

Problem Set 1

LING280 Experimental methods

Instructions

Submit your response to mwagers@ucsc.edu by January 15, 2pm. Please format it legibly and intelligently.

Background

In this problem set, you will be working with a data fragment from an acceptability study on island constraints. The materials were organized into four conditions by an experimental design that crossed two factors, with two levels each: *a*) the complementizer of the embedded clause: (*that, whether*); *b*) the site of extraction: (*matrix subject, embedded object*). An example item set from this experiment is given below.

COMP: <i>that</i> , EXTRACTION.SITE: <i>matrix subject</i>	Who thinks that John bought a car?
COMP: <i>that</i> , EXTRACTION.SITE: <i>embedded object</i>	What do you think that John bought?
COMP: <i>whether</i> , EXTRACTION.SITE: <i>matrix subject</i>	Who wonders whether John bought a car?
COMP: <i>whether</i> , EXTRACTION.SITE: <i>embedded object</i>	What do you wonder whether John bought?

142 undergraduates gave 2 judgments on each of the resulting 4 sentence types. They made their judgments on a 7 point scale.

Problems

Import the sample data set from the following CSV file: <http://people.ucsc.edu/~mwagers/ling280/data/islands.simple.csv>

1. Summarize the data in several different ways. For each of the four conditions in the design, determine its mean, standard deviation, median, and inter-quartile range. Include a code fragment with annotations. See Problem 6 for an example of an annotated code fragment.
2. Assuming the data are stored in a dataframe called `judge`, what does the following command do? Identify each function, argument or operator in the command and explain its role.

```
> apply( judge, 2, table) -> judge.tab
```

3. For each of the four conditions, what proportion of judgments were assigned a value of '6' or '7'? Include the annotated code fragment you used to compute this. Note, you should not build in any constant values: if I supply a data frame with different numbers of individuals or judgments, I should be able to run this fragment of code and get the correct answer.
4. Create a boxplot of the data. Make sure each condition is represented in the same plot (i.e., don't turn in 4 separate plots; but one plot containing 4 boxes.)
5. Describe the kind of representation of the data given by commands below. What is the difference between the two?

```
> hist(judge[, "whether.matrix"], breaks=7, freq=TRUE)
> hist(judge[, "whether.matrix"], breaks=7, freq=FALSE)
```

6. Consider the following code fragment. Before you execute it, read through it carefully and try to figure out what it does. After you execute it, examine the resulting boxplot. What is going on? In a paragraph, first explain what the code does. Then explain what can be deduced about the relationship between *a*) the number of observations made in an experiment and *b*) the precision and accuracy of that experiment's estimate of the random variable. Be specific and justify your claims with reference to the information conveyed by features of the boxplot.

```
### CODE FRAGMENT FOR QUESTION 6
# Create a matrix, rep.experiment, with 1000 rows and 5 columns
rep.experiment <- matrix(nrow=1000, ncol=5)

# Assign the data from the 'whether' embedded object condition to its own vector
judge[, "whether.embedded"] -> island.judgments

###
# SIMULATE THE 1000x REPLICATION OF AN EXPERIMENT
# Run a loop that iterates from 1 to 1000
# Each loop creates a replicant of 5 different experiments

for(i in 1:1000) {
  ## Simulate an experiment with 5 observations by ...
  # 1. Sampling 5 data points, with replacement, from island.judgments
  sample(island.judgments,5,replace=TRUE) -> n5

  # 2. Computing the mean of this sample and
  # 3. Storing it in row i, column 1 of rep.experiment
  mean(n5) -> rep.experiment[i, 1]

  ## Simulate an experiment with 10 observations
  sample(island.judgments, 10, replace=TRUE) -> n10;
  mean(n10) -> rep.experiment[i, 2]

  ## Simulate an experiment with 20 observations
  sample(island.judgments, 20, replace=TRUE) -> n20;
  mean(n20) -> rep.experiment[i, 3]

  ## Simulate an experiment with 40 observations
  sample(island.judgments, 40, replace=TRUE) -> n40
  mean(n40) -> rep.experiment[i, 4]

  ## Simulate an experiment with 80 observations
  sample(island.judgments,80,replace=TRUE) -> n80
  mean(n80) -> rep.experiment[i, 5]
}

### Visualize the location and spread of the means from the 5 different experiments

boxplot(rep.experiment,
  names=c("n = 5", "n = 10","n = 20","n = 40","n = 80"),
  ylim=c(1,7),
  ylab= "mean rating",
  xlab= "sample size",
  main= "Simulated outcome of 1000 experiments")
```

7. The following is an unannotated code fragment. Annotate each line.
What does this fragment compute? And what do you conclude from its output?

```
numbers <- seq(0, 7, by=0.05)

squared.deviation <- vector(length=length(numbers))

k <- 1

for(n in numbers){

  sum( (n - island.judgments)^2 ) -> sum.squares

  sum.squares / (length(island.judgments) - 1) -> squared.deviation[k]

  k <- k + 1

}

plot(numbers, squared.deviation, pch="*")

abline(h = min(squared.deviation))
```

Notes

You should get in the practice early of writing analysis scripts that are formatted in a consistent, legible style. Some aspects of style that affect legibility include principles for naming different data types (variables, functions, constants, etc), the kind of comments to include and how to include them, and the way whitespace is used to make the code readable. A straightforward option for R is to follow the guide given here <http://google-styleguide.googlecode.com/svn/trunk/google-r-style.html>. It is important to develop such good habits: it will both help you understand your analysis later (when you've forgotten exactly how you did what you did) and also make the analysis scripts appropriate for public consumption.