Experimentation

Michael Ernst

UPM workshop
Why do we run experiments?
An experiment answers questions

To design an experiment, first ask:

● What are you trying to establish?

What is the biggest concern or question that someone might have?

Is the experiment for you (you don’t know the answer) or for other people (you want to convince them)?
Know your question & state it clearly

A “fishing expedition” can be great for exploratory work
It is not useful as an experiment

Make your experiment:
- specific enough to be feasible and testable
- broad enough to generalize
Types of experiments

- Controlled experiments (quantitative evidence)
- Case studies (qualitative evidence)
  - Done by the researchers themselves
  - Done by other people

When is a case study better than a controlled experiment?

- Example: learning from the first users of a tool
- In HCI, 5 users is considered adequate
When not to do an experiment

Other types of evidence:

- surveys
- proofs
- natural experiment (observational experiment)
An experiment is a comparison

Observation O1 : process P in environment E1
Observation O2 : process P in environment E2
If O1 = O2, the environmental differences are irrelevant to the process
If O1 ≠ O2, the difference is caused by the environmental differences

It is not enough to report, “my technique does well”
You must compare to the state of the art
Minimize differences

E1 and E2 should be as similar as possible.
If many differences, which one caused $O_1 \neq O_2$?

E1 and E2 should be realistic of actual practice:
- real traces/logs, real development practices,
  ...

(Example: if you have 3 techniques, compare with "hold one out" rather than "just one".)
Aside: comparing enhancements

Suppose you implement 3 enhancements, which improves results

full = baseline+e1+e2+e3 > baseline

Which enhancement is best?

Wrong approach
compare:
baseline+e1 to baseline,
baseline+e2 to baseline,
baseline+e3 to baseline

Right approach
compare:
full - e1 to full,
full - e2 to full,
full - e3 to full

baseline+e2+e3
Treatment and effect

Treatment = input = independent variables
  ● We called this the “environment” earlier
  ● Minimize the number!
Effect = output = dependent variables

Subjects
Subjects

Experimental subjects:

- in social sciences, people
- in computer science, can be programs, etc.
  - it’s better to experiment on people when possible
Ethical considerations
(when experimenting on people)

- informed consent
- potential harm

Experiment is reviewed by the HSC (human subjects committee) or IRB (institutional review board)

Long turnaround: submit your application early!
Problem: People differ a lot

Medicine has it easy:
height differs by about a factor of 2
Programming skill differs by orders of magnitude
Individual ability/knowledge/motivation is an independent variable
Non-human subjects also have variation
College students vs. practitioners

You can learn a lot from students

- available
- homogeneous
- uncharacteristic? No evidence of this...
Example experiment

Programmers fix bugs, using 2 tools
s1: t1
s2: t2

If subject2 was faster, then tool2 is better

How can we fix this experiment?
Improvement 1: replication

Programmers fix bugs, using 2 tools
s1: t1  s3: t1  s5: t1  ...
s2: t2  s4: t2  s6: t2  ...

If on average subjects using tool2 are faster, then tool2 is better
(How many programmers do we need? Lots.)
Improvement 2: paired design

Programmers fix bugs, using 2 tools

s1:  t1xp1   t2xp2
s2:  t1xp1   t2xp2
s3:  t1xp1   t2xp2
s4:  t1xp1   t2xp2
...

If tool2 trials are faster on average, then tool2 is better

Needs fewer subjects: ~40 as a rule of thumb
Improvement 2b: paired design

Programmers fix bugs, using 2 tools

s1:  t1xp1  t2xp2
s2:  t1xp2  t2xp1
s3:  t1xp1  t2xp2
s4:  t1xp2  t2xp1

...  

If tool2 trials are faster on average, then tool2 is better

Avoids confounding effect, or tool-program interaction
Improvement 2c: paired design
(blocked/counterbalanced)

Programmers fix bugs, using 2 tools

s1:   t1xp1  t2xp2
s2:   t1xp2  t2xp1
s3:   t2xp1  t1xp2
s4:   t2xp2  t1xp1

... 

If tool2 trials are faster on average, then tool2 is better

Avoids another confound: learning/fatigue effects
Many other conflating factors exist

Example: self-selection
Combatting individual variation

- **Replication**
  - In a population, individual variation averages away (central limit theorem)

- **Randomization**
  - Avoid conflating effects

- **Statistics**
  - Indicates when a difference is large enough to matter
“There are three types of lies: lies, damned lies, and statistics.”

- Mark Twain
Choosing a statistical test

It’s best to consult an expert or take a course in experimental design or statistical methods (≠ a course in statistics)

When in doubt, use ANOVA

● “ANalysis Of VAriance”
False positive errors

False positive (or false alarm or Type I error): no real effect, but report an effect (through good/bad luck or coincidence)

– If no real effect, a false positive occurs about 1 time in 20
  ○ 5% is a convention; there is nothing magic about it

– If there is a real effect, a false positive occurs less often
A false positive

http://xkcd.com/882/
Lesson: don’t test too many factors
False negative errors

False negative (or miss or Type II error): real effect, but report no effect (through good/bad luck or coincidence)
– The smaller the effect, the more likely a false negative is
– How many die rolls to detect a die that is only slightly loaded?

The larger the sample, the less the likelihood of a false positive or negative
Correlation ≠ causation

Ice cream sales and murder rates are correlated

Lesson: you should always have an explanation for an effect

http://xkcd.com/552/
Statistical significance ≠ practical importance

After 1,000,000 rolls of a die, we find it is biased by 1 part in 10,000
Measurements (metrics to gather)

Decide methodology and measurements before you see the data.

A common error:
1. Observe what you see in the real world
2. Decide on a metric (bigger value = better)

For any observation, there is something unique about it. Example: dice roll
Don’t trust your intuition

• People have very bad statistical intuition
• It’s much better to follow the methodology and do the experiments
Digits of precision

2 digits of precision is usually enough:
1.2  34  2500  3.7x10^6

A difference of less than 1% doesn’t matter to readers

- The extra digits just distract

This is *not* consistent:  3.1  34.7  2594.6
Don’t report irrelevant measures

If a measurement is not relevant to your experimental questions, don’t report it.

- Don’t snow the reader with extra raw data
Biases

Your research is destined to succeed

- Hawthorne effect (observer effect)
- Friendly users, underestimate effort
- Sloppiness
- Fraud
  - (Compare to sloppiness)

Be as objective as possible
Threats to validity
Discuss them in the paper

- **Internal validity**
  whether an experimental treatment/condition makes a difference or not, and whether there is sufficient evidence to support the claim.

- **External validity**
  generalizability of the treatment/condition outcomes.
Another classification of threats to validity

- construct (correct measurements)
- internal (alternative explanations)
- external (generalize beyond subjects)
- reliability (reproduce)
Pilot studies (= prototypes)

Always do a pilot study
An experiment is costly
- your time, limited pool of subjects
You learn most from the first users
- you are certain to make users
If you change anything, do another pilot study
Reproducibility

Your experiments should be reproducible

- treat them like other software
  - version control, tests, ...

If there are subjective decisions, have them cross-checked

Publish your data!