The Sample Proportion and the Central Limit Theorem

To demonstrate that the large sample Normal distribution of $\widehat{P}$ (and therefore the Normal approximation to the Binomial) is an example of the Central Limit effect, we need only establish that $\widehat{P}$ is the mean (or average) of a random sample from some distribution.

The Binomial(n, p) distribution is the distribution of Y, the number of heads in n tosses of a biased coin, when the probability of getting a head on a single toss is p. We begin by setting up a separate random variable for each toss.

Let $$X_i = \begin{cases} 1, & \text{if the } i\text{th toss is a head;} \\ 0, & \text{if the } i\text{th toss is a tail.} \end{cases}$$

Now the total number of heads Y is given by $Y = X_1 + X_2 + \ldots + X_n$ (we count one more each time we get a head. Moreover, the X_i's are independent (as tosses are independent), and each X_i has the same distribution namely

x	0	1
$\text{pr}(X = x)$	$1-p$	p.

Thus, the X_i's constitute a random sample from this distribution. The distribution has

$$\text{E}(X) = 0 \times (1-p) + 1 \times p = p.$$

Since

$$\begin{aligned} \text{E}\left[(X-\mu)^2\right] &= (0-p)^2(1-p) + (1-p)^2 p = (1-p)\left[p^2 + p\,(1-p)\right] \\ &= p(1-p), \end{aligned}$$

we have

$$\begin{aligned} \text{sd}(X) &= \sqrt{\text{E}\left[(X-\mu)^2\right]} \\ &= \sqrt{p(1-p)}. \end{aligned}$$

Hence

$$\widehat{P} = \frac{Y}{n} = \frac{\Sigma X_i}{n} = \overline{X},$$

so that $\widehat{P}$ is the sample mean from a distribution with mean $\mu = p$ and standard deviation $\sigma = \sqrt{p(1-p)}$. The Central Limit Theorem tells us, therefore, that

$$\widehat{P} \text{ is approximately} \quad \text{Normal}\left(\mu_{\widehat{P}} = p, \sigma_{\widehat{P}} = \sqrt{\frac{p(1-p)}{n}}\right)$$

in large samples. Since the number of heads in n tosses is $Y = n\widehat{P}$, it follows that Y is also approximately Normally distributed. Thus, the Normal approximation to Binomial is a consequence of the Central Limit Theorem.

Since $\widehat{P}$ has been shown to be a sample mean you may think, "why not apply the formula given for se($\overline{x}$) in Section 7.2.3 of the text to the X_i's above to get a formula for se($\widehat{p}$)?" If you do this, it can be shown that you get our previous formula for se($\widehat{p}$) apart from a factor[1] of $1/\sqrt{1-(1/n)}$, which is essentially 1 for all reasonable sample sizes.

Adjustment for Sampling Without Replacement

As in the case of estimating means, we need to make an adjustment to the above theory when $\widehat{p}$ is obtained by sampling from a finite population without replacement, as in polls. This time, however, we can derive an expression from previous theory. When sampling without replacement, we should be using the Hypergeometric distribution for Y instead of the Binomial. Using the formula for sd(Y) for the Hypergeometric discussed in Chapter 5 on this web site,

$$\begin{aligned} \text{sd}(\widehat{P}) &= \text{sd}(\frac{1}{n}Y) \;=\; \frac{1}{n}\,\text{sd}(Y) \;=\; \frac{1}{n}\sqrt{np(1-p)\frac{N-n}{N-1}} \\ &\approx \sqrt{\frac{p(1-p)}{n}} \cdot \sqrt{1-f}. \end{aligned}$$

where $f = n/N$ is the sampling fraction. This is the usual formula apart from a finite population correction $\sqrt{1-f}$. The standard error formula is corrected in the same way. We can ignore the correction factor when $f < 0.1$.

There are very few practical situations in which finite population corrections are called for, although we did come across one recently in which a marketing research company polled 120 of New Zealand's top 400 business leaders ($n/N = 0.3$).

[1] Where each $x_i = 0$ or 1, it is not hard to show that $\sum(x_i - \overline{x})^2 = \sum x_i^2 - n\overline{x}^2 = n\widehat{p}(1-\widehat{p})$ and thus that $\frac{s_X}{\sqrt{n}} = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}/\sqrt{1-\frac{1}{n}}$.