THE COMPUTATION OF THE X-DISTRIBUTION

B. L. VAN DER WAERDEN
UNIVERSITY OF ZURICH

The X-test is a two-sample test, defined as follows. Let $x_1, \cdots, x_g$ and $y_1, \cdots, y_h$ be independent observed variables. Let $r_1, \cdots, r_g$ be the rank numbers of $x_1, \cdots, x_g$ among the x's and y's. Put $g + h = n$. Let Φ be the (cumulative) normal distribution function and $\Psi = \Phi^{-1}$ the inverse function. Put

$$a_r = \Psi\left(\frac{r}{n+1}\right), \qquad r = 1, \cdots, n. \tag{1}$$

The hypothesis H to be tested is: The x's have the same distribution as the y's. The test statistic is

$$X = \sum a_r, \tag{2}$$

the summation extending over the rank numbers $r_1, \cdots, r_g$ of the x's. If X exceeds a limit X_β depending on the level β, the hypothesis H is rejected. The two-sided test on the level 2β rejects when the absolute value $|X|$ exceeds the same limit X_β.

In my paper [1] I have proved that under the hypothesis H the statistic X is asymptotically normal for $g/h \to \infty$ or $h/g \to \infty$. Noether, in his review of my paper [2], pointed out that the asymptotic normality for $g + h \to \infty$ can also be proved when g/h and h/g remain bounded. A full proof for $g \to \infty$ and $h \to \infty$ was given by D. J. Stoker in his Amsterdam thesis [3].

For small g and h the exact limit X_β can be found by explicit computation of the largest X-values. Beyond $g = h = 10$, this computation becomes impracticable. The normal distribution may be used as an approximation, but the comparison with the exact values for $g = h = 8$ or 9 or 10 showed a systematic deviation. The normal approximation for X_β was always too large, so that the power of the test was diminished.

A closer examination showed that this deviation is mainly due to the rather large terms a_1 and a_n, which may or may not be included in the sum (2). An improved approximation could be obtained by separating these large terms from the sum (2).

Consider, for example, the case $g = h = 5$. The 10 terms a_r are, according to (1),

$$\begin{array}{lllll} a_1 = -1.34 & a_2 = -.91 & a_3 = -.60 & a_4 = -.35 & a_5 = -.11 \\ a_6 = +.11 & a_7 = +.35 & a_8 = +.60 & a_9 = +.91 & a_{10} = +1.34\,. \end{array} \tag{3}$$

The test statistic X is a sum of $g = 5$ terms a_r chosen at random from the 10 possible terms (3). Now if X were a sum of many terms, each having only a relatively small influence, the normal approximation would be very good. However, the terms a_1 and a_{10} are not small. Therefore they have to be considered separately.

Put $a_1 = -a$ and $a_n = +a$. Let X_{g-2}, X_{g-1} and X_g be sums of $g - 2$ or $g - 1$ or g terms chosen at random among the remaining terms $a_2, \cdots, a_{n-1}$.

We have to compute the distribution function $F(t)$ of the statistic X, that is, the probability of $X < t$. Let $F_{g-2}(t)$, $F_{g-1}(t)$ and $F_g(t)$ be the distribution functions of X_{g-2}, X_{g-1} and X_g. The probabilities that $a_1 = -a$ and $a_n = a$ both occur among the g terms of X, or that only one occurs and the other not, or that both are missing, are

$$\frac{g(g-1)}{n(n-1)}, \qquad \frac{gh}{n(n-1)}, \qquad \frac{h(h-1)}{n(n-1)}, \tag{4}$$

respectively. Hence the distribution function of X is

$$F(t) = \frac{g(g-1)}{n(n-1)} F_{g-2}(t) + \frac{gh}{n(n-1)} [F_{g-1}(t-a) + F_{g-1}(t+a)] + \frac{h(h-1)}{n(n-1)} F_g(t). \tag{5}$$

The most important case for the applications is $g = h = n/2$. In this case, neglecting terms of order n^{-1}, the probabilities (4) may be replaced by 1/4. Hence (5) simplifies to

$$4F(t) \sim F_{g-2}(t) + F_{g-1}(t-a) + F_{g-1}(t+a) + F_g(t). \tag{6}$$

Now replace F_{g-2}, F_{g-1} and F_g by normal distribution functions. The means of X_{g-2}, X_{g-1} and X_g are zero. The variances are

$$\sigma_{g-2}^2 = \sigma_g^2 = \frac{g(g-2)}{n-3} Q', \quad \text{and} \quad \sigma_{g-1}^2 = \frac{(g-1)^2}{n-3} Q', \tag{7}$$

where

$$Q' = \frac{1}{n-2} \sum_2^{n-1} a_r^2. \tag{8}$$

The difference between σ_{g-1} and σ_g is of order n^{-2} only and hence negligible. Thus we obtain from (6), replacing σ_{g-2} and σ_g by $\sigma_{g-1} = \sigma$,

$$4F(t) \sim 2\Phi\left(\frac{t}{\sigma}\right) + \Phi\left(\frac{t-a}{\sigma}\right) + \Phi\left(\frac{t+a}{\sigma}\right). \tag{9}$$

This is the required approximation.

It is an easy matter to determine t in such a way that the right member of (9) becomes equal to $1 - \beta$. The resulting $t = X_\beta$ is the asymptotic rejection limit.

For $g = h = 10$ and $\beta = .005$ (two-sided test on the 1 per cent level) the normal approximation leads to the asymptotic rejection limit 5.14. The improved approximation (9) leads to $X_\beta = 4.99$. This is much better, for the exact limit is 4.94.

Tables for the X-test have been computed and will be published in the form of a small book [4] by the author and E. Nievergelt.

REFERENCES

[1] B. L. VAN DER WAERDEN, "Ein neuer Test für das Problem der zwei Stichproben," *Math. Annalen*, Vol. 126 (1953), pp. 93–107.

[2] G. E. NOETHER, a review of [1], *Math. Reviews*, Vol. 15 (1954), p. 46.

[3] D. J. STOKER, *Oor 'n klas van toetsingsgroothede vir die probleem van twee steekproewe*, The Hague, Excelsior, 1955.

[4] B. L. VAN DER WAERDEN and E. NIEVERGELT, *Tables for Comparing Two Samples by X-Test and Sign Test*, to be published.