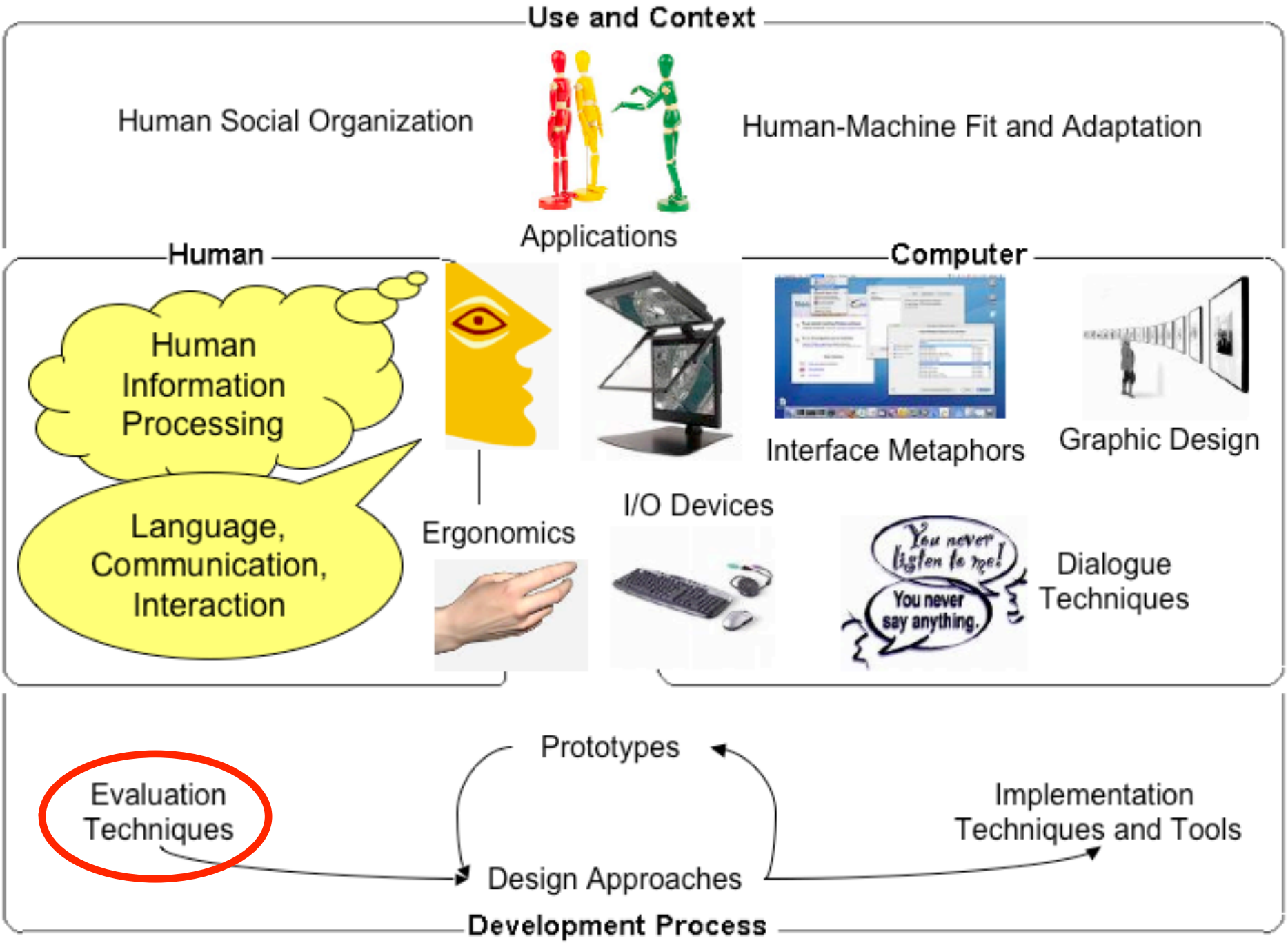
Quantitative Evaluations

Chapter 9

The slides in this lecture are partially based upon this from Drs. Vincent Ng and Saul Greenberg.

Lecture Overview

- What you'll learn in this lecture:
    - What is experimental design?
    - What is an experimental hypothesis?
    - How do I plan an experiment?
    - Why are statistics used?
    - What are the important statistical methods?


Use and Context
Human Social Organization
Human-Machine Fit and Adaptation
Applications
Human
Computer
Human Information Processing
Language, Communication, Interaction
Ergonomics
I/O Devices
Interface Metaphors
Graphic Design
You never listen to me!
You never say anything.
Dialogue Techniques
Prototypes
Evaluation Techniques
Implementation Techniques and Tools
Design Approaches
Development Process

Why Evaluate?

- “Without measurement, success is undefined; so any empirical evaluation begins with the definition of useful measures.” *Standards for Education and Psychological Testing*, 1985.

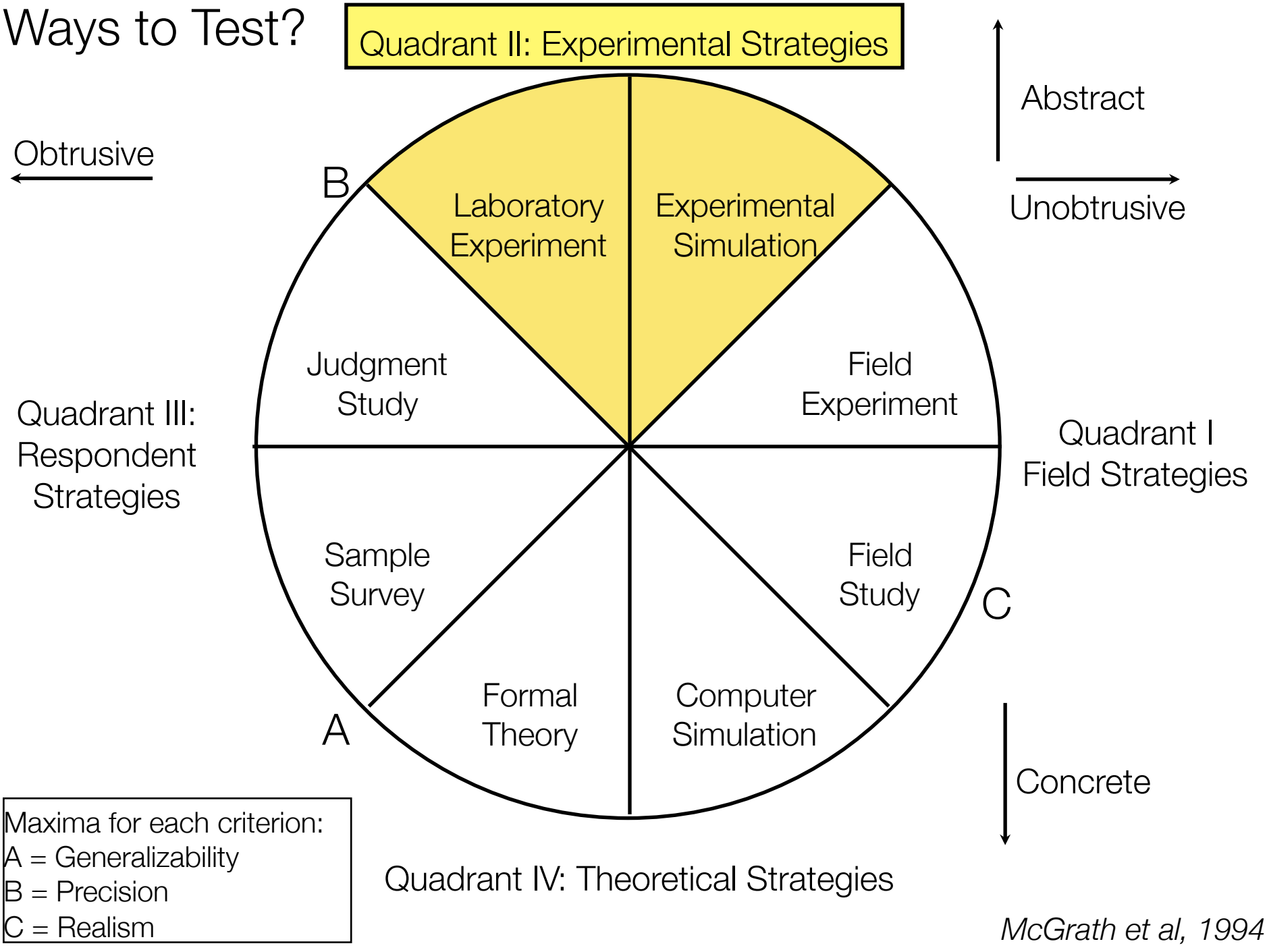
- We evaluate to:

    - Assess extent of system functionality

    - Assess effect of interface on user

    - Identify specific problems

Ways to Test?



McGrath et al, 1994

Modes of Evaluation

- Qualitative:
    - Measurements: subjective and relative
    - Gives you process data
- Quantitative:
    - Lots of experiments, lots of data, statistical analysis.
    - Measurements: precise, valid and reliable
    - Gives you “bottom-line” results

Bottom-line usability

- Situations in which numbers are useful
    - Time requirements for task completion
    - Successful task completion rate
    - Compare two designs on speed or number of errors
- Ease of measurement
    - Time is easy to record
    - Error or successful completion is harder; need to define in advance what these mean.

Collecting user performance data

- Data collected on system use (often lots of data)
- Exploratory:
  - hope something interesting shows up
  - but difficult to analyze
- Targeted
  - look for specific information, but may miss something
    - frequency of request for on-line assistance
      - what did people ask for help with?
    - frequency of use of different parts of the system
      - why are parts of system unused?
    - number of errors and where they occurred
      - why does an error occur repeatedly?
    - time it takes to complete some operation
      - what tasks take longer than expected?

User Performance Data

- Some common things we could measure:
  - Time: Easy to measure, suitable for statistical analysis (e.g. learning time, task completion time).
  - Errors: Shows where problems exist within a system. Suggest cause of difficulties.
  - Patterns of system use: Study the patterns of use in a given system. Preference and avoidance of particular sections.
  - Amount of work done in a given time.

Types of Measurements

- Four major scales of measurements
    - Nominal
    - Ordinal
    - Interval
    - Ratio

Nominal Scale

- Classification into named or numbered unordered categories
  - E.g. country of birth, user groups, gender…
- Allowable manipulations
  - Whether an item belongs in a category
  - Counting items in a category
- Statistics
  - Number of cases in each category
  - Most frequent category
  - No means, medians…
- Sources of error
  - agreement in labeling, vague labels, vague differences in objects
- Testing for error
  - agreement between different judges for same object

Ordinal Scale

- Classification into named or numbered ordered categories
  - no information on magnitude of differences between categories
  - e.g. preference, social status, gold/silver/bronze medals
- Allowable manipulations
  - as with interval scale, plus
  - merge adjacent classes
  - transitive: if A > B > C, then A > C
- Statistics
  - median (central value)
  - percentiles, e.g., 30% were less than B
- Sources of error
  - as in nominal

Interval Scale

- Classification into ordered categories with equal differences between categories
  - zero only by convention
  - e.g. temperature (C or F), time of day
- Allowable manipulations
  - add, subtract
  - cannot multiply as this needs an absolute zero
- Statistics
  - mean, standard deviation, range, variance
- Sources of error
  - instrument calibration, reproducibility and readability
  - human error, skill…

Ratio Scale

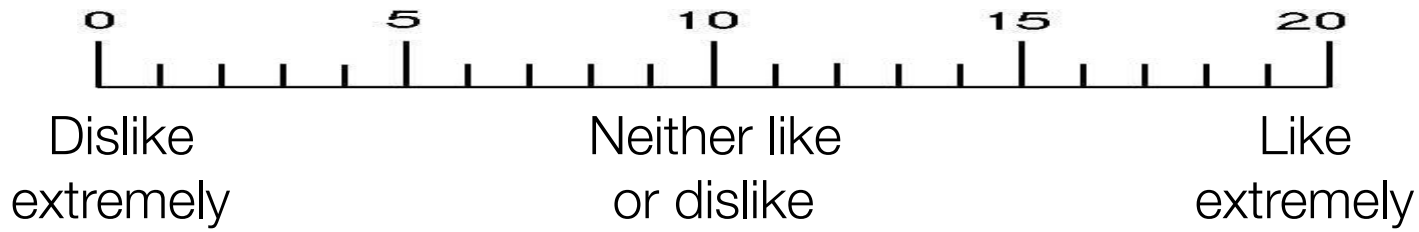
- Interval scale with absolute, non-arbitrary zero
    - e.g. temperature (K), length, weight, time periods
- Allowable manipulations
    - multiply, divide

Example: Let’s measure apples

- Nominal:
  - apple variety
  - Macintosh, Delicious, Gala…
- Ordinal:
  - apple quality
  - U.S. Extra Fancy
  - U.S. Fancy
  - U.S. Combination Extra Fancy / Fancy
  - U.S. No. 1
  - U.S. Early
  - U.S. Utility
  - U.S. Hail

Example: Let's measure apples (cont'd)

- Interval:
  - apple 'Liking scale'
    Marin, A. Consumers' evaluation of apple quality. Washington Tree Postharvest Conference 2002.
- After taking at least 2 bites how much do you like the apple?



- Ratio:
  - apple weight, size, …

How do we know if it's a good measure?

- Validity of measures:
  - A measure is valid if it makes sense: if the inferences based on it are sensible.
  - Most commonly used method: see if other people believe it!
- Reliability of measures:
  - A measure is reliable if repeated measurements under the same conditions yield the same or similar results.

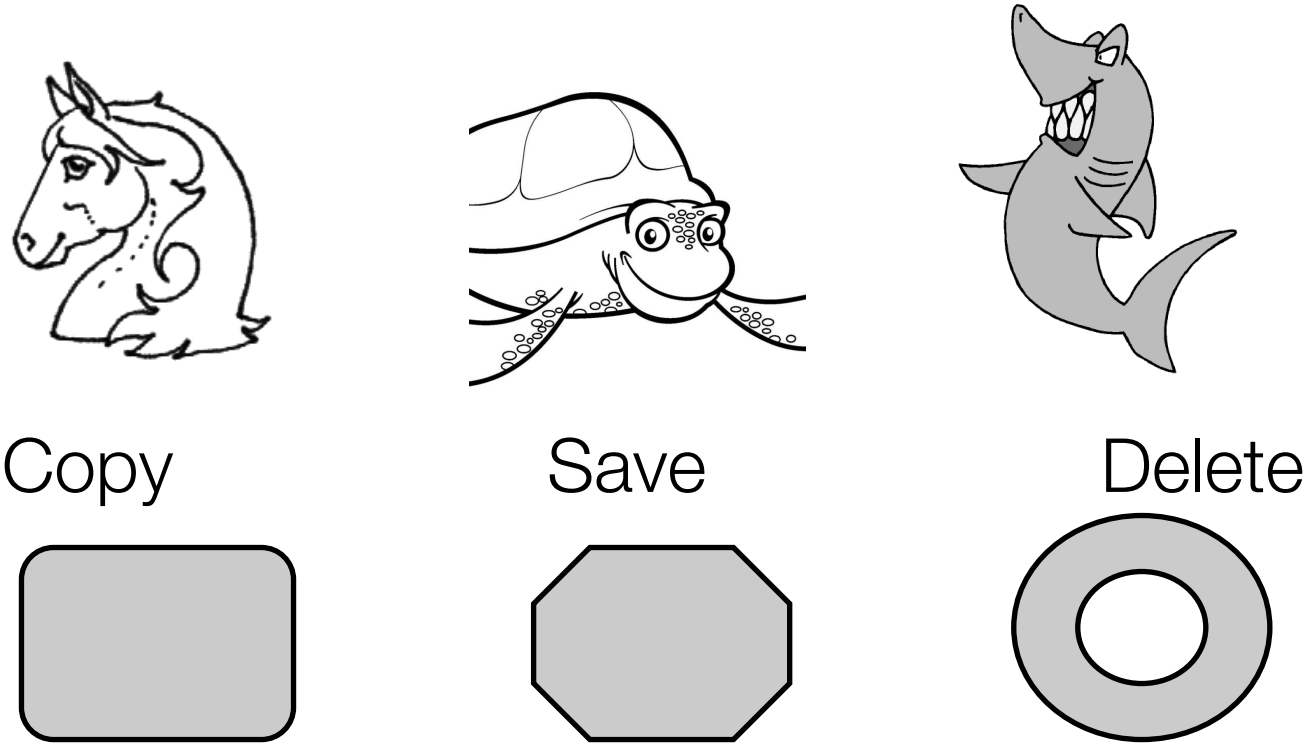
Measuring User Performance

- Time to complete specific tasks.
- Number of tasks completed within given time.
- Number of errors.
- Number of deviations (extra clicks) from optimal path.
- Accuracy (answer to question true or false).
- Ratio of successful interactions to errors.
- Time spent recovering from errors.
- Number of commands/features used.
- Number of features user can remember after test.
- How often help system used.
- Time spent using help.
- Ratio of positive to negative user comments.
- Number of times user sidetracked from real task.

What is an experiment?

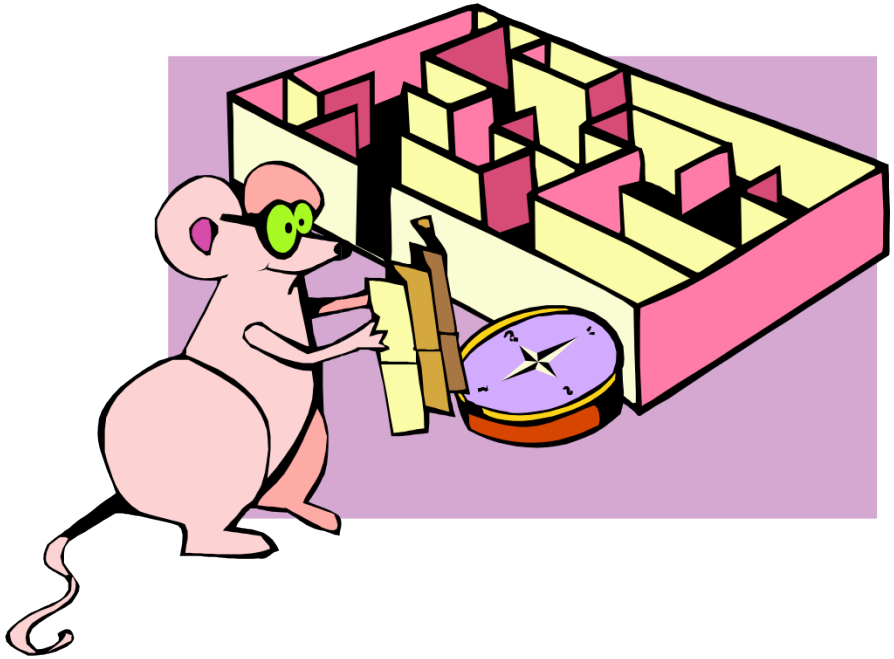
- An experiment is designed to test a hypothesis about the role of one variable (the independent variable) on another (the dependent variable).

An example



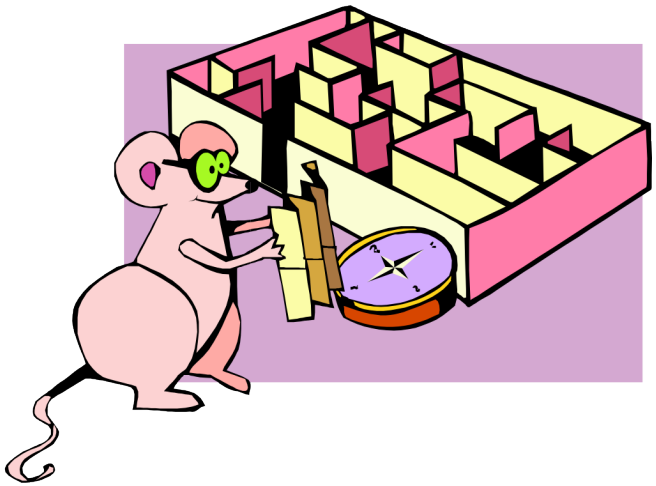
Question: Which set of icons work better?

How do we set up an experiment?



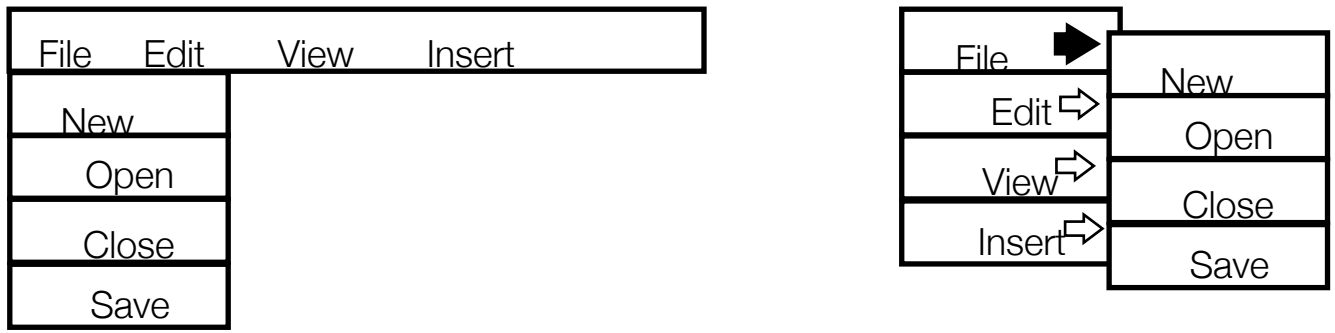
Controlled Experiments

- In all experiments, we need:
  - A lucid and testable hypothesis
  - Quantitative measurements
  - Statistics to measure confidence in the results
  - Experiments that we can replicate
  - Control of variables and conditions
  - Removal of experimenter bias



(A) A Lucid and Testable Hypothesis

- A hypothesis is a precise problem statement.
- Example 1:
  - There is no difference in the number of cavities in children and teenagers using crest and no-teeth toothpaste when brushing daily over a one month period
- Example 2:
  - There is no difference in user performance (time and error rate) when selecting a single item from a pop-up or a pull down menu of 4 items, regardless of the subject's previous expertise in using a mouse or using the different menu types"



(B) Quantitative Measurements

- The experimenter must systemically manipulate one or more independent variables in the domain under investigation.

- The manipulation must be made under controlled conditions, such that all variables that could affect the outcome of the experiment are controlled.

- The experimenter must measure some un-manipulated feature that changes, or is assumed to change, as a function of the manipulated independent variable.

What is a variable?

- Independent variables
  - Factors that are systemically varied by the experimenter.
  - Manipulated independent of the subject's behavior.
  - Determines a modification to the conditions the subjects undergo
  - May arise from subjects being classified into different groups
- For example:
  - Types of interfaces
  - Practice or training
  - Characteristics (gender, age) of users

Independent Variables

- In toothpaste experiment
  - toothpaste type: uses Crest or No-teeth toothpaste
  - age: <= 11 years or > 11 years
- In menu experiment
  - menu type: pop-up or pull-down
  - menu length: 3, 6, 9, 12, 15
  - subject type (expert or novice)

What is a variable?

- Dependent Variables
  - Measures to demonstrate the effect of independent variables.
  - Dependent upon the subject's behavior
  - The things that we set out to quantitatively measure/observe
- Properties
  - Readily observable
  - Stable and reliable so they do not vary under constant experimental conditions
  - Sensitive to the effects of the independent variables.
  - Readily related to some scale of measurement.
- For example:
  - Number of errors made
  - Time taken to complete a given task
  - Time taken to recover from an error

Dependent Variables

- In toothpaste experiment
  - Number of cavities
  - Frequency of brushing
  - Preference
- In menu experiment
  - Time to select an item
  - Selection errors made
  - Time to learn to use it to proficiency

(C) Subject Selection

- Judiciously select and assign subjects to groups
- Ways of controlling subject variability:
  - Reasonable amount of subjects
  - Random assignment
  - Make different user groups an independent variable
  - Screen for anomalies in subject group
    - Superstars versus poor performers

Novice

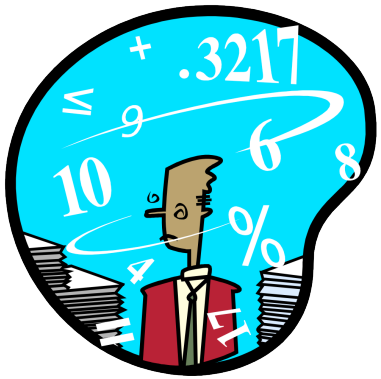
Expert

(D) Controlling Bias

- Control for bias
  - Unbiased instructions
  - Unbiased experimental protocols
    - Prepare scripts ahead of time
  - Unbiased subject selection

(E) Statistical Analysis

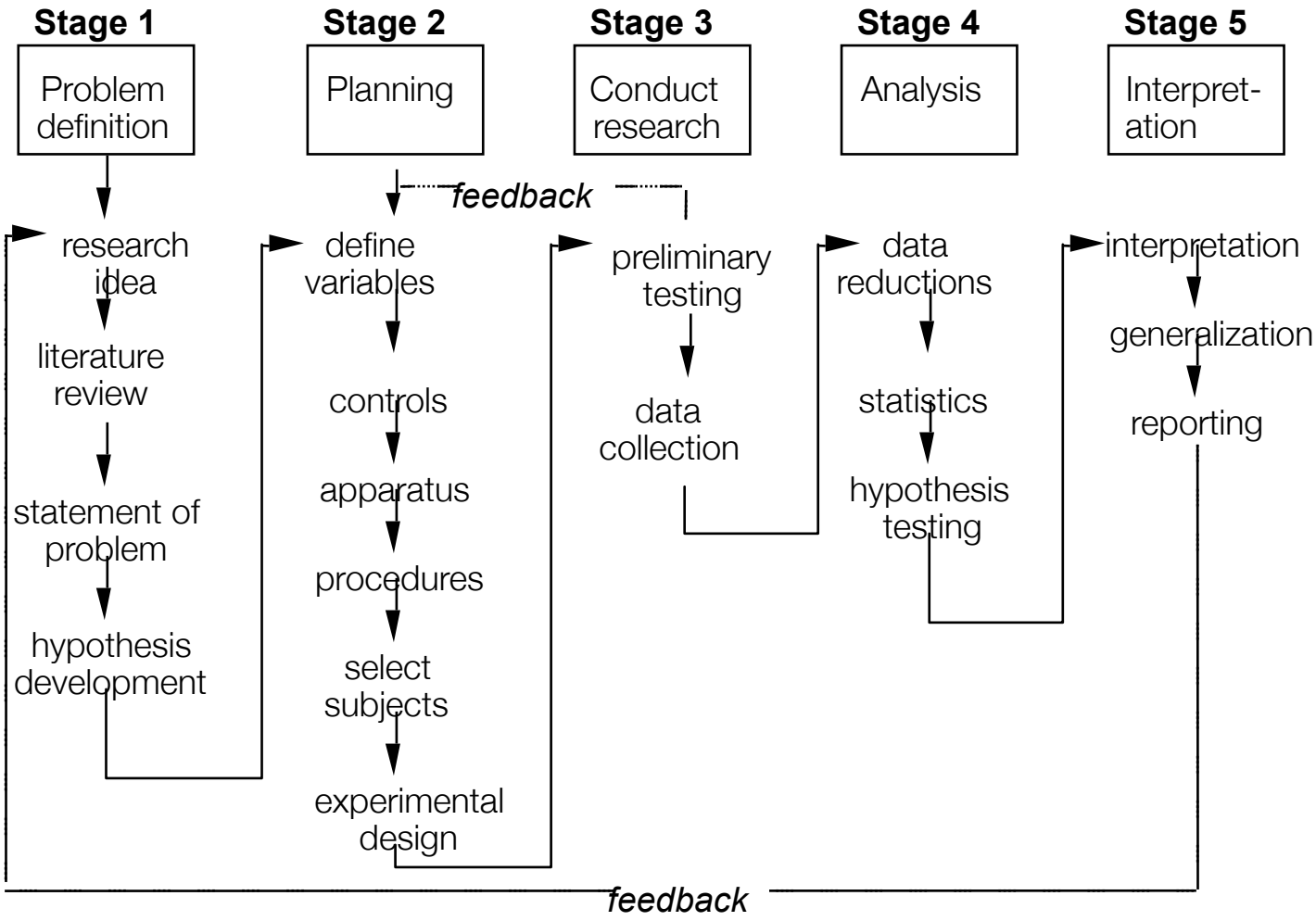
- Apply statistical methods to data analysis
- Confidence limits:
  - The confidence that your conclusion is correct
  - “The hypothesis that computer experience makes no difference is rejected at the .05 level” means:
    - A 95% chance that your statement is correct
    - A 5% chance you are wrong



(F) Interpretation

- Interpret your results
  - What you believe the results really mean
  - Their implications to your research
  - Their implications to practitioners
  - How generalizable they are
  - Limitations and critique

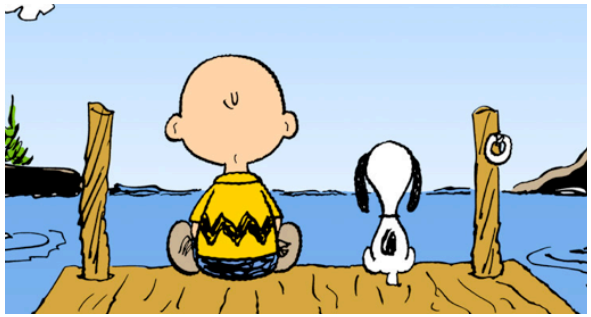
Planning flowchart for experiments

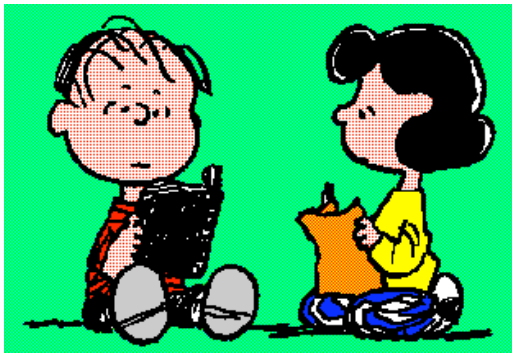


Controlling Variables

- In experiments,
    - Variables that are not of interest are held constant while independent variables are manipulated
    - The effects of the manipulation on the dependent variables are observed.

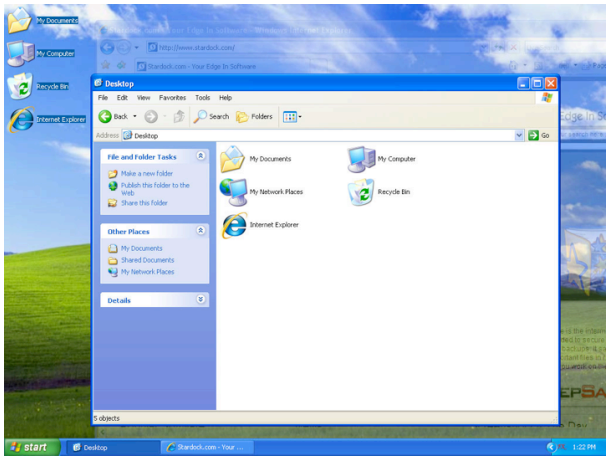
Example Experiment





Charlie Brown & Snoopy use one interface

Linus and Lucy use another interface

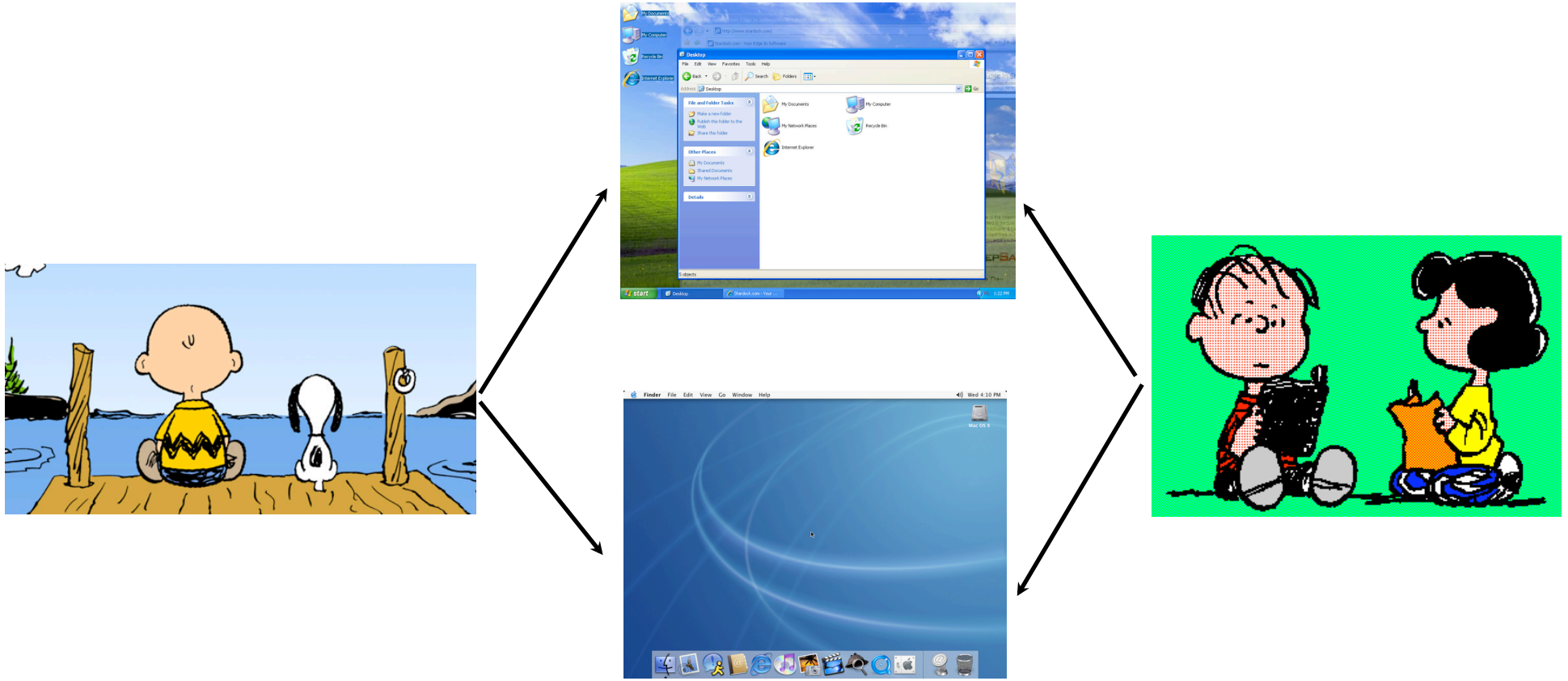


Example Experiment (cont'd)

- The illustrated experiment is an example of *between subjects design*.

Subjects using Interface A	Time (secs)	Subjects using Interface B	Time (secs)
Tom	11.7	Cindy	15.4
Dick	15.8	Joanna	16.3
Harry	12.3	Candy	10.2
Mary	18.9	Peter	11.7
Jane	9.7	Roy	7.5
Ida	10.2	Raymond	10.6
Carol	20.3	Victor	14.8
Nicholas	11.4	Jennifer	15.2
Tommy	12.6	Joyce	11.4
Jeanne	13.5	Jeremy	12.7
Florence	14	Ivan	13.7

Another Example Experiment



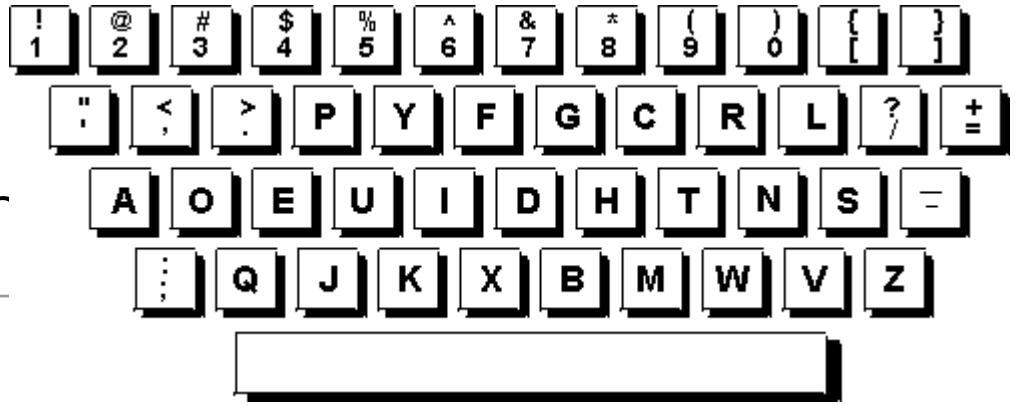
Everybody uses both interfaces

Another Example Experiment (cont'd)

- The illustrated experiment is an example of *within subjects design*.

Subjects on manual typewriter	Typing speed (wpm)	Subjects on electric typewriter	Typing speed (wpm)
Tom	35	Tom	40
Dick	42	Dick	45
Harry	32	Harry	35
Mary	25	Mary	30
Jane	30	Jane	32
Ida	30	Ida	35
Carol	30	Carol	40
Nicholas	36	Nicholas	37
Tommy	36	Tommy	42
Jeanne	30	Jeanne	34
Florence	34	Florence	37

Ordering and Confoundir



- Confounding
  - Happens when variables that were thought to be independent actually vary according to some other factor in the experiment.
- Ordering effects
  - The order in which experiments are done don't affect the results… or do they?
  - Example:
    - Learning to mouse type on any keyboard improves performance on the next keyboard
      - S1: Qwerty then Dvorak then Alphabetic
      - S2: Qwerty then Dvorak then Alphabetic
      - S3: Qwerty then Dvorak then Alphabetic
    - Result: Alphabetic > Dvorak > Qwerty performance — even if there really was no difference between keyboards!
  - Ordering of conditions is a variable that can confound the results!

Between vs. Within Subjects

- Within subjects
  - All participants try all conditions
    - Can isolate effect of individual differences
    - Requires fewer participants
    - Ordering and fatigue effects
- Between subjects
  - Each participant tries one condition
    - No ordering effects, less fatigue
    - User variation can bias results; cannot isolate effects due to individual differences.
    - Need more participants

Other Sources of Confounding

- Experience factors
  - People in one condition have more/less relevant experience than others.
- Experimenter/subject bias
  - Experimenter subconsciously treats subjects differently, or when subjects have different motivation levels.
- Uncontrolled factors
  - Time of day, system load.

Preventing Confounding: Randomization

- Control independent variable X's effects on dependent variable Y so that effects are distributed randomly among conditions (groups).
    - Makes the probability of any particular X influencing Y the same for all groups.
        - E.g. Randomly assign people to test Interface A first, or Interface B first
- Addresses the order effect.
    - But random assignments to conditions ensure that any effect due to unknown differences among subjects or conditions is random.

Preventing Confounding: Counterbalancing

- Mitigates order and carry-over problem
- Independent variables are varied across subjects in different orders
  - S1: Qwerty then Dvorak then Alphabetical
  - S2: Dvorak then Alphabetical then Qwerty
  - S3: Alphabetical then Dvorak then Qwerty...
- Distributes the order effect across all conditions, but does not remove them
- Fails if order effects are not the equal between conditions
  - E.g. People's performance improves when starting on Qwerty…
    … but worsens when starting on Dvorak

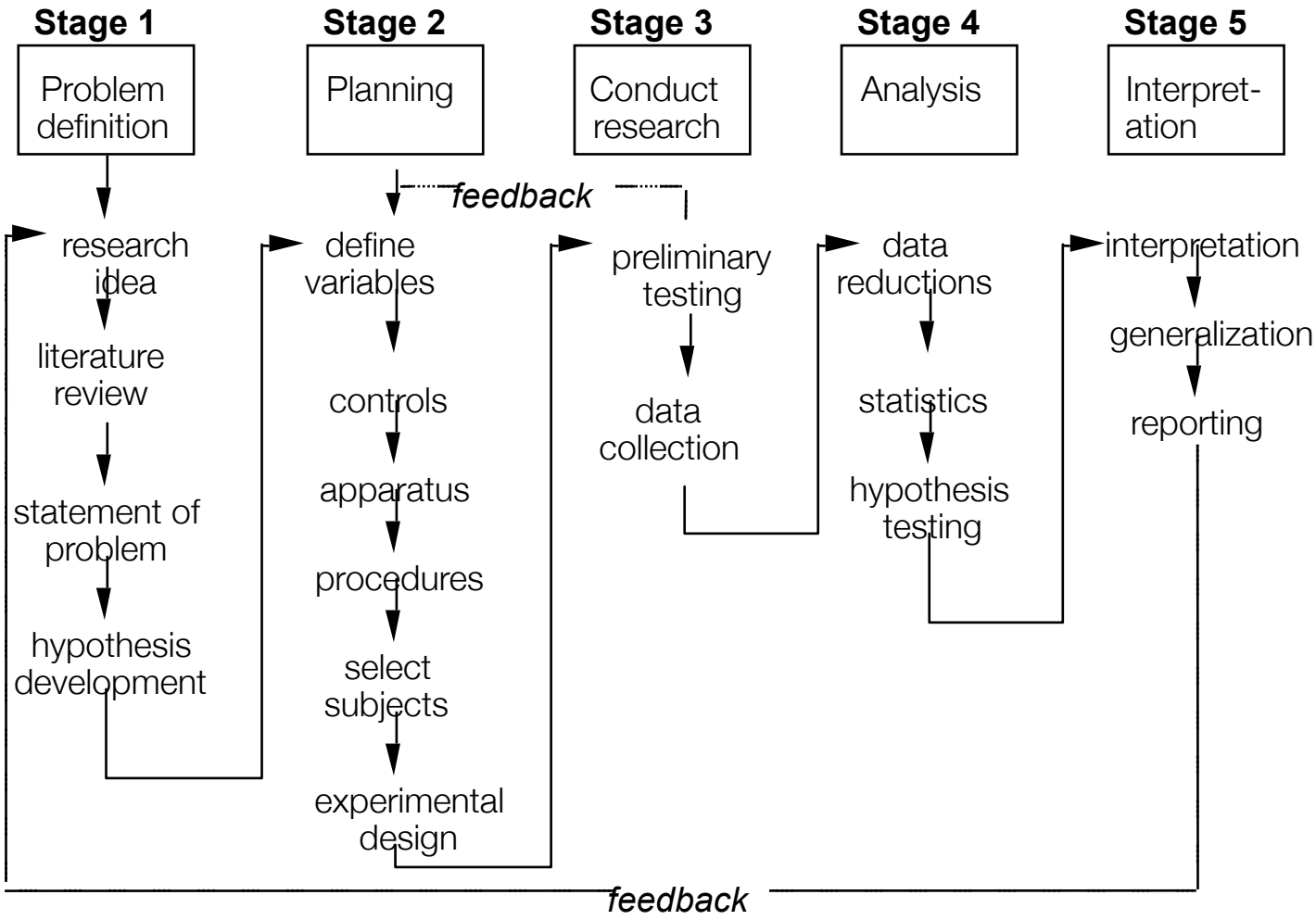
Preventing Confounding: Other methods

- Repeated Measures
    - Experiments are repeated under the same conditions and make order effects a systemically manipulated variable
- Matched Groups
    - To remove unwanted differences between different groups of subjects in different conditions.
    - E.g. Take N most experienced subjects and assign to N different groups. Then take next N, etc…

Preventing Confounding: Other methods (cont'd)

- Control conditions/groups
    - Used as a basis for comparing experimental groups.
- Experimenter/subject blind
    - Experimenter/subject does not know which condition is being tested.
    - Experimenter blind helps prevent experimenter from treating subjects differently because of preconceived bias.
    - Subject blind helps prevent subjects from acting differently because they know they are in specific conditions.

Planning Flowchart for Experiments



From an early ACM CHI tutorial

Other Kinds of Experiments

- Longitudinal Study
    - Same group of subjects are measured repeatedly over time
- Cross-sectional study
    - Different subjects are measured at different stages of the study

Hypotheses

- Prediction of the outcome of an experiment.
    - H0 (null hypothesis): there is no difference between the groups
        - Varying the independent variables cause no change.
    - H1 (Alternative): There are indeed differences between the groups.
- H0 is assumed to be true, and we want to prove it false using the experimental data.
- If we can prove that the differences are statistically significant, then we reject H0 and accept H1

Statistical Analysis

- Calculations that tell us
  - Mathematical attributes about our data sets
    - Mean, amount of variance, ...
  - How data sets relate to each other
    - Whether we are “sampling” from the same or different distributions
  - The probability that our claims are correct
    - “statistical significance”

Recap

- Mean of a set of data points:

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

- Variance of a set of data points:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$$

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right)$$

- Standard deviation of a set of data points = s = sqrt (variance)

Example: Investment

- Suppose you are running your business to sell food.
- We all know that the price of raw materials increase very much in the past year.
- You have two choices of ingredient A and B at the unit price of $10 and $11 respectively. Both requires your commitment for the coming year for a contract.
    - A may vary in price between $4 to $16, and B may vary in price between $10 to $12. Which one would you prefer?
- Take another example of investment in education, there are two sessions of COMP 5517. The mean GPA for both session A and B are 3.0 and there are many A's and many C's in session A but there are very few A's and few C's in session B (of course many B's). Which session would you take?

Example: Sampling

- To know whether your company would be successful, you need to interview potential customers with survey for the customer base. How many survey forms you need before you can get a reasonable picture?

- Hong Kong Population Census is asking 10% of population to provide more details, so that is a 10% sampling. Why 10% is sufficient? Why not 20% (more accurate)? Why not 5% (cheaper)?

- Statistical analysis provides us with the answers again.

Example: Differences between means

Condition one: 3, 4, 4, 4, 5, 5, 5, 6

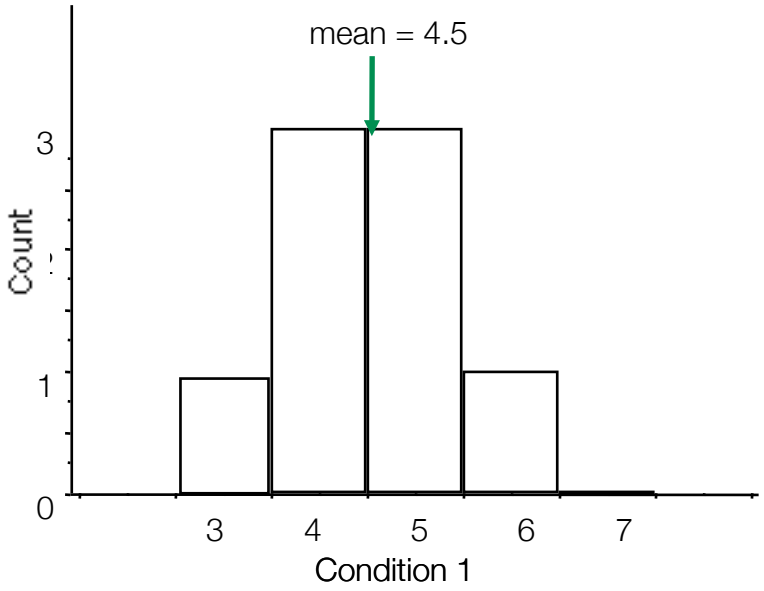
Condition two: 4, 4, 5, 5, 6, 6, 7, 7

- Given:
  - Two data sets measuring a condition
    - E.g. Height difference of males and females
    - Time to select an item from different menu styles ...
- Question:
  - Is the difference between the means of this data statistically significant?
- Null hypothesis: There is no difference between the two means
- Statistical analysis: can only reject the hypothesis at a certain level of confidence

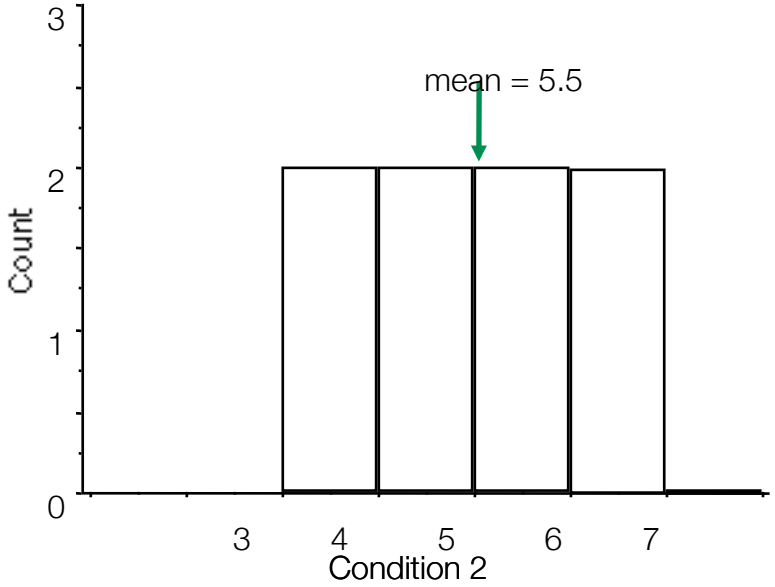
Example

- Is there a significant difference between these means?
- And how do you know?

Condition one: 3, 4, 4, 4, 5, 5, 5, 6



Condition two: 4, 4, 5, 5, 6, 6, 7, 7



Visual Inspection is Not Reliable

- Will almost always see variation in collected data
  - Differences between data sets may be due to:
    - Normal variation
    - E.g. Two sets of ten tosses with different but fair dice
    - Differences between data and means are accountable by expected variation
  - Real differences between data
    - E.g. two sets of ten tosses for with loaded dice and fair dice
    - Differences between data and means are not accountable by expected variation

Statistics

- To know the GPA of COMP 5517, you ask randomly 5 of your friends who have taken this subject before for their grades. If the average GPA is 3.02, and you know that you are always in the middle of the class, would you expect to get a B?

- Your friend also performs the same experiments by asking for another 5 friends, and get an average GPA of 3.21. What would you comment at your friend's result?

- Now you two combine the grades, and find that one friend is in common, so you take the 9 data points to get an average, which is 3.08. Would you trust this value? Why?

Law of Large Numbers

- Everyone knows that if you get more data points, you have a better picture of the situation.

- What is the relationship between accuracy and number of data points?

  - Simply more data => better accuracy is not good enough, we want some more concrete guarantees.

- Law of Large Numbers

  - Weak Law: The sample average converges in probability towards the expected value (mean of the population).

$$\overline{X}_n \xrightarrow{p} \mu \qquad \text{when } n \to \infty$$

  - Strong Law: The sample average converges almost surely to the expected value.

  -

$$\overline{X}_n \xrightarrow{a.s.} \mu \qquad \text{when } n \to \infty$$

Central Limit Theorem

- Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample of size n, drawn from independent and identically distributed random variables with expected values μ and variances σ^2.

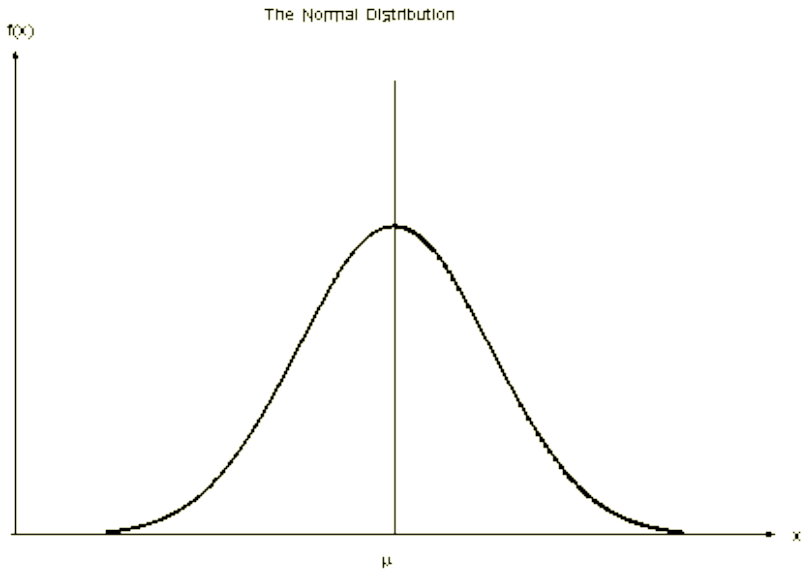
- The mean value of the random variables, $S_n = 1/n(X_1 + \ldots + X_n)$.

- Central limit theorem states that for large value of n, the distribution of S_n is approximately a **normal distribution** with **mean** μ and **variance** σ^2/n, regardless of the original statistical distribution that X_i comes from, i.e. **N**$(\mu,\sigma^2/n)$.

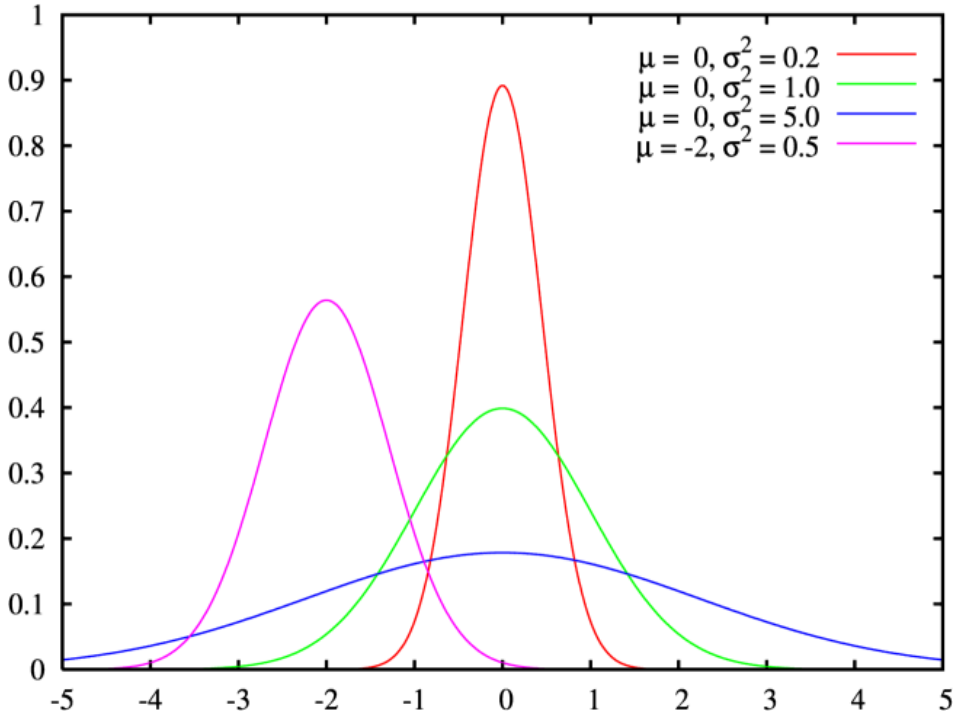
- In population census, assuming 2000000 families in Hong Kong, 10% means 200000. The average salary of each family will be accurate up to standard error of $\sigma/\text{sqrt}(n) = 0.0024\sigma$, where variance σ^2 can be approximated by the sample variance s^2.

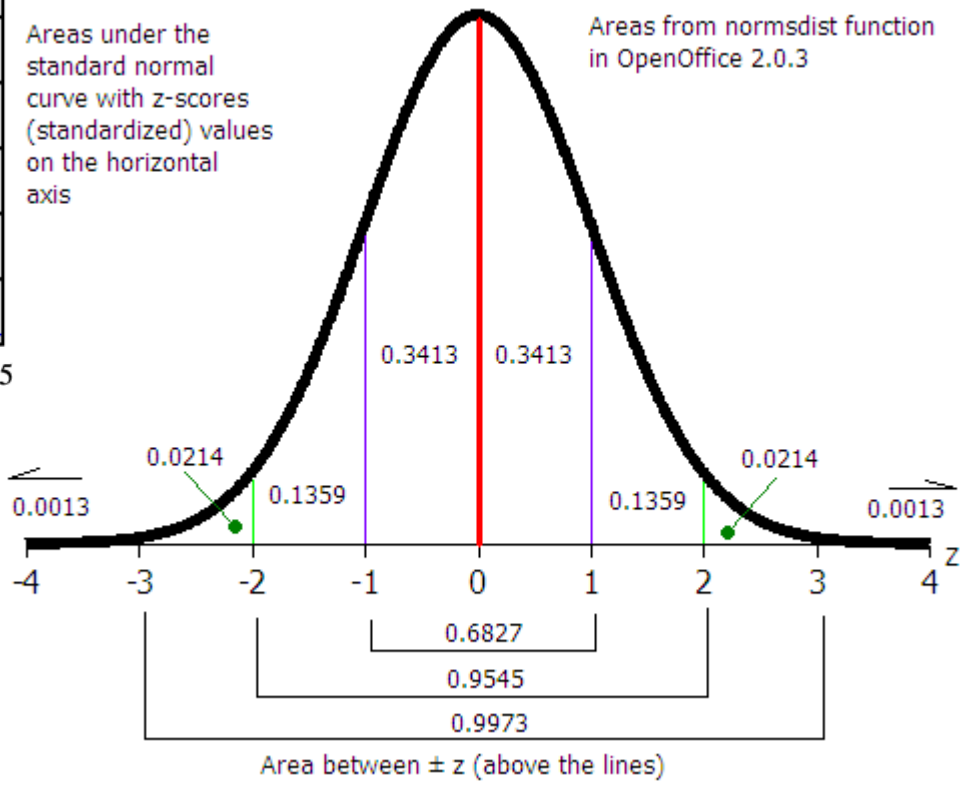
Normal Distribution

- A bell-shaped curve
  - $N(\mu,\sigma^2)$, with mean μ and variance σ^2 and thus SD of σ
- Occurs naturally for many datasets
- Generated naturally from the central limit theorem, when observing many data points.



Normal Distribution





Central Limit Theorem: Example

- Unless the distribution is bad, it is common sense that if n > 30, the sample size could often be considered large enough.

- The marks in a class like this one (over 40 students) are distributed normally. The beauty of normal distribution is that you can predict where you lie in the class, when given the mean and standard deviation of the marks.

- Suppose that the mean of the mid-term test is 66.3, and the standard deviation is 8.1, and you get 75, where would you be?

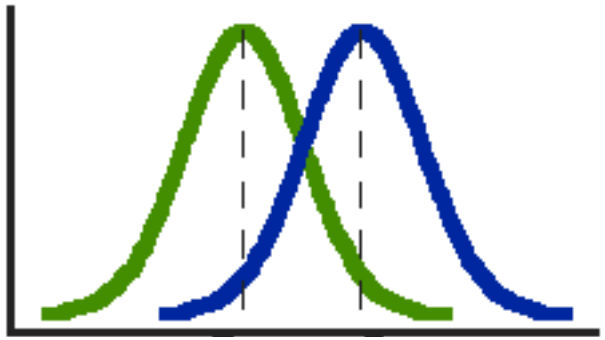
    - 75 is a little over mean + SD

    - In a normal distribution, 68% of marks will lie within mean ± SD, with 16% lower than mean – SD and 16% higher than mean + SD, so you are approximately at the top 16% of the class, so possibly be in the B+ or A range.

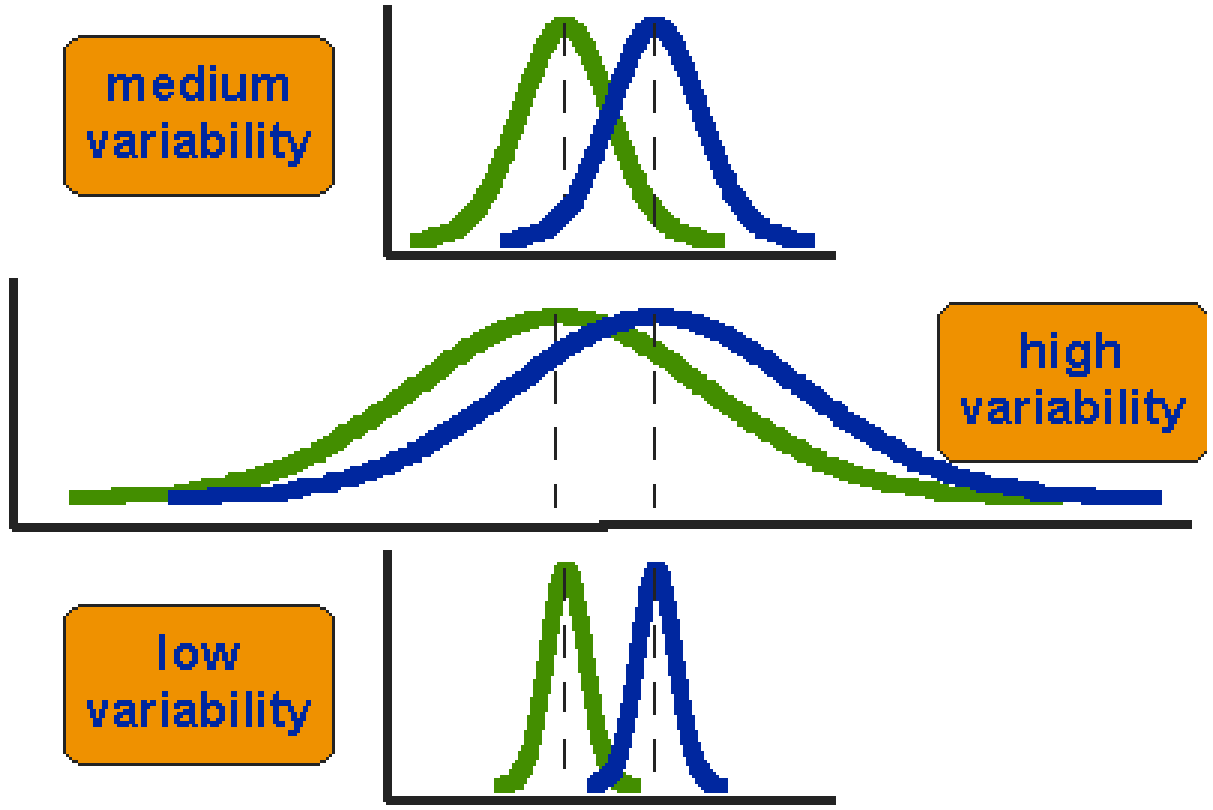
    - If you get mean + 2 SD (i.e. 83), you are within the top 2% and thus the A range.

T-Test

- A simple statistical test
  - Allows one to say something about differences between means at a certain confidence level
  - T-test has its foundation based on Central Limit Theorem.
- Null hypothesis of the t-test:
  - No difference exists between the means of two sets of collected data (or any difference is due to pure chance)
- Possible results:
  - I am 95% sure that null hypothesis is rejected
    - There is probably a true difference between the means
  - I cannot reject the null hypothesis
    - The means are likely the same



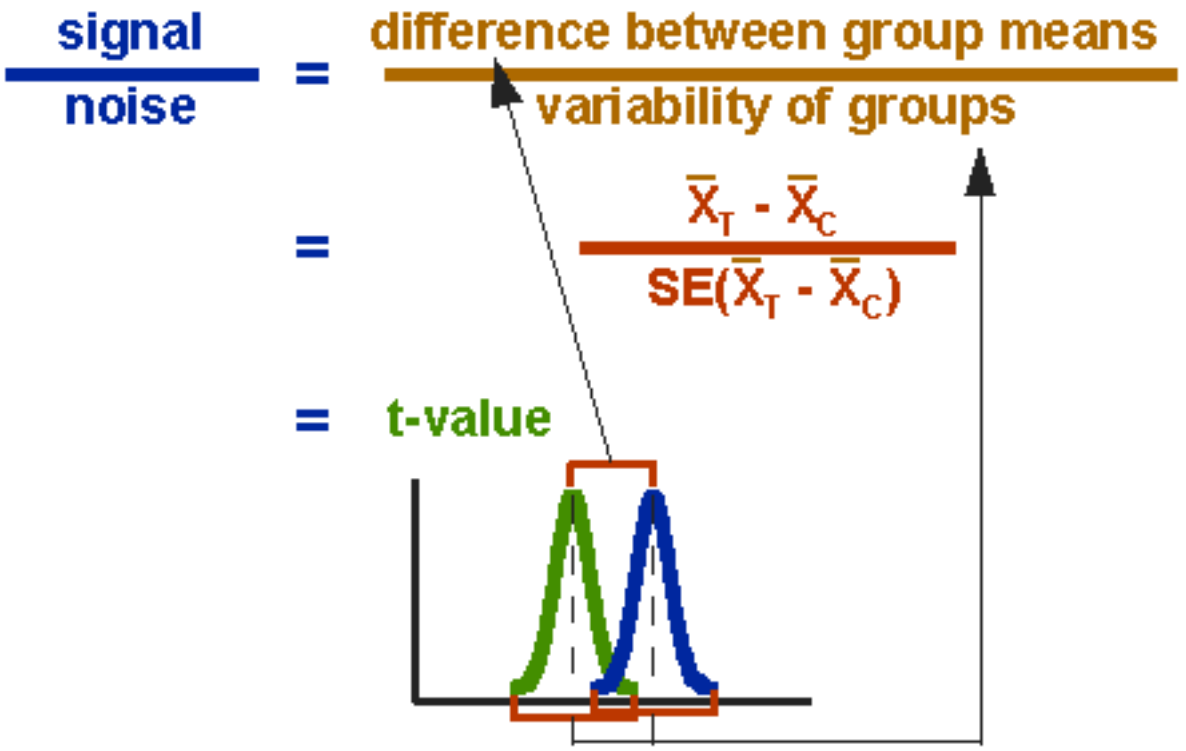
Any Differences?



T-Test Assumptions and Considerations

- Data points of each sample are normally distributed
  - Not always the case, but t-test very robust in practice
- Population variances are equal
  - T-test reasonably robust for differing variances
  - However, does deserve consideration
- Individual observations of data points in sample are independent
  - Must be adhered to!
- Significance level
  - Decide upon the level *before* you do the test!
  - Typically stated at the .05 or .01 level

The T-Test calculates a ratio



Calculating the T-Value (the actual formula)

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right)$$

1, 2	Group 1 and Group 2, respectively
n_i	Number of Data Points in Group i
$\bar{X}_i$	Mean of all elements in Group i
s_i	Standard Deviation of Group i
df	Degrees of freedom = $n_1 + n_2 - 2$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \text{ where } s^2_{\bar{X}_1 - \bar{X}_2} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{df}$$

pooled variance

Using the T-Test Ratio (or T-Value)

- The t-value is then compared against a table of significance.
- Tests whether the ratio is large enough (i.e. there is enough signal) that any difference we're seeing isn't just due to pure chance.
- We need to set a risk level (or alpha level)
    - Gives the level of confidence we want to have.
        - Usually set at 95%, or 0.05.
    - Also need the degrees of freedom
        - Related to number of independent pieces of information we're using.

Table of Significance for Two-Tailed T-Test

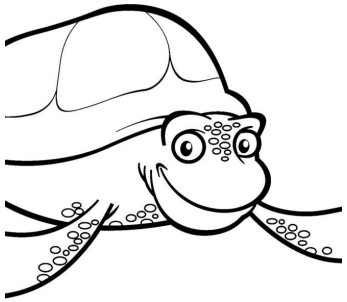
df	0.2	0.1	0.05	0.02	0.01
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.92	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.44	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.86	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.25
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.35	1.771	2.16	2.65	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.12	2.583	2.921
17	1.333	1.74	2.11	2.567	2.898
18	1.33	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845

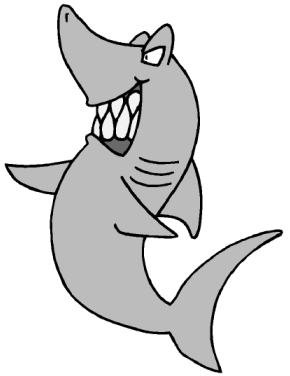
Table of Significance for Two-Tailed T-Test

Do you have any observation in the table values?

df	0.2	0.1	0.05	0.02	0.01
21	1.323	1.721	2.08	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.5	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.06	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.31	1.697	2.042	2.457	2.75
40	1.303	1.684	2.021	2.423	2.704
50	1.299	1.676	2.009	2.403	2.678
60	1.296	1.671	2	2.39	2.66
70	1.294	1.667	1.994	2.381	2.648
80	1.292	1.664	1.99	2.374	2.639
90	1.291	1.662	1.987	2.368	2.632
100	1.29	1.66	1.984	2.364	2.626

Example Experiment

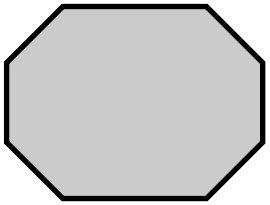


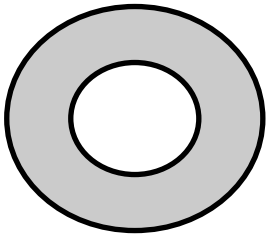


Copy

Save

Delete





Question: Is there a difference to the user?

Example Walkthrough

- First: What do we mean, difference?
    - User preference, or user performance?
    - How do we measure each one?
- Assume: measure user performance
    - Speed at which a user can accurately select an icon
- Experiment setup:
    - Two groups of users.
        - Group 1 uses the animal icons.
        - Group 2 uses the shape icons.

Example Walkthrough

- Results of experiment
    - Group 1: 3 4 4 4 5 5 5 6
    - Group 2: 4 4 5 5 6 6 7 7
- H0: There is no significant difference between the means at the 0.05 level
    - That is, there is no difference between the two groups of icons (where user performance is concerned)
- H1: There is a difference between the two groups.

Experiment Walkthrough: Step 1

Step 1: Calculating $s_{\bar{X}_1 - \bar{X}_2}$ and t

- $n_1 = 8, n_2 = 8$
- $\bar{X}_1 = 4.5, \bar{X}_2 = 5.5$
- $s_1^2 = 0.8571, s_2^2 = 1.429$
- $s^2_{\bar{X}_1 - \bar{X}_2} = 1.143$
- $df = 14$
- $t = -1.871$

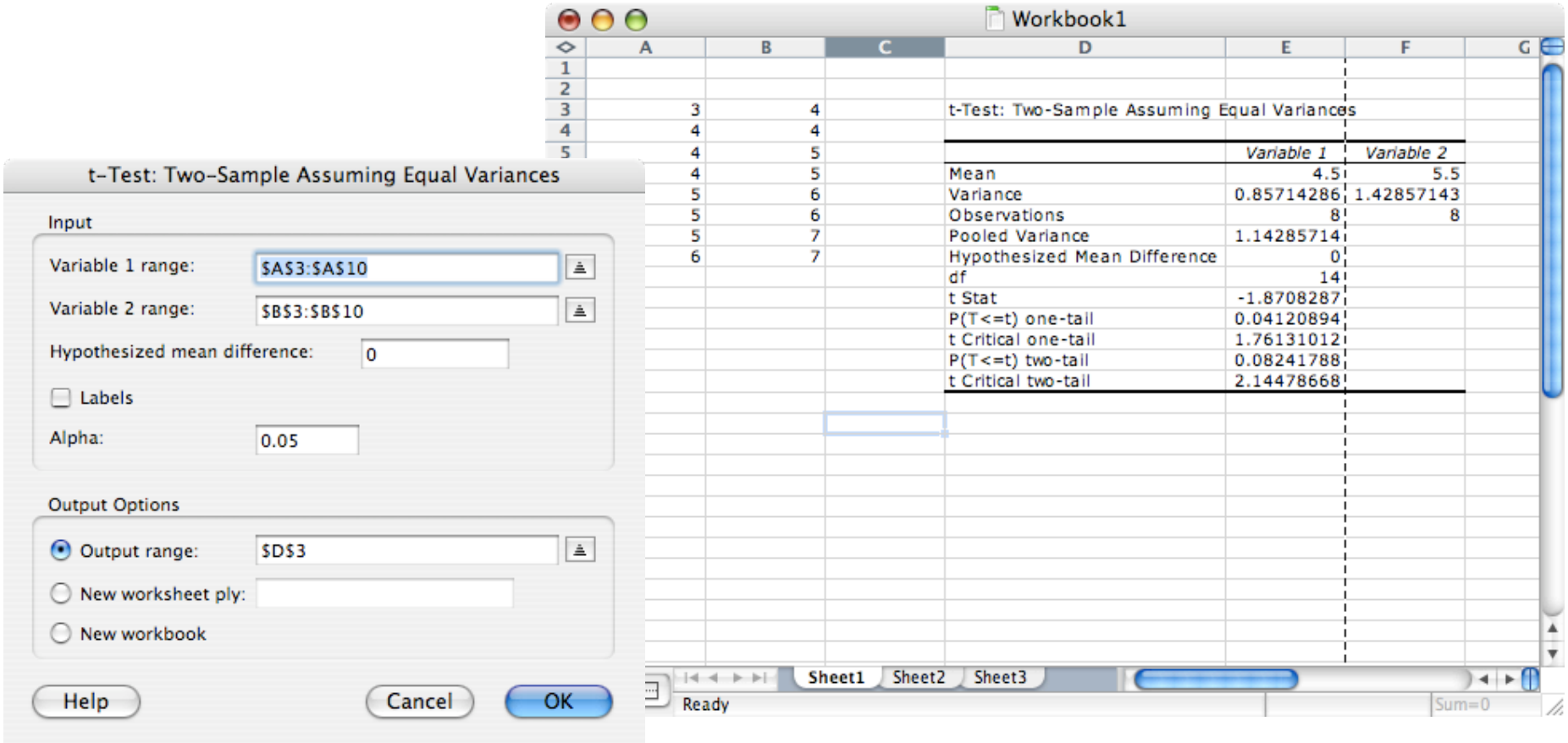
Example Walkthrough: Step 2

- Step 2: Look up critical value of t
  - Use the table for two-tailed t-test, df=14, p=0.05
  - Because t = 1.871 < 2.145, this means that we cannot reject the null hypothesis.
  - In other words, there is no difference between using the two sets of icons.
    - Or: we cannot prove that users perform faster with one set of icons.

df	0.2	0.1	0.05	0.02	0.01
13	1.35	1.771	2.16	2.65	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947

Example Walkthrough: The Lazy Way

- Or, use a statistics package (Excel has a simple one)



Different Kinds of T-Test

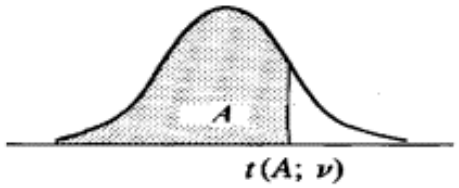
- Comparing two sets of independent observations
    - Usually different subjects in each group
    - Number per group may differ as well
        - E.g. Condition 1: S1-S20, Condition 2: S21–43
- Our example walkthrough was of this variety.

Watch Out

- Some t-statistic tables look different from our example table in the book and lecture note, so watch out for what the t value really means for your table!

- In this example, two-tailed test for $\alpha = 0.05$ should be equivalent to lookup for a value of 1-0.05/2 = 0.975, i.e. for df = 14, the value would be 2.145.

Entry is $t(A; \nu)$ where $P\{t(\nu) \leq t(A; \nu)\} = A$



	A						
ν	.60	.70	.80	.85	.90	.95	.975
1	0.325	0.727	1.376	1.963	3.078	6.314	12.706
2	0.289	0.617	1.061	1.386	1.886	2.920	4.303
3	0.277	0.584	0.978	1.250	1.638	2.353	3.182
4	0.271	0.569	0.941	1.190	1.533	2.132	2.776
5	0.267	0.559	0.920	1.156	1.476	2.015	2.571
6	0.265	0.553	0.906	1.134	1.440	1.943	2.447
7	0.263	0.549	0.896	1.119	1.415	1.895	2.365
8	0.262	0.546	0.889	1.108	1.397	1.860	2.306
9	0.261	0.543	0.883	1.100	1.383	1.833	2.262
10	0.260	0.542	0.879	1.093	1.372	1.812	2.228
11	0.260	0.540	0.876	1.088	1.363	1.796	2.201
12	0.259	0.539	0.873	1.083	1.356	1.782	2.179
13	0.259	0.537	0.870	1.079	1.350	1.771	2.160
14	0.258	0.537	0.868	1.076	1.345	1.761	2.145
15	0.258	0.536	0.866	1.074	1.341	1.753	2.131
16	0.258	0.535	0.865	1.071	1.337	1.746	2.120
17	0.257	0.534	0.863	1.069	1.333	1.740	2.110
18	0.257	0.534	0.862	1.067	1.330	1.734	2.101
19	0.257	0.533	0.861	1.066	1.328	1.729	2.093
20	0.257	0.533	0.860	1.064	1.325	1.725	2.086
21	0.257	0.532	0.859	1.063	1.323	1.721	2.080
22	0.256	0.532	0.858	1.061	1.321	1.717	2.074
23	0.256	0.532	0.858	1.060	1.319	1.714	2.069
24	0.256	0.531	0.857	1.059	1.318	1.711	2.064
25	0.256	0.531	0.856	1.058	1.316	1.708	2.060
26	0.256	0.531	0.856	1.058	1.315	1.706	2.056
27	0.256	0.531	0.855	1.057	1.314	1.703	2.052
28	0.256	0.530	0.855	1.056	1.313	1.701	2.048
29	0.256	0.530	0.854	1.055	1.311	1.699	2.045
30	0.256	0.530	0.854	1.055	1.310	1.697	2.042
40	0.255	0.529	0.851	1.050	1.303	1.684	2.021
60	0.254	0.527	0.848	1.045	1.296	1.671	2.000
120	0.254	0.526	0.845	1.041	1.289	1.658	1.980
∞	0.253	0.524	0.842	1.036	1.282	1.645	1.960

Different Kinds of T-Test

- Paired observations
  - Usually a single group studied under both experimental conditions
  - Data points of one subject are treated as a pair
    - Condition 1: S1-S20, Condition 2: S1-S20
    - Take the difference between the paired sample (here n = 20)
    - T-value = average of difference / sqrt(variance of difference/n)
    - df = n-1

Different Kinds of T-Test

- Non-directional vs directional alternatives
  - Non-directional (two-tailed)
    - No expectation that the direction of difference matters
      - E.g. If we want to know whether the grades given by Instructors A and B differ significantly.
  - Directional (one-tailed)
    - Only interested if the mean of a given condition is greater than the other
      - E.g. If we want to know whether Instructor A gives higher grades than Instructor B

Critical values for One-tailed T-Test

- The one-tailed critical values can be gotten from the two-tailed table:
  - If $p <= 0.5$, then the critical value for the one-tailed case is the same as the critical value for the two-tailed case with $p = 2p$.
  - If $p > 0.5$, then the critical value of the one-tailed case is the negative of the one-tailed critical value with $p = 1-p$
  - Which is the same as the negative of the two-tailed critical value with $p = 2(1-p)$.
- However, for one-tailed tests, the sign is important.
  - For positive one-tailed tests, we need t to be larger than the critical value.
  - For negative one-tailed tests, we need t to be smaller than the negative of the critical value.

One-Tailed T-Test Example

- Scenario: Sam the Mad Scientist wants to prove that people perform better with thumbs.
- Experiment: Two groups of 10 subjects each. Group 1 has their thumbs taped to their palm. The time it takes for each subject to tie a bow in a ribbon is measured.
    - Definition: Perform better == take less time to tie the ribbon.
    - H0: People with thumbs do not perform better than people without thumbs.
    - H1: People with thumbs perform better.
- Why is a one-tailed test appropriate?
- Which one-tailed test is appropriate? (positive or negative)

One-Tailed T-Test Example

- Group 1: 18.6, 12.8, 18.7, 14.2, 14.0, 15.5, 18.5, 28.5, 9.7, 19.1
- Group 2: 6.4, 5.7, 7.1, 8.2, 5.5, 5.7, 5.4, 11.4, 8.5, 6.4

Calculations

$$\bar{X}_1 = 16.96, \bar{X}_2 = 7.03$$

$$s_1^2 = 26.04, s_2^2 = 3.551$$

$$s_{\bar{X}_1 - \bar{X}_2} = 14.80$$

$$t = 5.773$$

$$df = 18$$

Looking up Critical Value

- t = 5.772, df = 18
- Critical value of 1-tail t-test can be derived from the 2-tail t-test table.
- In 1-tailed t-tests, we can accommodate twice the amount of error of the two-tailed t-test.

df	0.1	0.05	0.025	0.01	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.92	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.44	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.86	2.306	2.896	3.355

Looking up Critical Value

- t = 5.772, df = 18

- Therefore, one-tailed positive critical value for 0.05 at df=18 is equal to two-tailed critical value for 0.1 at df=18 = 1.734

- Since t is larger than the critical value, we say that we reject the null hypothesis (that people do not perform better with thumbs) with 95% confidence.

df	0.2	0.1	0.05	0.02	0.01
17	1.333	1.74	2.11	2.567	2.898
18	1.33	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861

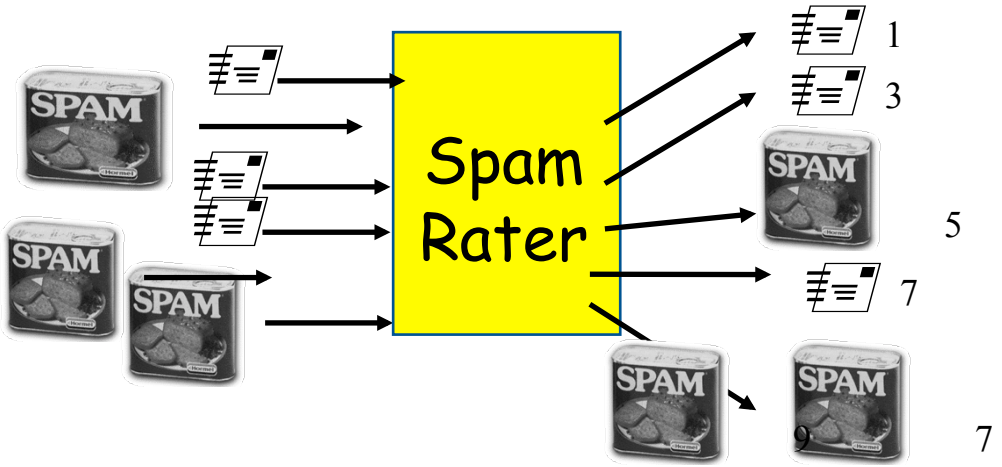
Significance Levels and Errors

- Statistics aren't always correct...
- Type 1 error
  - Rejecting the null hypothesis when it is, in fact, true
- Type 2 error
  - Accepting the null hypothesis when it is, in fact, false
- Effects of levels of significance
  - High confidence level (eg p<.0001) leads to greater chance of Type 2 errors
  - Low confidence level (eg p>.1) leads to greater chance of Type 1 errors

		Decision	
		False	True
"Reality"	True	Type I error	✓
	False	✓	Type II error

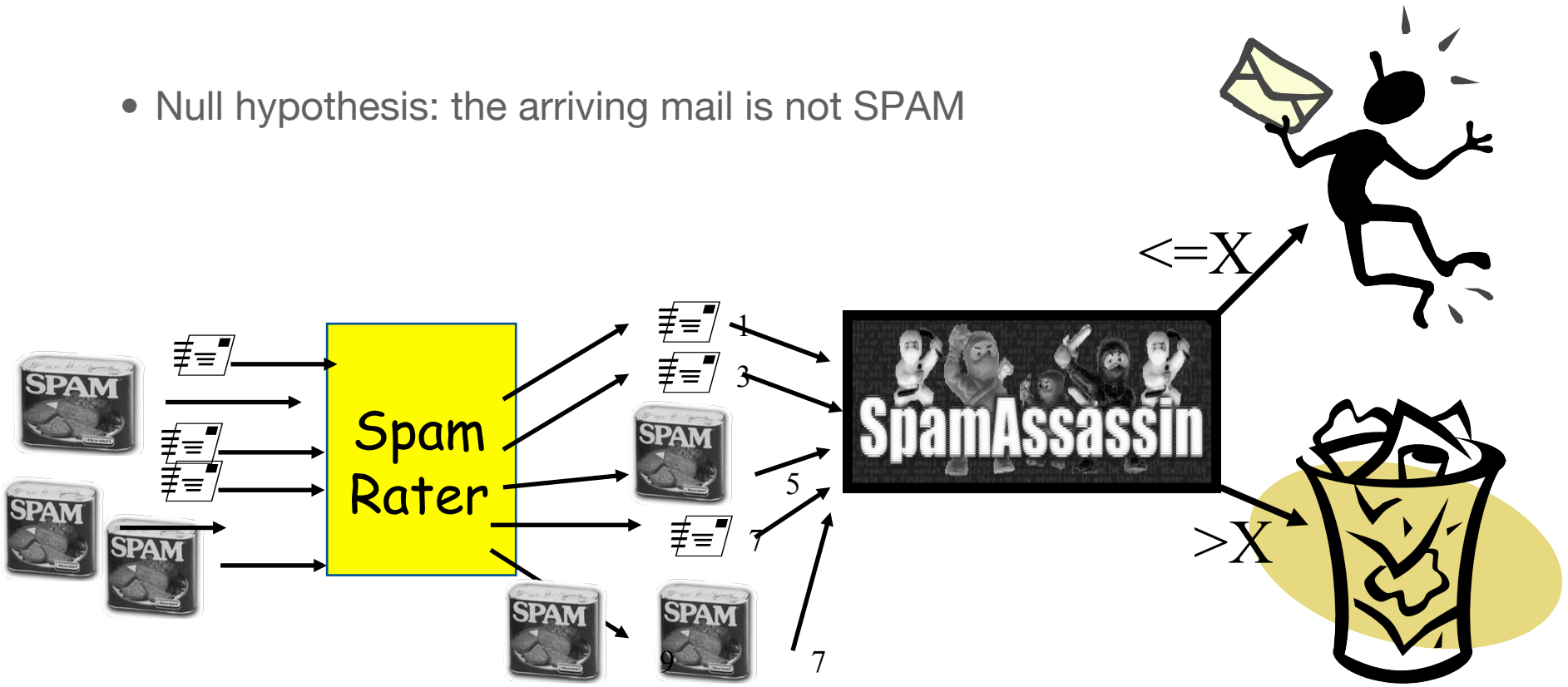
Example: The SpamAssasin Spam Filter

- A SPAM rater gives each email a SPAM likelihood
  - 0: definitely not SPAM
  - 1:..
  - 2:…
  - Etc
  - 10: definitely SPAM



Example: The SpamAssasin Spam Filter

- The SPAMAssassin deletes mail above a certain SPAM threshold
  - What should this threshold be?
  - Null hypothesis: the arriving mail is not SPAM



Example: The SpamAssasin Spam Filter

- Low threshold = many Type I errors
    - Many legitimate emails classified as spam, but you receive very few actual spams
- High threshold = many Type II errors
    - Many spams classified as email, but you receive almost all your valid emails
- Which is worse?

Different Kinds of Errors

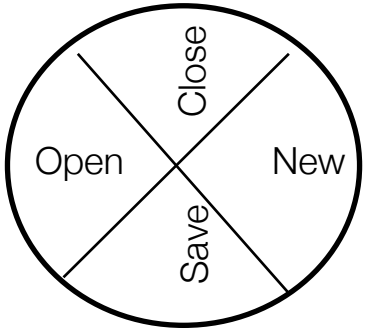
- Type I errors are generally considered worse because the null hypothesis is meant to reflect the incumbent (or default) theory.

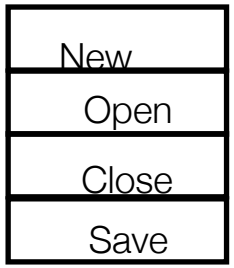
- BUT you must use your judgement to assess actual risk of being wrong in the context of your study.

- In other words, it depends on what you are trying to do, so people often prefer to use a higher confidence level.

Another Example: Menu Types

- H0: There is no difference between Pie and traditional pop-up menus
- What is the consequence of each error type?
- Consequence of Type 1 Error:
  - Extra work developing software
  - People must learn a new idiom for no benefit
- Consequence of Type 2 Error:
  - Use a less efficient (but already familiar) menu





- Which error type is preferable?
  - Redesigning a traditional GUI interface
    - Type 2 error is preferable to a Type 1 error
  - Designing a digital mapping application where experts perform extremely frequent menu selections
    - Type 1 error preferable to a Type 2 error

General Methods to Analyze Data

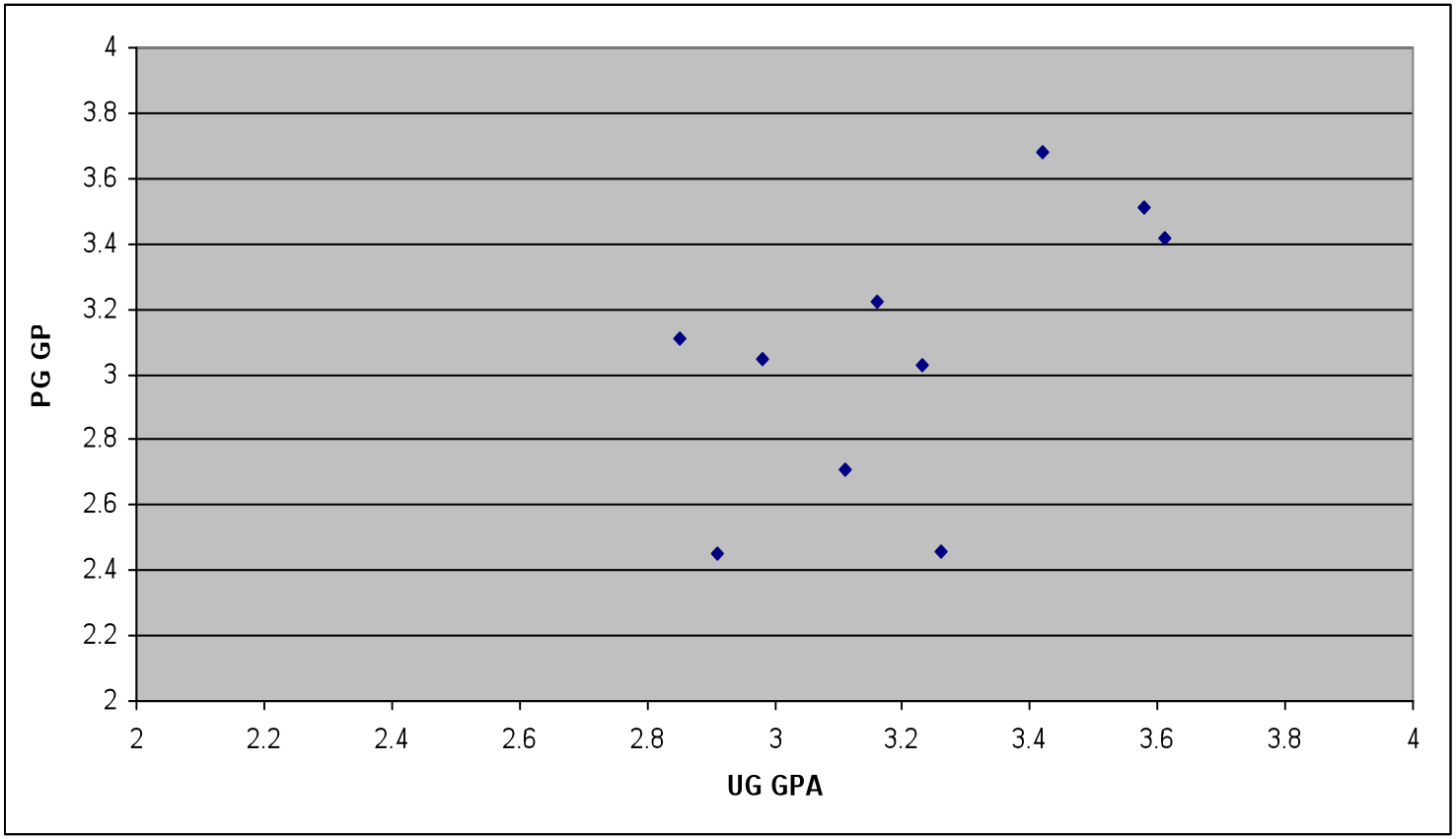
- Parametric Tests
  - Data of interval or (more commonly) ratio scale
  - Selection of subjects from the population random and independent
  - Observations drawn from normally distributed populations
  - Variance of each set of scores or group of scores must be comparable
  - Assume normal distribution, robust and powerful

Examples of Parametric Tests

- T-Test
    - Two groups of data and small sample size
- Correlation
    - Measures association
        - Example: Is there any relationship between the undergraduate GPA and the MSc GPA for a student? Is it that one with a higher undergraduate GPA will get a higher MSc GPA?
- Analysis of Variance and F-Test (ANOVA)
    - Test significant differences between groups (> 2)

Correlation: Example

UG GPA	3.16	3.23	2.91	2.85	3.58	3.26	3.11	2.98	3.61	3.42
PG GPA	3.22	3.03	2.45	3.11	3.51	2.46	2.71	3.05	3.42	3.68



Correlation: Example

- The measurement of correlation is Pearson's coefficient of correlation:

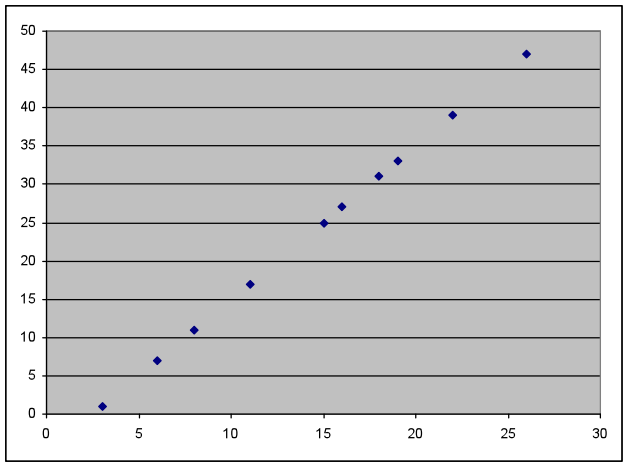
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}.$$

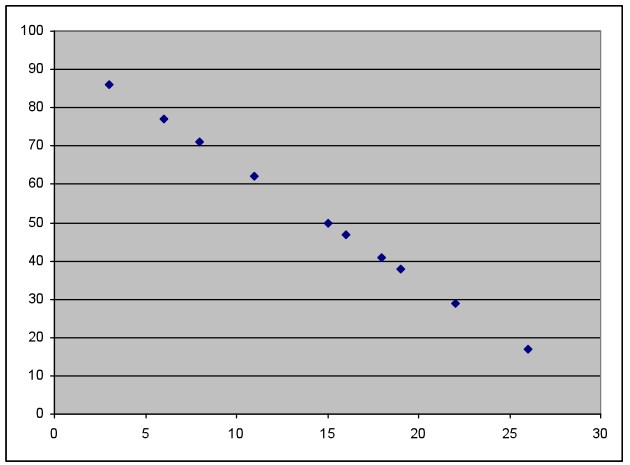
x	3.16	3.23	2.91	2.85	3.58	3.26	3.11	2.98	3.61	3.42	32.11
y	3.22	3.03	2.45	3.11	3.51	2.46	2.71	3.05	3.42	3.68	30.64
x2	9.9856	10.4329	8.4681	8.1225	12.8164	10.6276	9.6721	8.8804	13.0321	11.6964	103.734
y2	10.3684	9.1809	6.0025	9.6721	12.3201	6.0516	7.3441	9.3025	11.6964	13.5424	95.481
xy	10.1752	9.7869	7.1295	8.8635	12.5658	8.0196	8.4281	9.089	12.3462	12.5856	98.9894

r = (10*98.9894-32.11*30.64) / sqrt(10*103.734-32.11²)sqrt(10*95.481-30.64²)
= 0.602

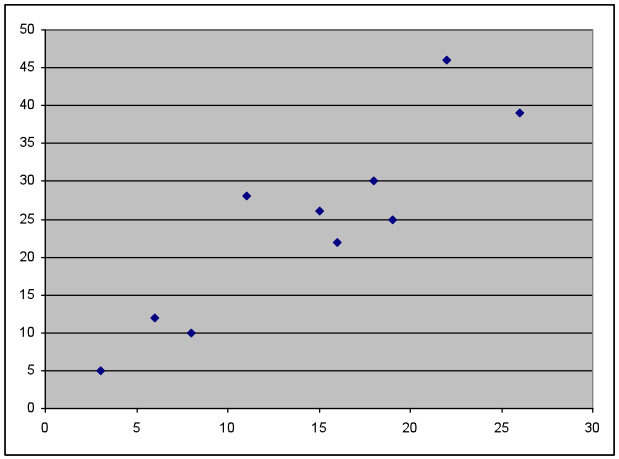
Correlation: Example



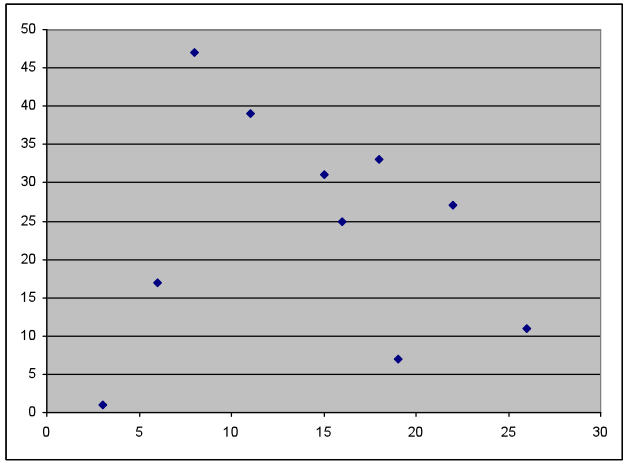
r = 1



r = -1



r = 0.9



r = 0

Methods to Analyze Data (cont'd)

- Nonparametric Tests
  - Data of nominal or ordinal scale
  - Selection of subjects random and independent
  - If scale is ordinal, then homogeneity of variance must be assumed
  - Does not assume normal distribution, less powerful, but more reliable

Examples of Non-Parametric Tests

- Mann-Whitney U Test
- Wilcoxon Matched Pairs Signed Rank Test
- Sign Test

What you should know now

- Controlled experiments can provide clear convincing result on specific issues
- Creating testable hypotheses are critical to good experimental design
- Experimental design requires a great deal of planning
- Statistics inform us about
  - Mathematical attributes about our data sets
  - How data sets relate to each other
  - The probability that our claims are correct
- There are many statistical methods that can be applied to different experimental designs
  - Parametric
  - Non-parametric