Homework 4 Solution

Mohsen Nabian

6/11/2015

Comments: Professor, I didn't notice your comments for the last three HWs and you commented me 3 times to use the direct PDF version, so I am very sorry about that. I read your instructions and I installed the LATEX, but unfortunately it is not working when I try to knit it with PDF. So I knited them HTML. I ll try to debug the issue (if I can) for the next homeworks.

Question 1:

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

based on question 3.d of homework 3 , mean=70, and sd=10. so my z score can be obtaind as follows:

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

b. What percentile are you? By assuming the grades are follwing the normal distribution as the question has assumed, we can simply calculate the percentile as :

```
pnorm(q=45,mean=70,sd=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.) we can simply calculate it in R: 2 is becuase we are calculating either directins.

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

print(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);
sum(smpl);
```

```
## [1] 669.9836
```

```
mean(smpl);
```

```
## [1] 74.44263
```

```
sd(smpl);
```

```
## [1] 10.7013
```

```
print(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.

$$sum = 95.42 + 71.09 + 77.47 + 63.64 + 65.74 + 86.83 + 65.18 + 69.56 + 74.99 = 669.98$$

$$mean = sum/n = 669.98/9 = 74.44$$

2c. Calculate by hand the sample standard deviation.

$$smpleSD = \sqrt{((1/(n-1)) \times (((95.42 - mean)^2 + (71.09 - mean)^2 + (77.47 - mean)^2 + (63.64 - mean)^2 + (65.74 - mean)^2 + (86.83 -}$$

d. Calculate by hand the standard error.

$$se = sd/\sqrt{n} = 10.73/3 = 3.58$$

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

$$p(xBar - 2se < miu < xBar + 2se) = 0.95$$

$$p(74.44 - 2 \times 3.58 < miu < 74.44 + 2 \times 3.57) = 0.95$$

$$p(67.28 < miu < 81.58) = 0.95$$

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

$$p(xBar - T(0.975, 8)) \times se < miu < xBar - T(0.975, 8) \times se) = 0.95$$

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

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

$$T = 2.30$$

$$p(74.44 - 2.30 \times 3.58 < miu < 74.44 + 2.30 \times 3.58)$$

$$p(66.21 < miu < 82.67) = 0.95$$

Question 3)

a. Explain why 2.e is incorrect:

for n<30, the sample mean distribution is not quite normal. We would be better to use T distribution which in fact takes into account sample number n in the distribution.

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.

$$1))3 \pm 2 \times 1.533$$

Incorrect.1) It used sd of samples rather than se. 2)It has used T(0.9,4),while the correct answer is T(0.95,3) .

$$2))3 \pm 1 \times 1.533$$

Incorrect since T(0.9,4) is used.

$$3))3 \pm 2 \times 1.638$$

Incorrect. 1.Since T(0.9,3) is wrongly used. 2.It used sd of samples rather than se.

$$4))3 \pm 1 \times 2.353$$

Correct!

$$5))3 \pm 1 \times 2.132$$

Incorrect.Since T(0.95,4) is wrongly 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.

Ans: According to question 2 and by using T distribution we ended up with the following Interval:

$$p(66.21 < miu < 82.67) = 0.95$$

what we intend as the interval length is :

$$(1/2) \times (82.67 - 66.21) = 8.23$$

as a result the anticipated interval can be calculated as follows:

$$(82.67 + 66.21)/2 = 74.44$$

$$8.23/2 = 4.11$$

so we should seek an n to satisfy the following interval:

$$p(70.33 < miu < 78.55) = 0.95$$

assuming the mean sample does not change,it is inferred that:

$$T(0.975, n - 1) \times s/sqrtn = 4.11$$

It used to be :

$$T(0.975, n-1) \times s/sqrtn = 8.22$$

assuming s would not change much when n increased, and T also does not change so much with increasing its dg of fdm, we could say: n1/n2=1/4. As a result, if we have four times the current number of samples, ie if we have 4*9=36 samples, we could aproxiately say we would be having half confidence interval.

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?

We assume that the standard deviation(s) and the mean of the sample does not change significantly by increasing n. Thus we would have the following equation:

$$2000 = 4se$$

$$se == 500 = s/sqrtn = 20000/sqrtn$$

so

$$n = 1600$$

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

similarly:

$$200 = 4se$$

$$se = 50 = s/\sqrt{n} = 20000/\sqrt{n}$$

so

$$n = 160,000!$$

Question 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
# 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
```

So results shows that with this method of random sampling we can make sure that the TRUE MEAN with 99% chance lies withing the range we specified with the sampling.