Confidence interval for population mean

Date: 2017-10-30

What it is

In general, confidence intervals use observed data to give a range of values where the population parameter is thought to exist. Confidence intervals for *mu* 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 $n \geq 30$).

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 in the US population. We can input the observed weight data 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 = d$weight, conf.level = 0.95)$conf.int
```

```
## [1] 173.6477 182.0323
## 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 173.6 and 182.0 pounds. In other words, the process used to create this confidence interval would include the true average weight in the population 95% of the time.

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 Central Limit Theorem. 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[2]. 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:

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

Thus, the 95% confidence interval for μ is

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

[2] See *t model* for more detail about the t distribution