A problem involving minuscule probabilities

R. G. D. Steel

Problem: Mutation rates in A and B lines of corn were observed as

	Non-mutants	Mutants	Total (approx.)
A	5×10^5	10	5×10^5
B	6×10^5	4	6×10^5

Is there a significant difference between lines?

Solution 1. Very small probabilities are involved and, in spite of the sample size, one might hesitate to use a χ^2-test. An obvious non-parametric test is Tchebycheff's Inequality:

$$P\left(|x - \mu| > k\sigma\right) \leq \frac{1}{k^2}$$

For $k = 5$, $\frac{1}{k^2} = .04$. Since this is close to the common probability of .05, $k = 5$ seems desirable. For $k = 3$, $\frac{1}{k^2} = .11$. Since small numbers are involved in two cells and since the inequality is valid for any continuous (we have a discrete) distribution with finite variance, the value $k = 3$ might be acceptable.

$$\mu = 0$$

$$x = \frac{10}{(5)10^5} - \frac{4}{(6)10^5} = \frac{40}{(30)10^5}$$

$$\hat{\sigma} = \sqrt{\frac{14}{(11)10^5}\left(\frac{(11)10^5 - 14}{(11)10^5}\right)\left(\frac{1}{(5)10^5} + \frac{1}{(6)10^5}\right)}$$

$$= \sqrt{\frac{14}{(11)10^5} \quad \frac{11}{(30)10^5}} \qquad \text{(approx.)}$$

I.e. We have had to estimate σ.

$$\text{Observed } k = \frac{|x - \mu|}{\hat{\sigma}} = \frac{40}{(30)10^5} \quad \frac{\sqrt{30}\,10^5}{\sqrt{14}}$$

$$= \frac{40\sqrt{420}}{30(14)} = 1.95$$

Since this implies a probability of about .26, we cannot reject the null hypothesis of no difference.

Solution 2. Cochran (e.g. 1) has shown that statisticians have been conservative in regard to small expected values in the common χ^2 tests. Hence, we might boldly apply the χ^2 test. This is equivalent to assuming that a binomial with a very small p can be approximated by a normal distribution if the sample size is sufficiently large. We will obtain, within rounding errors,

$$\chi^2 = k^2$$

$$= 3.80$$

so it is a matter of comparing the observed k with a value from the normal table.

$$P\ (|t| > 1.96) = .05$$

Hence we are almost exactly at the 5% point with k = 1.95.

As usually computed,

$$\chi^2 = \frac{\{4(5)10^5 - 10(6)10^5\}^2\,(11)10^5}{(11)10^5(14)(5)10^5(6)10^5}$$

$$= 3.81$$

Tabulated χ^2 (.05, 1 d.f.) = 3.841.

Solution 3. (Courtesy I. Blumen.) Since very small numbers of mutants are observed in very large samples, we may assume that the Poisson distribution offers a reasonable explanation of the data. The test is based on determining whether or not the 4:10 split is improbable in sampling a population where the true proportions are $6 \times 10^5 : 5 \times 10^5$ and we stop after 14 trials. We are seen to be dealing with a conditional Poisson distribution involving 14 mutants. See Steel (2). This may also be considered as approximating a ball and urn problem where sampling is with replacement from 6 + 5 balls and stops after 14 trials.

$$P\ (x \geq 10 \mid p = .45,\ n = 14) = \sum_{x=10}^{14} \binom{14}{x} .45^x\ .54^{14-x} = .0426$$

$$P\ (x \geq 10 \mid p = .46,\ n = 14) = .0500$$

so that

$$.0426 < P\ (x \geq 10 \mid p = 5/11,\ n = 14) < .0500$$

The result is seen to be virtually identical with that obtained by χ^2.

References

1. Cochran, W. G., Some methods for strengthening the common χ^2 tests, Biometrics 10 (1954), 417-451.

2. Steel, R. G. D., Relation between Poisson and multinomial distributions, BU-39-M (1953).