SML 201 Problem Set 3

Stuart Duffield and Kevin Hou

November 4, 2018

Problem set is due by 11:59pm on Thursday November 8. Please submit both a .Rmd and a .pdf file on Blackboard by the deadline **and** drop off a hard copy of the pdf file at 26 Prospect Avenue by 5pm of the **next day** of the due date. To look for the drop-off cabinet, after you enter the building turn to the left to enter the lounge area and the file cabinet is to your right with an open slot with the label "SML 201 Homework"; note that the building might be locked after 6pm and on the weekends. You are also welcome to bring your PDF copy to any lecture **before** the deadline and I will drop off the copy for you.

This problem set can be completed in groups of up to 2 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 instructors in person during office hours) from instructors for *problem sets*; however, please do not post code/solutions on Piazza on a public post.

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 the other group member). We expect that the work on any given problem set or project contains approximately equal contributions from both 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 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 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 and 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; the other person in your group should turn in the list of the people in your group in the *Text Submission* field on the submission page. This means that **everyone should make a submission**–either a file-upload or a text submission–regardless of whether you are working in a group or not.

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:

In order to receive full credits, please have sensible titles and axis labels for all your graphs and adjust values of the relevant graphical parameters so that your plots are informative. Also, all answers must be written in complete sentences.

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 about 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 are measured in millimeters (mm) while the total and tail lengths are in centimeters (cm).) You might not be reminded about the step of listing variables again in the next problem set or in 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.

Question 1

We would like to answer the question whether the gender of a possum is independent of the state in which the animal was trapped for possums in the dataset. 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).

```
library(DAAG)
Loading required package: lattice
```

Part a

What percentage of the trapped possums are female overall?

```
(tapply(rep(1, times = length(possum$sex)), possum$sex, sum)/length(possum$sex) *
    100)
       f        m
41.34615 58.65385
```

41.34615% of the trapped possums are female.

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?

```
(tapply(rep(1, times = length(possum[possum$Pop == "Vic", ]$sex)),
    possum[possum$Pop == "Vic", ]$sex, sum)/length(possum[possum$Pop ==
    "Vic", ]$sex) * 100)
       f        m
52.17391 47.82609

(tapply(rep(1, times = length(possum[possum$Pop == "other", ]$sex)),
    possum[possum$Pop == "other", ]$sex, sum)/length(possum[possum$Pop ==
    "other", ]$sex) * 100)
       f        m
32.75862 67.24138
```

52.17391% of the possums trapped in Victoria are female. 32.75862% of the possums trapped elsewhere are female.

Part c

If I randomly select a possum from the dataset, are the events that the possum is female and that the possum is from Victoria independent? Use the numerical values that you found in Parts a and/or b to support your argument.

The events that the possum is female and that the possum is from Victoria are not independent because the probability that the possum is female is not equal to the probability that the possum is female and from Victoria. Knowing that a possum is female will influence the probablility that the possume is from Victoria and vice versa.

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?

```
possum.percentageFemale = tapply(rep(1, times = length(possum$sex)),
    possum$sex, sum)/length(possum$sex)
sum(dbinom(10:20, 20, possum.percentageFemale[[1]]))
[1] 0.2855695
```

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

Part b

If I repeat the procedure described in Part a 10 times (i.e., repeatedly select 10 samples with replacement, and for each sample I draw 20 possums with replacement from the original dataset), what is the chance that at least 7 of these 10 samples have at least 50% females?

```
possum.partA = sum(dbinom(10:20, 20, possum.percentageFemale[[1]]))
sum(dbinom(7:10, 10, possum.partA))
[1] 0.007886783
```

0.789% chance that at least 7 of these 10 samples have at least 50% females

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 represents the female possums in your original dataset and 0 represents the male possums; make sure that the length of your vector matches the number of possums 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.

```
sex.bin <- as.numeric(possum$sex == "f")
set.seed(2018)
sim.possum <- sapply(1:1e+05, FUN = function(x) {
    sample(sex.bin, 20, replace = T)
})
sim.possum.mean <- sapply(1:length(sim.possum[1, ]), FUN = function(x) {
    mean(sim.possum[, x])
})
length(sim.possum.mean[sim.possum.mean >= 0.5])/length(sim.possum.mean) *
    100
[1] 28.388
```

By drawing 100,000 random samples of size twenty from the vector of genders, we find that 28.388% of these samples have atleast 50% females. This is consistent with my answer to part a.

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?

The population is all of the mountain bushtail possums in Victoria, New South Wales and Queensland. The sample is the 107 mountain bushtail possums captured and recorded in the dataset `possum`. The average head length of all the possums in the dataset is a statistic.

Part b

Since the dataset size (we are talking about the *absolute* size here) 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 number of possums in the dataset is very small *relative to* the number of all the possums in the locations of interest so the possums were selected independently.)

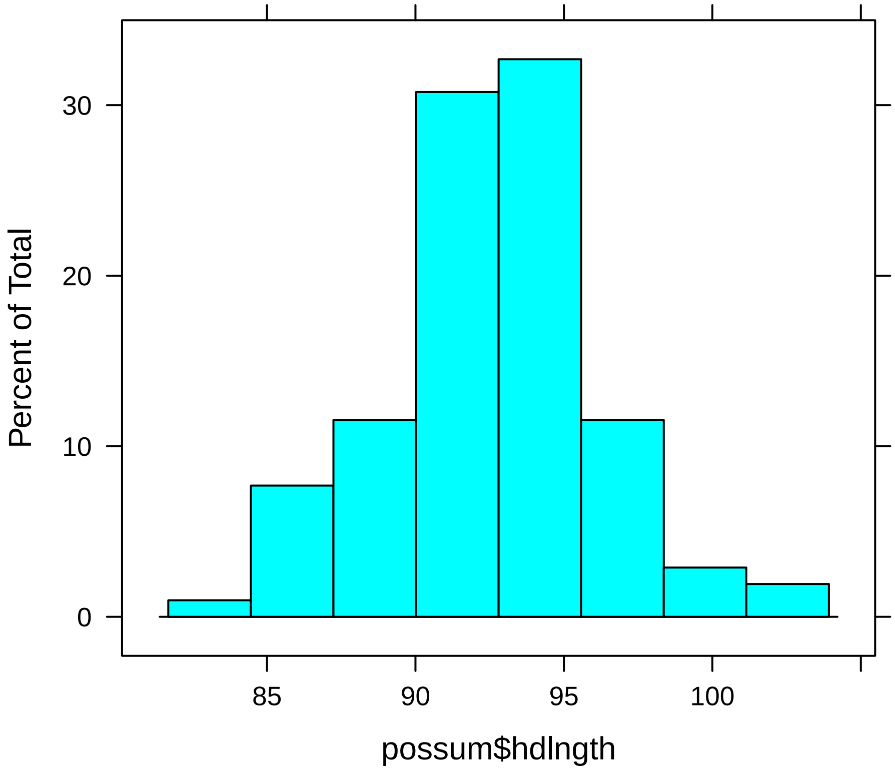
This assumption is based on the central limit theorem, which states that as the sample size gets bigger, the distribution of the sample means gets closer and closer to the normal distribution, where the mean is equal to mu and the standard deviation is equal to sigma/sqrt(n), so sample means themselves become better and better predictors of the population mean if the parameter we are interested in is normally distributed.

Part c

We want to check whether it is okay to assume that the Normal distribution is a good approximation for the head length distribution of *all* the possums at the locations of interest (i.e., in Victoria, New South Wales and Queensland). Since we do not have the data on all the possums in the locations of interest we will investigate the figures in our dataset to see if there is any evidence against this assumption.

Make a histogram for the possum head length measurements in the `possum` dataset. Observe the shape of the histogram; does the distribution look roughly symmetric? For a Normal distribution what percentages of the data points should fall within 1, 2, and 3 SD's of the mean of the distribution? Now, for the `possum` dataset what percentages of the head length measurements fall within 2 and 3 SD's of the head length average? (Side note: since the dataset covers only a small fraction of all the possums in the locations of interest, due to sampling variability even when the head length distribution of all the possums at the locations of interest follows a Normal distribution the percentage of the data falling within 1 SD of the mean often deviate quite a bit from the theoretical percentage; this is why we do not use the 1-SD-percentage to check here.)

```r
histogram(possum$hdlngth)
```



```
# Statistics
hdlngth.m = mean(possum$hdlngth)
hdlngth.sd = sd(possum$hdlngth)

possum.2.sd = possum$hdlngth[possum$hdlngth <= hdlngth.m + 2 * hdlngth.sd &
    possum$hdlngth >= hdlngth.m - 2 * hdlngth.sd]
possum.3.sd = possum$hdlngth[possum$hdlngth <= hdlngth.m + 3 * hdlngth.sd &
    possum$hdlngth >= hdlngth.m - 3 * hdlngth.sd]

length(possum.2.sd)/length(possum$hdlngth)
[1] 0.9326923
length(possum.3.sd)/length(possum$hdlngth)
[1] 1
```

The percentage of data points that fall within 2 and 3 SD's of the head length average is 93.27% and 100%, respectively.

Part d

What distribution do you think will be a good approximation for the distribution of the quantity $\frac{\bar{x}-\mu}{s/\sqrt{n}}$, where

- $\bar{x}$ is the average of the head length measurements in the dataset;

- s is the SD of the head length measurements in the dataset;
- μ is the average head length for all the possums in the locations of interest;
- n is the number of possums in the dataset.

Please provide the name of the distribution along with the parameter(s) of the distribution. (You might want to make some plots of the data to look at but you do not need to submit the plots for this part.)

The distribution is a standard normal distribution. This is a normal distribution with parameters $\mu = 0$ and $\sigma = 1$.

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.

```r
# 95% confidence interval
confidence = 0.95
sd = sd(possum$hdlngth)
n = length(possum$hdlngth)
m = mean(possum$hdlngth)

upperbound.z = qnorm(confidence + ((1 - confidence)/2))
lowerbound.z = qnorm((1 - confidence)/2)

upperbound.hdlngth = m + (upperbound.z * (sd/sqrt(n)))
lowerbound.hdlngth = m - (upperbound.z * (sd/sqrt(n)))

print(sprintf("%2.f%% confidence interval: (%2.2f, %2.2f)", confidence *
    100, lowerbound.hdlngth, upperbound.hdlngth))
[1] "95% confidence interval: (91.92, 93.29)"
```

Part f

Let's generalize what you did in Part e. Create a function named `make.CI` that will produce a confidence interval for its users. `make.CI` should take two input variables:

- `input.vector`: a vector of values to be used 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.

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.

In addition, 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 in appropriate units.

```
make.CI = function(input.vector, percentage.conf) {
    sd = sd(input.vector)
    n = length(input.vector)
    m = mean(input.vector)

    upperbound.z = qnorm(percentage.conf + ((1 - percentage.conf)/2))
    lowerbound.z = qnorm((1 - percentage.conf)/2)

    upperbound = m + (upperbound.z * (sd/sqrt(n)))
    lowerbound = m - (upperbound.z * (sd/sqrt(n)))

    return(c(lowerbound, upperbound))
}
```

```
# Same as part E
make.CI(possum$hdlngth, 0.95)
[1] 91.91612 93.28965

# 90% CI for average total length
make.CI(possum$totlngth, 0.9)
[1] 86.39321 87.78372

# 90% CI for average tail length
make.CI(possum$taill, 0.9)
[1] 36.69356 37.32567
```

Our function produces the same results as part E with a 95% confidence interval of 91.92 - 93.29mm average head length. The 90% confidence interval for average total lengths was 86.39 - 87.78mm. The 90% confidence interval for average tail length was 36.69 - 37.33mm.