Confidence interval for population proportion

Date: 2017-10-05

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 p use the sample proportion $\hat{p}$ and the sample size n to build an interval for p from the binomial distribution.

When to use it

The confidence interval for p requires the conditions of a binomial distribution (a fixed sample of independent binary outcomes). [1]

[1] Many textbooks teach the normal model approximation method, which requires a large sample size (usually defined as at least 10 success and at least 10 failures); however, the `binom.test()` R function used here does not employ the normal approximation and will work for any sample size.

How to use it

In R, the confidence interval is obtained most easily through the `binom.test()` function. Suppose we want to estimate the rate of obesity in the US population. First, we need to count the number of obese respondents and the total sample size:

```
table(d$bmicat)
```

```
##
## underweight      normal  overweight
##          12         167         180
##       obese
##         141
```

```
nrow(d)
```

```
## [1] 500
```

Then, we can use these values as inputs to `binom.test()`. Appending `$conf.int` will return only the confidence interval [2]:

[2] See *Hypothesis test for population proportion* for more detail about the output of `binom.test()`

```
binom.test(x = 141, n = 500, conf.level = 0.95)$conf.int
```

```
## [1] 0.2429479 0.3236520
## attr(,"conf.level")
## [1] 0.95
```

So, from our data we are 95% confident that somewhere between 24% and 32% of the US population is obese.

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

R uses a highly computational process to obtain the the 95% confidence interval above. For large samples (at least 10 success and at least 10 failures), the Central Limit Theorem suggests the Z model can be used instead. The formula for a 95% confidence interval using the Z approximation is:

$$\left(\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

In order to get a different confidence level, 1.96 needs to be changed to the appropriate value from the Z distribution. Some common choices are:

Confidence level	z
50%	0.67
80%	1.28
90%	1.64
98%	2.33
99%	2.58