UMAP 2012 Tutorial 2

Empirical Evaluation of User Modeling Systems

David N. Chin
chin@hawaii.edu

Univ. of Hawaii
Dept. of Information & Computer Sciences

Introduction

- Do UMs help/hinder your system?
  - Experiment design
  - How to run your experiments
  - Statistical data analysis
- No background in statistics needed

Agenda

I. Experiment Design
- A. Independent vs. dependent variables
- B. Nuisance variables
- C. Between-subjects vs. within-subjects designs
- D. Estimating sensitivity
- E. Factorial designs
- F. Caveats

II. Running Experiments
- A. Participants
- B. Controlling the environment
- C. Recording data

III. Experiment Analysis
- A. Means and variance
- B. Statistical tests
- C. ANOVA
- D. Explained variance

IV. Summary

Agenda

I. Experiment Design
- **A. Independent vs. dependent variables**
- B. Nuisance variables
- C. Between-subjects vs. within-subjects designs
- D. Estimating sensitivity
- E. Factorial designs
- F. Caveats

II. Running Experiments
- A. Participants
- B. Controlling the environment
- C. Recording data

III. Experiment Analysis
- A. Means and variance
- B. Statistical tests
- C. ANOVA
- D. Explained variance

IV. Summary

Independent Variables

- Conditions varied by experimenter
  - Absence or presence of a user model
  - User model A vs. user model B (vs. UM C)
  - Different levels of user modeling
  - Different UM parameter settings
  - Different user interfaces

Dependent Variables

- Response variables or recorded measures:
  - Frequency certain behaviors occur
  - Qualities of a behavior in a particular situation
  - Number of errors
  - Time to complete tasks
  - Quality of task results
  - Interaction patterns
  - Subjective evaluations

Covariant Variables

- Concomitant variables (covariates)
  - Not under experimental control
  - Age, gender, socioeconomic status, education, learning styles, previous experience, prior knowledge, aptitudes
  - Statistics: Analysis of covariance (ANCOVA)

26 June 2007 UM-07 tutorial 3: Chin 7

Covariant variables add noise to the measurements of dependent variables. For example, more computer-literate people may work faster in a web search task. The noise from this variability in task time may swamp the actual difference in mean search times with or without a UM helping the search. ANCOVA allows us to measure the covariate of computer literacy and use that to correct the search times to remove the noise added by differing degrees of computer literacy from the measured search time dependent variable. The next slides will show you commonly accepted measurements for certain covariates.

Cognitive Tests

- Kit of Factor-Referenced Cognitive Tests
  - Visualization, visual memory, memory span, perceptual speed, etc.
  - Ekstrom & French, Educational Testing Service
- Human Information Processing Survey
  - Left/right brain, integrated or mixed
  - Taggart & Torrance, Scholastic Testing Service

26 June 2007 UM-07 tutorial 3: Chin 8

The kit is a tool for studying reasoning, verbal ability, spatial ability, memory, and other cognitive processes. It contains 72 tests that have been demonstrated to be consistent markers in studies of 23 cognitive factors. The kit tests are intended for research use only. They should not be used for selection, counseling, or operational purposes. Information about the development of the 1976 edition of the kit may be found in: Ekstrom, R. B., French, J. W., & Harman, H. H. (1979). Cognitive factors: Their identification and replication. *Multivariate Behavioral Research Monographs, 79*(2). Buy from http://www.ets.org/research/policy_research_reports/monographs/kit_of_factor_referenced_cognitive_tests

The **Human Information Processing® Survey (HIP®)** is a training tool for human resource development. Individuals are assessed in terms of their processing preference: left-brain, right-brain, integrated, or mixed. The The **HIP® Strategy and Tactics Profiles** provide a description of a person's overall approach, as well as the specific tactics he or she uses in problem solving and decision making.
Professional Edition of the **HIP® Survey**, which can suggest how an individual may perform in the workplace, utilizes consumable, self-scoring survey forms and Strategy and Tactics Profiles. For university personnel and others studying human information processing, the Research Edition includes reusable survey forms, response sheets, and Strategy Profiles. Both editions of the **HIP® Survey** are time- and cost-effective methods of measuring the degree to which individuals think with either brain hemisphere. Buy from http://www.ststesting.com/2005gifthip.html

More Cognitive Tests

- Group Embedded Figures Test
  - Field independence
  - Witkin, Oltman, Raskin & Karp, mind garden
- Nelson-Denny Reading Test
  - Reading ability
  - Riverside Publishing

26 June 2007 UM-07 tutorial 3: Chin 9

From http://www.usd.edu/~ssanto/field.html:

Field independence and field dependence are sometimes referred to as "cognitive controls" in that they control the ways that individuals process information. Assessed by Group Embedded Figures Test, the idea behind field independence is that performance on perceptual/spatial tasks can diagnose an individual's ability to learn and perform on non-perceptual tasks.

Field independent students will prefer situations that allow them freedom in working toward their goals and solving problems. These learners like to work individually. Students who are field dependent may prefer group projects and need more assistance from the instructor. One way to help these students is to make sure that any diagrams and illustrations used as visual aids contain verbal information explaining them. In computer-based learning, software that enables the learner to flip and rotate the image, or slides showing different views of the same image, can be helpful. Buy from http://www.mindgarden.com/products/gefts.htm

The *Nelson-Denny Reading Test,* Forms G and H, is a reading survey test for high school and college students and adults. A two-part test, the Nelson-Denny measures vocabulary development, comprehension, and reading rate. Part I (Vocabulary) is a fifteen-minute timed test; Part II (Comprehension and Rate) is a twenty-minute test. The first minute of the Comprehension test is used to determine reading rate. Including the time needed to distribute materials, complete the name and information grids, and provide directions, the Nelson-Denny may be administered in forty-five minutes, or a single class period. A unique feature of the 1993 edition is the extended-time administration of the test to meet the needs of special populations, such as students with English as a second language or as a foreign language, or returning adults. Buy from http://www.riverpub.com/products/ndrt/index.html

Personality Tests

- Meyers-Briggs Type Indicator (MBTI)
  - Extraversion/Introversion
  - Sensing/Intuition
  - Thinking/Feeling
  - Judgment/Perception
  - CAPT
  - Must be trained to give MBTI

26 June 2007 UM-07 tutorial 3: Chin 10

MBTI has 16 personality types, a combination of (from http://www.infj.org/):

Extraversion/Introversion (E/I) describes how we are "energised": **extraverts** recharge and get energised from lots of interaction with other people, while **introverts** need to spend time alone to recharge their internal batteries.

Sensing/Intuition (S/N) describes whether we are more observant (sensing) or introspective (intuitive). **Sensates** pay more attention to the outside world, the current surroundings and its immediate needs, whereas intuitives heed the promptings of the inner world of thoughts and feelings. **Intuitives** are more likely to have their heads in the future or the past, exploring possibilities and pathways - Ns typically like to daydream. Note that this is not to be mistaken for introversion.

Thinking/Feeling (T/F) indicates whether our head or our heart rules us more. Contrary to popular belief, both thinking and feeling (in this context) are **rational** functions, used to make decisions and acting on them. A Feeling personality isn't illogical or irrational, despite what some may try to tell you! Feeling people cherish values more than principles -- so while they may follow rules, they will break them if it means helping somebody or being compassionate to others; the situation determines what the Feeler will do. Thinking types are more likely to stick to the principles and rules no matter what. They use logic to reach a conclusion and act on it.

Judging/Perceiving (J/P) determines how we run our lives. Perceivers prefer keeping their options open and would rather not be tied to a schedule. Note that this doesn't necessarily mean they are messy or disorganised people. With perceptive types, work doesn't have to be finished before play begins! Judgers are much more routine-oriented and orderly; they tend to have agendas, timetables, outlines, and so on. They would rather have closure than leave something unfinished, and prefer working towards a deadline. If they aren't on time, Js tend to get very nervous!

Buy from http://www.capt.org/

More Personality Tests

- Locus of Control
  - Attribution theory
  - Rotter, Queendom

26 June 2007 UM-07 tutorial 3: Chin 11

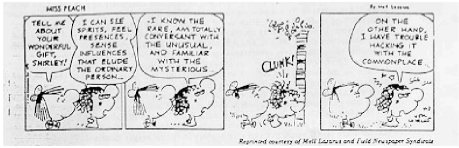
Do you control your destiny or are you controlled *by* it? This test assesses your locus of control orientation and your attribution style.

"A locus of control orientation is a belief about whether the outcomes of our actions are contingent on what we do (internal control orientation) or on events outside our personal control (external control orientation)." (Zimbardo, 1985, p. 275)Our attribution style determines which forces we hold responsible for our successes and failures. Both locus of control and attribution styles have great influence on our motivation, expectations, self-esteem, risk-taking behavior, and even on the actual outcome of our actions. What is your locus of control? And what forces are responsible for your successes and failures? Find out with the Locus of Control and Attribution Style Test. Examine the following statements and indicate how often you feel that way, to what degree you endorse the statement or how much it applies to you. After finishing the test, you will receive a detailed, personalized interpretation of your score that includes diagrams, information on the test topic and tips.

Buy from http://www.queendom.com/tests/access_page/index.htm?idRegTest=704

More Personality Tests

- Learning Style Inventory
  - Kolb, Hay Group



26 June 2007 UM-07 tutorial 3: Chin 12

From http://pss.uvm.edu/pss162/learning_styles.html:

The Kolb Learning Style Inventory (LSI) is a statistically reliable and valid, 12-item questionnaire and workbook, developed by David A. Kolb, Ph.D.

Experiencing: learning from specific experiences, being sensitive to feelings and people

Observation: observing before making judgments, viewing issues from different perspectives, looking for the meaning of things

Thinking: logically analyzing ideas, planning systematically, acting on an intellectual basis

Action: learning through 'hands on' activities, dealing with people and events through action

Buy from http://www.haygroup.com/leadershipandtalentondemand/enhancing/kolb.aspx

Agenda

I. Experiment Design
- A. Independent vs. dependent variables
- **B. Nuisance variables**
- C. Between-subjects vs. within-subjects designs
- D. Estimating sensitivity
- E. Factorial designs
- F. Caveats

II. Running Experiments
- A. Participants
- B. Controlling the environment
- C. Recording data

III. Experiment Analysis
- A. Means and variance
- B. Statistical tests
- C. ANOVA
- D. Explained variance

IV. Summary

26 June 2007 UM-07 tutorial 3: Chin 13

Nuisance Variables

- Make your data impossible to analyze
  - contribute unevenly to dependent variable values
- Major types of nuisance variables
  - Individual differences among participants
  - Environmental influences

26 June 2007 UM-07 tutorial 3: Chin 14

Imagine your dependent variable is which key is pressed on an electronic keyboard and your independent variable is the sound that you hear. Your participants are keyboards. Nuisance variables are individual differences in the programmed sound of the participating keyboards and environmental sounds like nearby construction noise. If the nuisance variables are too large, you might not even be able to hear the independent variable above the noise.

Individual Differences

- People differ
  - Intelligence, reading ability, perception (e.g., color blind, poor eyesight, poor hearing), spatial reasoning
  - Variability adds noise to measured variables
- Group experiments:
  - Interpersonal interactions can bias results
  - Leaders vs. followers
  - Personality clashes
  - Communication skills vary

26 June 2007 UM-07 tutorial 3: Chin 15

If you measure whether people do better with a user model in a between-subjects design, you may by chance end up with lots of people who are inherently better at the underlying task in the no-UM group than in the UM group.

In group experiments, especially among people who know each other, leaders (such as the group's boss) can often influence others strongly, sometimes just through body language.

Environmental Influences

- People are more tired
  - certain times of the day
  - certain days of the week
- Time sensitive influences
  - Construction jackhammers in afternoon only
  - Network slows at start of lab class
- Others (experimenter) bias the participants
  - Words, tone, body language

26 June 2007 UM-07 tutorial 3: Chin 16

It is a good idea to brainstorm about possible environmental influences on the dependent variables during the planning stage of your experiments. After you come up with a list, then you can think about mitigation.

Control of Nuisance Variables

- Randomization
  - "Average out" nuisance vars over ***many*** participants
- Blind: participant does not know if system has UM
  - So not influenced by which is "supposed to be better"
- Double-blind: experimenter does not know
  - So cannot inadvertently influence participant
  - Standard practice for drug trials

Caveats

- Non-random scheduling
  - Friendly, beautiful assistant runs no UM cases; rude, dirty assistant with bad body-odor runs UM cases
  - UM requiring Internet run with UM cases in the morning with high-load, no UM cases in afternoon

More Caveats

- In medical tests:
  - Placebos can lead to significant improvements (belief that UM/advanced tech. is being used)
  - So nicer computers, neater desks ⇒ bias
- In audio tests:
  - Imperceptibly louder (.1 dB) ⇒ better sounding
  - Experimenter body language biased participants, ***even when experimenters were trying NOT to***

26 June 2007 UM-07 tutorial 3: Chin 19

Typically about 35% of people are susceptible to the placebo effect where the idea that they are being treated (even though in reality they are not) leads to improvement in their condition.

In audio tests of which piece of equipment (e.g., an amplifier) sounds better, experimenters easily bias participants even when the experimenters were trying to be neutral. Medical studies have shown experimenter bias affects response variables when the experimenters became aware of the condition of specific patients due to known side-effects (or lack thereof) in the patients.

Experiment Rules

- Randomly assign enough participants to groups
- Randomly assign time slots to participants
- No distractions in test area (windows, noise)
- Experimenters should be blind
- Brainstorm about possible nuisance variables

26 June 2007 UM-07 tutorial 3: Chin 20

Random assignment is essential because it allows nuisance variables to "average out."

Agenda

I. Experiment Design
- A. Independent vs. dependent variables
- B. Nuisance variables
- **C. Between-subjects vs. within-subjects designs**
- D. Estimating sensitivity
- E. Factorial designs
- F. Caveats

II. Running Experiments
- A. Participants
- B. Controlling the environment
- C. Recording data

III. Experiment Analysis
- A. Means and variance
- B. Statistical tests
- C. ANOVA
- D. Explained variance

IV. Summary

Between Subjects Designs

- Different participants in experimental conditions
- Randomly assigned participants
- No learning effect
- More participants needed
- Individual differences can swamp measurements

Within Subjects Designs

- Participants exposed to several conditions
- Transfer of learning effects
  - Controlled by varying condition order
- Controls for variation among participants
- Fewer participants needed
- Effective for tasks that involve learning or changes over time

Agenda

Estimating Sensitivity

- Sensitivity, a.k.a. Power:
  - how easily an experiment can detect differences
  - officially: probability of rejecting a false null hypothesis
  - Less sensitive ⇒ more participants (sample size)
  - Less sensitive ⇒ lower significance
  - Smaller treatment effects ⇒ less sensitive
- Power (sensitivity) ∝ repeatability

Power Measure

Fraction of experiments for the given design, sample size and treatment effect would produce the given significance

- Power 0.5 ⇒ 1/2 experiments give non-significant results
- *Journal of Abnormal and Social Psychology* averages 0.5
- Should use power ≥ 0.8

  (80% of repeat experiments give significant results)

Why Power ≥ 0.8?

- High likelihood to successfully repeat experiment
- **If there is an effect, better chance of finding it**

Power Calculation

- Use pilot study to estimate effect size
- Best to use programs to calculate power:
  - G .G. Gatti & M. Harwell (1998), "Advantages of Computer Programs Over Power Charts for the Estimation of Power" In *Journal of Statistics Education* **6**(3).
  - UCSF's list of Power and Sample Size Programs
  - Statpages.org's list

The Gatti & Harwell paper is available online at http://www.amstat.org/publications/jse/v6n3/gatti.html

http://www.biostat.ucsf.edu/sampsize.html has a list of power and sample size calculating programs

http://statpages.org/#Power lists interactive websites for calculating power

Effect Size ω^2

Fraction of variance due to experimental treatment (UM)

- Aka treatment magnitude (η^2)
- $\omega^2 = \sigma_A{}^2 / (\sigma_A{}^2 + \sigma_{S/A}{}^2)$, where
  - $\sigma_A{}^2$ is the variance due to user modeling
  - $\sigma_{S/A}{}^2$ is the random variance among participants
- Typical ω^2 for social science effects:
  - .01 small, .06 medium, ≥ .15 large

Power Tradeoffs

- For better power: more participants or lower significance

	Effect Size (ω^2)					
	.01 (small)		.06 (medium)		.15 (large)	
	Significance Level		Significance Level		Significance Level	
Power	0.05	0.01	0.05	0.01	0.05	0.01
0.7	219	323	36	53	14	20
0.8	271	384	44	62	17	24
0.9	354	478	57	77	22	29

Agenda

I. Experiment Design
- A. Independent vs. dependent variables
- B. Nuisance variables
- C. Between-subjects vs. within-subjects designs
- D. Estimating sensitivity
- **E. Factorial designs**
- F. Caveats

II. Running Experiments
- A. Participants
- B. Controlling the environment
- C. Recording data

III. Experiment Analysis
- A. Means and variance
- B. Statistical tests
- C. ANOVA
- D. Explained variance

IV. Summary

26 June 2007 UM-07 tutorial 3: Chin 31

Factorial Designs

- Treatments combine levels of 2 or more factors
  - E.g., different interfaces, different UM parameters, different tasks, amount of UM feedback, etc.

26 June 2007 UM-07 tutorial 3: Chin 32

If you have more than one dependent variable, rather than run a separate experiment for each variable, it may be easier to combine them in a single experiment. Factorial designs allow you to do this more economically.

Why Factorial Designs?

- Advantages
  - Simultaneously study effects of all factors
  - Gives information about interaction among factors
- Disadvantages
  - Number of combinations large:
    2^n conditions for n factors of 2 levels each
  - Conducting experiments very detailed

Randomized Block Designs

- Homogeneous groups are called blocks
- Treatments are assigned randomly to blocks
- Reduces variability
- Common factorial designs:
  - Nested block design
  - Latin square design

Nested Block Design

- A block is broken up into sub-blocks
  - Based on a 2nd treatment or covariate variable
- Sub-blocks do not have every case of the 2nd var
  - So fewer participants are needed
    versus a fully cross-randomized block design
- More participants needed with more nesting levels
  - Exponentially more

Latin Square Design

- Not every block has every treatment
  - E.g., males get no UM and UM A, females get no UM and UM B
- Useful to vary treatment order evenly within-subjects

	OS Type		
Age	UNIX	Mac	Windows
elementary	B	A	C
high-school	A	C	B
college	C	B	A

Agenda

I. Experiment Design
- A. Independent vs. dependent variables
- B. Nuisance variables
- C. Between-subjects vs. within-subjects designs
- D. Estimating sensitivity
- E. Factorial designs
- **F. Caveats**

II. Running Experiments
- A. Participants
- B. Controlling the environment
- C. Recording data

III. Experiment Analysis
- A. Means and variance
- B. Statistical tests
- C. ANOVA
- D. Explained variance

IV. Summary

Caveats

- Failure to include a control group when needed
  - Missing no UM control group
- Experimental procedure itself generates a variable
  - Thinking aloud modifies problem solving strategy

More Caveats

- Contamination of data
  - Incorrect recording/transcription
- Unwarranted assumptions about scales
  - E.g., eye blink rates are not linearly related
- Confounding nuisance vars with relevant vars
  - LAN busy at start of hour during UM treatment

More Caveats 2

- Failure to take into account transfer of training
  - Participants who have used a similar system do better
- Insufficient observations for needed precision
- Tendency to favor one outcome over another

More Caveats 3

- Observer or experimenter bias
- Not recognizing the rarity of an event
  - Gambling wins ⇒ expectations of winning > actual odds
- Experimental procedure affects observed conditions
  - Knowledge of video camera affects behavior

Internal Validity

- Did the independent variables make a difference?
- Can you infer a cause and effect relationship?
- Did you control:
  - Extraneous variables?
  - Selection procedures?
  - Measurement procedures?
- Results hard to interpret without internal validity

Threats to Internal Validity

- History
  - Some other event affects the dependent variable
  - Time between pretest and posttest
  - The longer the time, the great the chance of history
- Maturation
  - Biological or psychological processes over time
  - Independent of external events

More Threats to Internal Validity

- Testing
  - Tendency to score higher on similar subsequent tests
- Instrumentation
  - Any change in observation (machines or judges)
- Statistical regression
  - Extreme score means tends to drift back to the middle

Consider trying to study people who get perfect SAT scores. The next time these people take an SAT test, they probably won't get a perfect score. Likewise if you want to study people who got everything wrong on a particular test, the next time these same people take the same or a similar test, they probably won't get all wrong again. This tendency of people with extreme scores to tend to drift back toward the middle is called statistical regression.

Other Internal Validity Threats

- Mortality
  - Loss of subjects between a pretest and a posttest
  - Drop-outs may differ from those who remain
  - Mean scores between the tests could differ
- Selection
  - Participants seek/do not seek exposure to the treatment
  - Likely differ in motivational levels, so don't compare

External Validity

- Can results be generalized?
- How representative are the results to:
  - Other populations?
  - Other variables?
  - Other situations?

Threats to External Validity

- Population
  - Experimentally accessible pop. differs from target pop.
  - Treatment effects interact w. participant characteristics
- Ecological
  - Incorrectly describing independent variable(s)
  - Incorrectly describing or measuring dependent variable(s)

26 June 2007 UM-07 tutorial 3: Chin 47

If you do not describe your independent variables correctly, then it becomes impossible for others to reproduce your experiment or sometimes even to understand your experiment.

More Ecological Validity Threats

- Multiple-treatment interference
- Interaction of history and treatment effects
- Interaction of time of measurement and treatment
- Pretest and posttest sensitization
- Hawthorne effect (expectation ⇒ improvement)
- Novelty and disruption effect
- Experimenter influence (Rosenthal/Pygmalion, Golem effects)

26 June 2007 UM-07 tutorial 3: Chin 48

Experiments at the Hawthorn Works factory found that any change in lighting led to a temporary improvement in productivity because workers expected the change to help. Robert Rosenthal and Lenore Jacobson studied the Pygmalion effect: random students that teachers were led to expect better performance from actually did do better. The Golem effect is for negative self-fulfilling prophecies.

Agenda

I. Experiment Design
- A. Independent vs. dependent variables
- B. Nuisance variables
- C. Between-subjects vs. within-subjects designs
- D. Estimating sensitivity
- E. Factorial designs
- F. Caveats

II. Running Experiments
- **A. Participants**
- B. Controlling the environment
- C. Recording data

III. Experiment Analysis
- A. Means and variance
- B. Statistical tests
- C. ANOVA
- D. Explained variance

IV. Summary

26 June 2007 UM-07 tutorial 3: Chin 49

Participants

- Participants must represent target population
- Participant sources
  - University laboratory schools
  - Introductory psychology participant pools
  - Public schools
  - Newspaper advertisements
  - Corporations
  - Internet sites

26 June 2007 UM-07 tutorial 3: Chin 50

A common problem with university-based experiments is that they typically use college students as participants and college students are **not** representative of the general population.

Participant Incentives

- Payment
- Gifts
- Class credit
- Desire to help state-of-the-art research

Incentives are often helpful to motivate participants. Unmotivated participants may drop out part way through the experiment (wasting your time and effort since you probably can't use their data) or work haphazardly or even semi-maliciously (e.g., just selecting random choices in a multiple-choice questionnaire).

Consent Agreement

- Participants should sign a consent form:
  - I have freely volunteered to participate
  - I have been informed about the tasks and the procedures
  - I have had a chance to ask questions about my concerns
  - I know that at any time I may discontinue participation in this experiment without prejudice
  - My signature below may be taken as an affirmation of all of the above, prior to participation

USA Federal Mandates

- Local institutional review board (IRB)
  - Required for all US institutions receiving federal funds
  - Approves all proposed human-subject studies ***beforehand***
  - Poor IRB oversight has led to Federal funding cutoffs

Agenda

Controlling the Environment

- Needed to control nuisance variables
- Factors include:
  - Room selection & preparation
  - Uniform instructions
  - Experimenter behavior

Room Selection & Preparation

- Select room to minimize distractions:
  - Audio: noise
  - Visual: no windows, posters, etc.
  - Isolate participants as much as possible
- Prepare computer area ergonomically
  - Anticipate different size participants
  - If network is used, avoid high load times

Uniform Instructions

- Written/taped instructions are more consistent
- Check instructions for clarity
- Debug instructions with pilot study
- Computer playback of instructions is very helpful
- Each experimenter runs equal #s of each treatment

Experimenter Behavior

- Strive for uniformity
  - ***Plan*** to minimize interactions with participants
  - All experimenters should be consistent in approach
  - Experimenters must be able to answer questions
- Interaction during experiment is bad
  - Strive to answer all questions beforehand
  - Pilot studies help catch unanticipated questions
  - Be prepared to discard participant data if necessary

Agenda

I. Experiment Design
- A. Independent vs. dependent variables
- B. Nuisance variables
- C. Between-subjects vs. within-subjects designs
- D. Estimating sensitivity
- E. Factorial designs
- F. Caveats

II. Running Experiments
- A. Participants
- B. Controlling the environment
- **C. Recording data**

III. Experiment Analysis
- A. Means and variance
- B. Statistical tests
- C. ANOVA
- D. Explained variance

IV. Summary

Recording Data

- Qualitative data
- Quantitative data

Qualitative Data

- Ethnographic field studies
  - Content analysis
- Case Studies
- Self reports
- Interviews

Qualitative Sources

- R.K. Yin (1988) *Case Study Research: Design and Methods*
- M.B. Miles & A.H. Huberman (1994) *Qualitative Data Analysis: A Sourcebook of New Methods*
- M. Meyers (ed.) *Qualitative Research in Information Systems*
- C. Marshall & G. Rossman (1989) *Designing Qualitative Research*
- D. Silverman (1993) *Interpreting Qualitative Data*
- R.P. Weber (1990) *Basic Content Analysis, 2nd edition*
- *Qualitative Research in Information Systems* journal and web links, www.qual.auckland.ac.nz

Sequential Data

- Think aloud tasks
- Video or audio taped records
- Recorded computer interactions
  - Record & replay GUI events
    (keystrokes, mouse movements, buttons, menus, etc.)
- Retroactive interview with playback records
- Eye movement monitors

Agenda

Experiment Analysis

- The simplest experiment has:
  - One independent variable w. 2 values (with/without UM)
  - Same # of participants in each group (with/without UM)
  - One dependent variable (e.g., task quality)
  - Analyze more dependent variables as if new experiment

Sample Dependent Variables

- Subjective evaluation of the system
  - Likert scale of 1 to 7 reduces biases of 1-5/1-10 scales
- Task speed
- Task quality (e.g., accuracy)
- Pupil dilation
  - Shown to be correlated with cognitive load

Mean and Variance

- *Mean* = average of dependent variable values
- *Variance* = average difference of values from mean
- There are two types of variance:
  - Between groups (due to the UM)
  - Within groups (due to "random" fluctuations)

Null Hypothesis

- Conjecture that the independent variable (e.g. UM/ no UM) makes no difference in the dependent variable(s) values
- Rejecting the null hypothesis depends on computing the likelihood that the difference in the means of the groups is not due to natural variations

Why Analysis?

- If the means of UM differs from no UM
  - So UM has a positive or negative effect
- Might this be caused by random fluctuations?
  - E.g., by chance more optimists were randomly assigned to the UM group, leading to higher subjective evaluations for the UM case

26 June 2007 UM-07 tutorial 3: Chin 69

Analysis allows one to determine the likelihood that a difference in means between 2 treatment groups is not due to random fluctuations. Without analysis, one's results are always questionable as due to random variations. Analysis allows one to quantify this probability. People generally accept that if the probability of the difference in means being due to random processes is less than .05, then the difference can be considered real.

Agenda

I. Experiment Design
- A. Independent vs. dependent variables
- B. Nuisance variables
- C. Between-subjects vs. within-subjects designs
- D. Estimating sensitivity
- E. Factorial designs
- F. Caveats

II. Running Experiments
- A. Participants
- B. Controlling the environment
- C. Recording data

III. Experiment Analysis
- A. Means and variance
- B. **Statistical tests**
- C. ANOVA
- D. Explained variance

IV. Summary

26 June 2007 UM-07 tutorial 3: Chin 70

Statistical Tests

- Non-parametric tests
  - Fewer assumptions about data
  - But less powerful
- Parametric tests
  - Preferred for data with normal (Gaussian) distribution
- Statpages.org's Choose the right test! list

26 June 2007 UM-07 tutorial 3: Chin 71

It is important to choose the right statistical test because the wrong test will give weaker or even incorrect results. Be sure to check not only the type of data for the test, but also check that the test's assumptions about its data is true of your own data.

http://statpages.org/#WhichAnalysis has a list of interactive websites for choosing the right statistical test.

Non-parametric Tests

- Assumptions:
  - Independent observations
  - Distribution free
  - Suitable for ordinal / ranked data

26 June 2007 UM-07 tutorial 3: Chin 72

User responses on Likert scale (or any other scale) subjective evaluations are ranked data because the difference between a 6 and a 7 is probably not the same as the difference between a 5 and a 6 or a 1 and a 2. The only thing you can safely say is that a 7 is higher than a 6. How much better cannot and should not be assumed. Therefore Likert scale responses should be analyzed with non-parametric tests. Parametric tests like ANOVA require that the data is actually linearly scalar. Unfortunately ANOVA is often wrongly used to analyze Likert scale responses.

Common Non-Parametric Tests

- Chi-square
  - Compares how each measure differs from expected
  - Goodness of fit and independence of random variables
- Median or Sign Test
  - Compares medians of two independent values
- Mann-Whitney U Test
  - Tests if 2 samples come from the same distribution
- Kruskal-Wallis 1-way ANOVA of Ranks
- Friedman 2-way ANOVA of Ranks

Parametric Tests of Significance

- Assumptions:
  - Independent observations
  - Observations from normal distribution
  - Homogeneity of variance in populations
  - Variables measured on equal unit interval scale
  - Null hypothesis tests for equal means or variances between independent samples

Common One/Two Sample Tests

- Difference from the mean (Z-test)
- Difference between 2 sample means (T-test)
- Variability differences in 2 samples (F-test)
- Analysis of Variance (ANOVA)
- Multivariate Analysis of Variance (MANOVA)
- Analysis of Covariance (ANCOVA)

26 June 2007 UM-07 tutorial 3: Chin 75

The Z-test compares the mean of a sample against the mean of a whole population to see if the difference is meaningful or just due to random selection. The T-test compares the difference in means between two samples (e.g., UM or no UM). The F-test compares the variances (standard deviations) of two samples.

Directional vs. Non-directional

- Directional (one-tail)
  - Hypothesis predicts direction of estimates
  - Easier to achieve significance
- Non-directional (two-tail)
  - No basis for deciding direction of the difference
  - GraphPad.com has a good faq on this

26 June 2007 UM-07 tutorial 3: Chin 76

http://www.graphpad.com/faq/viewfaq.cfm?faq=1318 gives a good description of how to determine if your test is one-tail or two-tail.

Agenda

I. Experiment Design
- A. Independent vs. dependent variables
- B. Nuisance variables
- C. Between-subjects vs. within-subjects designs
- D. Estimating sensitivity
- E. Factorial designs
- F. Caveats

II. Running Experiments
- A. Participants
- B. Controlling the environment
- C. Recording data

III. Experiment Analysis
- A. Means and variance
- B. Statistical tests
- **C. ANOVA**
- D. Explained variance

IV. Summary

ANOVA Assumptions

- Linear model
- Independence of scores
- Normal distribution
- Heterogeneity of variance

Linear Model

$Y_{ij} = \mu_T + \alpha_i + \varepsilon_{ij}$, where

- Y_{ij} is any observation of the dependent variable
- μ_T is the mean of all Y_{ij}
- α_i is the treatment (UM) effect (between group)
- ε_{ij} is the experimental error (within group, due to individual or environmental differences that hopefully have been randomly distributed among the Y_{ij})

Independence of Scores

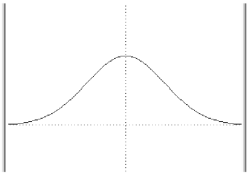
- The scores (Y_{ij}) are independent both within and between treatment groups (UM and no UM), i.e., each observation is not related in any way to any other observation
  - participants are randomly assigned to UM/no UM
  - participants are tested individually
  - participants do not discuss system with others (e.g., students in a class will talk, creating bias)

Normal Distribution

- Participant population is normally distributed
  - Verify by plotting Y_{ij} scores
  - Look for bell-shaped normal curve
    (x-axis = scores, y-axis = count of each score)
  - Symmetrical shapes with ≥ 12 participants are fine
  - Asymmetrical shapes require higher significance levels

Normal Curve Example

- One of a family of Normal curves



Homogeneity of Variance

- Suppose UM helps some but confuses others
  - If these occur equally frequently,

    then the mean is unchanged for UM vs. no UM
  - But the variance of Y_{ij} would be much higher for UM
- **Heterogeneity of variance invalidates analysis**

Variants of ANOVA

- Multivariate Analysis of Variance (MANOVA)
  - For multiple dependent variables and their interactions
- Kruskal-Wallis (one-way ANOVA) by ranks
  - Uses rank order rather than actual values
  - E.g., web search results by list order vs. similarity scores

Analysis of Covariance

- ANCOVA combines
  - Analysis of variance (ANOVA)
  - Regression analysis
- Allows reduction of error term ε_{ij}
  - Improves effect size relative to error (σ_A^2 vs. $\sigma_{S/A}^2$)
  - Improves power
- Corrects Y_{ij} using covariant variable(s)

ANCOVA Example

- UM system that hides less relevant hyperlinks
  - Independent variable: UM or no UM
  - Dependent variable: speed to find information
  - Covariant variable: participant reading speed
- ANCOVA corrects search times for reading speed, eliminating the variance due to reading speeds

ANCOVA Assumptions

- All ANOVA assumptions
- *Linear* regression
  - Dependent scores vary linearly with covariant variable
  - Equal population regression slopes for all groups
  - Unequal ⇒ ANCOVA cannot be used
    e.g., for whatever reason, the UM group did not improve search times as much for faster readers as the no UM group

ANCOVA Rules

- Gather covariate(s) ***before*** the experiment
  - Avoids UM/no UM affecting covariate
  - After is possible for "permanent" characteristics like IQ
- Test linearity and equal slope assumptions
  - By computer program ***and*** visually
- Different formulas for effect size and power
  - Use correct setup of computer programs

Agenda

I. Experiment Design
- A. Independent vs. dependent variables
- B. Nuisance variables
- C. Between-subjects vs. within-subjects designs
- D. Estimating sensitivity
- E. Factorial designs
- F. Caveats

II. Running Experiments
- A. Participants
- B. Controlling the environment
- C. Recording data

III. Experiment Analysis
- A. Means and variance
- B. Statistical tests
- C. ANOVA
- **D. Explained variance**

IV. Summary

Explained Variance

- Two possible analysis results: significant or not
- Significant results
  - Likelihood(difference in means is due to random fluctuations) < selected significance level (typically .05)
- Calculate and ***report***:
  - post-hoc probability
  - effect size
  - power

Non-Significant Results

- If calc. power is low, maybe need more participants
  - Use effect size to determine # of participants needed
  - If # too large, consider relaxing significance level to 0.1
  - Very difficult to prove effect does not exist
    (requires **very** many participants)

Interpreting Significant Results

- Statistically significant ≠ important differences
- Treatment effect may be increased variability
- Which 0.05 significance test is more impressive:
  A with 5 participants or B with 20?
  - A, because if A were increased to 20 participants,
    it would likely have better significance than B

Agenda

I. Experiment Design
- A. Independent vs. dependent variables
- B. Nuisance variables
- C. Between-subjects vs. within-subjects designs
- D. Estimating sensitivity
- E. Factorial designs
- F. Caveats

II. Running Experiments
- A. Participants
- B. Controlling the environment
- C. Recording data

III. Experiment Analysis
- A. Means and variance
- B. Statistical tests
- C. ANOVA
- D. Explained variance

IV. Summary

Summary

- Experiments require careful planning
  - Pilot studies prevent poorly designed main studies
- Experiments take a long time
  - Typically **months**
- Experiments are the only way

Where to Get More Information

- Books
- Web Sites
- People from your Psych. or Statistics depts. or human-factors group
- Software

Books

- G. Keppel (1991) *Design and Analysis: A Researcher's Handbook* (3rd ed.), Englewood Cliffs, NJ: Prentice-Hall.
- J. Stevens (1992) *Applied Multivariate Statistics for the Social Sciences* (2nd ed.), Hillsdale, NJ: Lawrence Erlbaum.
- J. Neter, W. Wasserman, M.H. Kutner (1985) *Applied Linear Statistical Models* (2nd ed.) Homewood, IL: Richard D. Irvin.
- D. Campbell, J. Stanley (1963) Experimental and Quasi-Experimental Designs for Research in *Handbook of Research on Teaching* N. L. Gage editor, Rand McNally & Co.
- S. Huck, W. Cormier, W. Bounds(1974) *Reading Statistics and Research*, New York, Harper & Row.

Web Sites

- Interactive Statistical Calculation Pages
- The World Wide Web Virtual Library: Statistics
- Electronic Textbook StatSoft
- OATIES (Online Analysis Tools in Excel Spreadsheets)
- Ball Aptitude Battery

26 June 2007 UM-07 tutorial 3: Chin 97

http://statpages.org/ has lots of good information besides listings of interactive calculation pages.

http://www.stat.ufl.edu/vlib/statistics.html lists statistics related pages around the world.

http://www.statsoft.com/textbook/stathome.html is a free electronic statistics textbook.

http://www.coventry.ac.uk/ec/~nhunt/oatbran/ is the Online Analysis Tools in Excel Spreadsheets.

http://www.careervision.org/About/BallAptitudeBattery.htm has a series of aptitude tests that may be useful for measuring covariates.

After Your Experiment

Publish in:

- *User Modeling and User-Adapted Interaction*
- Next UMAP Conference
- SIGCHI (ACM) Bulletin or Conference

26 June 2007 UM-07 tutorial 3: Chin 98

http://www.umuai.org/ is the website for *UMUAI*, the premiere journal in the user modeling field.

http://www.sigchi.org/ is the website for SIGCHI, the ACM Special Interest Group in Computer-Human Interaction.

Acknowledgements

- Sponsored by:
  - UMAP 2012, the 20th Conference on User Modeling, Adaptation, and Personalization, Montreal, Canada
  - User Modeling, Inc.
  - University of Hawaii

Your Copy

www2.hawaii.edu/~chin/UMAP2012/tutorial.pptx

www2.hawaii.edu/~chin/UMAP2012/tutorial-notes.pdf