INSH5301 Intro Computational Statistics

Ali Banijamali

02/03/2020

1.a. You get back your exam grade, and you got a 45. The class mean is 70 and the standard deviation 10. Assuming that the exam grades follow a normal distribution. What is your z score?

ANS.
Z = $\frac{45-70}{10}$ = -2.5

1.b. What percentile are you?

```
pnorm(45, mean=70, sd=10)
```

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

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

```
2*pnorm(45, mean=70, sd=10)
```

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

2.a. Write a script that generates a population of at least 10,000 numbers and samples at random 9 of them.

```
set.seed(1) # So that the following answer doesn't change on every run
pop <- rnorm(10000, mean=0, sd=10)
rand.sample <- sample(pop, size=9, replace=F)
rand.sample
```

```
## [1] -21.386093   3.926101  -9.772168  18.529964  -3.512498  -3.004396
## [7]  19.884856   2.008287  -8.468195
```

```
mean(rand.sample)
```

```
## [1] -0.1993492
```

```
sd(rand.sample)
```

```
## [1] 13.24666
```

2.b. Calculate by hand the sample mean. Please show your work using proper mathematical notation using latex.

ANS.
Sample Mean = $\frac{Sum(rand.sample)}{9}=\frac{(-21.386093+3.926101-9.772168+18.529964-3.512498-3.004396+19.884856+2.008287-8.468195)}{9}$=-
0.1993491

2.c. Calculate by hand the sample standard deviation.

ANS.
Sample Standard Deviation $= \sqrt{\frac{1}{9-1}}[(-21.386093-(-0.1993491))^2+(3.926101-(-0.1993491))^2+(-9.772168-(-0.1993491))^2+(18.529964-(-0.1993491))^2+(-3.512498-(-0.1993491))^2+(-3.004396-(-0.1993491))^2+(19.884856-(-0.1993491))^2+(2.008287-(-0.1993491))^2+(-8.468195-(-0.1993491))^2] = 13.24666$

Note: All of the above equation is under the square root! By extending the square root, the equation was getting out of the boundaris of the page, so I had to remove the top bar.

2.d. Calculate by hand the standard error.

ANS.
Sample Standard Error $= \frac{StandardDeviation}{\sqrt{n}} = \frac{13.24666}{\sqrt{9}} = 4.415553$

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

ANS.
The %95 Confidence Interval is: P$(\bar{x} - 1.96se \leq \mu \leq \bar{x} + 1.96se)$ Therefore,

$CI_{0.95} = [-0.1993491 - (1.96 \times 4.415553), -0.1993491 + (1.96 \times 4.415553)] = [-8.853833, 8.455135]$

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

ANS.
The %95 Confidence Interval is:
P$(\bar{x} - T \times se \leq \mu \leq \bar{x} + T \times se)$
for calculating T, degrees of freedom = sample size - 1 = 9 -1 = 8
$\alpha = \frac{1-0.95}{2}$=0.025
Looking up df on the table: for d.o.f.=8 and α=0.025, T=2.306
Therefore,
$CI_{0.95} = [-0.1993491 - (2.306 \times 4.415553), -0.1993491 + (2.306 \times 4.415553)] = [-10.38161, 9.982916]$

3.a. Explain why 2.e is incorrect

ANS.
The T-distribution method for calculating the confidence interval is prefered, because it takes into account the sample size. Also for Sample Size of smaller than 30, the distribution is not normal, so in these cases it is prefered to use the T-distribution method.

3.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.

ANS.
1. $3 \pm 2 \times 1.533$
False, Standard Error should have been used (Standard Deviation used here).

2. $3 \pm 1 \times 1.533$ False, Wrong T was read from the table. Degree of freedom is 4-1=3, T(0.95, 3) should have been used.

3. $3 \pm 2 \times 1.638$
False, T(0.9, 3) is wrong also, instead of Standard Error, Standard Deviation was used.

4. $3 \pm 1 \times 2.353$
   Correct

5. $3 \pm 1 \times 2.132$
   Wrong T used.

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

ANS.
I'll go by the T distribution solution where the confidence interval was between [-8.853833, 8.455135]. Looking at the formula for the confidence interval:
$\bar{x}$±T×se
Assuming negligible change in T and mean by increasing the sample size, we can halve the confidence interval approximately as following:

$$\frac{T \times \frac{S}{\sqrt{n_{new}}}}{T \times \frac{S}{\sqrt{n_{old}}}} = \frac{1}{2}$$

Therefore,
$\frac{n_{new}}{n_{old}} = 4$
So, approximately, by making the sample size four times larger, we can reach this goal.

4.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?

ANS.
We assume that we have enough samples to use the z score method:
Part A)
$1.96 \times \frac{sd}{\sqrt{n}} = 1000$
$1.96 \times \frac{20000}{\sqrt{n}} = 1000$
Therefore, n=1537
Part B) The same as Part A:
$1.96 \times \frac{sd}{\sqrt{n}} = 100$
$1.96 \times \frac{20000}{\sqrt{n}} = 100$
Therefore, n=153664

5. 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

(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.

(3) Your population distribution must be different from that used in the lesson, although anything else is fine, including any of the other continuous distributions we've discussed so far.

```r
# 1. Set how many times we do the whole thing
set.seed(1)
nruns <- 1000
```

```
# 2. Set how many samples to take in each run (20)
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){
    # 6. At random we get either a male or female beetle
    #    If it's male, we draw from the male distribution
    if(runif(1) < 0.5){
      sampler[i] <- runif(n=1,min=10,max=100)
    }
    #   If it's female, we draw from the female distribution
    else{
      sampler[i] <- runif(n=1,min=30,max=60)
    }
  }
  # 7. Finally, calculate the mean and 95% 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) - qt(0.995,19)*standard_error #T distribution
  sample_summary[j,3] <- mean(sampler) + qt(0.995,19)*standard_error #T distribution
}

counter = 0
for(j in 1:nruns){
  # If 15 is above the lower CI bound and below the upper CI bound:
  if(50 > sample_summary[j,2] && 50 < sample_summary[j,3]){
    counter <- counter + 1
  }
}
counter
```

```
## [1] 987
```