IntroComp_WeixuanChen_HW4

Weixuan Chen

2/4/2023

1

a

The fake grades have a normal distribution with mean 70 and standard deviation 10 from problem 3.d in HW3.

```
my_score <- 45
z_score <- (my_score - 70)/10
z_score
```

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

b

The quantile is equivalent to the probability P(X < x). We can see the two methods below are equivalent. The quantile is 0.62%.

```
pnorm(-2.5, mean = 0, sd = 1)
```

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

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

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

c

We have to calculate the probability of two tails. Since normal distribution is symmetric, we can use:

```
2*pnorm(-2.5, mean = 0, sd = 1)
```

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

2

a

```r
set.seed(1)
population <- rnorm(10000, mean = 0, sd = 1)
my_sample <- sample(population, 9, replace = FALSE)
my_sample
```

```
## [1] -2.1386093  0.3926101 -0.9772168  1.8529964 -0.3512498 -0.3004396  1.9884856
## [8]  0.2008287 -0.8468195
```

b

```r
mean(my_sample)
```

```
## [1] -0.01993492
```

$sum = -2.1386093 + 0.3926101 - 0.9772168 + 1.8529964 - 0.3512498 - 0.3004396 + 1.9884856 + 0.2008287 - 0.8468195 = -$
$sample\ mean = \frac{1}{N}\sum_{i=1}^{N} x_i = -0.1794143 \ / \ 9 = -0.01993492$

c

```r
sd(my_sample)
```

```
## [1] 1.324666
```

$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$ $s = \sqrt{\frac{1}{9-1}(-2.1386093 + 0.01993492)^2 + (0.3926101 - 0.01993492)^2 + (0.9772168 - 0.01993492)^2}$

d

```r
sd(my_sample)/sqrt(length(my_sample))
```

```
## [1] 0.4415553
```

$se = sd/\sqrt{n} = 1.324666/\sqrt{9} = 1.324666/3 = 0.4415553$

e

```r
mean(my_sample)-1.96*sd(my_sample)/sqrt(length(my_sample))
```

```
## [1] -0.8853832
```

```
mean(my_sample)+1.96*sd(my_sample)/sqrt(length(my_sample))
```

```
## [1] 0.8455134
```

$CI = (\bar{x} - 1.96 * se, \bar{x} + 1.96 * se) = (-0.8853832, 0.8455134)$ 1.96 here is the z-score for two-sided normal distribution at 97.5% percentile.

f

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

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

```
mean(my_sample)-t*sd(my_sample)/sqrt(length(my_sample))
```

```
## [1] -1.038163
```

```
mean(my_sample)+t*sd(my_sample)/sqrt(length(my_sample))
```

```
## [1] 0.9982933
```

$CI = (\bar{x} - 2.306004 * se, \bar{x} + 2.306004 * se) = (-1.038163, 0.9982933)$ 2.306004 here is the z-score for two-sided t-distribution at 97.5% percentile.

3

a

Since we only have sample size 9, which is less than 30, we cannot use normal distribution to approach the distribution of sample mean.

b

```
qnorm(0.95, 0, 1)
```

```
## [1] 1.644854
```

```
qt(0.90, 3)
```

```
## [1] 1.637744
```

```
qt(0.90, 4)
```

```
## [1] 1.533206
```

```
qt(0.95, 3)
```

```
## [1] 2.353363
```

```
qt(0.95, 4)
```

```
## [1] 2.131847
```

1. It uses the sd of sample, not se of the sample mean, and the degrees of freedom of t-score is 3, not 4, and the quantile should be 95% because it is two-sided, not 90%.
2. The degrees of freedom of t-score is 3, not 4
3. The quantile should be 95% because it is two-sided, not 90%. It is correct if we want one-sided CI, and it uses the sd of sample, not se of the sample mean
4. correct
5. The degrees of freedom of t-score is 3, not 4

4

a

Since the gap between the lower and upper bound of CI is exactly the $t_{0.975,8} * se = t_{0.975,8} * sd/\sqrt{n}$. If we want to shrink the CI by 1/2, we at lease have to shrink the gap by 1/2, which means to shrink $t_{0.975,8} * se = t_{0.975,8} * sd/\sqrt{n}$ by 1/2. So if we increase the sample size by 4 times, we have $t_{0.975,8} * se = t_{0.975,8} * sd/\sqrt{4n} = t_{0.975,8} * sd/(2 * \sqrt{n}) = 1/2 * t_{0.975,8} * sd/\sqrt{n}$. So now our sample size is 9, we at least have to increase our sample size to 4*9 = 36, the CI will shrink by 1/2.

b

Assume that our sample size is greater than 30 ($t - score \approx z - score$), we have $Interval = z_{0.975} * sd/(\sqrt{n})$. $1.96 * 20000/\sqrt{n} = 1000 \Longrightarrow n = (1.96 * 20000/1000)^2 = 1536.64 \approx 1537$.

```
round((1.96*20000/1000)^2)
```

```
## [1] 1537
```

$Interval = z_{0.975} * sd/(\sqrt{n})$. $1.96 * 20000/\sqrt{n} = 100 \Longrightarrow n = (1.96 * 20000/100)^2 = 153664$

```
(1.96*20000/100)^2
```

```
## [1] 153664
```

So for CI of ±1000, we need at least 1537 people, and for CI of ±100, we need at least 153664 people.

5

```r
# 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
counter <- 0
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=5,max=15)
                    }
    #   If it's female, we draw from the female distribution
    else{
      sampler[i] <- runif(n=1,min=15,max=25)
    }
  }
  # 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] 990
```

So our accuracy of CI is 98.5%.