More on CIs for μ

Where does the result of  having a t distribution come from? Remember that Y1, …, Yn need to be independent random variables each with a normal probability distribution that has E(Yi) = μ and Var(Yi) = σ2 for i = 1, …, n.

Suppose Z has a standard normal distribution and W has a chi-square distribution with ν degrees of freedom. Also, suppose Z and W are independent random variables. One can show that



has a t distribution with ν degrees of freedom.

The numerator corresponds to

,

where Z has an exact normal distribution with mean 0 and standard deviation 1 due to assumptions for Y1, …, Yn (we did not discuss this exact result in the past, but focused on the central limit theorem approximation).

The denominator corresponds to

,

where W has a chi-square distribution with ν degrees of freedom.

Going through the algebra, one can see that



when ν = n – 1.

Note that  and S2 can be shown to be independent random variables (Casella and Berger, 2002, p. 218). This allows Z and W to be independent.

Why do we make these mathematical assumptions?

Whenever doing research, it’s often best to start with a particular case, like a normal distribution here. One can look to generalize beyond this particular case to see if the assumptions were needed. Fortunately, statistical research has shown that T has an *approximate* t distribution in many, many other cases! This allows the CI to *work as expected*!

What does *work as expected* mean?

When a 100(1 – α)% confidence level is stated, the interval **actually** has a confidence level close to it. We can examine this more closely via Monte Carlo simulation!

Example: GPA and coverage (coverage\_GPA.R)

Suppose a sample of students was taken 1,000 times from the population of all students on campus. For each sample, student GPAs were recorded using a sample size of 20. Also, suppose the population can be characterized by the probability distribution shown below.



The PDF is



Without going into all of the details, this PDF is based on a multiplying by 4 a random variable W with a beta probability distribution with parameters α = 5 and β = 2. For more on this probability distribution, please see textbooks like Walpole et al. and Wackerly et al.

The mean and variance are μ = E(Y) = 2.8571 and σ2 = Var(Y) = 0.4082.

The data used here is the same as the example used when discussing the central limit theorem. Below are the first 6 samples of size 20 and the first 6 sample means:

> head(round(set1,2))

 [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]

[1,] 3.41 2.79 2.65 1.87 3.21 2.99 2.96 3.83 3.28 2.15

[2,] 3.01 3.15 3.83 2.25 3.75 3.76 2.22 1.99 2.50 3.29

[3,] 2.90 2.23 2.95 3.17 3.25 3.85 3.59 3.37 3.57 3.11

[4,] 3.06 2.59 2.50 2.60 2.33 2.19 2.01 2.29 2.94 2.89

[5,] 2.38 2.74 3.48 3.25 3.24 3.34 3.17 2.79 2.25 3.39

[6,] 2.46 3.51 2.21 2.88 3.75 2.40 3.50 3.04 2.63 2.98

 [,11] [,12] [,13] [,14] [,15] [,16] [,17] [,18] [,19]

[1,] 1.11 2.73 3.68 3.34 3.56 2.12 3.27 3.76 2.91

[2,] 3.64 3.55 3.45 2.45 3.57 3.26 3.04 2.77 3.33

[3,] 3.69 1.55 3.38 3.47 3.37 3.80 3.68 3.33 2.42

[4,] 3.34 2.33 3.81 3.72 2.83 3.27 3.74 2.10 2.95

[5,] 2.04 2.39 3.08 3.06 2.85 2.29 3.38 3.60 2.89

[6,] 2.00 1.90 3.98 2.51 2.41 1.74 3.20 2.60 2.17

 [,20]

[1,] 3.59

[2,] 3.34

[3,] 3.58

[4,] 2.79

[5,] 2.61

[6,] 2.81

Please see the program for how the data was simulated.

Next, we want to find 95% CIs for μ (n is set in the program).

> alpha <- 0.05 # Find 95% Cis

> means <- apply(X = set1, MARGIN = 1, FUN = mean)

> SDs <- apply(X = set1, MARGIN = 1, FUN = sd)

> lower <- means - qt(p = 1- alpha/2, df = n - 1) \*

 SDs/sqrt(n)

> upper <- means + qt(p = 1- alpha/2, df = n - 1) \*

 SDs/sqrt(n)

> head(data.frame(lower, upper))

 lower upper

1 2.631034 3.290285

2 2.843263 3.370958

3 2.943799 3.483025

4 2.557018 3.070460

5 2.695145 3.127585

6 2.442785 3.025248

> coverage <- mean(lower < mu & upper > mu)

> coverage

[1] 0.955

Coverage is the percentage of time that the intervals contain μ. It is estimating the stated confidence level of 95%. The coverage here is close to what we expected!

Notes:

* Coverage is also known as the estimated confidence level.
* A different set of 1,000 samples will most likely result in a coverage level different from 95.5%. However, all coverage levels will be relatively close to 95%.
* Plot of the first 50 intervals (see program for code)



The third interval is the first one that does not contain μ.

* Remember that the sample size is only 20 and the population does not resemble what would be expected for a normal probability distribution!

What about other sample sizes? The program provides the code. Below is a summary from running it.

|  |  |
| --- | --- |
| n | Coverage |
| 30 | 95.4% |
| 20 | 95.5% |
| 10 | 94.1% |
| 5 | 95.0% |

The CIs get wider though as n decreases. Why?

First 50 intervals for n = 5 and n = 20 using the same y-axis scale:



