$-\infty \leq v \leq \infty$. Thus neither y nor ω has an asymptotic normal distribution. It is, of course, this fact which makes the criterion of minimum variance illusory.

3. Other polynomial distribution functions. Let repeated samples of n independent values of x be drawn from a population characterized by $D(x) = \frac{k+1}{a^{k+1}} x^k$, $0 \leq x \leq a$, and k a positive integer or zero. It can be shown that the best linear estimate of the mean of the population is $y = \frac{(k+1)n+1}{n(k+2)} x_n$, where as before x_n is the largest item of the sample. The sampling distribution of y is easily obtained. It follows that

$$\sigma_y^2 = \frac{(k+1)a^2}{(k+2)^2[(k+1)n^2 + 2n]} = \frac{k+3}{n(k+1)+2} \sigma_{\bar{x}}^2,$$

where as usual $\bar{x}$ is the arithmetic mean of the sample. Again, if we write $u = \left(y - \frac{k+1}{k+2} a\right) \Big/ \sigma_y$, the limit of the distribution of u as n approaches infinity is, as before, e^{u-1}, $-\infty \leq u \leq 1$.

A NOTE ON TOLERANCE LIMITS

By Edward Paulson[1]

Columbia University

Among various statistical problems arising in the process of controlling quality in mass production, a rather important one appears to be the determination of tolerance limits when the variability of the product is known to be due to random factors. This problem was recently treated in a pioneer article by Wilks. This note will point out a relationship between tolerance limits and confidence limits (used in the sense of Neyman), and will use this concept to establish tolerance limits when the product is described by two qualities, the measurements on which are assumed to have a bivariate normal distribution.

For the case of a single variate, the problem of finding tolerance limits as stated by Wilks is to find a sample size n, and two functions $L_1(x_1 \cdots x_n)$ and $L_2(x_1 x_2 \cdots x_n)$ so that if $P = \int_{L_1}^{L_2} f(x)\, dx$ denotes the conditional probability of a future observation falling between the random variates L_2 and L_1, then

$$E(P) = \alpha, \quad \text{and Prob. } [\alpha - \Delta_1 \leq P \leq \alpha + \Delta_2] \geq \beta.$$

The relationship between confidence limits and tolerance limits will arise if confidence limits are determined, not for a parameter of the distribution, but for

[1] Work done under a grant-in-aid from the Carnegie Corporation of New York.

a future random observation (or for some function of the observations in a future independent sample). This is based on the following simple lemma: *If confidence limits $U_1(x_1 \cdots x_n)$ and $U_2(x_1 \cdots x_n)$ on a probability level $= \alpha_0$ are determined for g, a function of a future sample of k observations, and $P = \int_{U_1}^{U_2} \psi(g)\,dg$, then $E(P) = \alpha_0$.* For let $\psi(g)\,dg$ and $\varphi(U_1, U_2)\,dU_1\,dU_2$ denote the distribution of g and U_1, U_2 respectively, then by the definition of expected value

$$E(P) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left[\int_{U_1}^{U_2} \psi(g)\,dg\right]\varphi(U_1, U_2)\,dU_1\,dU_2\,.$$

This triple integral is however exactly the probability that g will lie between U_1 and U_2, which by the nature of confidence limits must equal α_0, which proves the lemma. In a similar manner it follows that if on the basis of a given sample an l dimensional confidence region is found for statistics $g_1, g_2, \cdots g_l$ derived from a future sample, and if P denotes the probability that $g_1 \cdots g_l$ all fall in the confidence region, then $E(P)$ in repeated sampling equals α. To establish tolerance limits, it is necessary in addition to $E(P)$ to also know the distribution of P, or at least σ_P^2, so the distribution of P can be approximated.

It appears, at least on an intuitive basis, that the "best" confidence interval can be used to determine the shape of the "most efficient" tolerance limits; this intuitive notion will gain additional support from the character of the tolerance region which will now be derived for an observation (x, y) from a distribution with probability density $f(x, y)$, where

$$f(x, y) = \frac{\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-m_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-m_x}{\sigma_x}\right)\left(\frac{y-m_y}{\sigma_y}\right) + \left(\frac{y-m_y}{\sigma_y}\right)^2\right]\right\}}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}$$

Suppose we have 2 independent samples

$$[(x_1, y_1)(x_2, y_2) \cdots (x_n, y_n)] \quad \text{and} \quad [(x, y)]$$

both from $f(x, y)$. Then it is known that

$$T^2 = \left(\frac{n}{n+1}\right)\frac{1}{1-r^2}\left\{\left(\frac{\bar{x}-x}{s_x}\right)^2 - \frac{2r}{s_x s_y}(\bar{x}-x)(\bar{y}-y) + \frac{(\bar{y}-y)^2}{s^2y}\right\}$$

where $\bar{x} = \sum_{i=1}^{n} x_i/n$, $s_x^2 = \sum_{1}^{n} (x_i - \bar{x})^2/(n-1)$, etc., has the distribution of Hotelling's Generalized Student Ratio [2]. A confidence region for a future observation (x, y) on the basis of a sample of n on a level of significance $= \alpha$ will be given by the elliptic region $T^2 \leq T_\alpha^2$ (in the x, y plane), where $T_\alpha^2 = 2(n-1)F_0/(n-2)$, where F_0 is the value of the F distribution (with $n_1 = 2$ and $n_2 = n - 2$ degrees of freedom) which is exceeded with probability $= 1 - \alpha$.

If P denotes the probability of a future observation falling in this ellipse, then $P = \iint_{T^2 \leq T_\alpha^2} f(x, y)\,dx\,dy$. By utilizing the fact [2] that T^2 is invariant under linear transformations, it is not difficult to see that the distribution of P will not involve any unknown parameters, so its distribution can be calculated under the assumption $m_x = m_y = \rho = 0$, $\sigma_x = \sigma_y = 1$. Then

$$P = F(\bar{x}, \bar{y}, s_x, s_y, r) = \iint\limits_{T^2 \leq T_\alpha^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}\, dx\, dy.$$

We know that $E(P) = \alpha$, and we will now calculate the variance of P by expanding P in a Taylor Series (to terms of the first order) about the point $\bar{x} = 0$, $\bar{y} = 0$, $r = 0$, $s_x = 1$, $s_y = 1$. P can clearly be put in the form

$$P = \frac{1}{2\pi} \int_{\bar{x}-T_\alpha\sqrt{\frac{n+1}{n}}s_x}^{\bar{x}+T_\alpha\sqrt{\frac{n+1}{n}}s_x} e^{-\frac{1}{2}x^2}\, dx \int_{y+rs_y\left(\frac{x-\bar{x}}{s_x}\right)-s_y\sqrt{(1-r^2)\left[T_\alpha^2\left(\frac{n+1}{n}\right)-\left(\frac{x-\bar{x}}{s_x}\right)^2\right]}}^{\bar{y}+rs_y\left(\frac{x-\bar{x}}{s_x}\right)+s_y\sqrt{(1-r^2)\left[T_\alpha^2\left(\frac{n+1}{n}\right)-\left(\frac{x-\bar{x}}{s_x}\right)^2\right]}} e^{-\frac{1}{2}y^2}\, dy$$

Taking derivatives and evaluating about the population values

$$\left[\frac{\partial P}{\partial \bar{x}}\right] = \left[\frac{\partial P}{\partial \bar{y}}\right] = \left[\frac{\partial P}{\partial r}\right] = 0,$$

$$\left[\frac{\partial P}{\partial s_x}\right] = \frac{e^{-\frac{1}{2}T_\alpha^2\left(\frac{n+1}{n}\right)}}{\pi} \int_{-T_\alpha\sqrt{\frac{n+1}{n}}}^{T_\alpha\sqrt{\frac{n+1}{n}}} \frac{x^2\, dx}{\sqrt{T_\alpha^2\left(\frac{n+1}{n}\right) - x^2}}$$

$$= \tfrac{1}{2}e^{-\frac{1}{2}T_\alpha^2\left(\frac{n+1}{n}\right)} T_\alpha^2\left(\frac{n+1}{n}\right)$$

$$\left[\frac{\partial P}{\partial s_y}\right] = \frac{e^{-\frac{1}{2}T_\alpha^2\left(\frac{n+1}{n}\right)}}{\pi} \int_{-T_\alpha\sqrt{\frac{n+1}{n}}}^{T_\alpha\sqrt{\frac{n+1}{n}}} \sqrt{T_\alpha^2\left(\frac{n+1}{n}\right) - x^2}\, dx$$

$$= e^{-\frac{1}{2}T_\alpha^2\left(\frac{n+1}{n}\right)} \cdot \tfrac{1}{2}T_\alpha^2\left(\frac{n+1}{n}\right).$$

So $$\delta P = e^{-T_\alpha^2\left(\frac{n+1}{n}\right)} \cdot \tfrac{1}{2}T_\alpha^2\left(\frac{n+1}{n}\right)[\delta s_x + \delta s_y],$$

and to terms of $0\left(\frac{1}{n}\right)$:

$$\sigma_P^2 = \frac{T_\alpha^4 e^{-T_\alpha^2}}{4n}.$$

Since for ordinary values of $\alpha(\alpha = .95$ or $.99)$ the distribution of P seems to approach normality very slowly, we will follow a suggestion of Wilks and suppose that a fairly close approximation to the distribution of P will be given by

$$\frac{\Gamma(u+v)}{\Gamma(u)\Gamma(v)} P^{u-1}(1-P)^{v-1}, \tag{1}$$

where
$$u = [\alpha^2(1 - \alpha) - \alpha\sigma_P^2]/\sigma_P^2$$
$$v = [\alpha(1 - \alpha)^2 - (1 - \alpha)\sigma_P^2]/\sigma_P^2.$$

This distribution can now be used to establish tolerance limits. For example, it follows from (1) that for a sample size $n \geq 214$, and a tolerance region given by the ellipse $T^2 = 9.21$, then $E(P) = .99$ and the Prob.$\{.985 \leq P \leq .995\} \geq .992$.

Care must be taken in the use of these and similar results, for if the distribution is not a bivariate normal one, a large error may be introduced which will not be eliminated with increasing n; however the error will probably be small when a tolerance region is found for the means $\bar{x}$, $\bar{y}$ of a future sample of k observations ($k \geq 20$) as contrasted with a tolerance region for a single observation. An exact treatment of the case when the bivariate distribution is unknown has been given by Wald in the present issue of the *Annals of Mathematical Statistics*.

REFERENCES

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

[2] Harold Hotelling, "A generalization of Student's ratio," *Annals of Math. Stat.*, Vol. 2 (1931), pp. 360–378.

A NEW APPROXIMATION TO THE LEVELS OF SIGNIFICANCE OF THE CHI-SQUARE DISTRIBUTION.

By Leo A. Aroian

Hunter College

Recent articles on the percentage points of the χ^2 distribution [1], [2], have directed my attention to a method proposed in my investigation of Fisher's z distribution [3], a method particularly useful and easily computed for n large. In addition, this method avoids interpolation. If $t = \dfrac{\chi^2 - n}{\sqrt{2n}}$, and $\alpha_3 = \sqrt{\dfrac{8}{n}}$, the measure of skewness for the χ^2 distribution, the following formulas give significance levels of t as quadratic functions of α_3, $t = a + b\alpha_3 + c\alpha_3^2$. The values of a, b, and c were found by the usual method of least squares, fitting each formula to the values of t [4] for $\alpha_3 = 0$, ± 0.1, ± 0.2, ± 0.3, and ± 0.4. Then the value of a in each instance was adjusted to give the proper value of t when $\alpha_3 = 0$: e.g. the constant term by the method of least squares for the 1 per cent point is 2.32633 which we change to 2.32635. The range $| \alpha_3 | \leqq .4$ corresponds to $n \geqq 50$, but the formulas are quite satisfactory for $n \geqq 30$. Formulas for t when $| \alpha_3 | > .4$ [3] are easily derived, but such results while more accurate in the range