Confidence Interval for Population Proportion

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, which include a sample of a fixed and known size, each observation an independent binary outcome. 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 for p is obtained most easily through the `binom.test()` function. Suppose we want to estimate the prevalence of diabetes among Missouri adults. First, we need to count the number of diabetic respondents and the total sample size:

```
table(mydat$diabetes, useNA='ifany')
```

```
##
##   Yes   No <NA>
##   264 1734    2
```

So, there are 264 positive responses out of 1998 ($= 264 + 1734$) valid responses. Then, we can use these values as inputs to `binom.test()`. Appending `$conf.int` will return only the confidence interval[1]:

```
binom.test(x=264, n=1998, conf.level=0.95)$conf.int
```

```
## [1] 0.1175814 0.1477715
## attr(,"conf.level")
## [1] 0.95
```

So, from our data we are 95% confident that somewhere between 12% and 15% of the Missouri adults have diabetes.

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.

[1] See *Hypothesis Test for Population Proportion* for more detail about the output of `binom.test()`

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. In particular,

$$\frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \sim Z$$

Since there is about a 0.95 probability that an observation will be within two standard deviations of the distribution mean[2], we have:

[2] And the Z distirbution has a standard deviation of 1

$$P\left(-2 < \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} < 2\right) = 0.95$$

Then we can solve the inequality in the probability to obtain just p on the inside[3]:

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

$$P\left(\hat{p} - 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 0.95$$

So, there is a 95% chance that p is between $\hat{p} - 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ and $\hat{p} + 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. Making the result a bit more compact, the normal approximation 95% confidence interval for p is

$$\hat{p} \pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$