Final Exam

Seth Metcalf

Table of Contents

Introduction

This is the final

Importing libraries.

```
library(ggplot2)
```

Due 11 May 2021.

Honor Pledge

Question 1

This exam is open book and open notes. You can look at anything on bcourses and you can use R both for calculation and to check answers for questions about R code. You can do the exam using R Markdown and knit from there if you want. You may NOT

(i) collaborate with, or seek help from, any other person.

(ii) search for answers to questions on the web from websites such as Chegg,

or any other site. Your exam should represent your own effort.

Please READ the following statement and then type your full name in the space provided. This is your electronic signature, indicating that you have read the statement and will abide by it:

By my honor, I affirm that, as as a member of the UC Berkeley community, I have acted with honesty, integrity, and respect for others; I have not shared or discussed any information about the exam or solutions with anyone; I have not looked for the answers online, and I have not violated the UC Berkeley Code for Student Conduct.

Signature (sign on your submitted exam to receive the point for this pledge, typed full name is sufficient):

Full Name

Seth Nijs Metcalf

Questions

Question 2

A box contains three red balls, three blue balls, and four green balls.

(a) Suppose seven draws are made with replacement. What is the chance that exactly two green balls are drawn?

```
dbinom(x = 2, size = 10, prob = 4/10)

## [1] 0.1209324
```

There is around a 12.1% chance that 2 green balls are drawn out of 7 draws.

(b) Write a line of code in R to find the chance in part a).

dbinom(x = 2, size = 10, prob = 4/10)

(c) In three draws without replacement, what is the chance the first two draws are the same color or the last two draws are the same color?

```
# Don't have to account for the chance that the same color is drawn for all
three draws because if that is the case then the scenario where the first two
are the same color is satisfied
red = (3/10)*(2/9) + (3/9)*(2/8) # (3/10)*(2/9) is for the chance that the
first two are red, (3/9)*(2/8) is for the last two are red, this is because
there is one less marble to worry about knowing that one was drawn
blue = (3/10)*(2/9) + (3/9)*(2/8)
green = (4/10)*(3/9) + (4/9)*(3/8)

red + blue + green

## [1] 0.6
```

Roughly a 60% chance that the first two or the last two balls in 3 draws are the same color

(d) In three draws without replacement, consider two events: A: the first draw is green B: the last two draws are the same color Are these events independent? Show relevant calculation for full credit.

```
red = (3/9)*(2/8)
blue = (3/9)*(2/8)
green = (3/9)*(2/8)

red + blue + green

## [1] 0.25
```

Because the first draw is green, then the chance of pulling greens in later draws has decreased, meaning that these two events are not independent. As we can see, the probability has gone down for the chance of pulling a green marble considering the first one is green.

Roughly a 25% chance of getting the second two draws the same color assuming the first is green.

Question 3

A stationery store wants to estimate the mean retail value of greeting cards that it has in its 250-card inventory based on a simple random sample of 30 cards. For parts a-c) of this problem, construct the confidence intervals as best you can. If you think there are any issues with the confidence intervals, explain the issue.

(a) Suppose the prices of the 30 cards in the simple random sample are in a data frame called cards in a column called price. Write code in R to make a 68% confidence interval for the average price of all 250 cards. You can assume there are no missing values, z=1 for a 68% CI, and you can use the sd() function in R even though it doesn't calculate the SD in the same way as the text. The last line of your R code should return the left end of the CI and the right end of the CI as a vector of length 2.

draws = 30 mean = mean(cards$price)se = sqrt(draws) * sd(cards$price) seper = se/sample

Upper_Bound = mean + seper Lower_Bound = mean - seper x = c(Lower_Bound, Upper_bound) x

(b) Suppose that on 3 of the cards in the sample, the first word on the card is the word "happy." Construct a 68% confidence interval for the percent of times the first word is the word "happy," for all 250 cards.

draws = 30 p = 0.10 sd = (1 - 0) * sqrt((p) * (1 - p)) se = sqrt(draws) * sd seper = se/sample

Upper_Bound = mean + seper Lower_Bound = mean - seper x = c(Lower_Bound, Upper_bound) x

(c) Suppose that all the data on prices is converted to standard units, resulting in a new list of numbers. (Note that you are not given enough information to actually find each value, but suppose the new list of numbers converted into standard units exists.) If possible, give the average and the SD of this new list of numbers and either explain briefly or show a calculation. If you need more information or you don't think it's possible, explain briefly.

If it has been converted to Standard Units, then the mean of the sample is 0 (because you subtract the mean from each number) while the standard deviation is the same as it was before (We can not find this exact value with the answer given).

Question 4

A statistics class has a midterm and final exam, both graded out of 100 points. The summary statistics on the exams are:

midterm: average = 70, SD = 12,

final: average = 65, SD = 16.

The scatter diagram is football-shaped and r =0.5

(a) (3 points) Find the equation for the regression line predicting final score from midterm score.

In order to find the regression line, we can use the equation (r * SD_y)/SD_x in order to find the slope. Knowing that if you have the average of x, you should get the average of y, we can plug in the avg of both values into y = mx + b (after calculating the slope) in order to get the y-intercept (b).

```
slope = (0.5 * 16)/12
slope

## [1] 0.6666667

b = 65 - slope*70
```

0.6667 is the slope of the regression line while 18.3333 is the y-intercept. Plugging it into the equation gives us the regression line of:

y = (0.6667) * x + 18.3333

(b) (5 points) Of all the people who were at the average on the midterm, about what percent got over the 80th percentile on the final?

```
qnorm(0.80, mean = 65, sd = 16)

## [1] 78.46594

# Score of 78.46594 is the 80th percentile for the final, roughly 0.84162125
SDs
```

(c) (3 points) Consider the following 3 people: Alice got 58 on the midterm and 59 on the final. Billy got 64 on the midterm and 71 on the final. Cathy got 76 on the midterm and 75 on the final. For each student, explain briefly why the student is or is not a good example of the regression effect. No calculation is required

Alice is not far off from being a good fit, has similar predicted score for the final given the midterm score. Billy is not a good fit at all. Given the correlation and the y-int, he scored far above what would have been predicted for him. Cathy is in between, she scored higher than what she would have been predicted, but it is not as far off as Billy;

Overall, Alice is the best fit according to the regression line.

(d) (5 points) Suppose there is a data frame in R called mydata with two columns named midterm and final. Write code in R to make a scatter plot with midterm on the x axis and final on the y axis. Include the regression line (without any shaded area around it). Also include a title, ”Midterm and Final Scores”.

midterm = mydata$midterm final = mydata$final

```
ggplot(mydata,aes(x=midterm, y=final)) + geom_point(alpha=0.7,color=“red”) +
geom_smooth(method=“lm”,se=F) + ggtitle(label = “Midterm and Final Scores”)
```

(e) (2 points) To do part (a) above, is it necessary (or at least better) for the scatter plot to be football-shaped? If so, explain why ONE property of being football-shaped is necessary or useful for the method to be meaningful. There may be more than one possible correct answer. If it’s not necessary or useful to be football-shaped, explain why not.

It is beneficial for it to be football-shaped as this allows the graph to be homoscedastic which limits the variance among the data points and allows the data to not only have better correlation, but a more accurate regression line; allows us to predict values (what score you got on the final given the score on the midterm) better.

(f) (2 points) Repeat the previous question, but for what is relevant for the answer to part (b). Do not use the same property as you used in the previous problem

It has to be football-shaped in order to use the rms formula in order to calculate the difference between the predicted value and the value that was observed.

Question 5

A scientist takes measurements of the weights of polar bears. The average weight of polar bears is believed to be 320 pounds, and the researcher is interested in whether weights are different on one island. Polar bear weights in general follow the normal curve fairly closely. A simple random sample of 5 polar bears on this island has the following weights: 301, 315, 298, 316, and 320.

(a) Do an appropriate hypothesis test to decide whether the polar bears on this island have an average weight which is lower than 320 pounds, or if this can be explained by chance.

Given the sample, we can put it into a vector called polarbears. Taking the mean of this vector will give us the mean of the weight of the polarbears, we subtract the EV from this (given to us by the problem of 320). We divide that by the SEavg by calculating the SD of the polarbears (using the SDplus formula) and dividing it by the total number of the sample to get the SE per polarbear. Finally, plugging the t-stat that was calculated into the pt() formula with 4 degrees of freedom (because there are 5 draws).

```
polarbears = c(301, 315, 298, 316, 320)
samplemean = mean(polarbears)
sd = sd(polarbears) * sqrt(5/(5-1))
se = sqrt(5) * sd
seavg = se/5
t=(samplemean-320)/seavg
pt(t, df = 4)

## [1] 0.05572807
```

There is about a 5.57% chance that the polar bears on the island have a mean weight of 320 given the data that was sampled. This means that we can not reject the null hypothesis, and this is most likely due to chance. A recommendation would be to increase the sample size of the problem as the low sample size could skew the calculated p-value.

(b) Suppose the data is in a data frame called data with a single variable called weight. Write R code to get the test statistic and also to get the p-value based on that test statistic.

Already used give or take the same code for the problem before just replaced the created vector with the dataframe column

samplemean = mean(data$weight)sd = sd(data$weight) * sqrt(5/(5-1)) se = sqrt(5) * sd seavg = se/5 t=(samplemean-320)/seavg pt(t, df = 5)

Question 6

(This problem is loosely based on data collected in lecture early this semester.) A simple random sample of 70 Stat 20 students is taken from a class of 350 students; the average number of hours spent on social media is 14, with an SD of 10. Last summer I taught a smaller class with only 100 students, and those 100 students reported an average number of hours on social media of 15, with an SD of 9.

Do a hypothesis test to decide whether the average number of hours on social media is different than the average for my class last summer

Have to calculate whether or not the total time spent on social media is the same. null is that the average number of hours spent in class 1 = the average number of hours spent in class 2. alternate is that the average number of hours spent in class 1 != the average number of hours spent in class 2. Therefore this is a two-tailed z-test.

First calculate the SE of both classes in order to calculate the difference and eventually the z value by subtracting the avgs from one another before dividing that by the SE of both.

```
class1_avg = 14
class1_size = 350
class1_SD = 10
class1_SE_sum = sqrt(class1_size) * class1_SD
class1_SE_avg = (class1_SE_sum/class1_size)

class2_avg = 15
class2_size = 100
class2_SD = 9
class2_SE_sum = sqrt(class2_size) * class2_SD
class2_SE_avg = (class2_SE_sum/class2_size)

class1SEsq = (class1_SE_avg * class1_SE_avg)
class2SEsq = (class2_SE_avg * class2_SE_avg)
SE_class1_class2_diff = sqrt(class1SEsq + class2SEsq)
classDiff = (class1_avg - class2_avg)
classZ = (classDiff - 0)/SE_class1_class2_diff

2 * pnorm(classZ) * 100
```

```
## [1] 33.94132
```

Given the two class samples. There is about a 34% that they have the same number of hours spent per week on social media given the data. This causes us to fail to reject the null hypothesis as this is most likely due to chance. The average hours spent on social media is most likely the same between the two classes.

Question 7

Some parts of this question have been simplified slightly to make it a reasonable exam question, but the basic results are accurate. CNN sponsored a public opinion poll April 21-26, 2021. There were 1000 US adults in the sample, and they asked them if they were college graduates. They also asked them “What effect do you think it would have on whether elections in the U.S. were conducted fairly if voters were always allowed to vote before Election Day including outside of normal business hours and on weekends?” Of the 1000 respondents, 40% said they were college graduates and 60% said they were not. Of the college grads, 75% said more fair, 7% said less fair, and 18% said about the same. For the non-college grads, about 60% said more fair, 11% said less fair, and 27% said about the same. You can assume the data come from a simple random sample with no bias of any kind.

Do a hypothesis test (using the full outline form) to see if college education and opinions on early voting (based on the answer to this question) are independent.

1)

1. null

a) Due to chance

b) There is an equal opinion between college-graduated educated adults and non-college graduated adults

2. alternative

a) Not due to chance

b) There is a *different* opinion between college-graduated educated adults and non-college graduated adults

(for z and t, should specify/imply one-sided vs two-sided)

3. test statistic (chi)

```
obs=c(300,28,72,360,66,162)
exp=c(400*c(660/1000,94/1000,234/1000), 600*c(660/1000,94/1000,234/1000))

chisq=sum((obs-exp)^2/exp)
pchisq(chisq,df=999)

## [1] 0
```

4. p-value = 0

5. conclusion

We reject the null hypothesis. Given the p-value, this is absolutely not due to chance. College educated pollers have a different opinion then those that are not college-educated.