Homework 04

Christina Lee

10/1/2021

1.

a. You get back your exam from problem 3.d of Homework 3, and you got a 45. What is your z score?

To answer this, plug what we know into the equation for z score. x is our score 45, with a mean of 70, and a standard deviation of 10. This gives us a z score of -2.5

$$z = \frac{x - \mu}{\sigma}$$

$$z = \frac{45 - 70}{10} = -2.5$$

b. What percentile are you?

To answer this, I used the pnorm() function to calculate the percentile where the first argument is the score, 70 is the mean and 10 is the SD.

```
pnorm(45,70,10)
```

```
## [1] 0.006209665
```

To verify, we can use the qnorm() function to see if the percentile we got corresponds to the score we got, which gives 44.87 and is close to our 45.

```
qnorm(0.006,70,10)
```

```
## [1] 44.87856
```

c. What is the total chance of getting something at least that far from the mean, in either direction? (Ie, the chance of getting 45 or below or equally far or farther above the mean.)

To answer this, I multiplied 2 (since we are asking for either directions) by pnorm() where 45 is the score, 70 is the mean and 10 is the SD.

```
totalchance <- 2*pnorm(45,70,10)
totalchance
```

```
## [1] 0.01241933
```

2.

a. Write a script that creates a vector of 10,000 integers generated from a poisson distribution with lambda = 10, and then draw a sample of 9 integers from that vector of 10,000 integers.

To answer this, I used the set.seed() to make sure the output of the random sample generator will stay the same in order for me to perform calculations for the next couple questions. I then used rpois() to create a vector with 10000 integers with a lambda of 10 and saved it in a variable called "randpois". Next, I used the sample() to draw a sample of 9 integers from "randpois" and saved the result in another variable called "newpois" in order for me to output the result.

```r
set.seed(1)
randpois <- rpois(10000,10)
newpois <-sample(randpois,9,replace=TRUE)
newpois
```

```
## [1] 17  8  9 13 12 10 10 10 10
```

b. Calculate by hand your sample's mean. Please show your work using proper mathematical notation using latex. After doing it by hand, verify your result with R.

$$17 + 8 + 9 + 13 + 12 + 10 + 10 + 10 + 10 = 99$$

$$99/9 = 11$$

$$\bar{x} = 11$$

```r
mean(newpois)
```

```
## [1] 11
```

c. Calculate by hand the sample standard deviation. Verify your result with R.

$$(17 - 11)^2 = 36$$

$$(8 - 11)^2 = 9$$

$$(9 - 11)^2 = 4$$

$$(13 - 11)^2 = 4$$

$$(12 - 11)^2 = 1$$

$$(10 - 11)^2 = 1$$

$$(10 - 11)^2 = 1$$

$$(10 - 11)^2 = 1$$

$$(10 - 11)^2 = 1$$

$$36 + 9 + 4 + 4 + 1 + 1 + 1 + 1 + 1 = 58$$

$$(\frac{1}{9}) * 58 = 6.\overline{4}$$

$$Sd = \sqrt{6.\overline{4}}$$

$$Sd = 2.54$$

```r
sd(newpois)
```

```
## [1] 2.692582
```

d. Calculate by hand the standard error. Verify your result with R.

$$SE = \frac{Sd}{\sqrt{n}}$$

$$SE = \frac{2.54}{\sqrt{9}} = 0.85$$

```
nsample <- 9
standard_error <- sd(newpois)/sqrt(nsample)
standard_error
```

```
## [1] 0.8975275
```

e. Calculate by hand the 95% CI using the normal (z) distribution. (You can use R or tables to get the score.)

$$P(\bar{x} - 1.96se \leq \mu \leq \bar{x} + 1.96se) = 0.95$$
$$\bar{x} - 1.96se = (11 - 1.96 * 0.85) = 9.334$$
$$\bar{x} + 1.96se = (11 + 1.96 * 0.85) = 12.\overline{6}$$

```
mean(newpois) - 1.96*standard_error # lower 95% CI
```

```
## [1] 9.240846
```

```
mean(newpois) + 1.96*standard_error # upper 95% CI
```

```
## [1] 12.75915
```

f. Calculate by hand the 95% CI using the t distribution. (You can use R or tables to get the score.)

$$P(\bar{x} - 2.306se \leq \mu \leq \bar{x} + 2.306se) = 0.95$$
$$\bar{x} - 2.306se = (11 - 2.306 * 0.85) = 9.04$$
$$\bar{x} + 2.306se = (11 + 2.306 * 0.85) = 12.96$$

```
mean(newpois) - 2.306*standard_error # lower 95% CI
```

```
## [1] 8.930302
```

```
mean(newpois) + 2.306*standard_error # upper 95% CI
```

```
## [1] 13.0697
```

3.

a. Explain why 2.e is incorrect.

2e is incorrect because the sample size "n", which in our case n= 9 is < 30. This means our sample mean is not normally distributed around the true mean miu. To deal with sample sizes that are smaller than 30, we need to use the T distribution and instead of using the z-table, we need to use the t-table –should not use 1.96 from the z-table to calculate the 95% CI, but need to figure out the degrees of freedom (n-1) and our t-score to calculate the 95% CI which is 2.306.

b. In a sentence or two each, explain what's wrong with each of the wrong answers in Module 4.4, "Calculating percentiles and scores," and suggest what error in thinking might have led someone to choose that answer. (http://www.nickbeauchamp.com/comp_stats_NB/compstats_04-04.html)

This is the wrong answer, first of all our standard error is not 2, it is 1. Second, our t score is not 1.533, it is 2.353 if we look at the the t table and look for 3 degrees of freedom and t0.05.

$$3 \pm 2 * 1.533$$

This is the wrong answer, because as mentioned above, our t score is not 1.533, it is 2.353.

$$3 \pm 1 * 1.533$$

This is the wrong answer, because our standard error is not 2 and our t score is not 1.638.

$$3 \pm 2 * 1.638$$

This is the CORRECT answer, because our sample mean is 3, with a standard error of 1 (se= sd/sqrt of n), and our t score is 2.353.

$$3 \pm 1 * 2.353$$

This is the wrong answer, because our t score is not 2.132.

$$3 \pm 1 * 2.132$$

4.

a. Based on 2, calculate how many more individuals you would have to sample from your population to shrink your 95% CI by 1/2 (ie, reduce the interval to half the size). Please show your work.

Assume the changes in the critical t value doesn't change much when nsample and degrees of freedom is increased –we can then focus on the sample size and denote everything else as "CI", and assume the width of the CI can be expressed as this equation below:

$$\frac{CI}{\sqrt{n}}$$

Now, assume the width of our current interval is:

$$Width1 = \frac{CI}{\sqrt{9}}$$

Since we want to half our CI, we are going to call this new sample size "n2" and stick it back to the equation and call it "Width2".

$$Width2 = \frac{CI}{\sqrt{n2}}$$

Again, we want Width2 to be half of Width1 so,

$$Width2 = \frac{1}{2} Width1$$

We can now plug our equation for Width1 and Width2 in, which gives:

$$\frac{CI}{\sqrt{n2}} = \frac{1}{2}\frac{CI}{\sqrt{9}}$$

Finally, we solve for n2 since it is what we are essentially looking for:

$$\frac{1}{\sqrt{n2}} = \frac{1}{2\sqrt{9}}$$

$$\frac{2}{\sqrt{n2}} = \frac{1}{\sqrt{9}}$$

$$(2\sqrt{9})^2 = (\sqrt{n2})^2$$

$$4 * 9 = n2$$

$$n2 = 36$$

Therefore, in order to "half" the interval the sample size has to be 36, which means you need to multiply 4 by the original sample size 9.

b. Say you want to know the average income in the US. Previous studies have suggested that the standard deviation of your sample will be $20,000. How many people do you need to survey to get a 95% cofidence interval of ± $1,000? How many people do you need to survey to get a 95% CI of ± $100?

From the problem we know:
width of interval is ± 1000 = 2000
1.96 x 2 = 3.92 = 4

Now, solve for standard error (se):

$$2000 = 4se$$

$$se = 500$$

Once we know the se, plug it in the equation to solve for n.

$$500 = \frac{Sd}{\sqrt{n}}$$

$$500 = \frac{20000}{\sqrt{n}}$$

$$(1)^2 = (\frac{40}{\sqrt{n}})^2$$

$$1 = \frac{1600}{n}$$

$$n = 1600$$

How many people do you need to survey to get a 95% CI of ± $100?

$$200 = 4se$$

$$50 = \frac{20000}{\sqrt{n}}$$

$$(1)^2 = (\frac{400}{\sqrt{n}})^2$$

$$1 = \frac{160000}{n}$$

$$n = 160000$$

5.

a. Write a script to test the accuracy of the confidence interval calculation as in Module 4.3. But with a few differences: (1) Test the 99% CI, not the 95% CI. (2) Each sample should be only 20 individuals, which means you need to use the t distribution to calculate your 99% CI. (3) Run 1000 complete samples rather than 100. (4) Your population distribution must be something other than a bimodal normal distribution (as used in the lesson), although anything else is fine, including any of the other continuous distributions we've discussed so far. Explain your result and how it validates the theoretical calculation of the 99% CI.

```
nruns <- 1000
# 2. Set how many samples to take in each run
nsamples <- 20
# 3. Create an empty matrix to hold our summary data: the mean and the upper and lower CI bounds.
sample_summary <- matrix(NA,nruns,3)
# 4. Run the loop
for(j in 1:nruns){
  sampler <- rep(NA,nsamples)
  # 5. Our sampling loop
```

```
  for(i in 1:nsamples){
    # Draw from a poisson distribution with 1 observation and a mean of 15.
      sampler[i] <- rpois(1,15)
  }
  # 7. Finally, calculate the mean and 99% CI's for each sample
  #    and save it in the correct row of our sample_summary matrix
  sample_summary[j,1] <- mean(sampler)  # mean
  standard_error <- sd(sampler)/sqrt(nsamples) # standard error
  sample_summary[j,2] <- mean(sampler) - 2.861*standard_error # lower 99% CI bound
  sample_summary[j,3] <- mean(sampler) + 2.861*standard_error # upper 99% CI bound

}
```

```
counter = 0
for(j in 1:nruns){
  # If 15 is above the lower CI bound and below the upper CI bound:
  if(15 > sample_summary[j,2] && 15 < sample_summary[j,3]){
    counter <- counter + 1
  }
}
counter/nruns
```

```
## [1] 0.991
```

This tells us that the sampler allows us to calculate 99% CI's that are right about 99% of the time –we are 99% confident that the true mean lies somewhere between our CI range even though we have a small sample size.