STAT323

Assignment 1 - Simple Random Sampling

March 15, 2019

Evaluating the quality of parameter estimates under SRS

In this assignment we will examine the effect of different quantities (sample size, variance, desired estimation error, confidence) and different data distributions (see Table 0.1 for details) on estimates of population parameters from a sample. I have generated some synthetic data from various different distributions (mostly fairly symmetric ones, but 2 in particular are quite skewed) for you to explore in this assignment.

Datasets

The file `Populations.zip` contains 12 .csv files consisting of populations of size 10,000 sampled from various distributions as described in Table 0.1. A 13th file `all.csv` contains the same data collated in a single .csv file where the i-th column of `all.csv` contains the i-th population listed in Table 0.1. Note that the parameters listed in Table 0.1 are those for the underlying data generator - the population parameters are slightly different and you should start by calculating the population mean, μ, and variance, σ^2, for each of these populations to obtain ground truth. The parameters of the various data generators are the theoretical values the population parameters should take, and are included to make it easy for you to sanity-check what you find the population parameters to be (I have already done this - none of them are very far away from their theoretical values).

i	File Name	Distribution	Theoretical Parameters
1	sampU.csv	Uniform $(-1,1)$	$\mu = 0,\ \sigma^2 = \frac{1}{3}$
2	sampU5.csv	Uniform $(-5,5)$	$\mu = 0,\ \sigma^2 = \frac{25}{3}$
3	sampN.csv	$\mathcal{N}(0,1)$	$\mu = 0,\ \sigma^2 = 1$
4	sampN2.csv	$\mathcal{N}(0,2)$	$\mu = 0,\ \sigma^2 = 2$
5	sampN5.csv	$\mathcal{N}(0,5)$	$\mu = 0,\ \sigma^2 = 5$
6	sampL.csv	Uniform $\{-2,-1,0,1,2\}$	$\mu = 0,\ \sigma^2 = 2$
7	sampBern.csv	Uniform $\{-1,1\}$	$\mu = 0,\ \sigma^2 = 1$
8	sampChi21.csv	χ^2_1	$\mu = 1,\ \sigma^2 = 2$
9	sampChi25.csv	χ^2_5	$\mu = 5,\ \sigma^2 = 10$
10	sampChi210.csv	χ^2_{10}	$\mu = 10,\ \sigma^2 = 20$
11	sampPoiss1.csv	Poisson (1)	$\mu = 1,\ \sigma^2 = 1$
12	sampPoiss5.csv	Poisson (5)	$\mu = 5,\ \sigma^2 = 5$

WHAT TO DO

The following tasks will be extremely onerous unless you use some statistical software to help you! I strongly suggest that you read through all of the tasks below before beginning to code, as you can save yourself time by coding sensibly to exploit the redundancy in the separate tasks 2–4. For example you can calculate the total and mean estimates at the same time and save them on each pass through the loop over the 50 samples.
Please come and see me ASAP if you are stuck before you start - for example, if you are not sure how to import the data files into your software, how to generate an SRS from these data, or how to carry out the estimation tasks listed below. My office hour is on Wednesday at 13.00, room G.3.31, or you can make an appointment at a mutually convenient time by emailing me. I have uploaded some commented demo R code to Moodle, `Assignment_1.R`, that you can use to get started if you wish but if you prefer to use software other than R (e.g. Minitab, Matlab, Julia, Python, ...) that is perfectly fine. The demo R code is also included as an appendix to this document.

For each population you need to carry out the following:

1. For each population calculate its mean, variance, and total.
2. For each population, $i = \{1,2,\ldots,12\}$, calculate the sample size n_i required such that:

   $$\Pr\left(|\mu_i - \bar{x}_i| > \frac{1}{4}\right) \leq 0.1$$

   i.e. so that the estimation error in the population mean is no more than ±1/4 with 90% confidence. You may use a normal approximation (i.e. the appropriate choice of $Z_{\alpha/2}$ and the population variance you calculated in 1) to estimate the required sample size.
3. Make 50 simple random samples of size n_i from the i-th population and calculate the sample mean and variance from each sample. Discuss the quality of your estimates relative to the corresponding population means and variances, and whether the calculated sample size seems adequate.

For example: Is the population mean you calculated in 1 within 1/4 of the sample mean for at least 45 of the samples in each case? For which populations were there a high number of bad estimates? Is there any obvious reason why this might be for those populations? For which populations were there a low number of bad estimates? Is there any obvious reason why this might be for those populations? For example, you could use your 50 sample means to estimate what the sampling distribution of the mean for the corresponding population looks like by, say, plotting a histogram of the mean estimates. What is the standard error of the mean? Is the normal assumption we made justified (which you could explore by e.g. calculating the population skewnesses, or with Q-Q plots)? Is there anything else that strikes you?

4. Repeat 2. and 3. for the population **total** allowing an error of ±50 in the total (with 90% confidence). Note that all of the samples sizes will be larger than before. Examine, consider, and describe your findings, in a similar way to how you did for the mean estimates.

5. Write a short report containing tables, plots (e.g. histograms, box and whiskers, Q-Q plots) and/or other graphics, as appropriate, describing what you found out. There is no page limit, but more than about 4–5 pages *of text* excluding appendices is probably excessive. Your report should be formally laid out – along the lines we covered in last Tuesday's lecture – and 40% of the marks for this assignment will be for good presentation. You need not include **all** the sections we discussed, for example abstract, TOC, background, are optional. Title, a brief introduction, graphs and tables, experimental protocol, results and discussion, conclusions, are not. You should not include the raw data in your report, and if you want to include your code it should be in an appendix.

Deadline

The deadline for this assignment is 12.00 noon on Monday 25th March 2019. Please hand your assignment to me at the start of the lecture. Submissions received up to a day late (i.e. Tuesday's lecture on 26th March) will be penalised with a 20% reduction in marks. Submissions received after 26th will not be marked except in exceptional circumstances.

Demo R Code

```
###Assign1 Stats323 Demo Code####
#set plot dimensions
par( mfrow = c( 4,3 ) )

#check working directory
getwd()
#set to location wanting to save files to
setwd("E:/Stat323/Assign1/")

##Load separate datasets
```

```
SampU.df = t(read.table("SampU.csv",header = FALSE,sep = ","))
SampU5.df = t(read.table("SampU5.csv",header = FALSE,sep = ","))
SampN.df = t(read.table("SampN.csv",header = FALSE,sep = ","))
SampN2.df = t(read.table("SampN2.csv",header = FALSE,sep = ","))
SampN5.df = t(read.table("SampN5.csv",header = FALSE,sep = ","))
SampL.df = t(read.table("SampL.csv",header = FALSE,sep = ","))
SampBern.df = t(read.table("SampBern.csv",header = FALSE,sep = ","))
SampChi21.df = t(read.table("SampChi21.csv",header = FALSE,sep = ","))
SampChi25.df = t(read.table("SampChi25.csv",header = FALSE,sep = ","))
SampChi210.df = t(read.table("SampChi210.csv",header = FALSE,sep = ","))
SampPoiss1.df = t(read.table("SampPoiss1.csv",header = FALSE,sep = ","))
SampPoiss5.df = t(read.table("SampPoiss5.csv",header = FALSE,sep = ","))
##put all datasets in to one matrix(could have used all file here instead)
matrix = cbind(SampU.df,SampU5.df,SampN.df,SampN2.df,SampN5.df,SampL.df,SampBern.df,
SampChi21.df,SampChi25.df,SampChi210.df,SampPoiss1.df,SampPoiss5.df)
#set column names for the populations
colnames(matrix) = cbind("SampU","SampU5","SampN","SampN2","SampN5","SampL","SampBern",
"SampChi21","SampChi25","SampChi210","SampPoiss1","SampPoiss5")

###QUESTION 1###

#declaring variables and setting to null (in case need to run again)
means = NULL
vars = NULL
totals = NULL

for(i in 1:12){
#calc mean var and totals for each population and put into vector
means = c(means,mean(matrix[,i]))
vars = c(vars,var(matrix[,i]))
totals = c(totals,sum(matrix[,i]))
# plot histograms of populations
hist(matrix[,i], main=substitute(paste(’Histogram of ’, matrix[a]),
list(a=colnames(matrix)[i])))
}
# sanity check the outputs
means
vars
totals

#create matrix of all stats
Q1 = cbind(means,vars,totals)
#set row names to stats for that population
row.names(Q1) = rbind("SampU","SampU5","SampN","SampN2","SampN5","SampL","SampBern",
"SampChi21","SampChi25","SampChi210","SampPoiss1","SampPoiss5")

#write csv file
write.csv(Q1,"Question1.csv",row.names=TRUE)

# sanity check the outputs. Compare head to top of csv file.
class(matrix)
dim(matrix)
head(matrix)


###QUESTIONS 2 and 3###

##declaring variables and assigning values
d = .25
z = 1.645
N = 10000
#variables to be used within loop
n0 = NULL
n = NULL
sampSize = NULL

#calculate req sample size for each population
for(i in 1:12){
#n = N/(1+((((d)^2)*N)/(Q1[i,2]*z)))
n0 = (Q1[i,2]*z^2)/(d^2)
n = n0*(1/(1+(n0/N)))
#created vector with required sample size
sampSize = rbind(sampSize,ceiling(n))
}
#row.names = rbind("SampU","SampU5","SampN","SampN2","SampN5","SampL","SampBern",
#SampChi21","SampChi25","SampChi210","SampPoiss1","SampPoiss5")
row.names(sampSize) = row.names(Q1)
write.csv(sampSize,"Question2.csv",row.names=TRUE)

sampMeans = mat.or.vec(50,12)
diffs = mat.or.vec(50,12)

# Estimate the sample means from samples of each of 12 populations, 50 times each
for(i in 1:12){
for(j in 1:50){
tempSample = sample(matrix[,i],sampSize[i],replace=FALSE,prob=NULL)
sampMeans[j,i] = mean(tempSample)
# get differences between pop and sample means
diffs[j,i] = abs(sampMeans[j,i] - means[i])
}
}

#For Question 4 you can make some straightforward modifications to the above code.
```