$\epsilon_{n,\alpha}$ so that the error committed by using $\tilde{\epsilon}_{n,\alpha}$ instead of $\epsilon_{n,\alpha}$ would be in the safe direction, and that this error becomes already very small for $n = 50$.

REFERENCES

[1] N. SMIRNOV, "Sur les écarts de la courbe de distribution empirique," *Rec. Math.* (*Mat. Sbornik*), N. S. Vol. 6 (48) (1939), pp. 3–26.

[2] A. WALD AND J. WOLFOWITZ, "Confidence limits for continuous distribution functions," *Annals of Math. Stat.*, Vol. 10 (1939), pp. 105–118.

ON THE ESTIMATION OF CENTRAL INTERVALS WHICH CONTAIN ASSIGNED PROPORTIONS OF A NORMAL UNIVARIATE POPULATION

BY G. E. ALBERT AND RALPH B. JOHNSON

University of Tennessee and Clemson Agricultural College

Summary. For samples of any given size $N \geq 2$ from a normal population, Wilks [1] has shown how to choose the parameter λ_p so that the expected coverage of the interval $\bar{x} \pm \lambda_p s$ will be $1 - p$. The present paper treats the choice of the minimal sample size N necessary to effect a certain type of statistical control on the fluctuation of that coverage about its expected value; a brief table of such minimal sample sizes is given.

1. Introduction. Let $F(y)$ denote the normal cumulative distribution function

$$F(y) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{y} e^{-(u-m)^2/(2\sigma^2)}\, du. \tag{1}$$

If p is any number in the range $0 < p < 1$, factors $\lambda(p)$ are well known such that the proportion

$$A = F(m + \lambda\sigma) - F(m - \lambda\sigma) \tag{2}$$

of the probability between $m \pm \lambda\sigma$ will equal $1 - p$.

If m and σ are unknown, it is natural to consider the random variable

$$A(\bar{y}, s; \lambda) = F(\bar{y} + \lambda s) - F(\bar{y} - \lambda s), \tag{3}$$

where $\bar{y} = \sum_{n=1}^{N} y_n/N$ and $s = \left\{\sum_{i=1}^{N} (y_i - \bar{y})^2/(N-1)\right\}^{\frac{1}{2}}$.

Obviously λ cannot be chosen to guarantee $A(\bar{y}, s; \lambda) = 1 - p$. S. S. Wilks [1] has shown that, for a random sample of size N, the expectation of (3) is $1 - p$,

$$EA(\bar{y}, s; \lambda) = 1 - p, \tag{4}$$

if the parameter λ is chosen as

$$\lambda = t_p \sqrt{\frac{N+1}{N}}. \tag{5}$$

In (5) t_p is such that for Student's t-distribution of $N - 1$ degrees of freedom

$$Pr[\,|\,t\,|\,\geq t_p] = p.$$

Wilks' study of the variability of $A(\bar{y}, s; \lambda)$ was based upon an approximate consideration of the variance of A. It is the purpose of this paper to present more precise results in this latter connection.

Let d_1, d_2 and α be assigned positive numbers satisfying the inequalities $0 \leq 1 - p - d_1 < 1 - p + d_2 \leq 1$, and $0 < \alpha < 1$. It is shown that if λ be chosen as in (5), the requirement

$$Pr[1 - p - d_1 \leq A(\bar{y}, s; \lambda) \leq 1 - p + d_2] \geq \alpha \tag{6}$$

places a lower bound on the sample size N. It is clear that if d_1 and d_2 are small and α near unity, (6) places a control on the variability of $A(\bar{y}, s; \lambda)$ about its expectation $1 - p$.

TABLE I

Smallest N for which (6) holds

p		.01		.05			.25			.50		
α		.95	.99	.80	.95	.99	.80	.95	.99	.80	.95	.99
d_1	d_2											
.075	.05	—	—	—	24	49	54	128	226	44	108	197
.05	.05	—	—	—	43	92	76	174	298	63	144	243
.025	.025	—	—	65	159	299	298	692	1194	245	567	975
.035	.015	—	—	107	274	510	420	1332	2628	337	1079	2184
.05	.01	12	27	196	640	1230	813	2991	5983	649	2488	4928
.025	.01	26	64	226	641	1230	907	2993	5983	725	2487	4928
.02	.01	37	88	254	657	1231	1025	3015	5982	825	2502	4928
.01	.01	110	319	428	1009	1750	1846	4319	7456	1507	3540	6084

Methods devised by Wald and Wolfowitz [2] are easily adapted to the approximate calculation of the probability (6).

Table I presents *minimal* values of the sample size N to effect the control (6) for various values of the constants p, d_1, d_2 and α. The indication is clear that the *prediction of probability intervals based upon the estimates $\bar{y}$ and s from small samples is not very reliable.*

2. The expectation of A and the probability (6). Writing $u = (\bar{y} - m)/\sigma$ and $v = s/\sigma$, $A(\bar{y}, s; \lambda)$ becomes

$$A^*(u, v; \lambda) = \frac{1}{\sqrt{2\pi}} \int_{u-\lambda v}^{u+\lambda v} e^{-\frac{1}{2}t^2}\, dt. \tag{7}$$

It is well known that the variables $u\sqrt{N}$ and $(N - 1)v^2$ are independently distributed, the first being normal with zero mean and unit variance and the second

being chi-square with $N - 1$ degrees of freedom. One readily derives (Wilks [1])

$$E(A) = Pr\left[|t| \leq \lambda\sqrt{\frac{N}{N+1}}\right],$$

where t has Student's distribution with $N - 1$ degrees of freedom. Setting this equal to $1 - p$, the choice (5) for λ is obtained.

To calculate the probability (6), one integrates the joint frequency function $f(u, v)$ over that portion of the half plane $-\infty < u < \infty, v > 0$ on which $1 - p - d_1 \leq A^* \leq 1 - p + d_2$. To perform the integration, one proceeds as in Wald and Wolfowitz [2] where a similar problem is solved. Define two functions

$$v_r = v_r(u), \qquad r = 1, 2 \tag{8}$$

by the equations

$$A^*(u, v_r; \lambda) = 1 - p + (-1)^r d_r, \qquad r = 1, 2, \tag{9}$$

where A^* is defined by (7) and λ is given by (5). The functions $v_r(u)$ are monotone increasing relative to $|u|$. It follows that

$$Pr\{1 - p - d_1 \leq A(\bar{y}, s; \lambda) \leq 1 - p + d_2\} = \sqrt{\frac{N}{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}Nu^2}P(u)du, \tag{10}$$

where

$$P(u) = Pr\{(N - 1)v_1^2(u) < \chi^2 < (N - 1)v_2^2(u)\}, \tag{11}$$

χ^2 being distributed as chi-square with $N - 1$ degrees of freedom.

The formulas (10) and (11) are too unwieldy for much computation. Following Wald and Wolfowitz [2] again, one can show that a good approximation for large N is

$$Pr\{1 - p - d_1 \leq A(\bar{y}, s; \lambda) \leq 1 - p + d_2\} \cong P(N^{-\frac{1}{2}}), \tag{12}$$

the right member being given by (11).

3. Computational procedure. For a given set of values of p, d_1, and d_2, one may now tabulate (12) against N by the following steps. Using λ as given by (5), the $v_r = v_r(N^{-\frac{1}{2}})$ defined by (9) are found by trial and error from a standard normal distribution table. Then (11) and (12) give the control probability. One easily picks out the minimal N for which (6) is satisfied. Tables of the incomplete gamma function [3] are available and the authors are in possession of graphs of the chi-square distribution prepared from these tables by the use of spline curves. The detail of the graphs is sufficient for three-decimal accuracy in reading probabilities. For small values of N and values beyond the range of tables, a variety of standard methods of approximation for (11) were used.

Lower and upper bounds for the interval (10) are easily devised using obvious approximate quadrature methods. See Wald and Wolfowitz [2] in this connection. The small values of N in Table I were checked by such a device. The

authors are confident that the computation was sufficiently accurate to make the table useful for practical purposes.

4. Generalization. The formulation of the problem discussed above may be generalized to the case in which the mean m of the distribution (1) depends linearly upon k sure variables $x_1^*, x_2, \cdots, x_k$. The N observations are then N $(k+1)$-tuples $(y_i; x_{i1}, x_{i2}, \cdots, x_{ik})$, $i = 1, 2, \cdots, N$, and the mean has the form

$$m = \alpha + \sum_{j=1}^{k} \beta_j (X_j - \bar{x}_j)$$

for an arbitrary set of values $X_1, X_2, \cdots, X_k$ of the sure variables. Referring to Cramér ([4], pages 551 and 552) for notations and formulas in order to save space here, one replaces the interval estimate $(\bar{y} \pm \lambda s)$ above by the interval from R_1 to R_2, where

$$R_r = \alpha^* + \sum_{j=1}^{k} \beta_j^* (X_j - \bar{x}_j) + (-1)^r \lambda^* \sigma^*, \qquad r = 1, 2,$$

$$\lambda^* = t_p \sqrt{\frac{N+M}{N-k-1}},$$

and

$$M = 1 + \sum_{i,j=1}^{k} \frac{L_{ij}}{L} (X_i - \bar{x}_i)(X_j - \bar{x}_j).$$

Here t_p is chosen as in (5) except that the degrees of freedom are now $N - k - 1$.

For this generalization, when N/M is large, the control probability (6) is approximated by $P(M^{\frac{1}{2}}/N^{\frac{1}{2}})$ where $P(u)$ is given by (11). Organized computation for this generalization does not seem feasible since the values of the quadratic form M may vary greatly from one application to another.

REFERENCES

[1] S. S. Wilks, "Determination of sample sizes for setting tolerance limits," *Annals of Math. Stat.*, Vol. 12 (1941), pp. 91–96.

[2] A. Wald and J. Wolfowitz, "Tolerance limits for a normal distribution," *Annals of Math. Stat.*, Vol. 17 (1946), pp. 208–215.

[3] K. Pearson, *Tables of the Incomplete Gamma Function*, Cambridge University Press 1922.

[4] H. Cramér, *Mathematical Methods of Statistics*, Princeton University Press, 1946.