Sample Sizes for Proportions

Gene Quinn

Sample Sizes

When designing a sample survey to estimate the proportion of a population that has some characteristic, an important question is what size sample to take.

Sample Sizes

When designing a sample survey to estimate the proportion of a population that has some characteristic, an important question is what size sample to take.

There are two competing interests in this problem.

First, we usually want to obtain the estimate with minimum cost, which argues for a small sample.

Second, we would like the estimate to be accurate, which argues for a large sample.

Sample Sizes

A common approach to this problem is to decide on an acceptable width for the confidence interval for p.

We decide in advance on a distance d so that we want the estimate x/n to have a probability of $1 - \alpha$ of lying within a distance of d of the population parameter p.

Sample Sizes

A common approach to this problem is to decide on an acceptable width for the confidence interval for p.

We decide in advance on a distance d so that we want the estimate x/n to have a probability of $1 - \alpha$ of lying within a distance of d of the population parameter p.

This means the width of the confidence interval for p is $2d$. Once we know this, we work backwards to determine the sample size n.

Sample Sizes

Of course, the confidence interval depends on the unknown parameter p.

Sample Sizes

Of course, the confidence interval depends on the unknown parameter p.

As with the margin of error, the usual approach is to assume the worst-case value for p, which is $1/2$.

Sample Sizes

Of course, the confidence interval depends on the unknown parameter p.

As with the margin of error, the usual approach is to assume the worst-case value for p, which is $1/2$.

This produces the widest possible confidence interval for any value of p, so regardless of the true value of p we can be sure that the probability our worst-case interval contains p is at least $1 - \alpha$.

Sample Sizes

Recall that the $100(1 - \alpha/2)\%$ confidence interval for the population proportion p is

$$\left(\frac{k}{n} - z_{\alpha/2} \sqrt{\frac{\frac{k}{n}(1 - \frac{k}{n})}{n}}, \quad \frac{k}{n} + z_{\alpha/2} \sqrt{\frac{\frac{k}{n}(1 - \frac{k}{n})}{n}} \right)$$

Sample Sizes

Recall that the $100(1-\alpha/2)\%$ confidence interval for the population proportion p is

$$\left(\frac{k}{n} - z_{\alpha/2}\sqrt{\frac{\frac{k}{n}(1-\frac{k}{n})}{n}}, \quad \frac{k}{n} + z_{\alpha/2}\sqrt{\frac{\frac{k}{n}(1-\frac{k}{n})}{n}}\right)$$

We would like the width of this interval (the difference between the upper and lower limits) to be no larger than $2d$:

$$2d = \left(\frac{k}{n} + z_{\alpha/2}\sqrt{\frac{\frac{k}{n}(1-\frac{k}{n})}{n}}\right) - \left(\frac{k}{n} - z_{\alpha/2}\sqrt{\frac{\frac{k}{n}(1-\frac{k}{n})}{n}}\right)$$

Sample Sizes

Combining terms gives

$$2d \;=\; 2z_{\alpha/2}\sqrt{\frac{\frac{k}{n}(1-\frac{k}{n})}{n}}$$

Sample Sizes

Combining terms gives

$$2d \;=\; 2z_{\alpha/2}\sqrt{\frac{\frac{k}{n}(1-\frac{k}{n})}{n}}$$

The worst case occurs when $\hat{p} = k/n = 1/2$, so we substitute $1/2$ for k/n to get

$$2d \;=\; 2z_{\alpha/2}\sqrt{\frac{\frac{1}{4}}{n}} \;=\; 2z_{\alpha/2}\sqrt{\frac{1}{4n}}$$

Sample Sizes

Solving for n, we get

$$\frac{2d}{2z_{\alpha/2}} = \sqrt{\frac{1}{4n}}$$

$$\frac{d^2}{z^2_{\alpha/2}} = \frac{1}{4n}$$

and finally

$$n = \frac{z^2_{\alpha/2}}{4d^2}$$

Sample Sizes

Solving for n, we get

$$\frac{2d}{2z_{\alpha/2}} = \sqrt{\frac{1}{4n}}$$

$$\frac{d^2}{z_{\alpha/2}^2} = \frac{1}{4n}$$

and finally

$$n = \frac{z_{\alpha/2}^2}{4d^2}$$