Software Usability

Course notes for CSI 5122 - University of Ottawa

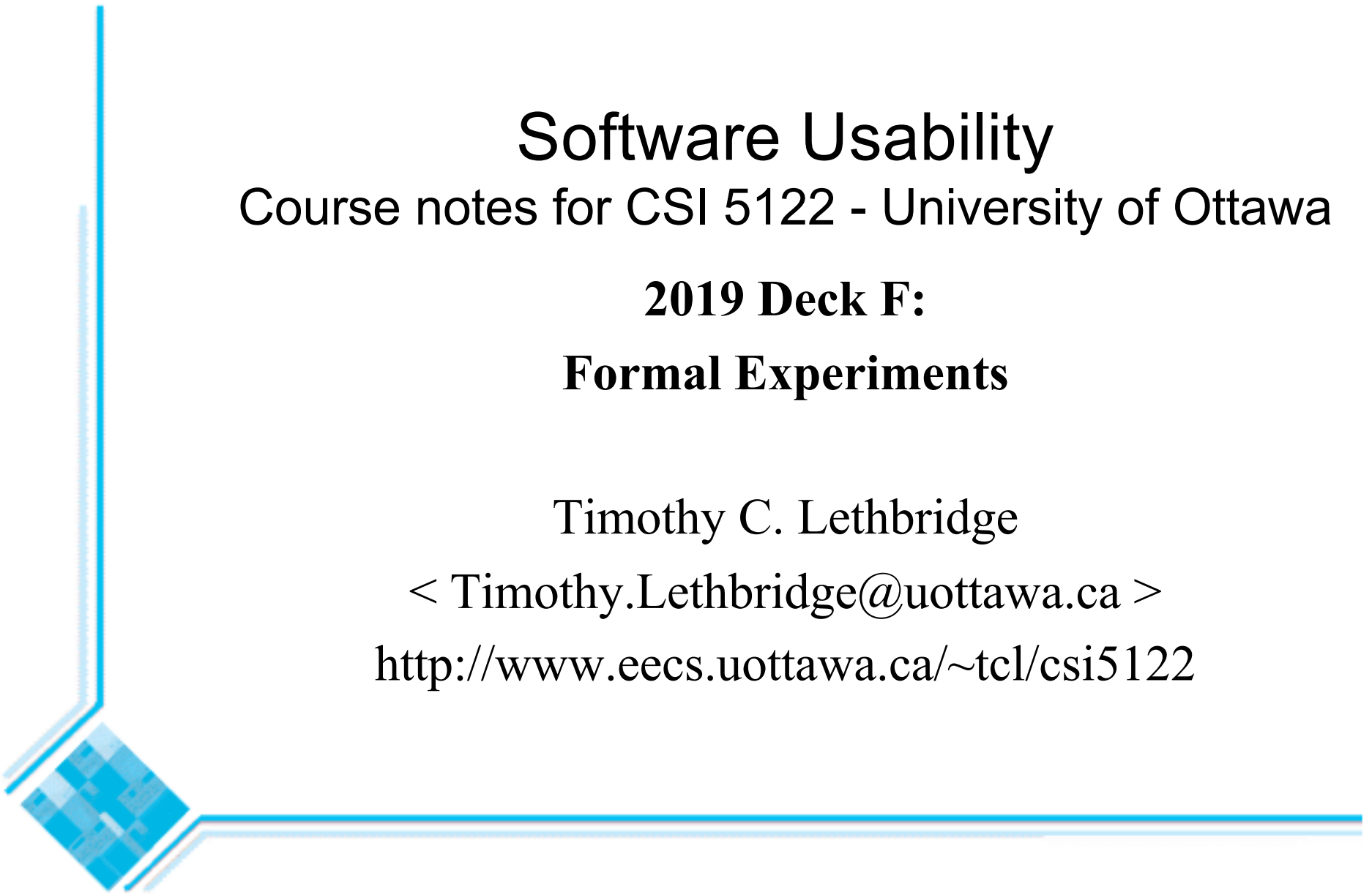
2019 Deck F:

Formal Experiments

Timothy C. Lethbridge

< Timothy.Lethbridge@uottawa.ca >

http://www.eecs.uottawa.ca/~tcl/csi5122



EXPERIMENTS: DETERMINING WHICH ALTERNATIVE IS BETTER

Steps in Experimentation

1. **Develop a hypothesis (at least one)**
2. **Understand who your participants (users) will be**
3. **Pick one or more independent variables to vary and dependent variables to measure, arising from your hypothesis**
4. **Design experiments to test hypotheses, focusing on:**
   - —precise tasks: Get users to use slightly different versions
   - —which users try which treatment, and
   - —in which order
5. **Choose real users and conduct experiment sessions**
6. **Statistically analyze results to draw conclusions**
7. **Decide what action to take based on conclusions**

Experimentation step 1:
Develop a hypothesis

A prediction of the outcome

- Some change in an *independent* variable causes some change in a *dependent* variable
- Or value 1 of independent variable (e.g, our new UI) is better than value 2 of the variable (e.g. old UI)

Aim of experiment is to show this is correct

Examples of things we might want to test:

- Can users successfully recognize icons?
- Which of two {icons, menu structures, screen designs etc.} is best?
  - —e.g. easiest to recognize, fastest to use, fewest mistakes

Case study live in class

Umpleonline: Creating a class diagram

- Testing user speed, preferences and error rate

What should the hypothesis be? More than one?

Example hypotheses for icon *recognition (1,2)*

Hypothesis Possibility 1: Users can successfully guess what icon I1 would do if it were to be clicked on

- **Without being given a list of actions**
- Advantage: Tests recall and/or best guesswork
- Key problem: Experimenter has to judge whether a user has successfully 'named' the correct action

Hypothesis Possibility 2: Users can successfully guess what icon I1 would do if it were to be clicked on

- When a **list of possible actions is presented**
- Key problem: This is now a *recognition* task, instead of recall

  —In the real world, recall is required.

Example hypotheses for icon *recognition* - (3,4)

Hypothesis Possibility 3: Users can successfully choose which of a set of icons will do action A1

- Key problem: Participants might be able to use a process of elimination

  —(if other icons are recognizable).

Hypothesis Possibility 4: Given a set of icons and a set of actions, users can successfully match icons to actions.

- Weaker because this is not something that users do in the real world

Example hypotheses to test *which icon is 'best'*

Hypothesis Possibility 5: Users are better at naming the action performed by Icon I1a than they are at naming the action performed by Icon I1b

- Key problem: Again, we have to subjectively quantify the names users choose

Hypothesis Possibility 6: Users are better matching the action performed by Icon I1a than they are at matching the action performed by Icon I1b

What does 'successfully' mean in Hypotheses 1-4?

- We have to define the success criterion
- Suggestion: “Novices get it right at least 80% of the time”

Don't confuse *success criterion* with *confidence level*

Confidence level is used to evaluate whether our results are statistically significant

- Usually 95%
- We want to be able to say: We are at least **95% confident** that we are correct to say "users get it right 80% of the time."
  - —i.e. Users might fail 20% of the time
  - —This conclusion might be invalid 5% of the time ($p < .05$)

Public opinion polls work the same way

- The pollster concludes that the Liberals will get 39% of the votes.
  - —This is accurate within +/- 4 percentage points, 95% of the time

Experimentation step 2:
What kind of users

Case study live in class

- Students?
- Expert modelers with the tool?
- Expert modelers with some other tool?

Experimentation step 3:
Pick variables to test

Variables can be

- Independent:
  - —Input, varied by the experimenter
  - —E.g. mode, system version, layout
  - —Each value of the variable is called a *treatment*
- Dependent:
  - —Output, measured by the experimenter
    - Speed, error rate, responses to likability questions in a survey
- Extraneous:
  - —In need of control
    - Age of users, expertise of users

Independent variables

Independent variables: Manipulated to produce different conditions

- Should not have too many
- They should not affect each other too much
- Make sure there are no hidden variables

In our icon examples:

- A The icons presented to the user

or

- B The tasks presented the user

Dependent variables

Measured value affected by independent variables

- A The task selected by the user
- B The icon selected by the user

- C Speed at doing a task
- D Number of errors made

Case study live in class

What would be the variables for the UmpleOnline experiment?

Experimentation step 4.

Design experiments to test hypotheses

Create a null hypothesis

- i.e. a change in independent variable causes no change in dependent variable

e.g.

- H0: Users cannot successfully state what an icon will do
- H0: Users cannot successfully select which icon performs a given task
- H0: There is no significant difference between screen designs, in terms of users' speed at filling in the data

Disprove null hypothesis!

Experiment design must be done carefully.

Three experiment designs:

Given two possible conditions to test

- e.g. Stop button is red vs. stop button is yellow.

Independent subject design

- Subjects are randomly distributed to the conditions

Matched subject design

- In order to control for another variable (e.g. age), a selection of all ages (etc.) is allocated to each condition

Repeated measures design

- Every subject tries every condition
- Effectively controls for all variables except increasing skill levels

Case study live in class

What should our experimental design be for the UmpleOnline experiment?

Examples of experimental designs

(remember -- always do pilot studies first)

H0: Users cannot successfully state what an icon will do

- Use 30 participants
- Tell each about the general purpose of the system (read a script)
- Have each look at 5 icons, and write down what they believe the icon will do
  - —one of the icons is the icon we are testing
  - —some of the others should be well-known and also unknown icons to server as controls

Examples of experimental designs - continued

H0: Users cannot successfully select which icon performs a given task

- Use 30 participants
- Tell each about the general purpose of the system (read a script)
- Have each look at 5 icons, and match the icons to a set of 10 tasks (5 tasks having no matching icon)

H0: There is no significant difference between screen designs, in terms of users speed at filling in the data

Repeated Measures Design

Use 7 participants (more if possible) who know about the domain

- Give each 4 lists of data items to fill in
  - —Randomly alter the order of the data items provided to subjects
- For two sets of data, each participant uses screen A
- For two sets of data, each participant uses screen B
- Screen A and B are alternated
  - —e.g. A then B, or B then A
  - —i.e. Randomly, some subjects start with A and some subjects start with B

Picking a set of participants (users)

A good mix to avoid biases

- See our earlier discussion of classes of users

For some experimental designs, we will needs lots of users

- but each will do very little!

Choose users that are representative of the users who will eventually use the system

- Who has to recognize the icons?

Find a sufficient number to get statistical significance

Remaining experimental steps

5. Conduct experiments (according to your design)

- Always have participants complete consent form first

6. Statistically analyze results to draw conclusions

- Discussed later
  - —e.g. using 't-tests'
  - —Use non-parametric statistics (based on rank) if the data is not 'normally distributed'
- Use ANOVA when there are multiple variables

7. Decide what action to take based on conclusions

Case Study: Text Selection Schemes

Early GUI research at Xerox on the Star Workstation

- Traditional experiments
- Results were used to develop Macintosh

Goal of study:

- Evaluate how to select text using the mouse

Case study - 2

Subjects

- Six groups of four
- In each group, only two are experienced in mouse usage

Variables

- Independent:
  - —Selection schemes: 6. strategically chosen patterns involving
    - Which mouse button (if any) could be double/triple/quad clicked to select character/word/sentence
    - -Which mouse button could be dragged through text
    - Which mouse button could adjust the start/end of a selection
- Dependent
  - —Selection time
  - Selection errors

Case study - 3

Hypothesis

- Some scheme is better than all others

Detailed experiment design

- Null hypothesis: No difference in schemes
- Assign a selection scheme to each group
  —matched subject
- Train the group in their scheme
- Measure task time and errors
- Each subject repeated 6 times
  —A total of 24 tests per scheme

Case study - 4

Analysis

- F-test used - scheme F found to be significantly better
  - —Point and draw through with left mouse
  - —Adjust with middle mouse

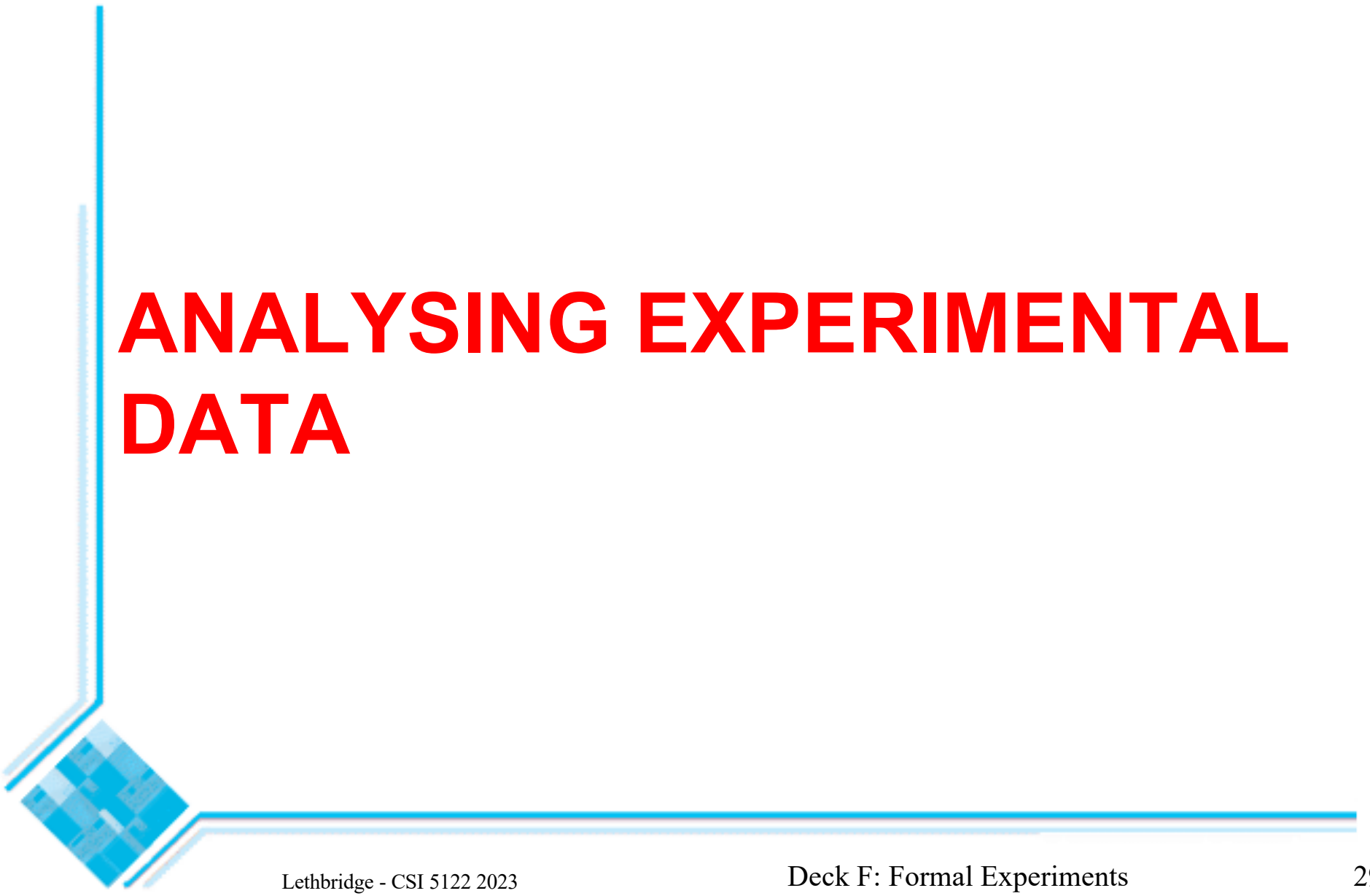
Action

- Try another combination similar to scheme F
  - —Conclusion after second experiment: Left mouse can be double-clicked

Questions to ask when reviewing experiments

Not all published experiments are done well!

- Were users adequately prepared?
- Were tasks complex enough to allow adequate evaluation?
- Did the task become boring to the users?
- Although effects ere found to be statistically significant, does that matter?
  - —Maybe not if a particular task is rarely performed
- Are there any other possible interpretations?
  - —Maybe users have learned to do better at task B because they did task A first!
- Are dependent variables consistent?
  - —e.g. users may prefer slower method
- Can results be generalized?
  - —Maybe selection results also apply to graphics, maybe not.



ANALYSING EXPERIMENTAL DATA

R is the best free tool to analyse data

There are other tools but to obtain R:

- http://cran.stat.sfu.ca/

For help

- http://www.statmethods.net

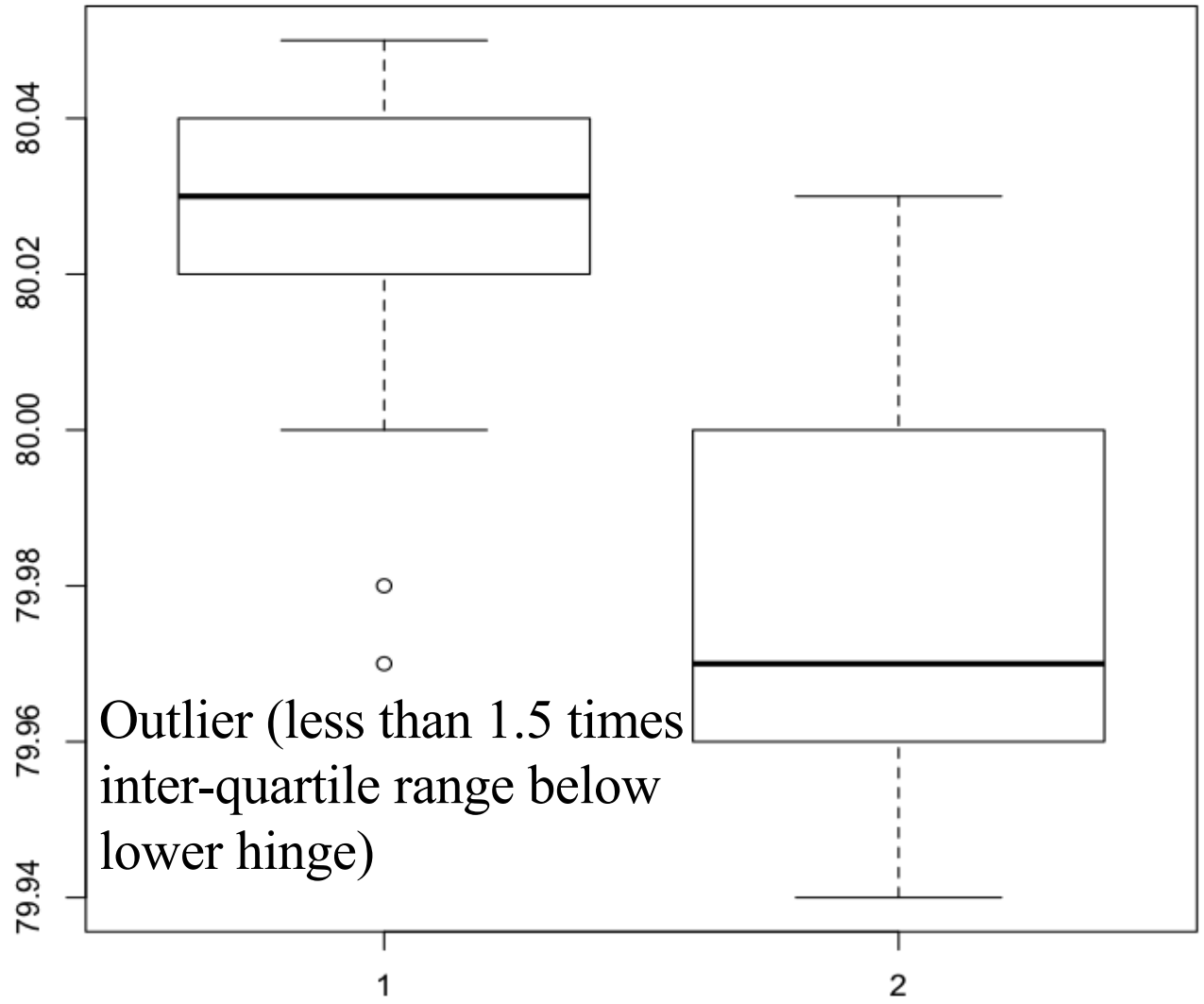
Simple way to enter and graph data in R (end with blank line)

```
A <- scan()
  79.98 80.04 80.02 80.04 80.03 80.03 80.04 79.97
  80.05 80.03 80.02 80.00 80.02

B <- scan()
    80.02 79.94 79.98 79.97 79.97 80.03 79.95 79.97

boxplot(A, B)
```

Boxplot generated by R



Outlier (less than 1.5 times inter-quartile range below lower hinge)

Maximum
(upper whisker)
Not counting outliers

3rd quartile,
75th percentile,
(upper hinge)

Median

1st quartile

Minimum
Not counting outliers

More diagrams

More sophisticated boxplot

```
boxplot(A, B, notch=TRUE, varwidth=TRUE)
```

Histogram

```
hist(A)
```

To learn more see

http://www.statmethods.net/graphs/boxplot.html

Hypothesis testing in R

t-Test: Test the equality of two means

- p-value is the probability of making a type I error
  - —Probability of rejecting a null hypothesis when it is true
  - —Probability of finding a difference when there isn't any
  - —We want it to be < .05 to be sure there is a significant difference
  - —The following is for 2-tailed, unequal variance

```
t.test(A, B)
```

  - —Warning: Avoid multiple t-tests since the more you do, the more chance that at least one of them has a type I error
  - —If you do this, you need to do a Bonferroni correction
    - Homework: Look this up

One-tailed t-test

Used when the alternative hypothesis is that one average is greater or less

- `t.test(A,B,alternative="greater")`

More on testing

F test: Test the equality of two variances

```
var.test(A, B)
```

- It is good if there is no significant difference

Classical t-test (when variances are same)

```
t.test(A, B, var.equal=TRUE)
```

Nonparametric testing

When you can't be sure that the data is normally distributed

- Wilcoxon / Mann-Whitney
- Uses ranked data and compares medians rather than means
- Becoming more common in published data, especially if samples size is low

```
wilcox.test(A, B)
```

Discussion of problems with the p-value

Read:

Regina Nuzzo,

"Scientific method: Statistical errors: *P* values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume."

Nature, 12 Feb 2014,

http://www.nature.com/news/scientific-method-statistical-errors-1.14700

To combat issues with inference testing:

- Calculate effect size (Cohen's d) … next slide
- Quote confidence intervals, not just p-values

Effect size: Cohen's D

```
install.packages("effsize")
library(effsize)
cohen.d(A,B)
```

Small effect: < 0.2

Medium effect (0.2 to 0.8)

Large effect > 0.8 : Worth considering

Using Excel to explore data

Live discussion in class

Ten rules to use statistics effectively

https://www.sciencedaily.com/releases/2016/06/160620191409.htm

REPORTING EXPERIMENT RESULTS

Reporting experimental results - 1

Describe the system briefly

- Show some screenshots of what you were testing

Explain your hypotheses and their rationale

- And state null hypotheses

Describe all aspects of the design

- Variables, subjects, etc.
- Ensure somebody else would be able to reproduce it

Reporting experimental results

In results, give

- Basic descriptive statistics (mean, median, standard deviation)
- Boxplots plus any other diagrams
- Results of hypothesis testing (P values)
- Effect size (this is new advice compared to years ago)
- Recommendations for how the results should be used
- Anything else you observed about the system

Explain threats to validity (discussed earlier in the course)

Wrap up report with a conclusion