Questions

- Final presentation (see wiki)
- Comments expecations

- Course evaluations
  www.courseevalum.umd.edu
- No office hours today

Qualitative vs. Quantitative

- Qualitative: Develop understanding of human experience
- Quantitative: Objectively measure human performance

- Less about more vs. more about less

When are each appropriate?

Quantitative Evaluation

- Gather (performance) measurements
- Methods
  - User interaction collection (i.e., logging)
    - *Mouse clicks, keys pressed,…*
    - *Data collected during system use*
      - Google, Amazon
  - Controlled experiments
    - *Set forth a testable hypothesis*
    - *Manipulate one or more* independent *variable*
    - *Observe effect on one or more* dependent *variable*
    - *Can be reproduced by others*

Controlled experiment

- State a lucid, testable hypothesis
- Identify independent and dependent variables
- Design the experimental protocol
- Choose the user population
- Apply for human subjects protocol review (IRB)
- Run some pilot participants
- Fix the experimental protocol
- Run the experiment
- Perform statistical analysis
- Draw conclusions
- Communicate results

Question Experiment

- Is it reliable?
  - Will you get the same result if someone else repeats the experiment?
  - Does the experiment take into account variations between subjects?
    - *Need for testing a sample of subjects*
- Is it valid?
  - Does the experiment reflects target use?
    - *Were users typical?*
    - *Were tasks typical?*
    - *Was the setting realistic?*
    - *Was the experience biased?*

Are results significant?

- Statistical significance
  - Chance that observed result is due to chance
  - Type I errors are the most disruptive

	Actual Situation	
Researcher's Decision	NO effect	Effect
NO effect	Correct decision	**Type II error**
Effect	**Type I error**	Correct decision

- Design significance?
  - 3.00s versus 3.05s?

Are results significant?

- Statistical significance
  - Comparing to the null hypothesis: "There is no effect"
  - Type I errors are the most disruptive

Researcher's Decision	Actual Situation: Null Hypothesis is True	Actual Situation: Null Hypothesis is False
Fail to reject the null hypothesis	Correct decision	**Type II error**
Reject the null hypothesis	**Type I error**	Correct decision

- Design significance?
  - 3.00s versus 3.05s?

Example

- Examine value of animation during scrolling [Klein & Bederson 2005] http://portal.acm.org/citation.cfm?doid=1056808.1057068
  - Various reading tasks
  - For various document types
  - While scrolling with varying kinds of animated transitions

State a lucid, testable hypothesis

"Animated scrolling will speed up reading and decrease errors, especially for plain documents."

Choose the variables

- Manipulate one or more *independent* variable (the thing you change)
  - Document type
  - Animation speed
- Observe effect on one or more *dependent* variable (the thing you measure)
  - Time to completion
  - Accuracy (i.e., error rate)

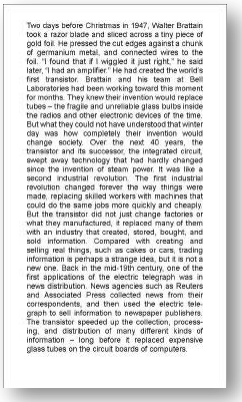
Design the experimental protocol

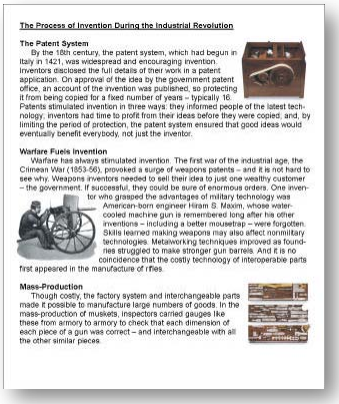
- Between or within subjects?
  - Between subjects: each subject runs one condition
    - *Need more subjects*
    - *Difference between subjects might introduce a bias*
    + *No learning effects across conditions*
  - Within subjects: each subject runs several conditions
    + *Need fewer subjects*
    + *No bias across participants*
    - *No learning effects across conditions*
  - Very important for the statistical analysis phase

- Which tasks?
  - Must reflect the hypothesis
  - Must avoid bias
    - *Instructions, ordering…*
    - *In doubt, always favor the null hypothesis*

Design the experimental protocol

- Running Example:
  - Reading while scrolling


Two days before Christmas in 1947, Walter Brattain took a razor blade and sliced across a tiny piece of gold foil. He pressed the cut edges against a chunk of germanium metal, and connected wires to the foil. "I found that if I wiggled it just right," he said later, "I had an amplifier." He had created the world's first transistor. Brattain and his team at Bell Laboratories had been working toward this moment for months. They knew their invention would replace tubes – the fragile and unreliable glass bulbs inside the radios and other electronic devices of the time. But what they could not have understood that winter day was how completely their invention would change society. Over the next 40 years, the transistor and its successor, the integrated circuit, swept away technology that had hardly changed since the invention of steam power. It was like a second industrial revolution. The first industrial revolution changed forever the way things were made, replacing skilled workers with machines that could do the same jobs more quickly and cheaply. But the transistor did not just change factories or what they manufactured, it replaced many of them with an industry that created, stored, bought, and sold information. Compared with creating and selling real things, such as cakes or cars, trading information is perhaps a strange idea, but it is not a new one. Back in the mid-19th century, one of the first applications of the electric telegraph was in news distribution. News agencies such as Reuters and Associated Press collected news from their correspondents, and then used the electric telegraph to sell information to newspaper publishers. The transistor speeded up the collection, processing, and distribution of many different kinds of information – long before it replaced expensive glass tubes on the circuit boards of computers.


The Process of Invention During the Industrial Revolution

The Patent System

By the 18th century, the patent system, which had begun in Italy in 1421, was widespread and encouraging invention. Inventors disclosed the full details of their work in a patent application. On approval of the idea by the government patent office, an account of the invention was published, so protecting it from being copied for a fixed number of years – typically 16. Patents stimulated invention in three ways: they informed people of the latest technology; inventors had time to profit from their ideas before they were copied; and, by limiting the period of protection, the patent system ensured that good ideas would eventually benefit everybody, not just the inventor.

Warfare Fuels Invention

Warfare has always stimulated invention. The first war of the industrial age, the Crimean War (1853-56), provoked a surge of weapons patents – and it is not hard to see why. Weapons inventors needed to sell their idea to just one wealthy customer – the government. If successful, they could be sure of enormous orders. One inventor who grasped the advantages of military technology was American-born engineer Hiram S. Maxim, whose water-cooled machine gun is remembered long after his other inventions – including a better mousetrap – were forgotten. Skills learned making weapons may also affect nonmilitary technologies. Metalworking techniques improved as foundries struggled to make stronger gun barrels. And it is no coincidence that the costly technology of interoperable parts first appeared in the manufacture of rifles.

Mass-Production

Though costly, the factory system and interchangeable parts made it possible to manufacture large numbers of goods. In the mass-production of muskets, inspectors carried gauges like these from armory to armory to check that each dimension of each piece of a gun was correct – and interchangeable with all the other similar pieces.

Confound!

Chose the user population

- Pick a well balanced sample
  - Novices, experts, average
  - Age group
  - Sex…
- Population group may be one of the independent variable
- Running example
  - Varied population, did not control

Run the experiment

- Always run pilots first!
  - There are always unexpected problem!
  - When the experiment has started you cannot pick and choose
- Use a check-list so that all subjects follow the same steps
- IRB - Don’t forget the consent form!
- Don’t forget to debrief each subject

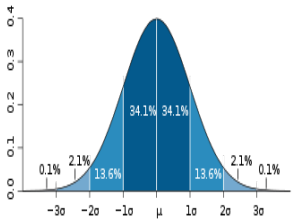
Run statistical analysis

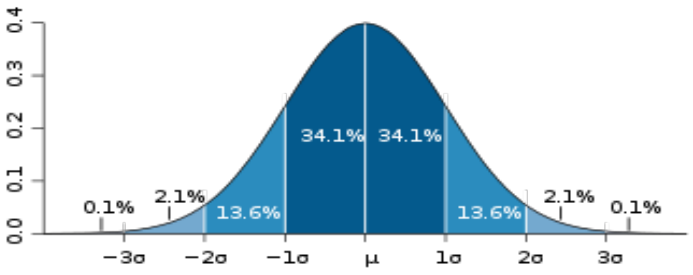
- Properties of our result data
  - Mean, variance…
- How different data sets relate to each other
  - Are we sampling from similar of different distributions?
- Probability that our claims are correct
  - Statistical significance:

    "The hypothesis that technique X is faster is accepted ($p < .05$)" means that there is a higher than 95% chance the hypothesis is true
  - Typical levels are .05 and .01 level
  - These levels are socially determined

Statistical tools I - Descriptive

- Mean
  - Average
- Median
  - Central value
- Standard Deviation
  - Measure of variation





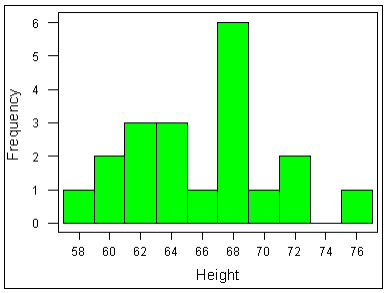
Statistical tools I - Descriptive

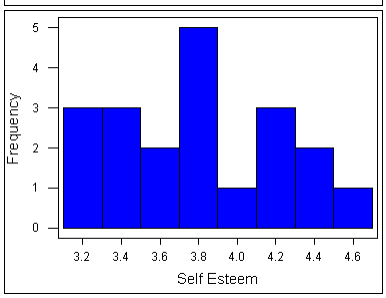
- Correlation
  - Measure the extent to which 2 concepts are related
  - Caveats
    - *Correlation does not imply cause and effect (hidden variable)*
      - Ice cream consumption and drowning
      - Third variable problem
      - Directionality problem
    - *Need a large enough group*

  - r = 0 (no correlation)
    r = 1 (positively correlated)
    r = -1 (negatively correlated)

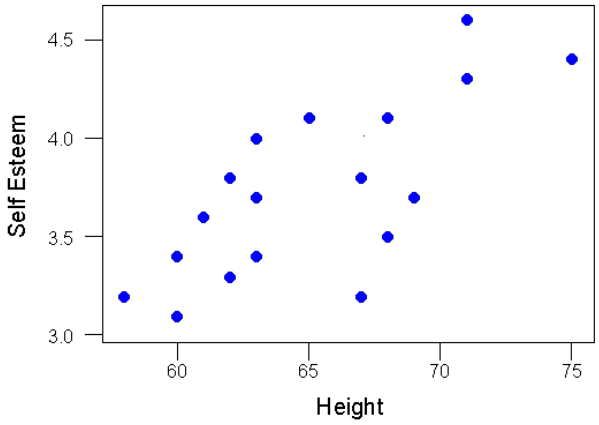
Example (random data)

Person	Height	Self Esteem
1	68	4.1
2	71	4.6
3	62	3.8
4	75	4.4
5	58	3.2
6	60	3.1
7	67	3.8
8	68	4.1
9	71	4.3
10	69	3.7
11	68	3.5
12	67	3.2
13	63	3.7
14	62	3.3
15	60	3.4
16	63	4.0
17	65	4.1
18	67	3.8
19	63	3.4
20	61	3.6
MEAN	**65.4**	**3.8**
SD	**4.4**	**0.4**





Example (continued)

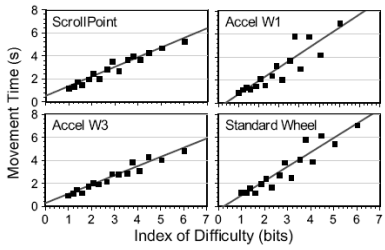


Correlation: r = 0.73

Statistical tools I - Descriptive

- Regression
  - Calculate the "best fit"

	R	R^2	Slope	Intercept (s)	IP (bps)
ScrollPoint	0.97	0.94	0.84	0.42	1.19
Accel W1	0.90	0.81	1.16	-0.51	0.86
Accel W3	0.97	0.95	0.80	0.18	1.25
Wheel Std	0.94	0.88	1.25	-0.42	0.80



Statistical tools II - Analytical

- T-test
  - Compare the mean of 2 populations
    - *Null hypothesis: no difference between means*
    - *Can only examine a single independent variable*
  - Assumptions
    - *Samples are normally distributed*
      - Very robust in practice
    - *Population variances are equal*
      - Reasonably robust for differing variances
    - *Individual observations in samples are independent*
      - Very important

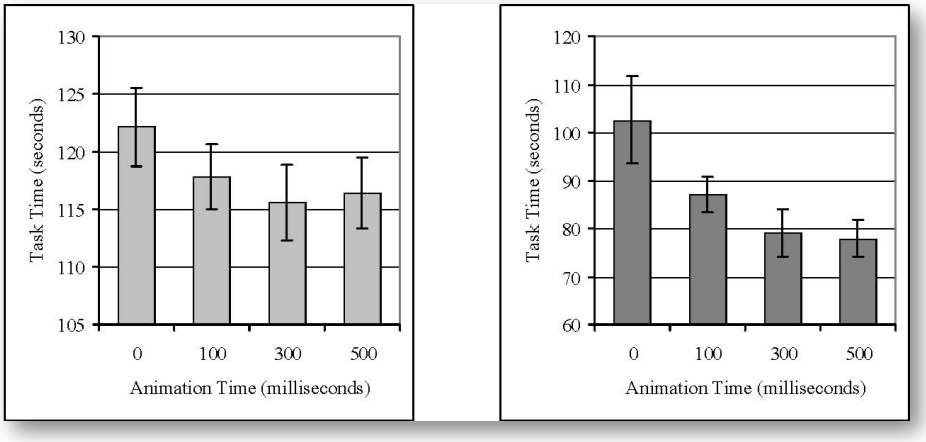
Statistical tool II - Analytical

- ANOVA
  - Single factor analysis of variance
    - *Compare three or more means*
  - Analysis of variance
    - *Compare relationship between many factors*
      - Beginners type at the same speed on all keyboards,
      - Touch-typist type fastest on the qwerty
- Your protocol influences the kind of test you can use
  - If in doubt, consult with a statistician before starting the experiment!

Reporting Results

- "This analysis revealed a significant main effect for Device, $F(2,15)=15.2$, $p<0.001$."
- "As one would expect, movement times increased as either W decreased or D increased (i.e., as the task got more difficult: for W, $F(2,25)=801$, $p<0.001$; and for D, $F(3,54)=1429$, $p<0.001$)."

Running example data



Error bars show +/- one standard deviation

Running Example Analysis

- Measured speed, error and RSD (relative subjective duration)
- Significant results for
  - reading time
  - reading error
  - counting time
- Animated scrolling
  - reduces reading errors by up to 54%
  - task time by up to 3%
  - task time by up to 24% for counting tasks
- 300ms seems to be best overall animation time for reading.
- Formatted documents had a higher base rate perf, and lower improvement – thus visual landmarks appear to be a good idea

In Class Experiment

- Compare mouse vs. keyboard
- Independent variable:
  - Interaction type (conditions: mouse, keyboard)
- Dependent variable:
  - Speed
- Experiment:
  - Bring up two text editor windows with 10 lines of text in one
  - Copy each line of text (one at a time) from one window to the other
  - Conditions:
    - Select text with mouse and press ctrl-c, then alt-tab, ctrl-v
    - Select text with mouse and selected edit->copy menu, then edit->paste
  - Manually measure entire task and divide by 10
  - Train on 3 lines
- Enter data at http://bederson.etherpad.com/4