**Confidence Intervals – Proportions – Wilson Score Interval**

Previously we developed a confidence interval for $p$ using $\hat{p}$:

$$\hat{p} \pm E = \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Where it is understood that the value of $z^*$ (referred to as a critical value) depends on the level of confidence. Here are some often used values of $z^*$.

| confidence level | z* |
|---|---|
| 80% | 1.282 |
| 90% | 1.645 |
| 95% | 1.960 |
| 99% | 2.576 |
| 99.9% | 3.291 |

This formula works quite well when the sample size is very large. However, for smaller samples, this formula can have a coverage area that varies quite a bit from what is expected. In other words, when we are expecting a 95% confidence interval to capture the population proportion $p$ 95% of the time, we might in practice be creating confidence intervals that only capture the population proportion 92% of the time. There is a more complicated formula that gives better (closer to expected) results.

Recall that in developing the formula

$$\hat{p} \pm E = \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

we reasoned that 95% of the values of $\hat{p}$ will be between these two values,

$$z = -1.96 = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}, \qquad z = +1.96 = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}$$

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

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

Putting these together (and replacing the 1.96 with z*) gives us

$$\hat{p} = p \pm z^* \sqrt{\frac{p(1-p)}{n}}$$

Let us solve this algebraically for $p$. Subtract $p$ from both sides.

$$\hat{p} - p = \pm z^* \sqrt{\frac{p(1-p)}{n}}$$

Square both sides.

$$(\hat{p} - p)^2 = z^2 \frac{p(1-p)}{n}$$

This is a quadratic equation in $p$ what we will try to get looking like $ap^2 + bp + c = 0$ and solve for $p$.

$$\hat{p}^2 - 2\hat{p}p + p^2 = \frac{z^2}{n}(p - p^2)$$

$$\hat{p}^2 - 2\hat{p}p + p^2 = \frac{z^2}{n}p - \frac{z^2}{n}p^2$$

$$p^2 + \frac{z^2}{n}p^2 - 2\hat{p}p - \frac{z^2}{n}p + \hat{p}^2 = 0$$

$$\left(1 + \frac{z^2}{n}\right)p^2 - \left(2\hat{p} + \frac{z^2}{n}\right)p + \hat{p}^2 = 0$$

This is a quadratic equation in $p$ where $a = \left(1 + \frac{z^2}{n}\right)$, $b = -\left(2\hat{p} + \frac{z^2}{n}\right)$ and $c = \hat{p}^2$. So

$$p = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} = \frac{\left(2\hat{p} + \frac{z^2}{n}\right) \pm \sqrt{\left(2\hat{p} + \frac{z^2}{n}\right)^2 - 4\left(1 + \frac{z^2}{n}\right)\hat{p}^2}}{2\left(1 + \frac{z^2}{n}\right)}$$

$$p = \frac{\left(2\hat{p} + \frac{2z^2}{2n}\right) \pm \sqrt{\left(2\hat{p} + \frac{2z^2}{2n}\right)^2 - 4\left(1 + \frac{z^2}{n}\right)\hat{p}^2}}{2\left(1 + \frac{z^2}{n}\right)} = \frac{2\left(\hat{p} + \frac{z^2}{2n}\right) \pm \sqrt{\left(2\left(\hat{p} + \frac{z^2}{2n}\right)\right)^2 - 4\left(1 + \frac{z^2}{n}\right)\hat{p}^2}}{2\left(1 + \frac{z^2}{n}\right)}$$

$$p = \frac{2\left(\hat{p} + \frac{z^2}{2n}\right) \pm \sqrt{4\left(\hat{p} + \frac{z^2}{2n}\right)^2 - 4\left(1 + \frac{z^2}{n}\right)\hat{p}^2}}{2\left(1 + \frac{z^2}{n}\right)} = \frac{2\left(\hat{p} + \frac{z^2}{2n}\right) \pm 2\sqrt{\left(\hat{p} + \frac{z^2}{2n}\right)^2 - \left(1 + \frac{z^2}{n}\right)\hat{p}^2}}{2\left(1 + \frac{z^2}{n}\right)}$$

$$p = \frac{\left(\hat{p} + \frac{z^2}{2n}\right) \pm \sqrt{\left(\hat{p} + \frac{z^2}{2n}\right)^2 - \left(1 + \frac{z^2}{n}\right)\hat{p}^2}}{\left(1 + \frac{z^2}{n}\right)} = \frac{\hat{p} + \frac{z^2}{2n} \pm \sqrt{\hat{p}^2 + 2\frac{z^2}{2n}\hat{p} + \left(\frac{z^2}{2n}\right)^2 - \hat{p}^2 - \frac{z^2}{n}\hat{p}^2}}{1 + \frac{z^2}{n}}$$

$$p = \frac{\hat{p} + \frac{z^2}{2n} \pm \sqrt{\frac{z^2}{n}\hat{p} + \left(\frac{z^2}{2n}\right)^2 - \frac{z^2}{n}\hat{p}^2}}{1 + \frac{z^2}{n}} = \frac{\hat{p} + \frac{z^2}{2n} \pm \sqrt{\frac{z^2}{n}(\hat{p} - \hat{p}^2) + \frac{z^4}{4n^2}}}{1 + \frac{z^2}{n}}$$

$$p = \frac{\hat{p} + \frac{z^2}{2n} \pm z\sqrt{\frac{1}{n}(\hat{p} - \hat{p}^2) + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}} = \frac{\hat{p} + \frac{z^2}{2n} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}}$$

The Wilson Score Interval

$$p = \frac{\hat{p} + \frac{z^2}{2n} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}}$$

Notice that as $n \to \infty$ the quantities $\frac{z^2}{2n}$ and $\frac{z^2}{4n^2}$ both approach 0, and the quantity $1 + \frac{z^2}{n}$ approaches 1, thus our formula approaches

$$p = \frac{\hat{p} + 0 \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + 0}}{1} = \hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

which was our previous (simpler) formula. Here are some test cases to show the relationship between these two formulas (at 95% confidence).

| $\hat{p}$ | $\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ | $\frac{\hat{p} + \frac{z^2}{2n} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}}$ |
|---|---|---|
| $\hat{p} = \frac{10}{50} = 0.2$ | $(0.0891, 0.3109)$ | $(0.1124, 0.3304)$ |
| $\hat{p} = \frac{100}{500} = 0.2$ | $(0.1649, 0.2351)$ | $(0.1673, 0.2373)$ |
| $\hat{p} = \frac{1000}{5000} = 0.2$ | $(0.1889, 0.2111)$ | $(0.1891, 0.2113)$ |
| $\hat{p} = \frac{10000}{50000} = 0.2$ | $(0.1965, 0.2035)$ | $(0.1965, 0.2035)$ |

As you can see, the simpler formula (which is taught in most introductory classes) is a good approximation to the Wilson formula when the sample size is quite large. But with technology in our hand, there is less need for the simpler formula since technology (or code) will handle the Wilson formula with ease.