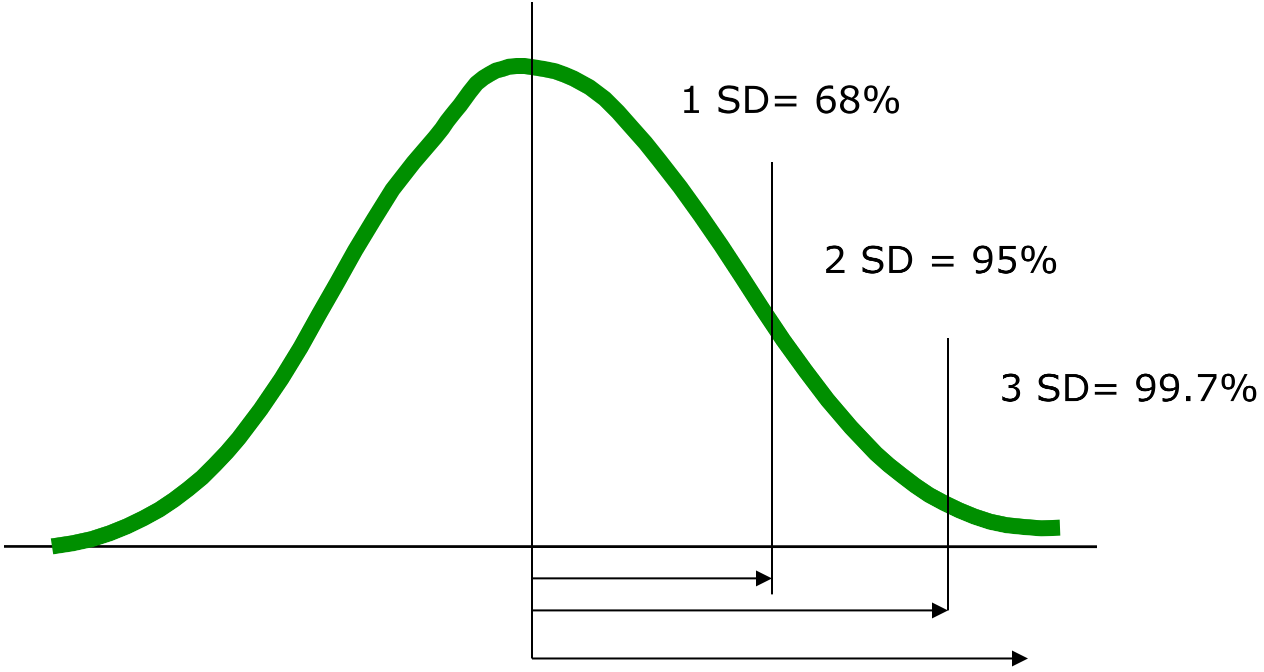
- Describe the properties of a specific data set
- Measures of central tendency (single variable)
  - Frequency distribution (e.g., of errors)
  - Mean (average), median (middle value), mode (most frequent value in a set)
- Measures of spread (single variable)
  - Amount of variance from the mean, standard deviation
- Relationships between pairs of variables
  - Scatterplot
  - Correlation
- Sufficient to make meaningful recommendations for most tests

Using Descriptive Statistics to Summarize Performance Data E.g., Task Completion Times

- Mean time to complete – rough estimate of group as a whole
  - Compare with original benchmark: is it skewed above/below?
- Median time to complete – use if data very skewed
- Range (largest value – smallest value) spread of data
  - If small spread then mean is representative of the group
  - A good measure
- Standard Deviation (SD) is the square root of the variance
  - How much variation or "dispersion" is there from the average (mean or expected value) in a normal distribution
  - If small, then performance is similar, if large, then more analysis is needed
  - Influence by outliers possible, so rerun without them as well

Software Engineering
Rochester Institute of Technology

Normal Curve and Standard Deviation



Summarizing Performance Data (Cont.)

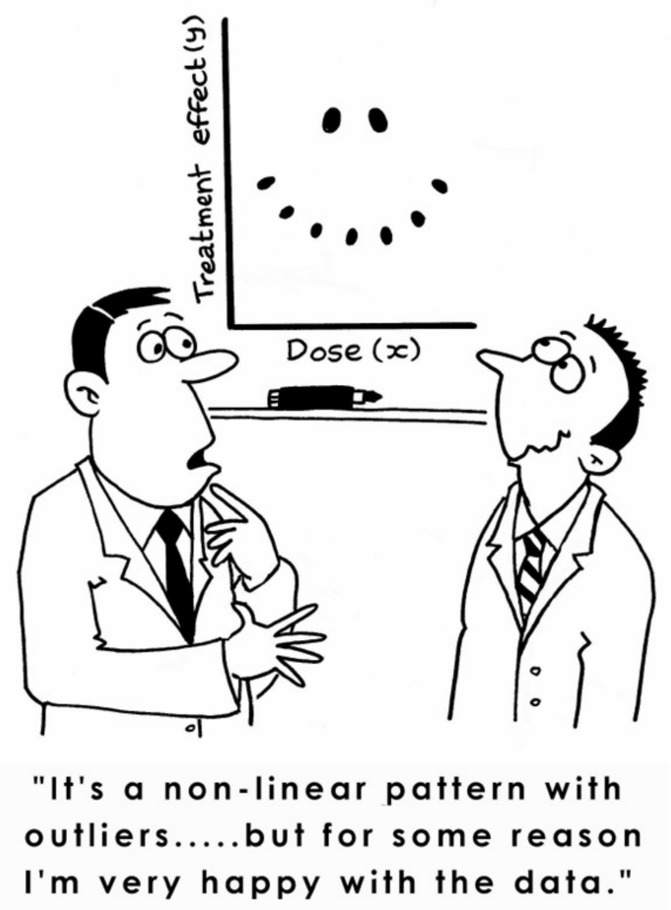
- Interquartile range (IQR) – another measure of statistical spread
  - Find the three data points (quartiles) that divide the data set into four equal parts, where each part has one quarter of the data
  - Difference between the upper (Q_3) and lower (Q_1) quartile points is the IQR
  - IQR = Q3 - Q1 (“middle fifty”)
  - Find outliers - below Q_1 - 1.5(IQR) or above Q_3 + 1.5(IQR)

i	x[i]	Quartile
1	102	
2	104	
3	105	Q_1
4	107	
5	108	
6	109	Q_2 (median)
7	110	
8	112	
9	115	Q_3
10	116	
11	118	

Software Engineering
Rochester Institute of Technology

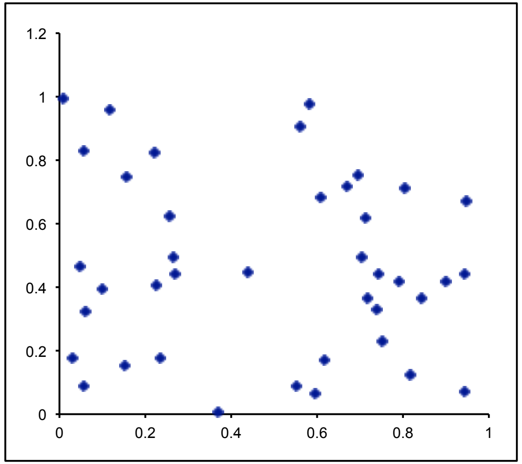
Correlation

- Allows exploration of the strength of the linear relationship between two continuous variables
- You get two pieces of information; direction and strength of the relationship
  - Direction
    - +, as one variable increases so does the other
    - -, as one variable increases, the other variable decreases
  - Strength
    - Small: .01 to .29 -.01 to -.29
    - Medium: .3 to .49 -.3 to -.49
    - Large: .5 to 1 -.5 to -1

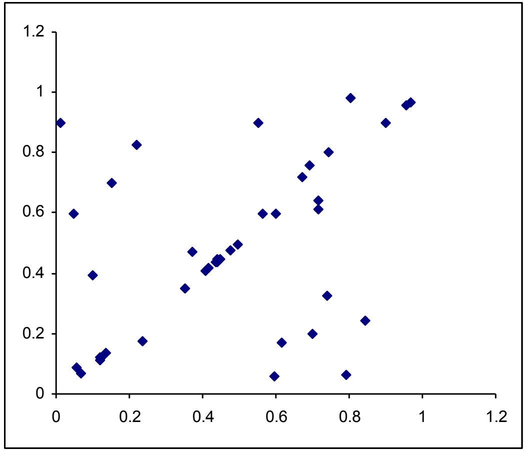


Scatterplots

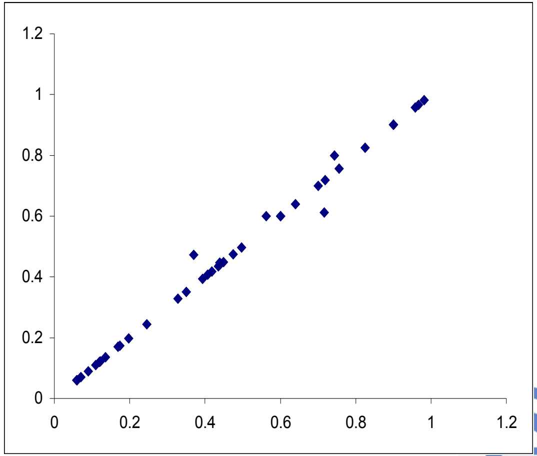
- Need to visually examine the data points
- Scatterplot – plot (X,Y) data point coordinates on a Cartesian diagram



r = .00



r = .40



r = .99

Errors in Testing

- Sample is not big enough
- The sample is biased
  - You have failed to notice and compensate for factors that can bias the results
- Sloppy measurement of data.
- Outliers were left in when they should have been removed
  - Is an outlier a fluke or a sign of something more serious in the context of a larger data set?

Software Engineering
Rochester Institute of Technology

Data Analysis Activity

- See the Excel spreadsheet “Sample Usability Data File” under “Assignments and In-Class Activities” in myCourses
- Follow the directions
- Submit to the Activity dropbox “Data Analysis”

Supplemental Information
Inferential Statistics

Inferential Statistics

- Infer some property or general pattern about a larger data set by studying a *statistically significant sample* *(*large enough to obtain repeatable results*)*
  - In expectation the results will generalize to the larger group
  - Analyze data subject to random variation as a sample from a larger data set
- Techniques:
  - Estimation of descriptive parameters
  - Testing of statistical hypotheses
- Can be complex to use, controversial
  - Keep Inferential Statistics Simple (KISS 2.0)

Statistical Hypothesis Testing

- A method for making decisions about statistical validity of observable results as applied to the broader population
- Based on data samples from experiments or observations
- Statistical hypothesis – (1) a statement about the value of a population parameter (e.g., mean) or (2) a statement about the kind of probability distribution that a certain variable obeys

Software Engineering
Rochester Institute of Technology

Establish a Null Hypothesis (H_0)

- The **null hypothesis H_0** is a simple hypothesis in **contradiction** to what you would like to prove about a data population
- The **alternative hypothesis H_1** is the **opposite**
  - what you would like to prove
- For example: I believe the mean age of this class is greater than or equal to 20.7
  - H_0 - the mean age is < 20.7
  - H_1 – the mean age is ≥ 20.7

Does the Statistical Hypothesis Match Reality?

	STATE OF NATURE	
DECISION	H_0 is true	H_0 is false
Accept H_0:	satisfactory	Type II error
Reject H_0:	Type I error	satisfactory

- Two types of errors in deciding whether a hypothesis is true or false
  - Note: a **decision** about what you **believe** to be true or false about the hypothesis, **not a proof**
- **Type I error** is considered **more serious**

Null Hypothesis

- Null hypothesis (H_0) – hypothesis stated in such a way that a **Type I error occurs** if you believe the **hypothesis is false and it is true**
- **In any test of H_0** based on **sample observations** open to **random variation**, there is a **probability of a Type I error**
    - **P(Type I Error) = α**
    - Called the "significance level"
- Essential idea - limit, to the small value of α, the likelihood of incorrectly reaching the decision to reject H_0 when it is true
    - As a result of experimental error or randomness

Software Engineering
Rochester Institute of Technology

How It Works

- Establish H_0 (and H_1)
- Establish a relevant test statistic and distribution for the sample (e.g., mean, normal distribution)
- Establish the maximum acceptable probability of a Type I error - the significance level α (0.05)
- Describe an experiment in terms of …
  - Set of possible values for the test statistic
  - Distribute the test statistic into values for which H_0 is rejected (critical region) or not
  - Threshold probability of the critical region is α
- Run the experiment to collect data and compute the test statistic **p**
- If **p** > α reject H_0

Simple Example

- I believe the mean age of this class is ≥ 20.7
- Establish H_0
  - The mean age in this class is less than 20.7 years
- Establish a relevant test statistic and distribution for the sample
  - Mean, assume normal distribution from 17 to 26 of all undergraduate SE students
- Establish the significance level α
  - 0.05 by convention
- Distribute the test statistic into values for which H_0 is rejected (critical region)
  - Let's say 19 and above
  - Run the test with a sample size of 10, compute the mean μ and the probability p of that mean value occurring from a sample size of 10 in the general population
- If $p > \alpha$, reject H_0