A FORMULA FOR SAMPLE SIZES FOR POPULATION TOLERANCE LIMITS

By H. Scheffé and J. W. Tukey

Princeton University

In a paper to appear in a later issue of this journal dealing with various results on non-parametric estimation, we shall discuss in detail an approximate formula for the numerical calculation of sample sizes for Wilks' population tolerance limits. Because of the practical usefulness of this formula, it seems desirable to make it available without delay. Its accuracy is adequate for all direct applications.

An interval I is said to cover a proportion π of a univariate population with cumulative distribution function $F(x)$ if $\int_I dF = \pi$. Let $X_1, X_2, \cdots, X_n$ be a random sample from the population, and $Z_1 \leq Z_2 \leq \cdots \leq Z_n$ be a rearrangement of $X_1, X_2, \cdots, X_n$. Define $Z_0 = -\infty$, $Z_{n+1} = +\infty$, and consider the proportion B of the population covered by the random interval (Z_k, Z_{n-m+1}). Then $Pr\{B \geq b\}$ is independent of $F(x)$ if $F(x)$ is continuous[1], and equals $1 - I_b(n - r + 1, r)$, where $r = k + m$ and $I_x(p, q)$ is K. Pearson's notation for the incomplete Beta function.

Choose a confidence coefficient $1 - \alpha$, a pair of positive integers k, m, and a fraction b. The sample size n for which we can make the statement "the probability is $1 - \alpha$ that the random interval (Z_k, Z_{n-m+1}) cover a proportion b or more of the population" is then determined by the equation

$$I_b(n - r + 1, r) = \alpha, \tag{1.1}$$

where $r = k + m$. Our approximate solution is

$$n \doteq \tfrac{1}{4}\chi_\alpha^2(1 + b)/(1 - b) + \tfrac{1}{2}(r - 1), \tag{1.2}$$

where χ_α^2 is the 100α percent point on the χ^2-distribution with $2r$ degrees of freedom. The required values of χ_α^2 may be found for $\alpha = .1, .05, .025, .01, .005$ in Catherine Thompson's table [2]. For this range of α, and for $b \geq .9$, extensive numerical calculations indicate that the error of (1.2) is less than one tenth of one percent, and is always positive, that is, n is slightly overestimated by (1.2). We have not yet obtained an analytic proof of this statement, which refers to the difference from the exact (and, in general, non-integral) solution of (1.1).

As explained elsewhere [1], formula (1.2) may be used for Wald's solution of the multivariate case.

REFERENCES

[1] H. Scheffé, *Annals of Math. Stat.*, Vol. 14 (1943), p. 324.

[2] C. M. Thompson, *Biometrika*, Vol. 32 (1941), p. 189.

[1] That the theory is valid in this case we show later. Previous proofs have required the continuity of $F'(x)$.