

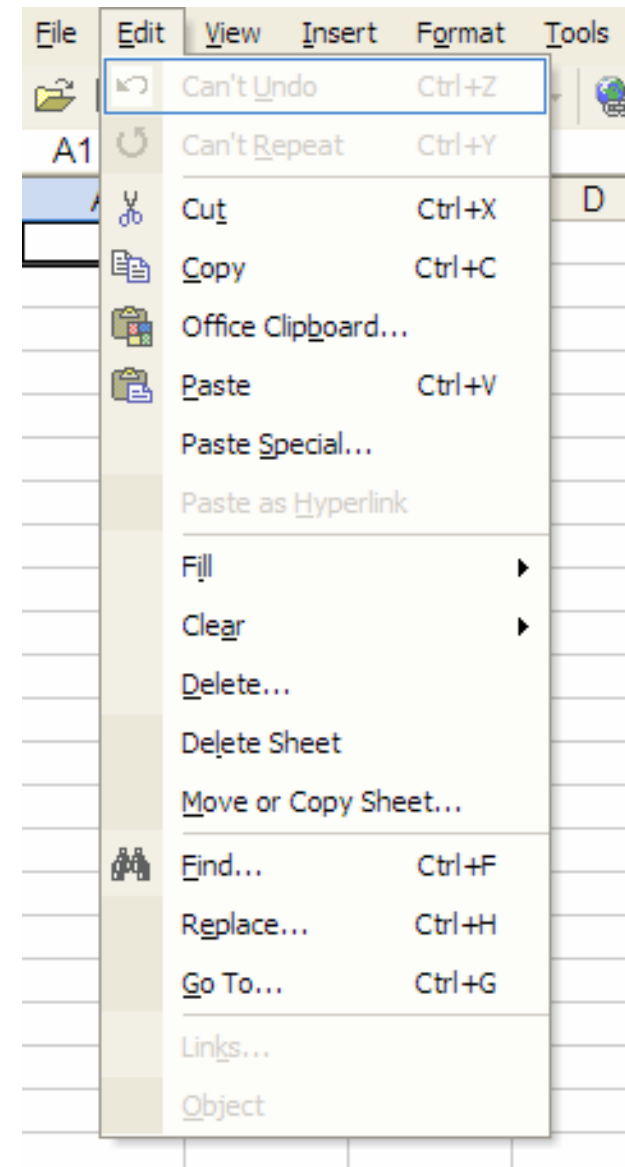
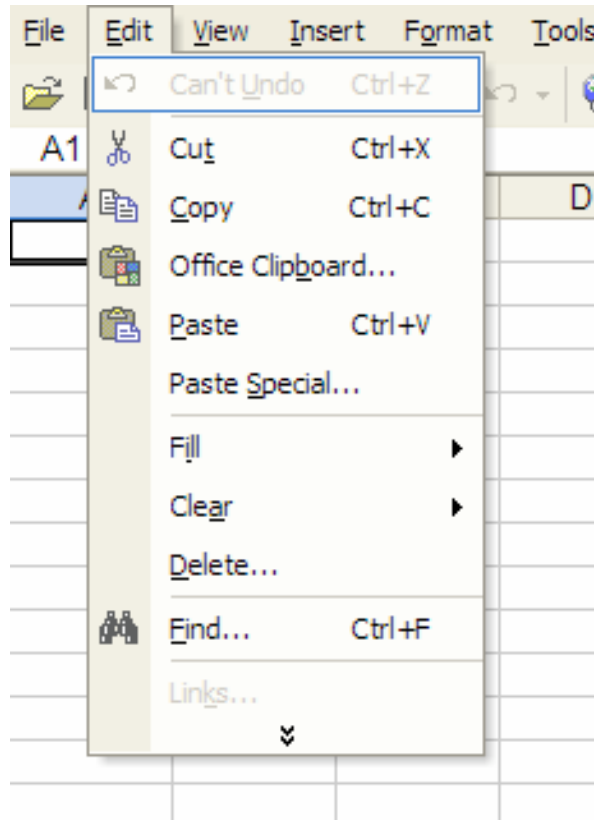
User Interface Design

Dr. Oliver Obst

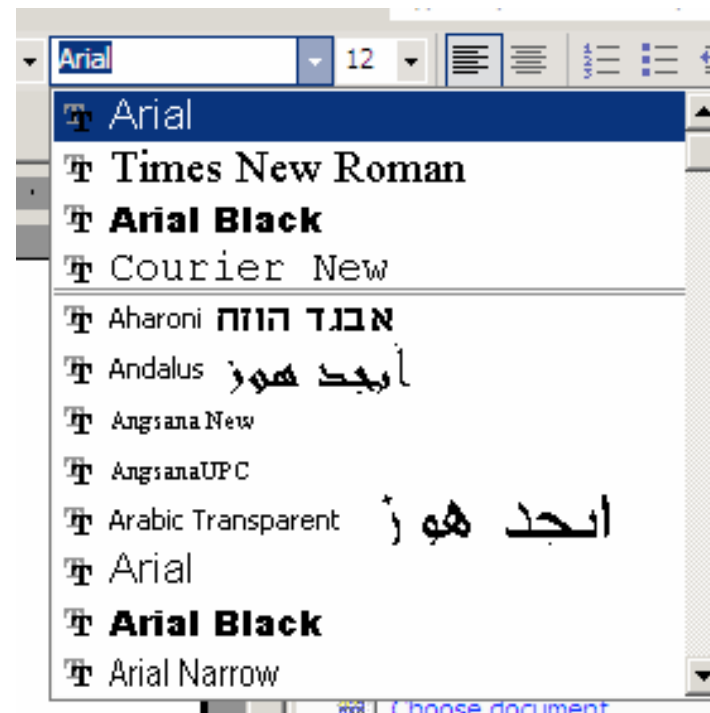
Experiment Design & Analysis

Lecture 12, part 2

UI Hall of Fame/Shame?



UI Hall of Fame/Shame?



Today's Topics

- Experiment design
- Error bar analysis
- Statistical testing

Controlled Experiment

- Start with a testable **hypothesis**
 - e.g. Mac menu bar is faster than Windows menu bar
- Manipulate **independent variables**
 - different interfaces, user classes, tasks
 - in this case, y-position of menubar
- Measure **dependent variables**
 - times, errors, satisfaction
- Use statistical tests to accept or reject the hypothesis

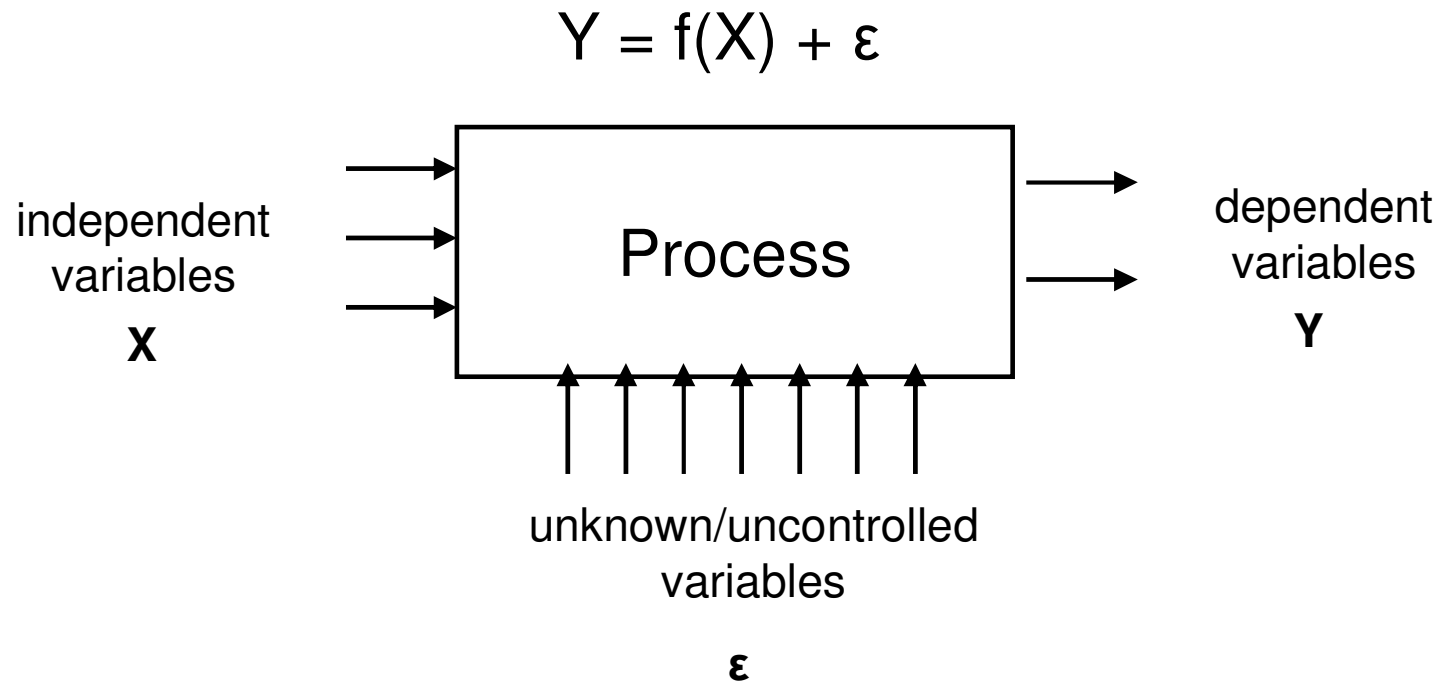
Design of the Menubar Experiment

- Users
 - Windows users or Mac users?
 - Age, handedness?
 - How to sample them?
 - Within-or between-subjects?
- Implementation
 - Real Windows vs. real Mac
 - Artificial window manager that lets us control menu bar position
- Tasks
 - Realistic: word processing, email, web browsing
 - Artificial: repeatedly pointing at fake menu bar
- Measurement
 - When does movement start and end?
- Ordering
 - of tasks and interface conditions
- Hardware
 - mouse, trackball, touchpad, joystick?
 - PC or Mac? which particular machine?

Design of the Menubar Experiment

- Users
 - Windows users or Mac users?
 - Age, handedness?
 - How to sample them?
 - Within-or between-subjects?
- Implementation
 - Real Windows vs. real Mac
 - Artificial window manager that lets us control menu bar position
- Tasks
 - Realistic: word processing, email, web browsing
 - Artificial: repeatedly pointing at fake menu bar
- Measurement
 - When does movement start and end?
- Ordering
 - of tasks and interface conditions
- Hardware
 - mouse, trackball, touchpad, joystick?
 - PC or Mac? which particular machine?

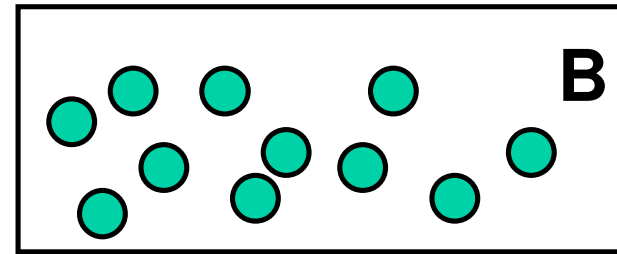
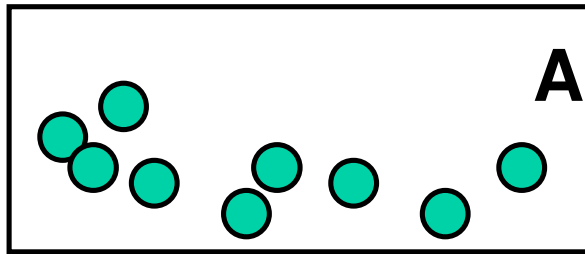
Schematic View of Experiment Design



Concerns Driving Experiment Design

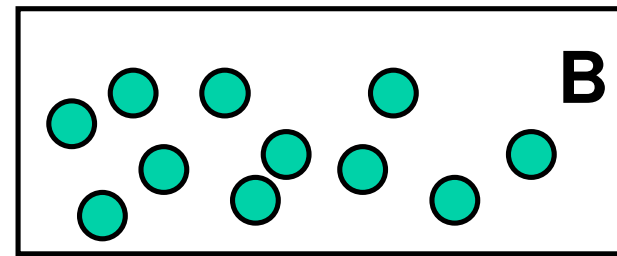
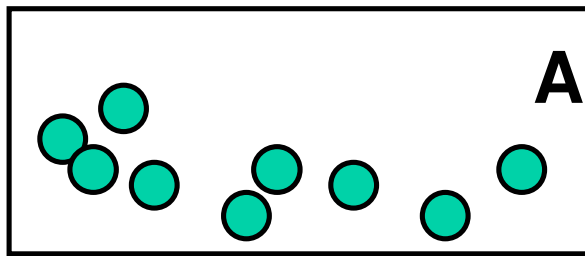
- Internal validity
 - Are observed results actually **caused** by the independent variables?
- External validity
 - Can observed results be **generalised** to the world outside the lab?
- Reliability
 - Will consistent results be obtained by **repeating** the experiment?

How Many Tennis Balls in Each Box



- Hypothesis: box A has a different number of balls than box B
- Reliability
 - Counting the balls manually is reliable only if there are few balls
 - Repeated counting improves reliability
- Internal validity
 - Suppose we weigh the boxes instead of counting balls
 - What if an A ball has different weight than a B ball?
 - What if the boxes themselves have different weights?
- External validity
 - Does this result apply to all boxes in the world labelled A and B?

How Many Tennis Balls in Each Box



- Hypothesis: box A has a different number of balls than box B
- Reliability
 - Counting the balls manually is reliable only if there are few balls
 - Repeated counting improves reliability
- Internal validity
 - Suppose we weigh the boxes instead of counting balls
 - What if an A ball has different weight than a B ball?
 - What if the boxes themselves have different weights?
- External validity
 - Does this result apply to all boxes in the world labelled A and B?

Threats to Internal Validity

- **Ordering effects**

- People learn, and people get tired
- Don't present tasks or interfaces in same order for all users
- Randomise or counterbalance the ordering

- **Selection effects**

- Don't use pre-existing groups (unless group is an independent variable)
- Randomly assign users to independent variables

- **Experimenter bias**

- Experimenter may be enthusiastic about interface X but not Y
- Give training and briefings on paper, not in person
- Provide equivalent training for every interface
- Double-blind experiments prevent both subject and experimenter from knowing if it's condition X or Y
 - ▶ Essential if measurement of dependent variables requires judgement

Threats to Internal Validity

- **Ordering effects**
 - People learn, and people get tired
 - Don't present tasks or interfaces in same order for all users
 - Randomise or counterbalance the ordering
- **Selection effects**
 - Don't use pre-existing groups (unless group is an independent variable)
 - Randomly assign users to independent variables
- **Experimenter bias**
 - Experimenter may be enthusiastic about interface X but not Y
 - Give training and briefings on paper, not in person
 - Provide equivalent training for every interface
 - Double-blind experiments prevent both subject and experimenter from knowing if it's condition X or Y
 - ▶ Essential if measurement of dependent variables requires judgement

Threats to External Validity

- Population

- Draw a random sample from your real target population

- Ecological

- Make lab conditions as realistic as possible in important respects

- Training

- Training should mimic how real interface would be encountered and learned

- Task

- Base your tasks on task analysis

Threats to Reliability

- Uncontrolled variation

- Previous experience
 - ▶ Novices and experts: separate into different classes, or use only one class
- User differences
 - ▶ Fastest users are 10 times faster than slowest users
- Task design
 - ▶ Do tasks measure what you're trying to measure?
- Measurement error
 - ▶ Time on task may include coughing, scratching, distractions

- Solutions

- Eliminate uncontrolled variation
 - ▶ Select users for certain experience (or lack thereof)
 - ▶ Give all users the same training
 - ▶ Measure dependent variables precisely
- Repetition
 - ▶ Many users, many trials
 - ▶ Standard deviation of the mean shrinks like the square root of N (i.e., quadrupling users makes the mean twice as accurate)

Threats to Reliability

- Uncontrolled variation

- Previous experience
 - ▶ Novices and experts: separate into different classes, or use only one class
- User differences
 - ▶ Fastest users are 10 times faster than slowest users
- Task design
 - ▶ Do tasks measure what you're trying to measure?
- Measurement error
 - ▶ Time on task may include coughing, scratching, distractions

- Solutions

- Eliminate uncontrolled variation
 - ▶ Select users for certain experience (or lack thereof)
 - ▶ Give all users the same training
 - ▶ Measure dependent variables precisely
- Repetition
 - ▶ Many users, many trials
 - ▶ Standard deviation of the mean shrinks like the square root of N (i.e., quadrupling users makes the mean twice as accurate)

Blocking

- Divide samples into subsets which are more homogeneous than the whole set
 - Example: testing wear rate of different shoe sole material
 - Lots of variation between feet of different kids
 - But the feet on the same kid are far more homogeneous
 - Each child is a block
- Apply all conditions within each block
 - Put material A on one foot, material B on the other
- Measure difference within block
 - $\text{Wear}(A) - \text{Wear}(B)$
- Randomise within the block to eliminate internal validity threats
 - Randomly put A on left foot or right foot

Between Subjects vs. Within Subjects

- “Between subjects” design
 - Users are divided into two groups:
 - ▶ One group sees only interface X
 - ▶ Other group sees only interface Y
 - Results are compared between different groups
 - ▶ Is $\text{mean}(x_i) > \text{mean}(y_j)$?
 - Eliminates variation due to ordering effects
 - ▶ User can't learn from one interface to do better on the other
- “Within subjects” design
 - Each user sees both interface X and Y (in random order)
 - Results are compared within each user
 - ▶ For user i , compute the difference $x_i - y_i$
 - ▶ Is $\text{mean}(x_i - y_i) > 0$?
 - Eliminates variation due to user differences
 - ▶ User only compared with self

Experiment Analysis

- Hypothesis: Mac menubar is faster to access than Windows menubar
- Design: between-subjects, randomised assignment of interface to subject

Windows	Mac
400	360
220	210
560	500
340	305

Statistical Testing

- Compute a statistic summarising the experimental data
 - mean(Win)
 - mean(Mac)
- Apply a statistical test
 - t test: are two means different?
 - ANOVA (ANalysis Of VAriance): are three or more means different?
- Test produces a p value
 - p value = probability that the observed difference happened purely by chance
 - If $p < 0.05$, then we are 95% confident that there is a difference between Windows and Mac

Hypothesis Testing

- Our hypothesis: position of menubar matters
 - i.e., $\text{mean}(\text{Mac times}) < \text{mean}(\text{Windows times})$
 - This is called the alternative hypothesis (also called H_1)
- If we're wrong: position of menu bar makes no difference
 - i.e., $\text{mean}(\text{Mac}) = \text{mean}(\text{Win})$
 - This is called the null hypothesis (H_0)
- We can't really disprove the null hypothesis
 - Instead, we argue that the chance of seeing a difference **at least as extreme** as what we saw is very small if the null hypothesis is true

Statistical Significance

- Compute a statistic from our experimental data
 $X = \text{mean}(\text{Win}) - \text{mean}(\text{Mac})$
- Determine the probability distribution of the statistic assuming H_0 is true
 $\text{Pr}(X=x | H_0)$
- Measure the probability of getting the same or greater difference
 $\text{Pr}(X > x_0 | H_0)$ one-sided test
 $2 \text{Pr}(X > |x_0| | H_0)$ two-sided test
- If that probability is less than 5%, then we say
 - “We reject the null hypothesis at the 5% significance level”
 - equiv.: “difference between menubars is statistically significant ($p < .05$)”
- Statistically significant does not mean scientifically important

Statistical Significance

- Compute a statistic from our experimental data
 $X = \text{mean}(\text{Win}) - \text{mean}(\text{Mac})$
- Determine the probability distribution of the statistic assuming H_0 is true
 $\text{Pr}(X=x \mid H_0)$
- Measure the probability of getting the same or greater difference
 $\text{Pr}(X > x_0 \mid H_0)$ one-sided test
 $2 \text{Pr}(X > |x_0| \mid H_0)$ two-sided test
- If that probability is less than 5%, then we say
 - “We reject the null hypothesis at the 5% significance level”
 - equiv.: “difference between menubars is statistically significant ($p < .05$)”
- Statistically significant does not mean scientifically important

T test

- T test compares the means of two samples
- Two-sided:
 - H_0 : means are equal
 - H_1 : means are different
- One-side:
 - H_0 : means are equal
 - H_1 : $\text{mean}(A) < \text{mean}(B)$
- Assumptions:
 - samples A & B are independent (between-subjects, randomized)
 - normally distributed
 - equal variance

T test

- T test compares the means of two samples
- Two-sided:
 - H_0 : means are equal
 - H_1 : means are different
- One-side:
 - H_0 : means are equal
 - H_1 : $\text{mean}(A) < \text{mean}(B)$
- Assumptions:
 - samples A & B are independent (between-subjects, randomized)
 - normally distributed
 - equal variance

Paired T Test

- For within-subject experiments
- Uses the mean of the differences (each user against themselves)
- H_0 : mean of differences is zero
- H_1 : mean of differences is nonzero (two-sided test)

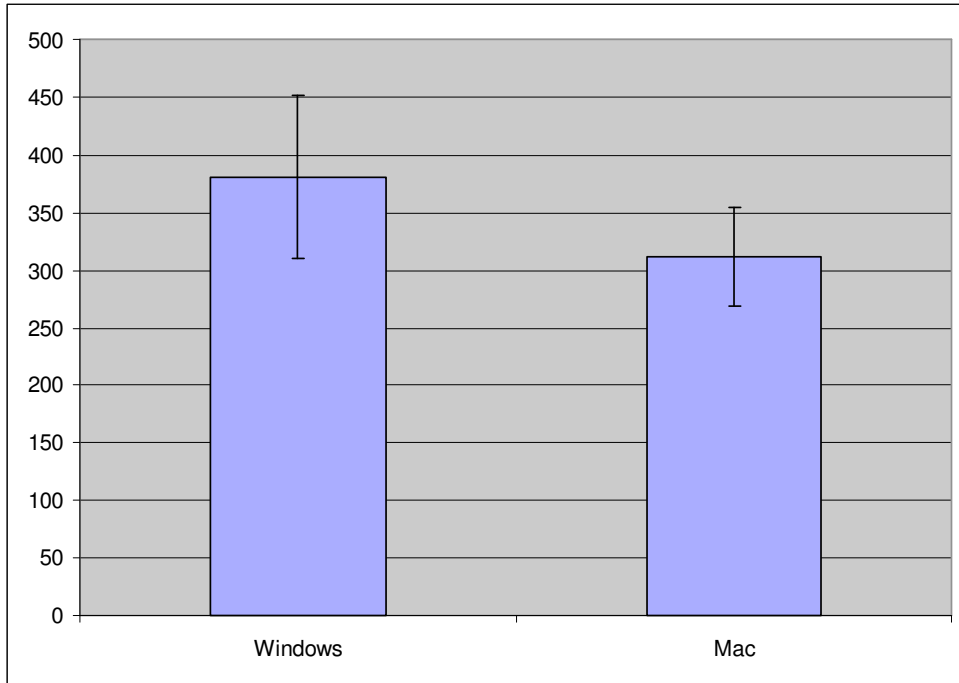
Analysis of Variance (ANOVA)

- Compares more than 2 means
- One-way ANOVA
 - 1 independent variable with $k \geq 2$ levels
 - H_0 : all k means are equal
 - H_1 : the means are different (so the independent variable matters)

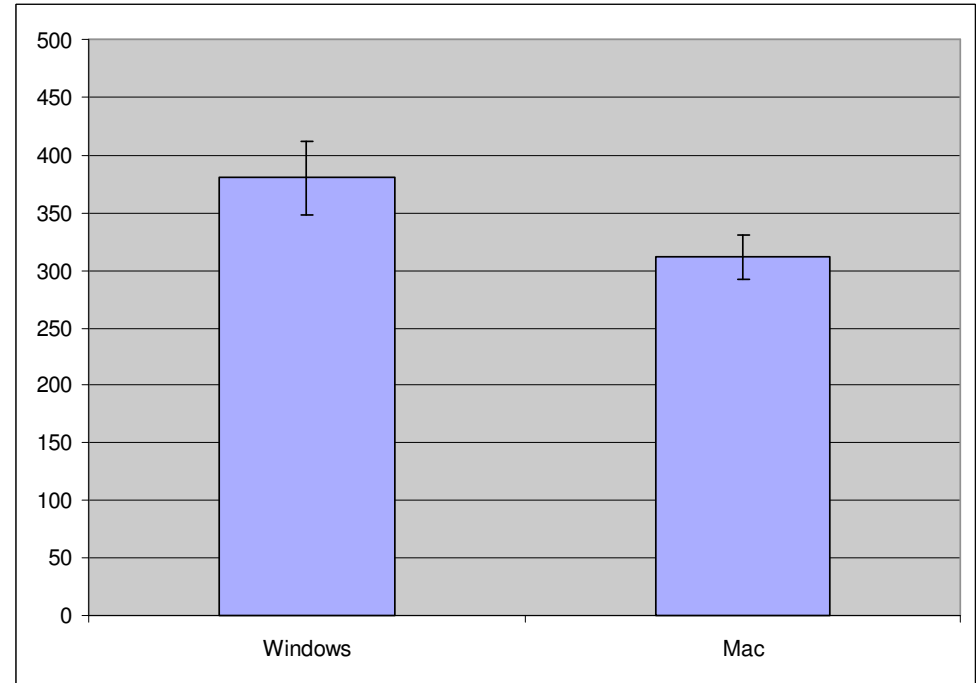
Two-Way ANOVA

- 2 independent variables with j and k levels, respectively
- Tests whether each variable has an effect independently
- Also tests for interaction between the variables

Standard Error of the Mean

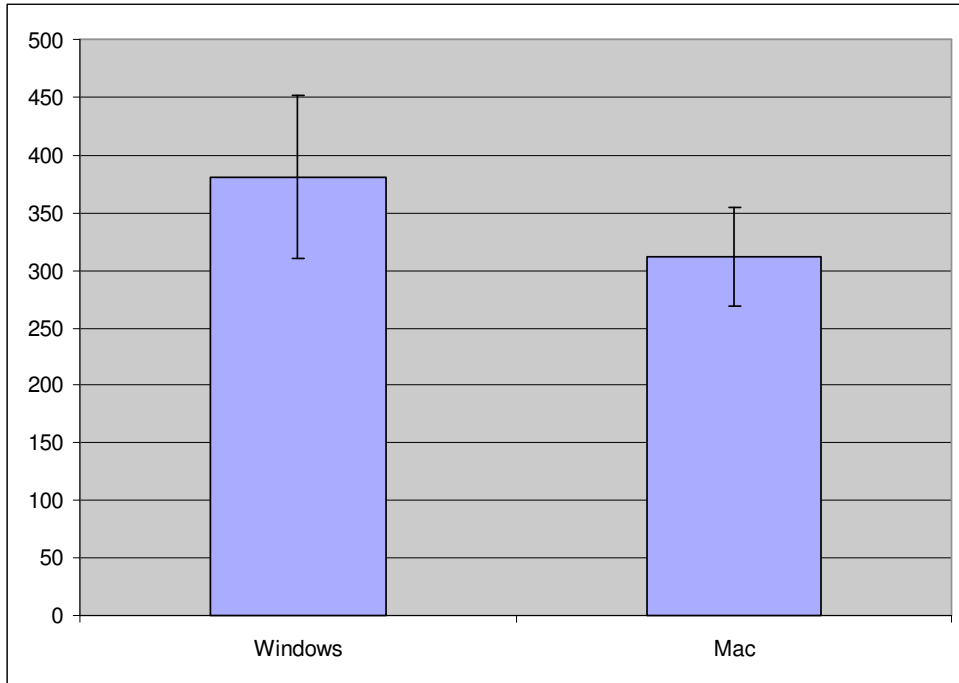


4 users for each condition:
Error bars overlap, so can't
conclude anything

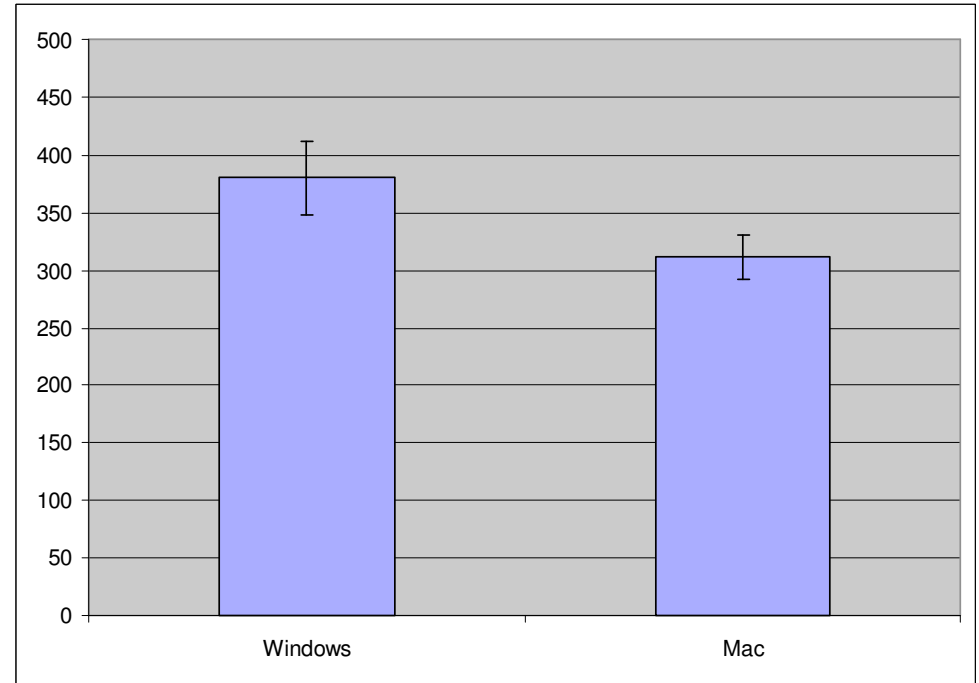


16 users for each condition:
Error bars are disjoint, so
Windows is significantly different
from Mac

Standard Error of the Mean



4 users for each condition:
Error bars overlap, so can't
conclude anything



16 users for each condition:
Error bars are disjoint, so
Windows is significantly different
from Mac

