HW_04_Gupta_S

Sumit Gupta

October 3, 2017

Question 1.

```
Sys.setenv(PATH=paste(Sys.getenv("PATH"),"C:/Program Files/MiKTeX
2.9/miktex/bin/x64/",sep=";"))
```

a. You get back your exam from problem 3.d, and you got a 45. What is your z score? $z = (x - \mu)/\sigma$

```
z <- (45-70)/10
z
```

```
## [1] -2.5
```

b. What percentile are you?

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

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

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.)

```
Total_Chance<- 2*pnorm(q=45,mean=70,sd=10)
Total_Chance
```

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

Question 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)
  population<-rnorm(n=10000,mean=75,sd=10)
  smpl<-sample(population,9, replace=FALSE)
  smpl
```

```
## [1] 95.42559 71.09582 77.47009 63.64816 65.74687 86.83901 65.18924 69.56930
## [9] 74.99957
```

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

$sum = 95.42 + 71.09 + 77.47 + 63.64 + 65.74 + 86.83 + 65.18 + 69.56 + 74.99 = 669.98$ $mean = sum/n = 669.98/9 = 74.44$

c. Calculate by hand the sample standard deviation. $SampleSD(s) = \sqrt{((1/(n-1)) \times (((95.42 - mean)^2 + (71.09 - mean$ 10.73013

d. Calculate by hand the standard error.

$std.error = sd/\sqrt{n} = 10.73/3 = 3.58$

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

$p(\bar{x} - 2se < \mu < \bar{x} + 2se) = 0.95$ $p(74.44 - 2 \times 3.58 < \mu < 74.44 + 2 \times 3.57) = 0.95$ $p(67.28 < \mu < 81.58) = 0.95$

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

$p(\bar{x} - T(0.975, 8)) \times se < \mu < \bar{x} - T(0.975, 8) \times se) = 0.95$

```
T <- qt(0.975,8)
T
```

```
## [1] 2.306004
```

$p(74.44 - 2.30 \times 3.58 < \mu < 74.44 + 2.30 \times 3.58)$ $p(66.21 < \mu < 82.67) = 0.95$

Question 3.

a. Explain why 2.e is incorrect.

- Since here n<30, t-distribution is more appropriate to use.

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.

1. Incorrect: Since it uses std deviation instead of std error
2. Incorrect: It uses T(0.9,4) instead of T(0.95,3)
3. Incorrect: It uses std deviation of samples rather than std error. Also, T(0.9,3) is used
4. Correct
5. Incorrect: T(0.95,4) should not be used.

Question 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.

- With T-distribution we had: $p(66.21 < \mu < 82.67) = 0.95$

The intended interval is: $(1/2) \times (82.67 - 66.21) = 8.23$

Therefore, the interval can be calculated as: $(82.67 + 66.21)/2 = 74.44$ $8.23/2 = 4.11$

Thus, $p(70.33 < \mu < 78.55) = 0.95$ Therefore it will be: $T(0.975, n-1) \times s\sqrt{n} = 4.11$

which was earlier: $T(0.975, n-1) \times s\sqrt{n} = 8.22$

Here, assuming when n is increased s and T do not change: n1/n2 = 1/4. So, if we have 4 times the number of samples, then we could say we have 1/2 CI.

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?

- Assuming the std. deviation and the mean of sample doesnot change much by increasing n, we have:

$2000 = 4se$ $se = 500 = s\sqrt{n} = 20000/\sqrt{n}$ which gives $n = 1600$

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

- $200 = 4se$ $se = 50 = s\sqrt{n} = 20000/\sqrt{n}$ $n = 160000$

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% 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 different from that used in the lesson, although anything else is fine, including any of the other continuous distributions we've discussed so far.

```
# 1. Set how many times we do the whole thing
nruns <- 1000
# 2. Set how many samples to take in each run
nsamples <- 20
```

```r
# 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 r andom 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=6,max=14)
                    }
    #   If it's female, we draw from the female distribution
    else{
      sampler[i] <- runif(n=1,min=16,max=24)
    }
  }
  # 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 # lower 95% CI bound
  sample_summary[j,3] <- mean(sampler) + qt(0.995,19)*standard_error # lower 95% 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
  }
}
print(counter)
```

```
## [1] 988
```