Example:

A random sample of 500 people in a certain country which is about to have a national election were asked whether they preferred “Candidate A” or “Candidate B”.

From this sample, 320 people responded that they preferred Candidate A.

Construct an approximate 95% confidence interval for the true proportion of people in the country who prefer Candidate A.

Let p be the true proportion of the people in the country who prefer Candidate A.

We have an estimate

$$\hat{p} = \frac{320}{500} = \frac{16}{25}$$

The estimator is

$$\hat{p} = \frac{\text{\# in the sample who like A}}{\text{\# in the sample}}$$

The Model:

Take a random sample of size n.

Record $X_1, X_2, \ldots, X_n$ where

$$X_i = \begin{cases} 1 & \text{person i likes Candidate A} \\ 0 & \text{person i likes Candidate B} \end{cases}$$

Then $X_1, X_2, \ldots, X_n$ is a random sample from the Bernoulli distribution with parameter p.

The Model:

Note that, with these 1's and 0's,

$$\hat{p} = \frac{\text{\# in the sample who like A}}{\text{\# in the sample}}$$

$$= \frac{\sum_{i=1}^{n} X_i}{n} = \overline{X}$$

By the Central Limit Theorem, $\hat{p} = \overline{X}$ has, for large samples, an approximately normal distribution.

The Model: $\hat{p} = \overline{X}$

$$E[\hat{p}] = E[X_1] = p$$

$$Var[\hat{p}] = \frac{Var[X_1]}{n} = \frac{p(1-p)}{n}$$

So, $$\hat{p} \overset{approx}{\sim} N\left(p, \frac{p(1-p)}{n}\right)$$

The Model: $\hat{p} = \overline{X}$

$$\hat{p} \overset{\text{approx}}{\sim} N\left(p, \frac{p(1-p)}{n}\right)$$

In particular,

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

behaves roughly like a N(0,1) as n gets large

The Model: $\hat{p} = \overline{X}$

What does “large” mean?

“n>30” is a rule of thumb to apply to all distributions, but we can (and should!) do better with specific distributions.

- $\hat{p}$ lives between 0 and 1.
- the normal distribution lives between $-\infty$ and ∞

- $\hat{p}$ lives between 0 and 1.
- The normal distribution lives between $-\infty$ and ∞.
- However, 99.7% of the area under a N(0,1) curve lies between -3 and 3,

i.e. “99.7% of the probability for a normal distribution is within 3 standard deviations of it’s mean

$$\hat{p} \overset{\text{approx}}{\sim} N\left(p, \frac{p(1-p)}{n}\right)$$

$$\Rightarrow \quad \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

However, this quantity is unknown to us.

In practice, we approximate it with

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

Go forward using normality if the interval

$$\left(\hat{p} - 3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + 3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

is completely contained within [0,1].

$$-z_{-\alpha/2} < \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} < z_{\alpha/2}$$

$$-z_{-\alpha/2} < \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} < z_{\alpha/2}$$

Although it looks difficult to isolate p "in the middle", it can be done.

It is equivalent to saying that

$$\frac{(\hat{p} - p)^2}{\frac{p(1-p)}{n}} < z^2_{\alpha/2}$$

i.e. $$(\hat{p} - p)^2 - z^2_{\alpha/2}\,\frac{p(1-p)}{n} < 0$$

$$-z_{-\alpha/2} < \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} < z_{\alpha/2}$$

However, it is far more common to just plug $\hat{p}$ in for the p's in the denominator to get

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

as an approximate $100(1-\alpha)\%$ confidence interval for p.

Back to the example:

Let p be the true proportion of the people in the country who prefer Candidate A. Find.a 95% confidence interval for p.

We have $n = 500, \quad \hat{p} = \dfrac{16}{25}$

Check

$$\left(\hat{p} - 3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + 3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right) = (0.5756, 0.7044)$$

95% $\Rightarrow$ $z_{0.025} = 1.96$

$n = 500, \quad \hat{p} = \dfrac{16}{25}$

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

gives

$$(0.5979, 0.6821)$$