7.3

Confidence intervals for proportions

We know that a histogram is a good description of how the observations of a random sample are distributed and given information about the accuracy of those relative frequencies (or, percentages) associated with the various classes.

Let, Y be the frequency of measurements in the interval out of the n observations, so that (under the assumptions of independence & constant probability p) Y has a binomial distribution $b(n,p)$. Thus, the problem is to determine the accuracy of the relative frequency $\frac{Y}{n}$ as an estimator of p. We solve this problem by finding, for the unknown p, a confidence interval based on $\frac{Y}{n}$.

For a $b(n,p)$ and $\frac{Y}{n}$ is an unbiased estimator for p,

$$\frac{Y-np}{\sqrt{np(1-p)}} = \frac{\frac{Y}{n}-p}{\sqrt{\frac{p(1-p)}{n}}}$$

has an approximate normal distribution $N(0,1)$, provided that n is large enough.

Thus, for a given probability $1-\alpha$, we can find a $z_{\alpha/2}$ such that,

$$P\left[-z_{\alpha/2} \le \frac{\frac{Y}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} \le z_{\alpha/2}\right] \simeq 1-\alpha \qquad (1)$$

$$\Rightarrow P\left[\frac{Y}{n} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \le p \le \frac{Y}{n} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right] \simeq 1-\alpha$$

Unfortunately, the unknown parameter p appears in the endpoints of this inequality. There are two ways out of this dilemma. First, we could make an additional approximation, namely, replacing p with $\frac{Y}{n}$ in $p(1-p)$. That is, if n is large enough, it is still true that

$$P\left[\frac{Y}{n} - z_{\alpha/2}\sqrt{\frac{\frac{Y}{n}(1-\frac{Y}{n})}{n}} \le p \le \frac{Y}{n} + z_{\alpha/2}\sqrt{\frac{\frac{Y}{n}(1-\frac{Y}{n})}{n}}\right] \simeq 1-\alpha \qquad (2)$$

Thus, for large n, if the observed Y equals y, then the interval

$$\left[\frac{y}{n} - z_{\alpha/2}\sqrt{\frac{\frac{y}{n}(1-\frac{y}{n})}{n}},\ \frac{y}{n} + z_{\alpha/2}\sqrt{\frac{\frac{y}{n}(1-\frac{y}{n})}{n}}\right]$$

serves as an approximate $100(1-\alpha)\%$ confidence interval for p. That is, p is within $z_{\alpha/2}\sqrt{\frac{\frac{y}{n}(1-\frac{y}{n})}{n}}$ of $\hat{p} = \frac{y}{n}$.

A second way to solve for p in the inequality in ① can be written as,

$$\frac{\left|\frac{Y}{n} - p\right|}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{\alpha/2}$$

② is equivalent to

$$H(p) = \left(\frac{Y}{n} - p\right)^2 - \frac{z_{\alpha/2}^2\, p(1-p)}{n} \leq 0 \quad \text{—— ③}$$

we have to find those values of p for which $H(p) \leq 0$. Let, $\hat{p} = \frac{Y}{n}$ and $z_0 = z_{\alpha/2}$

$\therefore$ ③ $\Rightarrow$

$$H(p) = \left(1 + \frac{z_0^2}{n}\right)p^2 - \left(2\hat{p} + \frac{z_0^2}{n}\right)p + \hat{p}^2$$

Find ~~out~~ $p = ?$

| $ax^2 + bx + c = 0$
| $\Rightarrow x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$

These values give the endpoints for an approximate $100(1-\alpha)\%$ confidence interval for p. If n is large, $\frac{z_0^2}{2n}$, $\frac{z_0^2}{4n^2}$ and $\frac{z_0^2}{n}$ are small. Thus, the confidence intervals given by ② & ③ are approximately equal when n is large.

Example: 7.3-1 :

Here,

sample size, $n = 40$

$$\frac{y}{n} = \frac{8}{40} = 0.2$$

$$1-\alpha = 90\% = 0.9 \Rightarrow \alpha = 0.1$$

$$\Rightarrow \alpha/2 = 0.05$$

$$\therefore z_{\alpha/2} = z_{0.05} = 1.645$$

$\therefore$ the approximate 90% confidence interval for ~~...~~, the fraction p is,

$$\left[\frac{y}{n} - z_{\alpha/2}\sqrt{\frac{\frac{y}{n}(1-\frac{y}{n})}{n}}, \frac{y}{n} + z_{\alpha/2}\sqrt{\frac{\frac{y}{n}(1-\frac{y}{n})}{n}}\right]$$

$$= \left[0.2 - 1.645\sqrt{\frac{.2 \times .8}{40}}, 0.2 + 1.645\sqrt{\frac{.2 \times .8}{40}}\right]$$

$$= [0.096, 0.304].$$

or $[9.6\%, 30.4\%]$ Ans

#

If we use the interval for ③, we get, $[0.117, 0.321]$ because of small sample size.

For $n = 400$, we get both intervals are $[0.167, 0.233]$ and $[0.169, 0.235]$ respectively, which differ very little.

Example: 7.3.2: Here, $y = 185$, $n = 351$

$$\therefore \frac{y}{n} = 0.527$$

and, $1-\alpha = 95\% = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow \frac{\alpha}{2} = 0.025$

$$\therefore z_{\alpha/2} = z_{0.025} = 1.96$$

$\therefore$ The approximate 95% confidence interval for the fraction p of the voting population who favor the candidate is,

$$\left[0.527 - 1.96\sqrt{\frac{0.527\times 0.473}{351}},\ 0.527 + 1.96\sqrt{\frac{0.527\times 0.473}{351}}\right]$$

$$= [0.475,\ 0.579]$$

The one-sided confidence interval for p given by,

$\left[0, \frac{y}{n} + z_\alpha\sqrt{\frac{\frac{y}{n}(1-\frac{y}{n})}{n}}\right]$ provides an upper ~~bound~~ bound for p.

and $\left[\frac{y}{n} - z_\alpha\sqrt{\frac{\frac{y}{n}(1-\frac{y}{n})}{n}}, 1\right]$ provides a lower bound for p.

~~Exercise: 7.3 - 1-6, 8~~

Exercise: 7.3: $\boxed{1-6,8}$