THE VARIANCE OF THE PROPORTIONS OF SAMPLES FALLING WITHIN A FIXED INTERVAL FOR A NORMAL POPULATION

By G. A. Baker

University of California, Davis

Suppose that we have a normal population

$$y = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-m)^2}{2\sigma^2}\right\} \tag{1}$$

and we draw samples of N from this population. We wish to estimate the proportion, p, of the population between two fixed limits, $m + \lambda\sigma$ and $m + \mu\sigma$. One way to make this estimate is simply to count the number of observed x's which fall in this interval. We shall denote this number by n. Then the ratio

$$n/N \tag{2}$$

is an estimate of p. If this is done the variance of p is well known to be

$$\frac{p(1-p)}{N}. \tag{3}$$

The method of estimating p by counting the number in a definite interval is nonparametric and requires no assumption of normal or other specified type of sampled population for validity. However, if we know that the sampled population is normal then we may make use of this knowledge in estimating p and possibly obtain an improved estimate.

Another way to estimate p which makes use of the form of the sampled population is to compute

$$\begin{aligned} \bar{x} &= \frac{1}{N}\sum_{i=1}^{N} x_i \\ s^2 &= \frac{1}{N}\sum_{i=1}^{N} (x_i - \bar{x})^2, \end{aligned} \tag{4}$$

and hence the integral

$$\int_{m+\lambda\sigma}^{m+\mu\sigma} \frac{e^{-(x-\bar{x})^2/(2s^2)}}{s\sqrt{2\pi}}\, dx. \tag{5}$$

It is implied in elementary texts that (5) is a better estimate of p than is (2) although this point is not discussed.

It is the purpose of the present note to discuss the variance of the estimate (5) and compare this variance with (3).

Now (5) is a function of the first two moments of the sample and it follows from an application of a theorem stated by H. Cramér [1] that (5) is asymptotically normal with mean p and variance given by

$$\sigma_p^2 = \frac{1}{2\pi N}\left[\frac{(\lambda e^{-\frac{1}{2}\lambda^2} - \mu e^{-\frac{1}{2}\mu^2})^2}{2} + (e^{-\frac{1}{2}\lambda^2} - e^{-\frac{1}{2}\mu^2})^2\right]. \tag{6}$$

To compare the relative efficiency of the counting method with (6) in complete detail would be somewhat tedious. The referee suggests a brief discussion of the cases $\lambda = -\infty$, where we are counting the proportion less than some known value, and $\lambda = -\mu$, where a portion out of the middle of the distribution is being counted. These cases are of particular practical interest.

If $\lambda = -\infty$, then (6) becomes

$$\sigma_p^2 = \frac{e^{-\mu^2}}{2\pi N}\left[\frac{\mu^2}{2} + 1\right]. \tag{7}$$

We choose values of μ as indicated below:

μ	p	Relative Efficiency of (3)
−2.3263	0.01	0.27
−1.2816	0.1	0.56
−0.8416	0.2	0.66
−0.5244	0.3	0.75
−0.2533	0.4	0.64
0.0000	0.5	0.64

We get values of the relative efficiency of (3) that are low for small p and somewhat higher for larger values of p.

If $\lambda = -\mu$, then (6) becomes

$$\sigma_p^2 = \frac{\mu^2 e^{-\mu^2}}{\pi N}. \tag{8}$$

We choose values of μ as indicated below:

μ	p	Relative Efficiency of (3)
1.2816	0.8	0.63
0.8416	0.6	0.46
0.2533	0.2	0.12

We see that the relative efficiency of (3) ranges from close to 0.75 to rather small values.

Other choices of λ and μ yield relative efficiencies of about the same order of magnitude as those illustrated.

REFERENCE

[1] HARALD CRAMÉR, "*Mathematical Methods of Statistics*," Princeton University Press, 1946, section 28.4, pp. 366–367.