

Problem Set 3: Simulations and Sampling Distributions

For this assignment, I want you to get familiar with the concept of simulating a sampling distribution and trying to understand how sample sizes are related to our ability to make accurate inference about some quantity.

You will need to use R for this homework, though there will be no need to read in any external data for this assignment. For some of the assignment, I have attached some code that will get you on your way. Code contained in slide sets as well as R script posted to the website will also be of use to you. Your grade will be based on the analytical content of your answers. Apart from turning in your word-processed answers, please submit the R code you use to complete this assignment. This assignment is due Wednesday, Nov. 12.

Directions:

In the recent 2008 California elections, controversial ballot measure Proposition 8 passed by a margin of (essentially) 52 percent to 48 percent (I'm not accounting for fractional differences). Thus, the "true" proportion of support in the state was .52 for "yes" and .48 for "no." As analysts, we may be interested in estimating what the proportion is from sample data. Since we cannot canvas the population, we must sample the population, usually with very small samples. For this assignment, I want you to use R's simulation capabilities to sample. These are the following tasks I want you to accomplish/questions I want you to answer:

Task 1: Create a population of 1 million voters. Each voter is either a "1" (supporter of Prop. 8) or a "0" (an opponent of Prop. 8). Since we are omnipotent, ensure the population proportion for supporters is .52 and for opponents is .48.

Task 2: Draw samples of size 10, 100, 500, 1000, and 10000 and in a table, report what the estimated sample proportion is along with the difference between this value and the known population value of .52 (i.e. if I draw a sample size of 10 and find the proportion is .40, the difference between my sample and the true value is $.40 - .52 = -.12$).

Question 1: Describe the pattern of results in your table. What seems to be the relationship between sample size and your estimate for the population proportion? What do the differences across sample sizes look like? (10 points)

Task 3: To generate a sampling distribution, I want you now to generate 1000 samples of size 500 from your population. From these 1000 samples, I want you to compute the mean (i.e. the average estimated proportion of "supporters" from the samples).

Question 2: What does this number represent? How does its value compare to that of the known population proportion of .52 "supporters"? Also, provide a definition of a sampling distribution in terms that would be transparent to a general

audience. How/why has what you have done in Task 3 a sampling distribution? (10 points)

Task 4: Compute an 80 percent, 90 percent, and 95 percent confidence interval based on your sampling distribution.

Question 3: Provide an interpretation of each of these intervals (that is, what do they describe? what do they tell you?). Further, describe the basic differences between the level-of-confidence and the width of the confidence interval. (10 points)

Task 5: Plot the sampling distribution as a histogram and include the plot in this assignment.

Question 4: Describe the basic features of the histogram (where is its mode or peak?). How does this distribution relate back to the known population value of “supporters”? On the plot, hand draw the range of your confidence intervals computed in Task 4. (10 points)

Question 5: Suppose someone says “you can’t trust a sample of 500!” Based on this exercise, what would you tell that individual? (10 points)

R Shell Code

Adapt this code to this assignment. This code is adapted from some examples in Verzani, which is required reading. To answer the question, you will have to go beyond just looking at lecture slides. Note that you will have to insert numbers where I ask you to!

```
#Creating a population of 0s and 1s

pop <- rep(0:1, c(insert a number for the number of 0s
out of 1 million, insert a number for the number of 1s out of 1 million))

#Creating an object that saves your sample proportion. Here, you are doing a SRS of size n from
#your population of 1 million. Whatever number you insert in the sample statement will
#return a sample size that large. So if I put in 10, I'll get a sample size of 10
#and the object sampleprop will give me the sample proportion for this sample.

set.seed(insert some number here)
sampleprop <- mean(sample(pop, insert sample size here), replace=FALSE)
sampleprop

#To resample, this code will work. I'm using a "for" loop to
#repeatedly sample (1000 times) from the population samples of size
# 500. This code can be used "as is" for the assignment. The object
#"sampledistmean" gives you the mean of all of the samples.

set.seed(insert some number here)
resample <- c()
for(i in 1:1000) resample[i]=mean(sample(pop, 500))

sampledistmean<-mean(resample)

#This code will return an 80 percent confidence interval. You
#need to add code to return a 90 and 95 percent c.i. Note that .1 and .9
#are used to bound the central 80 percent region; hence an 80 percent
#c.i.

quantile(resample, c(.1, .9)) #80 percent confidence interval

#This code gives a histogram

hist(resample, br=16, xlim=range(.52), xlab="label this yourself", ylab="label this yourself", main=
"label this yourself")
abline(v = sampledistmean)
```