Confidence interval for population mean

2020-03-29

What it is

In general, confidence intervals use observed data to give a range of values where the population parameter is thought to exist. They provide the range of potential parameter values which care *compatible* with the observed data. Confidence intervals for the population mean μ use the sample mean $\bar{y}$, standard deviation s, and sample size n to build an interval for μ from the t distribution.

When to use it

The confidence interval for μ requires the conditions of the Central Limit Theorem. First, this means the data needs to be a random sample from the population. Additionally, the population should be normally distributed or the sample should be "large". The general rule of thumb is $n \geq 30$. If n is less than 30 (or larger but not by much) a QQ plot can help assess how close to normal the data are and whether this method is appropriate.

How to use it

In R, the confidence interval for μ is obtained most easily through the `t.test()` function. Suppose we want to estimate the average weight of all adult Californians. We can input the observed weight data from our sample to `t.test()`. Appending `$conf.int` will return only the confidence interval [1]:

[1] See *Hypothesis test for population mean* for more detail about the output of `t.test()`

```
t.test(x = dat$weight, conf.level = 0.95)$conf.int
```

```
## [1] 165.9942 173.1498
## attr(,"conf.level")
## [1] 0.95
```

So, from our data we are 95% confident that the average weight of U.S. adults is somewhere between 166 and 173 pounds. In other words, the true mean being any value within that range would be compatible with the data we observed.

We can change the confidence level by changing `conf.level` in the R code. For example, using `conf.level=0.90` will give a 90% confidence interval.

Why it works

The confidence interval for μ requires the t distribution, which is a modification of the normal distribution. Under the conditions of the CLT (large, random sample), we know that $\bar{y}$ is approximately normally distributed, or, when standardized:

$$\frac{\bar{y}-\mu}{\sigma/\sqrt{n}} \sim Z$$

However, σ is the *population* standard deviation, which we do not know. We can replace σ with its estimate, s, but this adds extra uncertainty and variability to our statistic. The new statistic is no longer a Z distribution, but a t distribution. Ultimately, we have

$$\frac{\bar{y}-\mu}{s/\sqrt{n}} \sim t^{(n-1)}$$

where $n-1$ are the degrees of freedom of the t distribution.

We obtain a 95% confidence interval by starting with the probability statement

$$P\left(-t_{0.025}^{(n-1)} < \frac{\bar{y}-\mu}{s/\sqrt{n}} < t_{0.025}^{(n-1)}\right) = 0.95$$

Then we can solve the inequality in the probability to obtain just μ on the inside[2]:

$$P(\bar{y} - t_{0.025}^{(n-1)}\frac{s}{\sqrt{n}} < \mu < \bar{y} + t_{0.025}^{(n-1)}\frac{s}{\sqrt{n}}) = 0.95$$

[2] Steps: (1) multiply by $\frac{s}{\sqrt{n}}$; (2) subtract $\bar{y}$; (3) multiply by -1, remembering to switch the direction of the inequality signs

So, there is a 95% chance that μ is between $\bar{y} - t_{0.025}^{(n-1)}\frac{s}{\sqrt{n}}$ and $\bar{y} + t_{0.025}^{(n-1)}\frac{s}{\sqrt{n}}$. Making the result a bit more compact, the 95% confidence interval for μ is

$$\bar{y} \pm t_{0.025}^{(n-1)}\frac{s}{\sqrt{n-1}}$$