SML 201 Problem Set 4

Tyler Campbell

April 26, 2018

Problem set 4 is due by 11:59pm on Thursday April 26 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.

Please feel free to skip the step of listing variables and variable descriptions for this problem set.

Question 1

We will use the `possum` dataset (from the `DAAG` package) that we used in Problem Set 3. Just as for Problem Set 3 you can assume that the mountain brushtail possums in the dataset are a simple random sample chosen from a large population; thus, the possums in the dataset are approximately independent of each others.

In this question we would like to predict the average ear conch length of the female mountain brushtail possums in Victoria with a 95% confidence interval. Since we do not know the population SD and since the sample size is not very large (only 24), we need to check if it is reasonable to assume that the population (i.e., the ear conch lengths of all the female possums in Victoria) is Normal; if the population is Normal then the standardized sample mean of the ear conch lengths follows a t-distribution with degree of freedom 23 (see week7b_notes or week8a_notes). However, if the population is not Normal the standardized sample mean does not necessarily follow a t-distribution, and in that case we will need to use Bootstrap sampling to construct the 95% confidence interval.

Note that the ear conch lengths were measured in mm.

Part a

From `possum$earconch` extract the ear conch length measurements that correspond to the female mountain brushtail possums trapped in Victoria and store these values in a vector called `f.vic.earconch`. There should be 24 values in `f.vic.earconch`.

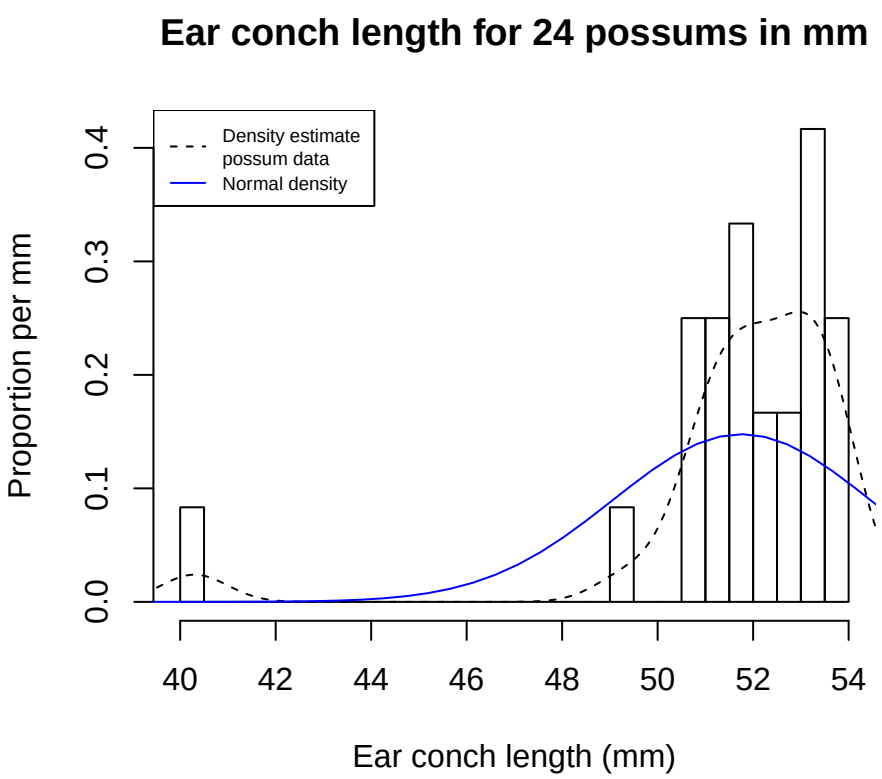
(i) Draw a histogram for these 24 values and superimpose the density curve of a Normal distribution with the same mean and SD values as the sample.

(ii) Make a quantile-quantile plot to investigate the relationship between the sample quantiles and the Normal(0, 1) quantiles; also, add the qq line on the graph for reference.

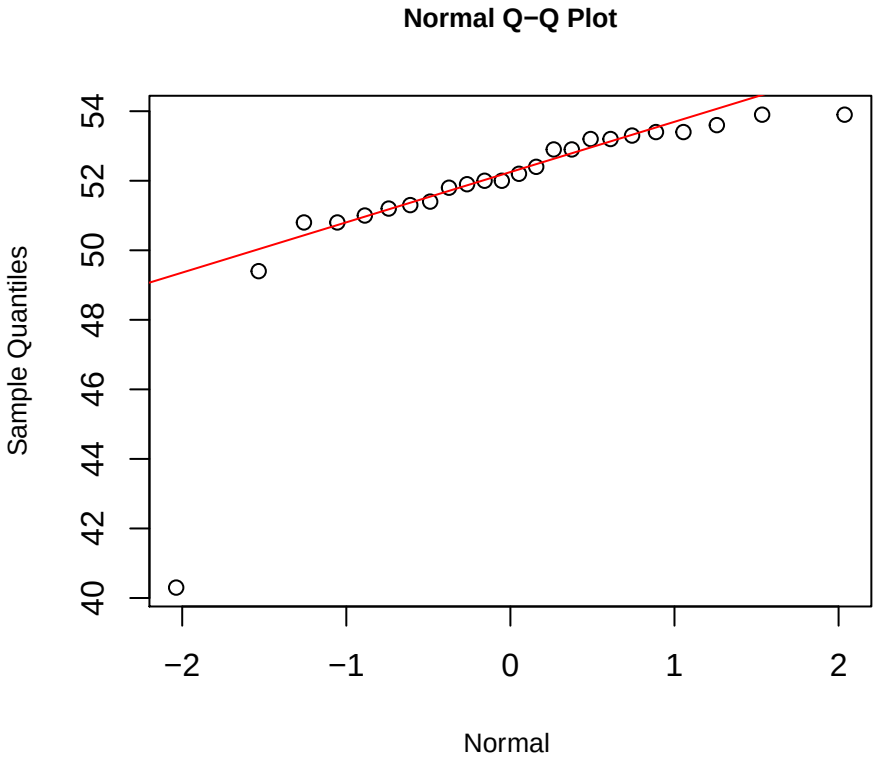
Is it reasonable to assume that the population distribution is Normal? Justify your answer.

It is not reasonable to assume the distribution is normal as the density doesn't look symmetric or normal, as well as the qq-plot does not form a striaght-line.

```r
library(DAAG)
Loading required package: lattice
# extract the ear conch length measurements that correspond to the
# female mountain brushtail possums trapped in Victoria
f.vic.earconch = possum[possum$sex == "f" & possum$Pop == "Vic", ]$earconch
# histrogram of ear conch length
hist(f.vic.earconch, breaks = 20, freq = F, xlab = "Ear conch length (mm)",
    main = "Ear conch length for 24 possums in mm", ylab = "Proportion per mm")
# density of histogram
lines(density(f.vic.earconch), lty = 2)
# normal distributions
xseq = seq(from = min(f.vic.earconch) * 0.9, to = max(f.vic.earconch) *
    1.1, length = 50)
lines(x = xseq, y = dnorm(x = xseq, mean = mean(f.vic.earconch), sd = sd(f.vic.earconch)),
    col = "blue")
legend(x = "topleft", y = 0.3, legend = c("Density estimate \npossum data",
    "Normal density"), lty = c(2, 1), col = c("black", "blue"), cex = 0.6)
```




```r
# Compare to the quantiles of Normal(0,1) distribution
qqnorm(y = f.vic.earconch, main = "Normal Q-Q Plot", xlab = "Normal ",
    ylab = "Sample Quantiles", cex.lab = 0.8, cex.main = 0.8)
qqline(y = f.vic.earconch, col = "red")
```



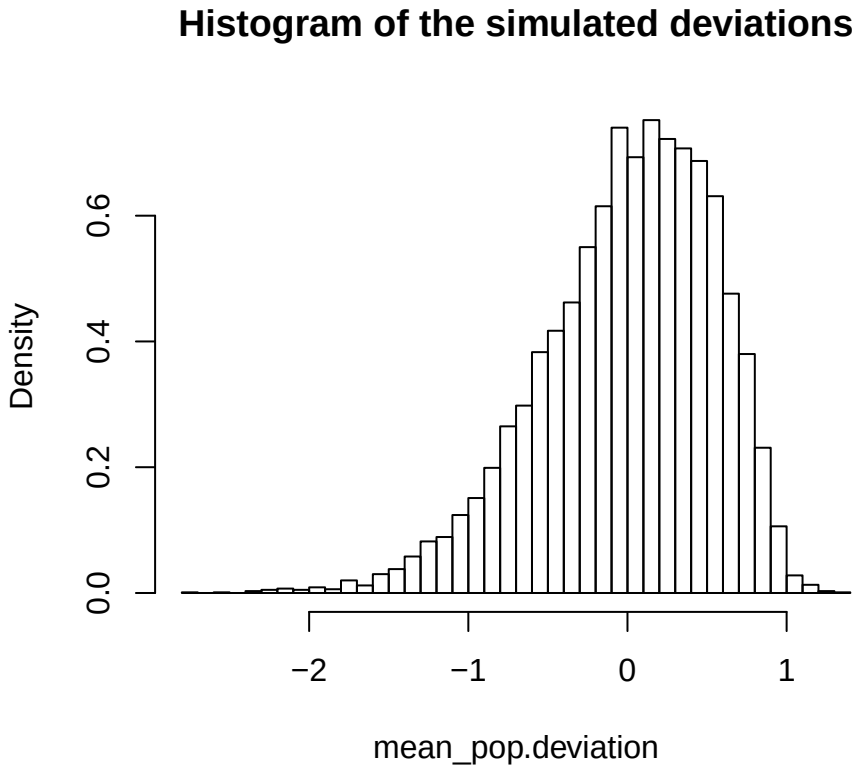
Part b

Use Bootstrap sampling to construct a 95% confidence interval for the average ear conch length of the female mountain brushtail possums in Victoria. Follow the steps that we did in lecture (see week8a_notes) and use `set.seed(5792)`. Approximate the distribution of the deviations between sample means and the population mean (i.e., sample mean - population mean) by using 10,000 bootstrap samples drawn from the original sample with replacement. Plot a histogram of the simulated deviations. Report the 95% Bootstrap confidence interval. 95% CIs is (50.90417,52.99167)

```
set.seed(5792)
# Create Bootstrap samples
mat = matrix(sample(f.vic.earconch, size = length(f.vic.earconch) *
    10000, replace = T), ncol = 10000)
# Calculate the deviations of Bootstrap samples
mean_pop.deviation = apply(mat, MAR = 2, FUN = mean) - mean(f.vic.earconch)

dev = quantile(mean_pop.deviation, prob = c(0.95 + (1 - 0.95)/2, (1 -
    0.95)/2))
# 95 % CIs
mean(f.vic.earconch) - dev
   97.5%     2.5%
50.90417 52.99167
```

```r
# histogram of simulated devations
hist(mean_pop.deviation, breaks = 30, main = "Histogram of the simulated deviations",
    freq = F)
```



Part c

Select the correct answer and explain why this is expected based on shape of the histogram in part b: The Bootstrapping CI created in Part b has

(A.) longer left arm;
(B.) longer right arm;
(C.) equal length for both left and right arms.

Histrogram skew adn arm length are opposite, from the histogram CI has longer right arm as the histogram is left skewed.

Question 2

(Hypothetical) Run the code chunk below. `age.dist` lists the proportions of New Jersey Mercer county citizens (only persons 18 and older are considered) in different age categories; e.g., the fraction of the adult citizens in Mercer county that are between 18 and 30 years old

is about .2857, and the fraction of the adult citizens that are between 30 and 40 years old is about .1020, etc. One study of grand juries in Mercer County compared the demographic characteristics of the jurors with the general population to see if the jury panels were representative of the population. The results for juror ages are shown in the vector `jurors`.

```
age.dist = c(0.2857, 0.102, 0.2245, 0.2449, 0.1429)
names(age.dist) = c("[18,30]", "(30,40]", "(40,50]", "(50,60]", "(60,90+)")
age.dist
 [18,30]  (30,40]  (40,50]  (50,60] (60,90+)
  0.2857   0.1020   0.2245   0.2449   0.1429
jurors = c(22, 52, 52, 26, 35, 60, 40, 52, 26, 58, 72, 22, 47, 47,
    31, 45, 80, 80, 18, 58, 80, 46, 52, 80, 60, 79, 58, 46, 43, 58,
    26, 60, 75, 80, 52, 45, 52, 26, 22, 22, 43, 22, 28, 48, 46, 18,
    41, 58, 58, 60, 46, 58, 46, 79, 58, 79, 20, 47, 52, 60, 66, 66,
    39, 46, 67, 37, 22, 48, 70, 54, 53, 36, 19, 43, 72, 76, 36, 68,
    38, 46, 65, 61, 53, 43, 28, 20, 59, 45, 65, 31)
```

The investigators wanted to find out if there is any evidence showing that the group of jurors might not be a simple random sample drawn from the adult citizens in Mercer county.

Part a

Test the hypothesis that the fraction of people that are 30 years old or under in the population where the jurors were drawn from is 0.2857; do this test at the $\alpha = .05$ significance level.

(i) State the H_0 and H_1 hypotheses.

H_0: The fraction of people that are 30 years old or under in the population where the jurors were drawn from is equal to 0.2857

H_1: The fraction of people that are 30 years old or under in the population where the jurors were drawn from is not equal to 0.2857

(ii) Calculate the p-value and state whether you will reject the H_0 or not. p-value = 0.02188437, will reject H_0

```
# recreate sample data and 2 sided t test
recreate = rep(c(0, 1), times = table(jurors < 31))
t.test(recreate, mu = 0.2857)$p.value
[1] 0.02188437
```

(iii) Interpret the result in terms of the p-value and α; i.e., state the meaning/definition of the p-value in the context of this problem. State the meaning of α. Also, explain how you arrived at the conclusion of whether you would reject H_0 with your p-value and α.

The p-value is the probability finding the observed results when the null hypothesis is true. In the context of this problem, the p-value is the probability that the proportion of [18,30] years olds from a sample of jurors is equal to .2857. α is at a statistically significant level and is the amount of error willing to tolerate when rejecting the null hypothesis. The p-value is less than the signifcance level α therefore reject the null hypothesis.

Part b

Based on your sample will a 95% CI for the proportion of people that are 30 years old or under in the population cover 0.2857? Construct this 95% CI to verify your answer. The 95% CI will not cover 0.2857. Verified by interval being (0.1064482, 0.2713296).

```
# t test to find 95% CI
t.test(recreate, mu = 0.2857)$conf.int
[1] 0.1064482 0.2713296
attr(,"conf.level")
[1] 0.95
```

Part c

Test the hypothesis that the ages of the people in the population (where the 90 jurors were selected from) follow the distribution shown in `age.dist`. Do this test at the $\alpha = .05$ significance level. State the H_0 and H_1 hypotheses. Report the p-value and state whether you will reject the H_0 or not. H_0: The data does follow the distributions of age.dist

H_1: The data does not follow the distributions of age.dist

p-value = 0.1309545, will not reject hypothesis

```
# observed data
ob = c(table(jurors < 31)[2], table(jurors > 30 & jurors < 41)[2],
    table(jurors > 40 & jurors < 51)[2], table(jurors > 50 & jurors <
        61)[2], table(jurors > 60)[2])
# expected data
ex = age.dist * 90
# x-squared
x2 = sum((ob - ex)^2/ex)
# p value
1 - pchisq(x2, length(ob) - 1)
[1] 0.1309545
chisq.test(ob, p = age.dist)$p.value
[1] 0.1309545
```

Part d

For the test in part c, what is the critical value and what are the rejection and acceptance regions of the test?

critical value = 9.487729 acceptance region is (-∞, 9.487729) rejection region is [9.487729,∞)

```
# find critical value
qchisq(0.95, df = 4)
[1] 9.487729
```

Part e

Between the two tests in Part a and Part c which test is more appropriate for answering the investigators' question: is there any evidence showing that the group of jurors might not be a simple random sample drawn from the adult citizens in Mercer county? Explain why briefly.

Part c as it looks at all 5 distributions of ages rather than just 1.

Question 3

An experiment is performed to see whether calculators help students do word problems. The subjects are a group of 500 thirteen-year-olds in a certain school district. All the subjects work the problem below. Half of them are chosen at random and allowed to use calculators; the others do the problem with pencil and paper. In the calculator group, 18 students get the right answer; in the pencil-and-paper group, 59 do. Can this difference be explained by chance? What do you conclude? Construct a hypothesis test at $\alpha = .01$ level to answer this question–remember to state your hypotheses. Report your test statistic and p-value and state if you are going to reject the null or not. [1]

The problem: An army bus holds 36 soldiers. If 1128 soldiers are being bussed to their training site, how many buses are needed?

Note: 1128/36 = 31.33, so 32 buses are needed. However, 31.33 was a common answer, especially in the calculator group; 31 was another common answer.

H_0 : There is no difference between those who used calculators and those who used paper and pencil.

H_1 : There is a difference between those who used calculators and those who used paper and pencil.

p-value = 3.057e-07 and test statistic = 5.2058

p-value is less than $\alpha = .01$ level therefore reject the null hypothesis.

The difference is real not by chance. Can conclude that the use of a calculator does not help students ability to solve word problems.

```
# recreate samples and t test
calc = rep(c(0, 1), times = c(250 - 18, 18))
no.calc = rep(c(0, 1), times = c(250 - 59, 59))
school = matrix(c(18, 232, 59, 191), nrow = 2, ncol = 2)
t.test(no.calc, calc, alternative = "two.sided")

    Welch Two Sample t-test

data:  no.calc and calc
t = 5.2058, df = 411.26, p-value = 3.057e-07
```

[1] problem was taken and modified based on a problem from *Statistics* by Freedman, Pisani, Purves (3rd Edition) W.W. Norton & Company.

```
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1020727 0.2259273
sample estimates:
mean of x mean of y
    0.236     0.072
```

Question 4

Use the dataset `census` in the `openintro` package. We will test whether gender is independent of the marital status based on the dataset.

```
# laod packages
library(openintro)
library(dplyr)
```

Part a

The dataset `census` records the marital status of 500 people. Create a data frame `gender.mstatus` that contains the two columns, `sex` and `maritalStatus`, of `census`. We will use `gender.mstatus` for the rest of question. With the data frame `gender.mstatus` create a table that shows the total number of people falling in each category of the marital status for the 500 people; make sure that you print this table out in your code chunk. What is the proportion of the people in the `Separated` category?

.006 of the people are in the Separated category

```
# dataframe with sex and maritalStatus
gender.mstatus = select(census, sex, maritalStatus)
# table of each maritalStatus
t = table(gender.mstatus$maritalStatus)
t

             Divorced  Married/spouse absent Married/spouse present
                   38                     14                    192
  Never married/single              Separated                Widowed
                  222                      3                     31
t["Separated"]/length(gender.mstatus$maritalStatus)
Separated
    0.006
```

Part b

Since the proportion of people in the `Separated` category is very small, this will result in low expected counts (think about how you would calculate the expected counts) for the cells

related to the `Separated` category when we do the test. We will combine the two categories `Married/spouse absent` and `Separated` this way:

(i) Replace all the `Separated` elements in `gender.mstatus$maritalStatus` with `Married/spouse absent`;

(ii) Drop the level `Separated` for `gender.mstatus$maritalStatus`;

(iii) Check that the level `Separated` was dropped by creating a 2 x 5 contingency table for the variables `sex` and `maritalStatus`. Make sure that you print this table out in your code chunk.

```
# replace separated
gender.mstatus$maritalStatus[gender.mstatus$maritalStatus == "Separated"] = "Married/spouse abse
# drop level Separated
gender.mstatus$maritalStatus = factor(gender.mstatus$maritalStatus)
# create contingency table for sex and maritalStatus
ct = table(gender.mstatus$sex, gender.mstatus$maritalStatus)
ct
```

```
        Divorced Married/spouse absent Married/spouse present
  Female       21                     6                     92
  Male         17                    11                    100

        Never married/single Widowed
  Female                   93      20
  Male                    129      11
```

Part c

Test the hypothesis that gender and marital status are independent at the significance level $\alpha = .05$. Make sure that you go through the entire procedure for testing (i.e., state your null and alternative hypotheses, report the value of your statistic and the p-value, and state whether you will reject the null hypothesis at the significance level $\alpha = .05$).

H_0 : Gender and marital status are independent

H_1 : Gender and marital status are dependent

p-value = 0.08708 and statistic (X-squared) = 8.12584

p-value of 0.08708 is greater than the significance level $\alpha = .05$ therefore we do not reject the null hypothesis.

```
# independent test
chisq.test(ct, correct = FALSE)
```

```
    Pearson's Chi-squared test

data:  ct
X-squared = 8.1258, df = 4, p-value = 0.08708
```

Question 5

We have a simple random sample of 200 people from a population and with the data on these people we would like to fit a linear model predicting the height of someone from the population based on the person's weights and gender. The model that we would like to fit is

$$Height_i = \beta_0 + \beta_{male} 1_{male,i} + \beta_{\{male.Weight\}} 1_{male,i} Weight_i + \beta_{\{female.Weight\}} 1_{female,i} Weight_i + error_i$$

where the 1's are dummy variables and the errors are independent and identically distributed with mean 0 and constant sd.

After we fit the model we have the following estimates:

- $\hat{\beta}_0 = 147$
- $\hat{\beta}_{male} = 2$
- $\hat{\beta}_{\{male.Weight\}} = .3$
- $\hat{\beta}_{\{female.Weight\}} = .37$

All weights are in kilograms and heights are in centimeters.

Part a

With the estimated coefficients above provide two equations, one for the males only and the other for the females, that one can use to calculated the estimated height for a person from this population if the weight and the gender of this person are provided.

Male Height = 149 + .3 Weight

Female Height = 147 + .37 Weight

Part b

Use your equations above to predict the height (in cm) of a man from this population who weights 86 kilograms. Similarly, predict the height of a woman from this population who weights 52 kilograms.

174.8 cm for male weighing 86 kg 166.24 cm for female weighing 52 kg

```
# male weight to height calculation
149 + 0.3 * 86
[1] 174.8
# female weight to height calculation
147 + 0.37 * 52
[1] 166.24
```

Part c

For your two equations in Part a, do they share the same y-intercept? Do they have the same slope?

No, they do not share the same y-intercept or slope.