SML 201 Problem Set 3

Tyler Campbell

April 4, 2018

Problem set 3 is due by 11:59pm on Wednesday April 4 on Blackboard. Please submit both a .Rmd and a .pdf file on Blackboard. If the due date falls on a date that has a lecture on the next day please bring a hard copy of the pdf file to the first lecture after the due date; otherwise, please drop off the pdf copy at 26 Prospect Avenue (see the *Submitting Problem Sets and Projects* section under *Problem Sets and Projects* on the Syllabus for detailed instructions) by 5pm on the next day of the due date.

Make sure that you have all your digital signatures along with the honor pledge in each of these documents (there should be more than one signature if you work in groups).

This problem set can be completed in groups of up to 3 students. Unlike for projects, you can work with whoever you prefer for problem sets. It is okay to work by yourself, if this is preferable. You are welcome to get help (you can either ask questions on Piazza or talk to the instructors in person during office hours) from instructors for *problem sets*; however, please do not post code on a public post on Piazza.

When working in a group it is your responsibility to make sure that you are satisfied with all parts of the report and the submission is on time (e.g., we will not entertain arguments that deficiencies are the responsibility of other group members). We expect that the work on any given problem set or project contains approximately equal contributions from all members of the group; we expect that you each work independently first and then compare your answers with each other once you all finish or you all work together. Failing to make contributions and then putting your name on a project will be considered a violation of the honor code. Also, please do not divide work among your group mates.

For all parts of this problem set, you **MUST** use R commands to print the output as part of your R Markdown file. You are not permitted to find the answer in the R console and then copy paste the answer into this document.

If you are completing this problem set in a group, please have only **one** person in your group turn in the .Rmd and .pdf files; other people in your group should turn in the list of the people in your group in the *Text Submission* field on the submission page.

Please type your name(s) after "Digitally signed:" below the honor pledge to serve as digital signature(s). Put the pledge and your signature(s) at the beginning of each document that you turn in.

> I pledge my honor that I have not violated the honor code when completing this assignment.

Digitally signed: Tyler Campbell

In order to receive full credits, please have sensible titles and axis labels for all your graphs and adjust values for all the relevant graphical parameters so that your plots are informative. Do not round off values at an intermediate step and avoid hand-code in values. Also, all answers must be written in complete sentences.

Just a friendly reminder: Please remember to annotate your code and have answers in the write up section, not in the code chunks.

In this problem set we will use the `possum` dataset from the `DAAG` package. You will need to install and load the package in order to access the dataset. The dataset consists of measurements on 104 mountain brushtail possums, trapped at seven sites from Southern Victoria to central Queensland in Australia. (In my opinion these brushtail possums look a lot cuter than the possums that I encountered in California (http://www.arkive.org/common-brushtail-possum/trichosurus-vulpecula/image-G39813.html)). For the purpose of this problem set, you can assume that the mountain brushtail possums in the dataset are a simple random sample chosen from a large population; thus, the the possums in the dataset are approximately independent of each others.

Remember to look up the information on the dataset on the help manual before you start working on the questions. You should also check the size of the dataset and the data types of the variables in the dataset.

In your report list the variable names, the units the variables are in, and the variable descriptions shown on the R help manual for the variables used in this problem set. (Note: This piece of information is not shown on the help manual but the head and ear conch lengths in the dataset were measured in millimeters (mm) and the total and tail lengths were in centimeters (cm).) You might not be reminded about the step of listing variables again in the next problem set or future projects since by now you are expected to have formed a habit of including variable descriptions in your reports–this step is crucial since without the variable descriptions your readers will not be able to fully understand what you are trying to convey in a report.

possum is a dataframe that is 104 rows by 14 columns

case: observation number "numeric"

site: one of seven locations where possums were trapped "numeric"

Pop: a factor which classifies the sites as Vic Victoria, other New South Wales or Queensland "factor"

sex: a factor with levels f female, m male "factor"

age: age "numeric"

hdlngth: head length (in mm) "numeric"

skullw: skull width "numeric"

totlngth: total length (in cm) "numeric"

taill: tail length (in cm) "numeric"

footlgth :foot length "numeric"

earconch: ear conch length (in mm) “numeric”

eye: distance from medial canthus to lateral canthus of right eye “numeric”

chest: chest girth (in cm) “numeric”

belly: belly girth (in cm) “numeric”

```
library("DAAG")
Loading required package: lattice
# gain information about possum dataset class
class(possum)
[1] "data.frame"
# column classes
sapply(possum, class)
     case      site       Pop       sex       age   hdlngth    skullw
"numeric" "numeric"  "factor"  "factor" "numeric" "numeric" "numeric"
 totlngth     taill  footlgth  earconch       eye     chest     belly
"numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric"
# size
dim(possum)
[1] 104  14
```

Question 1

Using this dataset we would like to answer the question whether the gender of a possum is independent of the state in which it was trapped. With the vector `possum$Pop` you can find out whether a possum was trapped in the state Victoria or not (see details on the help manual).

Part a

What percentage of the trapped possums are female overall?
41.35% of the trapped possums are female

```
# find percentage of females trapped
perc = table(possum$sex)[1]/length(possum$sex) * 100
percf = perc/100
round(perc, digits = 2)
    f
41.35
```

Part b

Among the possums that were trapped in Victoria, what percentage of them are female? Among the possums that were trapped in other states (New South Wales or Queensland), what percentage of them are female?

Among the possums that were trapped in Victoria, 52.17% are female. Among the possums that were trapped in other states (New South Wales or Queensland), 32.76% are female.

```
# percentage of females trapped in Victoria
victoria = possum[possum$Pop == "Vic", ]
perc = table(victoria$sex)[1]/length(victoria$sex) * 100
round(perc, digits = 2)
    f
52.17
# percentage of females trapped in Other
other = possum[possum$Pop == "other", ]
perc = table(other$sex)[1]/length(other$sex) * 100
round(perc, digits = 2)
    f
32.76
```

Part c

Based on the information provided by the dataset, would you say that the event that a possum is female and the event that a possum is from Victoria are more likely to be independent or not? Use the numerical values that you found in parts a and/or b to support your argument.

The probability that a possum captured is a female is 41.35%. The probability that a possum captured is a female given the event that the possum is captured in Victoria is 52.17%. Since the chance that the possum is a female is greater given the possum is captured in Victoria, the event that a possum is female and the event that a possum is from Victoria are not independent.

Question 2

Part a

If I randomly select 20 possums with replacement from the dataset `possum`, what is the chance that at least half of the possums are females?

There is a 28.56% chance that at least half of the possums are females.

```
# binom distrubtion to find probability of 50% females
p = sum(dbinom(x = 10:20, size = 20, prob = percf))
p
[1] 0.2855695
# alternatively
1 - pbinom(q = 9, size = 20, prob = percf)
[1] 0.2855695
```

Part b

If I repeat the procedure described in Part a 10 times (i.e., repeatedly select 10 samples of 20 with-replacement-drawns), what is the chance that at least 7 of these 10 samples have at least 50% females?

There is a 0.79% that at least 7 of these 10 samples have at least 50% females.

```
# binom distrubtion to find probability that 7 out of 10 trials
# are 50% females
sum(dbinom(x = 7:10, size = 10, prob = p))
[1] 0.007886783
# alternatively
1 - pbinom(q = 6, size = 10, prob = p)
[1] 0.007886783
```

Part c

Verify your answer in part a with a simulation outlined in the following steps:

a) Set the seed of your simulation with set.seed(2018) and simulate 100,000 samples, each of size 20;

b) the simulated samples should be drawn with replacement from a vector of 0's and 1's, where 1 means "female" and 0 means "male"; the length of the vector should match the number of possums and the number of 1's in the vector should match the number of females in the `possum` dataset;

c) use either the function `apply()` or `sapply()` to find out the percentage of females in each of the simulated samples;

d) calculate the percentage of simulated samples that have at least 50% females.

From the simulation the percentage is 28.39% which is very close to 28.56% found in part 2A.

```
set.seed(2018)
fm = (as.numeric(possum$sex) - 2) * -1
m = 1e+05
# calculate percentage of females
tmean = sapply(1:m, FUN = function(x) mean(sample(fm, size = 20, replace = T)))
# calculate the percentage of simulated samples that have at least
# 50% females
sum(tmean >= 0.5)/m
[1] 0.28388
```

Question 3

With the `possum` dataset we would like to predict the average head length of all the mountain brushtail possums in Victoria, New South Wales and Queensland with a 95% confidence interval.

Part a

What is the population and what is the sample in this case? How big is the sample size? Is the average head length of all the possums in the dataset `possum` a parameter or a statistic?

Population is all the mountain brushtail possums in Victoria, New South Wales and Queensland. The sample is the possum dataset. The sample size is 104. It is a statistic because it comes from a sample.

Part b

Since the dataset size is large enough we can assume that the histogram for the possum head lengths in the dataset is a good approximation for the histogram for the head lengths of all the possums in Victoria, New South Wales and Queensland; which principle of probability is our assumption based on? You can assume that the sample size is very small *relative to* the population size.

This assumption is based on the Law of Large Numbers (LLN).

Part c

If the histogram for the head length measurements in the sample is close to a Normal distribution, approximately what percent of the head length measurements in the sample should fall within 3 SD's of the sample mean? Compare this figure to the actual percentage in your sample.

If the histogram for the head length measurements in the sample is close to a Normal distribution then 99.7% of the head length measurements in the sample falls within 3 sd's of the mean.

The actual percentage in the sample is 100%.

```
# calculate % of head length measurements in the sample within 3
# sd's of mean
headlength = possum$hdlngth
m = mean(headlength)
s = sd(headlength)
min = m - 3 * s
max = m + 3 * s
count = length(headlength >= min & headlength <= max)
```

```
count/length(headlength) * 100
[1] 100
```

Part d

What distribution do you think will be a good approximation for the distribution of the quantity $\frac{sample.mean - population.mean}{sample.sd/\sqrt{n}}$, where *sample.mean* and *sample.sd* are the sample mean and sample SD of the head length measurements in the sample, respectively, *population.mean* is the average head length for the population and n is the sample size? Please provide the name of the distribution along with the parameter(s) of the distribution. (You should make some plots of the data to look at but you do not need to submit the plots for this part.)

Since we do not know the population SD, the sample size is relatively small and it is a reasonable assumption that the population follows a normal distribution (based on that the dots roughly fall on a straight line from qqplot and the Sample distribution is roughly symmetric) t-distrubution should be used with a parameter that is the degrees of freedom. This is equal to n-1 which is 103 in this case. ## Part e

Construct a 95% confidence interval to predict the average head length of all the mountain brushtail possums in Victoria, New South Wales and Queensland.

95% confidence interval for the population mean of head length is 91.91 - 93.30.

```
# a 95% CI for the mean of the population
p = 0.95
mean(headlength) + c(-1, 1) * qt(p + (1 - p)/2, df = length(headlength) -
    1) * sd(headlength)/sqrt(length(headlength))
[1] 91.90796 93.29781
```

Part f

Create a function named `make.CI` that produces a confidence interval for its users. `make.CI` should take two input variables:

- `input.vector`: a vector of values that the user will provide for making the confidence interval;
- `percent.conf`: the percentage of confidence for the resulting confidence interval.

`make.CI` should output the two endpoints of the confidence interval.

First, test your function by using it to construct a 95% confidence interval for the average possum head length and compare your output with your answer in Part e to make sure that your function behaves the way you expect.

Next, use your function to construct 90% confidence intervals for the average total length and the average tail length for the possums in Victoria, New South Wales and Queensland; report the figures for the confidence intervals.

Values from function and Part E are consistent.

90% confidence interval for the population mean of total length is 86.39 - 87.79.

90% confidence interval for the population mean of tail length is 36.69 - 37.33.

```
# a p% CI for the mean of the population from sample v given that
# p = perc.conf and v is input.vector
make.CI = function(input.vector, percent.conf) {
    ci = mean(input.vector) + c(-1, 1) * qt(percent.conf + (1 - percent.conf)/2,
        df = length(input.vector) - 1) * sd(input.vector)/sqrt(length(input.vector))
    print(ci)
}

make.CI(possum$hdlngth, 0.95)
[1] 91.90796 93.29781
make.CI(possum$totlngth, 0.9)
[1] 86.38690 87.79003
make.CI(possum$taill, 0.9)
[1] 36.69069 37.32854
```