STAT20 Homework #7

Seth Metcalf

Table of Contents

Introduction

This is Homework #7, which contains questions from Chapters 21, 23, and 26, and also contain an R problem. *Due 2 April 2021.*

Chapter 21

1. Ch 21 C3:

Refer to exercises 3 and 4 above (Ch 21 A7 and A8). For each one of the three people described in the last exercise, compute a 95%-confidence interval for the percentage of 1's in the box. Which of the three intervals cover the population percentage, that is, the percentage of 1's in the box described in exercise 3 (Ch 21 A7)? Which do not? (Remember, the three people who take the samples do not know the contents of the box; but you do.)

Explanation

P = 72/400 = 0.18 SE% = sqrt((P * (1-P))/n) * 100% = sqrt((0.18 * 0.82)/400) * 100% = 0.0192 * 100% = 1.92% 18 + 2(1.92) = 21.84 18 - 2(1.92) = 14.16 A 95% confidence interval for this percentage of 1's is between

P = 84/400 = 0.21 SE% = sqrt((P * (1-P))/n) * 100% = sqrt((0.21 * 0.79)/400) * 100% = 0.0204 * 100% = 2.04% 21 + 2(2.04) = 25.08 21 - 2(2.04) = 16.92 The first person estimates the percentage of 1 's in the box as 21% , and figures this estimate is likely to be off by 2.04% or so.

P = 98/400 = 0.245 SE% = sqrt((P * (1-P))/n) * 100% = sqrt((0.245 * 0.755)/400) * 100% = 0.0215 * 100% = 2.15% 24.5 + 2(2.15) = 28.8 24.5 - 2(2.15) = 20.2 The first person estimates the percentage of 1 's in the box as 24.5% , and figures this estimate is likely to be off by 2.15% or so.

The first two contain the percentage that was listed in A7 (20%) but the third one does not.

2. Ch 21 Rev 2:

The Residential Energy Consumption Survey found in 2001 that 47% of American households had internet access. A market survey organization repeated this study in a certain town with 25,000 households, using a simple random sample of 500 households: 239 of the sample households had internet access.

a) The percentage of households in the town with internet access is estimated as ___ ; this estimate is likely to be off by ___ or so.

Sample% = 239/500 = 47.8% SD = (1 - 0) * sqrt(0.478 * 0.522) = 0.4995 SE = sqrt(500) * 0.4995 = 11.1692 SE% = 11.1692/500 * 100% = 0.022 * 100% = 2.2%

The percentage of households in the town with internet access is estimated as *47.8%*; this estimate is likely to be off by *2.2%* or so.

b) If possible, find a 95%-confidence interval for the percentage of all 25,000 households with internet access. If this is not possible, explain why not.

Sample% +- 2 SE's gives the 95% confidence interval. 47.8% + 2(2.2%) = 52.2% 47.8% - 2(2.2%) = 43.4% We are 95% confident that the percentage of the 25,000 households with internet access is between the percents of 43.4% and 52.5%.

Chapter 23

3. Ch 23 B6:

A university has 30,000 registered students. As part of a survey, 900 of these students are chosen at random. The average age of the sample students turns out to be 22.3 years. and the SD is 4.5 years.

a) The average age of all 30,000 students is estimated as ___ . This estimate is likely to be off by ___ or so.

avgAge = 22.3 yrs SEsum = sqrt(900) * 4.5 = 30 * 4.5 = 135 SEavg = 135/900 = 0.15 The average age of all 30,000 students is estimated as *22.3 years old*. This estimate is likely to be off by *0.15 years* or so.

b) Find a 95%-confidence interval for the average age of all 30,000 registered students.

Sample% +- 2 SE's gives the 95% confidence interval. 22.3 + 2(0.15) = 22.6 years old 22.3 - 2(0.15) = 22 years old We are 95% confident that the average age of the 30,0000 students between the ages of 22 years old and 22.6 years old.

4. Based on Ch 23 Rev 3,4:

A real estate office wants to make a survey in a certain town, which has 50,000 households, to determine how far the head of household has to commute to work. A simple random sample of 1,000 households is chosen, the occupants are interviewed, and it is found that on average, the heads of the sample households commuted 8.7 miles to work; the SD of the distances was 9.0 miles. (All distances are one-way; if someone isn't working, the commute distance is defined to be 0.)

a) The average commute distance of all 50,000 heads of households in the town is estimated as ___ , and this estimate is likely to be off by ___ or so.

avgDist = 8.7 miles SEsum = sqrt(1000) * 9.0 = 284.6 miles SEavg = 284.61/1000 = .2846 miles The average commute distance of all 50,000 heads of households in the town is estimated as *8.7 miles*, and this estimate is likely to be off by *.2846 miles* or so.

b) If possible, find a 95%-confidence interval for the average commute distance of all heads of households in the town. If this isn't possible, explain why not.

Sample% +- 2 SE's gives the 95% confidence interval. 8.7 + 2(.2846) = 9.27 years old 8.7 - 2(.2846) = 8.13 years old We are 95% confident that the average commute distance of the 50,000 households is between 8.13 and 9.27 miles.

c) The real estate office interviewed all persons age 16 and over in the sample households; there were 2,500 such persons. On the average, these 2,500 people commuted 7.1 miles to work, and the SD of the distances was 10.2 miles. (Again, if someone isn't working, the commute distance is defined to be 0; and all distances are one-way.) If possible, find a 95%-confidence interval for the average commute distance for all people age 16 and over in this town. If this isn't possible, explain why not.

avgDist = 7.1 miles SDsum = 10.2 miles SEsum = sqrt(2500) * 10.2 = 510 SEavg = 510/2500 = 0.204 miles 7.1 + 2(0.204) = 7.508 7.1 - 2(0.204) = 6.896 We are 95% confident that the average commute distance of the 2,500 people age 16 or over is between 6.896 and 7.508 miles.

5. Ch 23 D6:

One hundred draws are made at random with replacement from a box. The sum of the draws is 297. Can you estimate the average of the box? Can you attach a standard error to your estimate, on the basis of the information given so far? Explain briefly.

Explanation

We can estimate the average of the box knowing that the sum of draws is 297 with 100 draws; the estimation would be 297/100 or 2.97 as the average. However, we can not attach a standard error because of the fact that there is no way to get the SD of the box with the information that is given – a vital piece in finding the SE.

6. Ch 23 Rev 10:

A survey organization takes a simple random sample of 625 households from a city of 80,000 households. On the average, there are 2.30 persons per sample household, and the SD is 1. 75. Say whether each of the following statements is true or false, and explain.

a) The SE for the sample average is 0.07.

EVavg = 2.30 SEsum = sqrt(625) * 1.75 = 43.75 SEavg = 43.75/625 = 0.07 True, the SE for the sample average is 0.07 according to the data we are given.

b) A 95%-confidence interval for the average household size in the sample is 2.16 to 2.44.

False, the confidence interval refers to the population average and not the sample average.

c) A 95%-confidence interval for the average household size in the city is 2.16 to 2.44.

EVavg +- 2(SEavg) 2.30 + 2(0.07) = 2.44 2.30 - 2(0.07) = 2.16 True, according to calculations from the data that we are given the is the correct confidence interval for the population.

d) 95% of the households in the city contain between 2.16 and 2.44 persons.

False, this statement is about individual household size, not the average household size.

e) The 95%-confidence level is about right because household size follows the normal curve.

False. If it follows the normal curve if around all data is within 3 SDs of the mean. Since below 0 is impossible to occur in real life and 2 SDs away from the mean is already below 0, we know that the household size does not follow the normal curve.

f) The 95%-confidence level is about right because, with 625 draws from the box, the probability histogram for the average of the draws follows the normal curve.

True, since the more draws that occur, the more that the the probability distriction becomes normal and because of this our 95% confidence interval is right.

Chapter 26

7. Ch 26 C7, C8:

Many companies are experimenting with "flex-time," allowing employees to choose their schedules within broad limits set by management. Among other things, flex-time is supposed to reduce absenteeism. One firm knows that in the past few years, employees have averaged 6.3 days off from work (apart from vacations). This year, the firm introduces flex-time. Management chooses a simple random sample of 100 employees to follow in detail, and at the end of the year, these employees average 5.5 days off from work, and the SD is 2.9 days. Did absenteeism really go down, or is this just chance variation?

a) Do a full hypothesis test as done in lecture (null, alternative, z statistic, p-value, and conclusion).

mean = 6.3 SDsum = 2.9 SEsum = sqrt(100) * 2.9 = 29 SEavg = 29/100 = 0.29

null = The average of the box is equal to the average that is given to us: 6.3 days. alt = The average of the box is less than the average that is given to us: 6.3 days. zscore = (5.5 - 6.3)/0.29 = -2.75

```
#P Value
pnorm(-2.75) * 100
```

```
## [1] 0.2979763
```

Conclusion: The P-value is less than or equal to 1% meaning that we reject the null hypothesis at the 1% level. It doesn't appear to be due to chance. This shows that the absenteeism day average appears to have gone down.

b) Repeat for a sample average of 5.9 days and an SD of 2.9 days. You can skip doing the null and alternative since they are the same, but redo the z, p, and conclusion.

mean = 5.9 SDsum = 2.9 SEsum = sqrt(100) * 2.9 = 29 SEavg = 29/100 = 0.29

zscore = (5.5 - 5.9)/0.29 = -1.38

```
#P Value
pnorm(-1.38) * 100
```

```
## [1] 8.379332
```

Conclusion: The P-value is more than 5%. This means that we can not reject the null due to it being able to occur due to chance. This shows that absenteeism average probably didn't change it was just a rare sample average.

8. Ch 26 E10:

A colony of laboratory mice consisted of several hundred animals. Their average weight was about 30 grams, and the SD was about 5 grams. As part of an experiment, graduate students were instructed to choose 25 animals haphazardly, without any definite method.

The average weight of these animals turned out to be around 33 grams, and the SD was about 7 grams. Is choosing animals haphazardly the same as drawing them at random? Or is 33 grams too far above average for that? Formulate the null hypothesis as a box model; compute z and P. (There is no need to formulate an alternative hypothesis about the box; you must decide whether the null hypothesis tells you the SD of the box: if not, you have to estimate the SD from the data.)

Hypothesis testing

avgSample = 33 sdBox = 7 avgBox = 30 draws# = 25

null = That the 33 grams is due to chance and the average of the box still is 30. alt = That the average of the box has gone up since last time that it was recorded and is no longer 30 grams.

SEsum = sqrt(25) * 7 = 35 SEavg = 35/25 = 1 mean = 30 x = 33 zscore = (33-30)/1.4 = 2.14285714

```
#P Value
(1 - pnorm(2.14285714)) * 100
```

```
## [1] 1.606229
```

Conclusion: Since the p-value is between the percentage between 1% and 5%, it is statistically significant and we can reject the null hypothesis at the 5% level. It doesn't seem like chance, but it could be. It might mean that the average has gone up since it was last recorded, but it could just be to chance.

9. Ch 26 Rev 2:

With a perfectly balanced roulette wheel, in the long run, red numbers should turn up 18 times in 38. To test its wheel, one casino records the results of 3,800 plays, finding 1,890 red numbers. Is that too many reds? Or chance variation? Do a full hypothesis test including null, alternative, z, p, and conclusion.

Hypothesis testing

draws# = 3800 observed = 1890

null = That the roulette is still balanced and that it is still an 18/38 chance that red shows up. alt = That the roulette wheel is no longer balanced and that it has become a higher chance to get a red number on the roulette wheel.

EVsum = 1800 SD = (1 - 0) * sqrt((18/38) * (20/38)) = 0.4993 SEsum = sqrt(3800) * 0.4993 = 30.7789 x = 1890 zscore = (1890-1800)/30.7789 = 2.90

```
#P Value
(1 - pnorm(2.90)) * 100
```

```
## [1] 0.1865813
```

Conclusion: The null hypothesis gets rejected at the 1% level. This is highly significant and probably means that the roulette table is no longer balanced as it should be. It is unlikely due to chance and the roulette table should be fixed.

R Question

10. R Problem:

R problem, continuing previous problem Ch 26 Rev 2. Do a simulation of 3800 plays of a fair roulette wheel. You can use the lecture code from 3/18 (card guessing example) with minor changes. Treat this as a box with 0's and 1's, counting the number of reds in 3800 plays. Do this simulation 100,000 times. For a), b), and c), your code should be written so that it produces the answers in your Markdown output file.

a) What proportion of the times is the number of reds 1890 or more? Note that there isn't a "right" answer to this since it's based on your simulation.

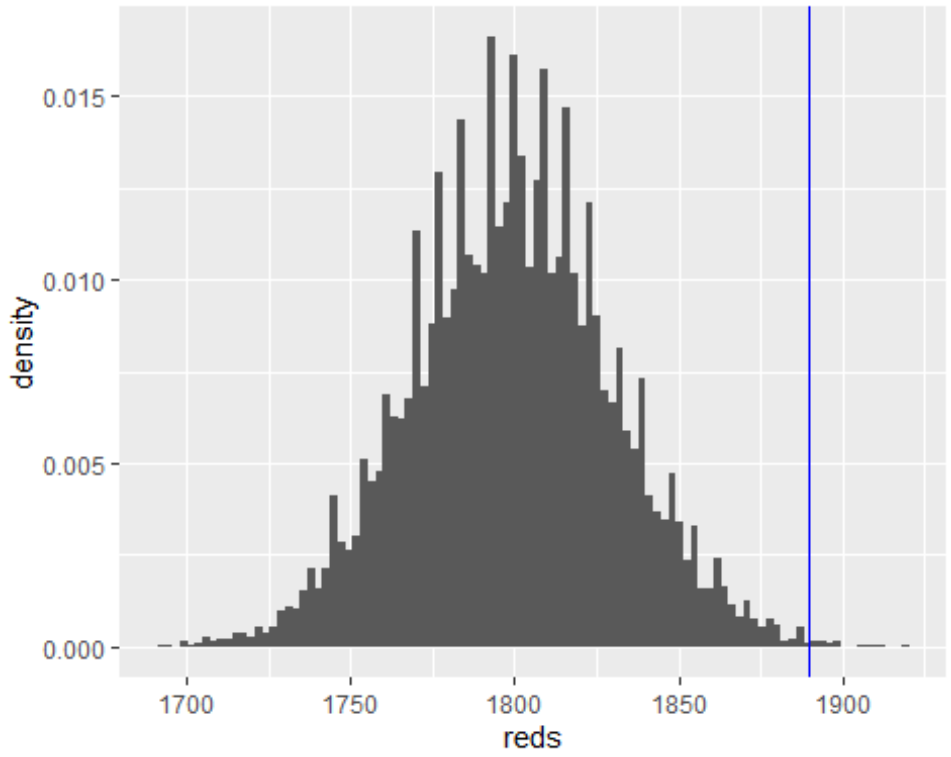
```
library(ggplot2)

B = 10000
reds = replicate(B,sum(sample(0:1,3800,repl=T,prob=c(0.526315789,0.473684211)
)))
mean(reds >= 1890)

## [1] 0.0017
```

b) Make a histogram of the number of reds, with a vertical line through 1890. (Again, see lecture code, make minor changes.)

```
reddata=data.frame(reds)
ggplot(reddata,aes(x=reds,y=..density..)) + geom_histogram(bins = 100) + geom
_vline(xintercept = 1890, col="blue")
```



c) Get the p-value according to the binomial distribution (use the pbinom function).

```
(1 - pbinom(1889, 3800, 18/38))
```

```
## [1] 0.001829548
```

d) Explain the differences between these three methods (normal approx, simulation, and binomial). Recall that the p-value is the chance of getting data like we got (1890 reds), or more extreme, given that the null (fair roulette wheel) is true. Which of the 3 methods gives us the most accurate value for the p-value (the chance of getting data like we got, or more extreme, given that the null is true)?

Through normal approximation, we use the z-score value that is found and use the normal distribution to estimate the probability that the number of observations occur. Simulation is different every time as it relies on doing the draws itself – which is not consistent. Lastly, binomial is completely based off the probability that a certain number of observations occurs given a probability.`