Homework Assignment 04

Juan Cruz Ferreyra

2024-10-04

Section 1

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

The grades in the exam are distributed $N \sim (70, 10)$. The Z-score for 45 is calculated the following way:

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

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

b. What percentile are you?

We can verify our result in the previous excercise by comparing the percentile obtained in the standard normal distribution, with the percentile corresponding to the original value in a normal distribution with the given parameters.

```
paste0(
  "The percentile for -2.5 value in the standard normal is: ",
  round(pnorm(-2.5, 0, 1), 4)
  )
```

```
## [1] "The percentile for -2.5 value in the standard normal is: 0.0062"
```

```
paste0(
  "The percentile for 45 value in the normal with original parameters is: ",
  round(pnorm(45, 70, 10), 4)
  )
```

```
## [1] "The percentile for 45 value in the normal with original parameters is: 0.0062"
```

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.)

Since the normal distribution is symmetrical, the chance of getting 45 or a more extreme value is two times the percentile calculated in the previous excercise.

```r
paste0(
  "The chance of getting a 45, or a value farther from the mean is: ",
  round(pnorm(45, 70, 10) * 2, 4)
  )
```

```
## [1] "The chance of getting a 45, or a value farther from the mean is: 0.0124"
```

Section 2

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

```r
set.seed(42)

popu <- sample.int(100, size = 10000, replace = TRUE)
n_sample = 9

popu_sample <- sample(popu, n_sample, replace = FALSE)
popu_sample
```

```
## [1] 39 56 52 98 16 50 89 82 55
```

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

The sample mean is calculated the following way:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

where x_i corresponds to the ith element in our sample.

$$\bar{x} = \frac{39 + 56 + 52 + 98 + 16 + 50 + 89 + 82 + 55}{9} = \frac{537}{9} = 59.\widehat{6}$$

c. Calculate by hand the sample standard deviation.

The sample standard deviation is calculated the following way:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where x_i corresponds to the ith element in our sample.

$$s = \sqrt{\frac{(39-\bar{x})^2 + (56-\bar{x})^2 + (52-\bar{x})^2 + (98-\bar{x})^2 + (16-\bar{x})^2 + (50-\bar{x})^2 + (89-\bar{x})^2 + (82-\bar{x})^2 + (55-\bar{x})^2)}{8}}$$

where $\bar{x} = 59.\widehat{6}$

$$s = \sqrt{\frac{427.11 + 13.44 + 58.78 + 1469.44 + 1906.78 + 93.44 + 860.44 + 498.78 + 21.78}{8}}$$

$$s = \sqrt{\frac{5350}{8}} = 25.86$$

```
# Auxiliar code
cat(paste0("(x_i - x_bar)^2 results:\n"))
```

```
## (x_i - x_bar)^2 results:
```

```
result <- 0
for (x_i in popu_sample) {
  partial_result <- (x_i - (537/9))^2
  cat(paste0(round(partial_result, 2), "\n"))

  result <- result + partial_result
}
```

```
## 427.11
## 13.44
## 58.78
## 1469.44
## 1906.78
## 93.44
## 860.44
## 498.78
## 21.78
```

```
cat(paste0("Summation result: ", result, "\n"))
```

```
## Summation result: 5350
```

```
cat(paste0("Sample standard deviation: ", round(sqrt(result / 8), 2), "\n"))
```

```
## Sample standard deviation: 25.86
```

d. Calculate by hand the standard error.

The standard error, that is the measure of the spread of $\bar{x}$s around the truth μ, is calculated the following way:

$$se = \frac{s}{\sqrt{n}} = \frac{25.86}{\sqrt{9}} = \frac{25.86}{3} = 8.62$$

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

The confidence interval for the parameter μ is constructed by forming an interval around our estimate $\bar{x}$. For a given α, in this case $\alpha = 0.05$, we use our estimate \$se = \$ for the standard error, and the central limit theorem to contruct an (approximate) $1 - \alpha$ confidence interval as following:

$$[\bar{x} - z\ se,\ \bar{x} + z\ se]\ ,$$

where z is obtained from the relation:

$$\Phi(z) = 1 - \frac{\alpha}{2}$$

Then, if $\alpha = 0.05$, we use R to find z for $\Phi(z) = 1 - \alpha/2 = 0.975$:

```
round(qnorm(p = 0.975), 2)
```

```
## [1] 1.96
```

Now we can construct an approximate 95% confidence interval of the form:

$$[59.66 - 1.96\ 8.62,\ 59.66 + 1.96\ 8.62]$$

$$[42.76,\ 76.56]$$

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

We can repeat the steps in the previous exercise, but using the t-distribution instead of the normal distribution to find t for $T(t, v) = 1 - \alpha/2 = 0.975$, where $v = n - 1 = 8$:

```
round(qt(p = 0.975, df = n_sample - 1), 2)
```

```
## [1] 2.31
```

Now we can construct an approximate 95% confidence interval of the form:

$$[59.66 - 2.31\ 8.62,\ 59.66 + 2.31\ 8.62]$$

$$[39.75,\ 79.57]$$

Section 3

a. Explain why 2.e is incorrect.

It is incorrect because we don't know the real standard deviation of the population. Even more, although we don't know the real standard deviation of the population, if the sample size were bigger, 30 samples at least, the t-distribution would approximate the normal distribution and we could just use it almost safely. With a sample size equal 9, using the normal distribution is not accurate enough.

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)

a. INCORRECT:

- It used the standard deviation of the sample mean instead of the standard error.
- It entered the t-table with $df = n$ instead of $df = n - 1$.
- It entered the t-table with T_α instead of $T_{\alpha/2}$.

b. INCORRECT:

- It entered the t-table with $df = n$ instead of $df = n - 1$.
- It entered the t-table with T_α instead of $T_{\alpha/2}$.

c. INCORRECT:

- It used the standard deviation of the sample mean instead of the standard error.
- It entered the t-table with T_α instead of $T_{\alpha/2}$.

d. CORRECT

e. INCORRECT:

- It entered the t-table with $df = n$ instead of $df = n - 1$.

Section 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.

In the expresion $\bar{x} \pm t\ se$, the term that defines the size of the confidence interval is $t\ se$, where $se = s/\sqrt{n}$. If we want to shrink our confidence interval to the half by tweaking n, we can calculate the relationship between the original n and the new one the following way:

$$2(t_{new}\ \frac{s_{new}}{\sqrt{n_{new}}}) = t_{old}\ \frac{s_{old}}{\sqrt{n_{old}}}$$

As we know, we have three elements in our equation that depends on the value of n. Particularly t and s depends on n based on the concept of degrees of freedom. We can put aside that for a moment, and imagine that t and s remains constant with the changes on n to simplify our calculations. We can do that based on that s is our estimation for the standard deviation of the original distribution, which true value does not depend on n; and that the impact of n on the t-score is considerably lower than the appearance of n in the calculation of the standard error and including it in our calculations can increase the complexity of the equation considerably. Then we have:

$$2(t\ \frac{s}{\sqrt{n_{new}}}) = t\ \frac{s}{\sqrt{n_{old}}}$$

$$\frac{t\ \frac{s}{\sqrt{n_{new}}}}{t\ \frac{s}{\sqrt{n_{old}}}} = \frac{1}{2}$$

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

$$n_{new} = 4\ n_{old}$$

So, considering just the imapct of n in the standard error calculation for the construction of the confidence interval, n_{new} should be 4 times n_{old} to shrink the interval to the half.

Our original n was 9, so in order to quadruplicate the sample size we need 27 more individuals.

But, what happens to the t-score once we increase n_{old} from 9 to 36?

Our degrees of freedom increases from 8 to 35. Then, our t-score decreases too.

```
old_t <- round(qt(p = 0.975, df = n_sample - 1), 3)
cat(paste0("Original t-score with n=9: ", old_t))
```

```
## Original t-score with n=9: 2.306
```

```
new_t <- round(qt(p = 0.975, df = n_sample * 4 - 1), 3)
cat(paste0("Original t-score with n=36: ", new_t))
```

```
## Original t-score with n=36: 2.03
```

To avoid solving the complex equation by hand, we can iteratively change the value for n, adding integers from 1 to 27, until we satisfy the following condition:

$$2(t_{new} \frac{s_{new}}{\sqrt{n_{new}}}) < t_{old} \frac{s_{old}}{\sqrt{n_{old}}}$$

As we noted earlier, s is an estimate for the standard deviation of the population, being σ a parameter that does not rely on n. Then we can simplify it from both sides of the inequation:

$$2\frac{t_{new}}{\sqrt{n_{new}}} < \frac{t_{old}}{\sqrt{n_{old}}}$$

So we proceed:

```
old_t <- qt(p = 0.975, df = n_sample - 1)
old_interval_scale <- old_t / sqrt(n_sample)

for (i in 1:27) {
  new_n_sample <- n_sample + i
  new_t <- qt(p = 0.975, df = new_n_sample - 1)

  new_interval_scale <- new_t / sqrt(new_n_sample)

  if (2 * new_interval_scale < old_interval_scale) {
    cat(paste0("New sample size should be ", new_n_sample, " to shrink confidence interval by 2"))
    break
  }
}
```

```
## New sample size should be 29 to shrink confidence interval by 2
```

We can verify our result by calculating the new confidence interval:

```
old_sample_mean <- 59.66
old_sample_std <- 25.86
new_n_sample <- 29

new_sample_se <- old_sample_std / sqrt(new_n_sample)

new_t_score <- qt(p = 0.975, df = new_n_sample - 1)

new_ci <- round(new_t_score * (old_sample_std / sqrt(new_n_sample)), 2)

cat(paste0(
  "The new confidence interval for the same sample mean std, and sample size ",
  new_n_sample,
  " is:\n", "[",
  old_sample_mean - new_ci,
  ", ",
  old_sample_mean + new_ci,
  "]"
  ))
```

```
## The new confidence interval for the same sample mean std, and sample size 29 is:
## [49.82, 69.5]
```

So, we need to sample 20 more individuals in our specific case to reduce the confidence interval to the half, when we use the t distribution tables.

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% confidence interval of ± $1,000? How many people do you need to survey to get a 95% CI of ± $100?

In this problem we use the normal distribution table to calculate the z-score due to the high number of samples. Then, as we saw in a previous excercise, for $\Phi(z) = 1 - \alpha/2 = 0.975$ we have $z = 1.96$.

We can use that information of our calculation of a confidence interval of ± $1,000, where we have a sample standard deviation of $20,000 and an unknown sample size.

$$\bar{x} \pm z \ \frac{s}{\sqrt{n}} = \bar{x} \pm 1,000,$$

$$1.96 \ \frac{20,000}{\sqrt{n}} = 1,000$$

$$\frac{39.200}{1,000} = \sqrt{n}$$

$$39.2^2 = n$$

$$n = 1,536.64 \approx 1,537$$

So, for a confidence interval of ± $1,000 or less the sample size should be of at least 1,537.

To reduce 10 times the confidence interval, that is, to obtain a confidence interval of ± $100, we just replace the corresponding term in our previous calculations:

$$\frac{39.200}{100} = \sqrt{n}$$

$$392^2 = n$$

$$n = 153,664$$

We need 153,664 people to achieve a confidence interval of ± $100. Notably, to reduce 10 times the confidence interval we need to increase 10^2 times our sample size.

Section 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.

```
set.seed(42)

nruns <- 1000
nsample <- 20

lambda <- 4

sample_summary <- matrix(NA, nruns, 3)
for(j in 1 : nruns){

  X <- rpois(nsample, lambda)

  sample_summary[j,1] <- mean(X)  # mean

  standard_error <- sd(X) / sqrt(nsample) # standard error
  tscore <- qt(p = 0.995, df = nsample - 1)
  sample_summary[j,2] <- mean(X) - (tscore * standard_error) # lower 99% CI bound
  sample_summary[j,3] <- mean(X) + (tscore * standard_error) # lower 99% CI bound
}

sum(lambda > sample_summary[, 2] & lambda < sample_summary[, 3])
```

```
## [1] 986
```