Confidence Interval

Chirantan Ganguly

29/04/2020

##Confidence Interval With Known SD We know that in a normally distributed phenomenon, 95% of cases will fall within 1.96 standard deviations above and below the mean. Let's see what that would look like. Imagine we magically know that the world population mean for happiness has a value of 36.5, with a standard deviation of 7. Let's find out where 95% of the people in the world lie.

```
#Upper limit
36.5+1.96*7
```

```
## [1] 50.22
```

```
#Lower limit
36.5-1.96*7
```

```
## [1] 22.78
```

In the last question we demonstrated how 95% of a population fall between 1.96 standard deviations above and below the population mean. Let's pretend we have psychic knowledge that the standard deviation of sadness in the world is 8, but we need to find out what the mean is. We take a sample of 300 people. Let's estimate where the population mean is likely to lie using this sample.

If you remember, the formula for calculating the confidence interval is the sample mean +/- 1.96 * standard deviation. In this case, the standard deviation is the population standard deviation, divided by the square root of the sample size.

Sample mean=30.40301

```
m<-30.40301
s<-8/(300^0.5)
# Upper Confidence interval
m+1.96*s
```

```
## [1] 31.3083
```

```
m-1.96*s
```

```
## [1] 29.49772
```

Calculating a Confidence Interval Without The Population Standard Deviation

Unfortunately in reality we usually don't know a population standard deviation, and thus must rely on sample standard deviations and T-scores. T-scores come from T-distributions, which help us account for error that occurs when we sample from a population. We use a different T-distribution to calculate cumulative probabilites depending on our degrees of freedom.

Lets say we conducted another study on how often people get angry when they're driving (known as 'road rage') using a sample of 200 people chosen at random, saved in your console as rrage. Let's calculate the 95% confidence interval for where the population mean lies.

This time we must use a slightly different formula: sample mean +/- t value * standard error. The standard error is calculated as the population standard deviation, divided by the square root of the sample size. The T-score for a df of 199 is 1.9720.

```
rrage<-c(66.277594, 41.797057, 74.996510,31.343688 ,43.748065, 57.974486,  9.543953,  37.833743, 43.536162, 42.222427, 49.718105, 47.894649, 41.206225, 68.525092, 28.701092, 43.876010, 63.588086, 50.395117, 48.799049, 55.449002, 53.313623,47.977837, 60.111216, 55.178696, 46.611072, 29.391114, 41.653766, 33.946198, 15.355090, 41.410880, 46.971639, 72.262041, 71.786006, 65.518313, 34.637214, 23.951520, 37.209896, 67.437965, 51.229966, 55.928747, 38.541238, 28.404050, 57.782350, 44.833594, 35.337264, 50.254519, 62.346285, 62.346061, 64.367405, 44.661671, 65.392093, 49.567160, 52.664752, 68.692891, 60.227733, 50.786827,68.610279, 44.286321, 38.544401, 55.025239, 50.822025, 49.027888, 44.612572,37.759472, 66.748629, 26.743206, 66.312969, 35.024732, 62.724370, 31.554896,59.176019, 40.754794, 63.882131, 40.434870, 65.698497, 55.439718, 47.889265, 43.476327, 30.952781, 49.585799, 46.629259, 76.906593, 51.359901, 51.291071, 52.639557, 41.071144, 49.498104, 51.570746, 66.953171, 70.165667, 80.918207, 22.447639, 57.818351, 46.968827, 49.051107, 53.728587, 68.001528, 65.414863, 41.662799, 52.172216, 57.672603, 46.841512, 67.039624, 59.812886, 62.767586, 51.533235, 48.460107, 43.917176, 57.714567, 28.789255, 61.852921, 74.120044, 36.691157, 28.293658, 41.697073, 36.967216, 29.479138, 76.827016, 48.883280, 44.749039, 60.560773, 53.886944, 64.718698, 72.714644, 61.365603, 55.128359, 59.600245, 24.955837, 40.465069, 36.061126, 34.631620, 66.381891, 28.976661, 73.367924, 28.881501, 41.785756, 40.679771, 54.134942, 34.705822, 62.976621, 45.603066, 45.079440, 54.064360, 69.583507, 46.552308, 51.158751, 53.021142, 76.793681, 43.327627, 44.632754, 65.045930, 37.998647, 61.454759, 27.185464, 66.167633, 56.540480, 44.855085, 47.961152, 53.965006, 71.728019, 64.519242, 54.628256, 68.592178, 54.677434, 63.571984, 24.897728, 45.006112, 60.263968, 71.723088, 54.551264, 56.721914, 57.198945, 58.249535, 42.179689, 32.786745, 30.674381, 29.288461, 55.328910, 68.850470, 44.509630, 51.836206, 56.065891, 52.150104, 74.620056, 65.550220, 69.007400, 69.862008, 54.091566, 53.325092, 33.122630, 59.514454, 33.663525, 27.247643, 34.028506, 51.056060, 25.157728, 53.034807, 61.950945, 34.881075, 64.447628)
```

```
m<-mean(rrage)
s<-sd(rrage)/(200^0.5)
m+1.972*s
```

```
## [1] 52.60419
```

```
# Calculate the upper 95% confidence interval
m-1.972*s
```

```
## [1] 48.72093
```

```
# Calculate the lower 95% confidence interval
```

The data from this study was from a sample of 200, and the results are saved in your console as rrage if you need them. Now let's try finding the 90% confidence interval (corresponding to a T score of 1.6525), and comparing what happens when we use these different intervals.

```
# Calculate the range of the 95% confidence interval
2*1.9720*s
```

```
## [1] 3.883264
```

```
# Calculate the range of the 90% confidence interval
2*1.6525*s
```

```
## [1] 3.254104
```

Calculating A Confidence Interval for a Proportion

Instead of measuring road rage as a continuous variable, you ask a sample to simply answer "yes" or "no" to the question "do you have road rage?". The outcome is saved in your console as roadrage. Let's find what proportion of your sample do have road rage.

```
roadrage<-c("no",  "no",  "yes", "no",  "no",  "no",  "yes", "yes", "yes", "no",  "no",  "no" , "yes", "no"
,  "yes", "no",  "yes",
     "yes", "yes", "no",  "no",  "yes", "no",  "no",  "yes", "yes", "no",  "no",  "yes", "no", "yes", "no",
"no",  "yes",
     "no",  "no",  "no",  "no",  "no",  "yes", "no",  "yes", "yes", "no",  "no",  "yes", "no",  "no",  "yes
", "no",  "yes",
     "no",  "yes", "no",  "yes", "no",  "no",  "yes", "yes", "yes", "yes", "yes", "no",  "no",  "no",  "no"
,  "no",  "no" ,
     "no",  "no",  "yes", "no",  "yes", "no",  "no",  "yes", "yes", "yes", "no",  "yes", "yes", "no",  "yes
", "no",  "no" ,
     "yes", "no",  "yes", "yes", "no",  "no",  "no",  "no",  "no",  "no",  "no",  "no",  "no",  "no",  "no"
,  "no",  "no" ,
     "yes", "no",  "no",  "yes", "no",  "no",  "no",  "no",  "no",  "no",  "no",  "no",  "yes", "no",  "no"
,  "no",  "no" ,
     "no",  "no",  "yes", "yes", "no",  "yes", "yes", "no",  "no",  "no",  "no",  "no",  "yes", "no",  "no"
,  "yes", "no" ,
     "no",  "no",  "no",  "no",  "yes", "yes", "no",  "no",  "no",  "yes", "no",  "no",  "yes", "no",  "no"
,  "no",  "yes",
     "no",  "no",  "yes", "no",  "no",  "yes", "yes", "no",  "no",  "no",  "no",  "no",  "no",  "no",  "yes
", "no",  "no" ,
     "no",  "no",  "yes", "no",  "no",  "no",  "no",  "no",  "no",  "yes", "no",  "no",  "no",  "yes", "no"
,  "no",  "yes",
     "yes", "yes", "no",  "yes", "no",  "yes", "yes", "yes", "yes", "yes", "no",  "no",  "yes")
```

```
c<-roadrage=="yes"
c<-roadrage[c]
p<-length(c)/200
p
```

```
## [1] 0.35
```

In your study you found that a proportion p of 0.35 of your sample said they have road rage. The standard error of this proportion is found through square root of: p multiplied by 1 - p, divided by n. Let's try this!

```
se<-sqrt((p*(1-p))/200)
se
```

```
## [1] 0.03372684
```

So you've done most of the hard word already because you have already calculated p and the standard error. Let's finalise this by calculating the upper and lower ends of the 95% confidence interval for your road rage study.

```
# Calculate the upper level of the confidence interval
p+1.96*se
```

```
## [1] 0.4161046
```

```
# Calculate the lower level of the confidence interval
p-1.96*se
```

```
## [1] 0.2838954
```

The last confidence interval you calculated was the 95% confidence interval for the proportion of people who said they had road rage. Now let's try finding the 99% confidence interval (corresponding to a Z score of 2.58), and comparing what happens when we use these different intervals.

```
# Report the range of the 95% confidence interval
2*1.96 * se
```

```
## [1] 0.1322092
```

```
# Report the range of the 99% confidence interval
2*  2.58 * se
```

```
## [1] 0.1740305
```

Sample Size

You're interested in looking at how many days in the week students drink alcohol, and need to know what kind of sample size to use.You have established an estimated mean of 3.5, and standard deviation of 1.25. You want to calculate a confidence interval of 95%, with a margin of error of 0.2. Let's input these values into our equation to estimate our required sample size!

```
ss<-1.25*1.25
# Assign the standard deviation squared to new object "ss"
zs<-1.96*1.96
# Assign the value of the Z-score squared to new object "zs"
ms<-0.2*0.2
# Assign the value of the margin of error squared to the new object "ms"
ss*zs/ms
```

```
## [1] 150.0625
```

```
# Calculate the neccessary sample size
```

Now you're conducting a study on what proportion of students drink alcohol and want to know what sample size to use for a confidence interval of 95%, with a margin of error of 0.05.

The sample size will be p multiplied by 1-p multiplied by the Z-score squared, divided by the margin of error squared. Let's try to find this using the 'safe approach' for p. This is the value above which the output p*(1-p) cannot get any larger

```
p<-0.5*(1-0.5)
z<-1.96*1.96
ms<-0.05*0.05
p*z/ms
```

```
## [1] 384.16
```