NOTES

This section is devoted to brief research and expository articles, and notes on methodology.

A NOTE ON THE BEST LINEAR ESTIMATE

By Allen T. Craig

University of Iowa

1. Introduction. Let the chance variable x be subject to the distribution function $D(x)$ and as usual let $E[g(x)]$ denote the mathematical expectation of the function $g(x)$. If $x_1, x_2, \cdots, x_n$ constitute a sample of n independent values of x, the function $y = c_1x_1 + c_2x_2 + \cdots + c_nx_n$ is frequently called the best linear estimate of $E(x)$ when the c's are so chosen that $E(y) = E(x)$, and $E[y - E(x)]^2 = \sigma_y^2$ is a minimum. It is the purpose of this note to give an example of an estimate y, best in the sense defined, yet such that, y' being another estimate,

$$Pr[E(x) - \delta \leq y \leq E(x) + \delta] \leq Pr[E(x) - \delta \leq y' \leq E(x) + \delta],$$

for every $\delta > 0$.

2. The rectangular distribution. Consider $D(x) = 1/a$, $0 \leq x \leq a$, and let the n items of each sample be arranged in ascending order of magnitude so that $x_1 \leq x_2 \leq \cdots \leq x_n$, $n \geq 2$. The generating function $G(t)$ of the moments of the distribution of $y = c_1x_1 + c_2x_2 + \cdots + c_nx_n$ is

$$G(t) = E(e^{ty}) = \frac{n!}{a^n}\int_0^a \int_0^{x_n} \int_0^{x_{n-1}} \cdots \int_0^{x_2} e^{t(c_1x_1 + \cdots + c_nx_n)}\, dx_1\, dx_2 \cdots dx_n.$$

Thus

$$E(y) = G'(0) = \frac{a}{n+1}[c_1 + 2c_2 + 3c_3 + \cdots + nc_n],$$

and

$$\begin{aligned} E(y^2) = G''(0) = \frac{a^2}{(n+1)(n+2)} & [1\cdot 2c_1^2 + 2\cdot 3c_2^2 + \cdots + n(n+1)c_n^2 \\ & + 2\{1\cdot 3c_1c_2 + 1\cdot 4c_1c_3 + \cdots + 1\cdot(n+1)c_1c_n \\ & + 2\cdot 4c_2c_3 + \cdots + 2(n+1)c_2c_n \\ & \vdots \\ & + (n-1)(n+1)c_{n-1}c_n\}]. \end{aligned}$$

From $E(y) = E(x) = a/2$, we have

$$c_1 = \tfrac{1}{2}(n+1) - 2c_2 - \cdots - nc_n.$$

Thus $\sigma_y^2 = G''(0) - a^2/4$ with c_1 in $G''(0)$ replaced by $\frac{1}{2}(n+1) - 2c_2 - \cdots - nc_n$. From $\frac{\partial \sigma_y^2}{\partial c_j} = 0$, $j = 2, 3, \cdots, n$, we obtain the following system of $n - 1$ non-homogeneous linear equations in $n - 1$ unknowns:

$$\begin{aligned}
4c_2 + 6c_3 + \cdots + 2nc_n &= n + 1 \\
6c_2 + 12c_3 + \cdots + 4nc_n &= 2(n + 1) \\
8c_2 + 16c_3 + \cdots + 6nc_n &= 3(n + 1) \\
\vdots \qquad\qquad &\qquad \vdots \\
2nc_2 + 4nc_3 + \cdots + 2n(n-1)c_n &= (n - 1)(n + 1).
\end{aligned}$$

Since the determinant of the coefficients is not zero, the solution $c_2 = c_3 = \cdots = c_{n-1} = 0$, $c_n = (n+1)/2n$, is unique. Further, we see that $c_1 = 0$ so the best linear estimate of the mean of the rectangular population is $y = (n+1)x_n/2n$, where x_n is the largest item in the sample.

The distribution function of y is readily found to be

$$D(y) = n\left[\frac{2n}{a(n+1)}\right]^n y^{n-1}, \qquad 0 \le y \le \frac{n+1}{2n}\,a.$$

From this, it follows that $\sigma_y^2 = \dfrac{a^2}{4n(n+2)}$.

It has long been known[1] that the sampling distribution of the statistic $\omega = \frac{1}{2}(x_1 + x_n)$, where x_1 and x_n are respectively the smallest and largest items in samples of size n from a rectangular population, has a smaller variance than does that of the arithmetic mean $\bar{x}$ of all n items. The distribution function of ω is

$$\begin{aligned}
D(\omega) &= \frac{2^{n-1} n \omega^{n-1}}{a^n}, & 0 \le \omega \le \tfrac{1}{2}a, \\
&= \frac{2^{n-1} n}{a^n} (a - \omega)^{n-1}, & \tfrac{1}{2}a \le \omega < a,
\end{aligned}$$

so that $E(\omega) = \frac{1}{2}a$ and $\sigma_\omega^2 = \dfrac{a^2}{2(n+1)(n+2)}$. Thus $\sigma_y^2 = \frac{1}{2}\sigma_\omega^2$, approximately. Yet Pittman has recently proved that for every $\delta > 0$, $Pr[E(x) - \delta \le \omega \le E(x) + \delta]$ exceeds the probability that any other estimate, including y, will fall in this interval of length 2δ about the mean $a/2$.

If we write $u = \dfrac{y - a/2}{\sigma_y}$ and $v = \dfrac{\omega - a/2}{\sigma_\omega}$, then the limits of $D(u)$ and $D(v)$ as n approaches infinity are respectively e^{u-1}, $-\infty \le u \le 1$, and $\dfrac{1}{\sqrt{2}}\, e^{-\sqrt{2}|v|}$,

[1] R. A. Fisher, "Theoretical foundations of mathematical statistics," *Phil. Trans. Roy. Soc. London*, Series A, Vol. 222 (1921), pp. 309–368.

$-\infty \leq v \leq \infty$. 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 $L_2(x_1x_2 \cdots x_n)$ 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

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

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.