NOTES

This section is devoted to brief research and expository articles on methodology and other short items.

NOTE ON ASYMPTOTIC VALUE OF PROBABILITY DISTRIBUTION OF SUM OF RANDOM VARIABLES WHICH ARE GREATER THAN A SET OF ARBITRARILY CHOSEN NUMBERS

By Bradford F. Kimball

New York Public Service Commission, New York City

The purpose of this note is to present the following theorem which the author needed in connection with a computational problem. Since the theorem has general implications for statistical theory which do not seem to have been brought out heretofore, readers of this journal may find it of interest.

Theorem: *In Euclidean space of n dimensions with coordinates x_i ($i = 1, 2, \cdots n$) let a_i be n constants whose sum is u_0, and consider a hyperplane*

$$x_1 + x_2 + \cdots + x_n = u, \qquad u \geq u_0. \tag{1}$$

Let k denote a positive constant, and $I(u)$ the $n - 1$ fold integral defined by

$$I(u) = \int \cdots \int \exp.\left[-k\left(\sum_{i=1}^{n} x_i^2\right)\right] dx_1\, dx_2 \cdots dx_{n-1} \tag{2}$$

taken over that part of the hyperplane for which $x_i \geq a_i$. Then

$$\lim_{u \to \infty} \sqrt{n}\, e^{ku^2/n} I(u) = (\pi/k)^{(n-1)/2}. \tag{3}$$

Proof: The integral may be reduced to another integral by reduction of the quadratic form

$$\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n-1} x_i^2 + \left(u - \sum_{i=1}^{n-1} x_i\right)^2$$

to a sum of squares y_i^2 such that y_i does not involve x_j for $j < i$. Dropping the superscript $n - 1$ in the notation for Σ and letting the subscript on Σ denote the least value of i involved in the sum, the expansion of this form may be written

$$2\sum_{1} x_i^2 - 2u\sum_{1} x_i + 2\sum_{1} x_i x_j + u^2, \qquad i < j.$$

The transformation may be performed progressively as follows:

$$(\sqrt{2}\, x_1)^2 + 2(x_2 + x_3 + \cdots - u)\, x_1 + [(u - \sum_{2} x_i)/\sqrt{2}]^2 = y_1^2$$

giving

$$y_1 = \sqrt{2}\, x_1 - (u - \sum_{2} x_i)/\sqrt{2}.$$

Using the remainder of the terms involving x_2, one can complete the square with terms not involving x_2, and the value of y_2 is

$$y_2 = x_2\sqrt{3}/\sqrt{2} - (u - \sum_3 x_i)/\sqrt{6}.$$

Continuing until only terms involving x_{n-1} and u remain,

$$y_{n-1} = x_{n-1}\sqrt{n}/\sqrt{n-1} - u/\sqrt{n(n-1)}.$$

The remaining term in u will be found to be u^2/n.

Making the above transformation, the integral becomes

$$(4) \quad I(u) = [e^{-ku^2/n}/\sqrt{n}] \int \cdots \int \exp.\left[-k\left(\sum_{i=1}^{n-1} y_i^2\right)\right] dy_1\, dy_2 \cdots dy_{n-1}.$$

In order to fix the limits of integration on y_i it will be noted that the projection of the critical region of the hyperplane (1) upon the $n-1$ dimensional space in the original variables x_i delineates a region in that $n-1$ space bounded by the $n-1$ hyperplanes

$$x_i = a_i, \qquad i = 1, 2 \cdots, n-1$$

and the hyperplane

$$x_1 + x_2 + x_3 + \cdots + x_{n-1} = u - a_n.$$

Hence, if (2) is considered as an iterated integral with the integration performed in the order of the subscripts of x_i, the intervals of integration are

$$\begin{aligned}
a_1 &\leq x_1 \leq u - a_n - \sum_2 x_i \\
a_2 &\leq x_2 \leq u - a_n - a_1 - \sum_3 x_i \\
&\cdots\cdots \\
a_r &\leq x_r \leq u - \sum_{r+1} x_i - (u_0 - \sum_{r+1} a_i) \\
&\cdots\cdots \\
a_{n-1} &\leq x_{n-1} \leq u - u_0 + a_{n-1},
\end{aligned}$$

where it is recalled that $\sum_{r+1}$ denotes $\sum_{i=r+1}^{n-1}$.

Now transforming to y_i and using the general transformation equation

$$(5) \qquad y_r = x_r\sqrt{r+1}/\sqrt{r} - (u - \sum_{r+1} x_i)/\sqrt{(r+1)r}, \qquad 1 \leq r \leq n-1$$

where $\sum_n x_i = 0$ under definition of summation symbol noted above,

Lower limit of $y_r = -(u - \sum_{r+1} x_i)/\sqrt{(r+1)r} + a_r\sqrt{r+1}/\sqrt{r}$,

Upper limit of $y_r = (u - \sum_{r+1} x_i)\sqrt{r}/\sqrt{r+1} - (u_0 - \sum_{r+1} a_i)\sqrt{r+1}/\sqrt{r}$,

since

$$\sqrt{r+1}/\sqrt{r} - 1/\sqrt{(r+1)r} = \sqrt{r}/\sqrt{r+1}.$$

It is not difficult to show that

$$\sum_{r+1} x_i = C_{r+1}(y_i) + (n - 1 - r)u/n$$

where $C_{r+1}(y_i)$ denotes a linear combination of y_i for $i \geq r + 1$, which does not involve the variable u. In other words

$$u - \sum_{r+1} x_i = (r + 1)u/n - C_{r+1}(y_i).$$

Making this substitution, the limits on y_r are found to be

$$\text{Lower limit of } y_r = -(u/n)(\sqrt{r+1}/\sqrt{r}) + C_{r+1}(y_i)/\sqrt{(r+1)r} + a_r\sqrt{r+1}/\sqrt{r},$$

$$\text{Upper limit of } y_r = (u/n)(\sqrt{r(r+1)} - C_{r+1}(y_i)\sqrt{r}/\sqrt{r+1} - (u_0 - \sum_{r+1} a_i)\sqrt{r+1}/\sqrt{r}.$$

It will now be clear that as u becomes infinite, the limit of the integral in (4), will be the $n - 1$ fold integral taken over the whole of $n - 1$ space. This latter integral is easily evaluated to give $(\pi/k)^{(n-1)/2}$, and the theorem follows.

The following two corollaries, which are restatements of the theorem in less general form, bring out the implications for statistical theory.

COROLLARY 1. *With $k = 1/2$ define $F_n(u)$ by*

$$F_n(u) = (2\pi)^{-n/2} I(u). \tag{6}$$

The differential $F_n(u)du$ represents then the probability that n random variables x_i taken from a normally distributed population with zero mean and unit standard deviation, fall into a region

$$a_i \leq x_i \tag{7}$$

and have a sum with value in the neighborhood $(\pm\frac{1}{2}\, du)$ of u.

Recalling that u_0 is the value of u when each x_i has value a_i

$$P[a_i \leq x_i] = \int_{u_0}^{\infty} F_n(u)\, du \tag{8}$$

is the probability that all values of x fall into the region (7). *Denoting the normal probability function by $\phi(t)$, corollary 1 implies that*

$$\lim_{u \to \infty} \sqrt{n}\, F_n(u)/\phi(u/\sqrt{n}) = 1. \tag{9}$$

Since $\phi(u/\sqrt{n})du/\sqrt{n}$ represents the probability distribution of the sum of the n random variables when the condition (7) is removed, certain implications for the theory of statistics emerge. One of these is noted in the example given below. Corollary 1 can be stated in a different form as follows:

COROLLARY 2. *If $p_A(u)du$ denotes the probability that the sum of n random varia-*

bles from a normally distributed population be in the neighborhood ($\pm\frac{1}{2}\,du$) *of* u, *and* $p_B(u)$ *denotes the probability that the sum of* n *random variables from the same population that do fall in region* (7) *have a value in this neighborhood of* u, *it follows from* (9) *and the nature of the functions integrated that for arbitrarily small positive* δ, *a value of* u, *say* u' *can be found, sufficiently large, such that for all* $u \geq u'$,

$$(10) \qquad (1 - \delta)\, p_A(u) \leq p_B(u)\, P[a_i \leq x_i] \leq p_A(u).$$

Rate of convergence. The author having had occasion to compute $F_n(u)$ for values of n from 2 to 5, and $a_i = 0$, a table showing the rate of convergence of $F_n(u)$ to its limit for this range of n (and $a_i = 0$) is shown below. In this table the values of the ratios of the minimum values of u ($= u'$) to the standard deviation of u ($= \sqrt{n}$) are shown for a sequence of values of δ which approaches zero.

Rate of Convergence of $\sqrt{n}\, F_n(u)/\phi(u\sqrt{n})$ *to Unity for* $a_i = 0$

	Critical Ratio $u'/\sqrt{n}$			
δ	$n = 2$	$n = 3$	$n = 4$	$n = 5$
.5	0.67	1.34	1.92	2.46
.25	1.15	1.95	2.65	3.27
.1	1.64	2.60	3.40	4.11
.05	1.96	3.02	3.89	4.65
.01	2.58	3.85	4.87	5.76
.001	3.29	4.84	6.03	7.08

Example. An example showing the possible bearing of this theorem upon practical considerations of sampling is the following: If in a quality control problem, samples of size 4 were used, and a follow-up of samples which showed a large deviation of sample mean were pursued, a case of a particularly large deviation such as 4 would possibly receive special attention. With sample mean equal to $u/4$, the ratio of this mean to its standard deviation is $u/2$. Turning to the table, in column $n = 4$ it will be noted that the value of δ for $u'/2 = 4$ is somewhat less than .05. Hence from (9) and (10), for $u \geq u'$ ($u' = 8$, $n = 4$),

$$.95\, \phi(u/\sqrt{n})/\sqrt{n} < F_n(u) \leq \phi(u/\sqrt{n})/\sqrt{n}.$$

It follows that

$$.95 \int_{u'}^{\infty} \phi(u/\sqrt{n})\, du/\sqrt{n} < \int_{u'}^{\infty} F_n(u)\, du \leq \int_{u'}^{\infty} \phi(u/\sqrt{n})\, du/\sqrt{n}.$$

If one now considers a set of random samples of size 4, the last integral on the right represents the expected proportion of the set which falls into the sub-set A for which $u \geq u'$ (and hence with deviation of mean relative to standard deviation of mean greater than $u'/\sqrt{n}$). The middle integral represents the expected proportion of the original set which falls into a sub-set B for which $u \geq u'$ *and* $x_i \geq 0$. It follows from the inequality on the left that the expected proportion of the sub-set A which falls into the sub-set B is greater than 95 per cent. Hence

one infers that the probability is greater than .95 that for a sample showing such a large deviation from the mean ($u/\sqrt{n} = 4, n = 4$) *all* the constituent elements will have deviations on the same side of the population mean. Thus if all the elements of the sample investigated are found to have deviations on the same side of the population mean, this could *not* be construed as *additional evidence* that the sample indicated an abnormal condition.

This conclusion is weaker than the facts of the example warrant, since it is based upon the *integral* of $F_n(u)$ from u' to infinity. Unfortunately the author does not have data available on the rate of convergence of these integrals.

NOTE ON A MATRIC THEOREM OF A. T. CRAIG

By Harold Hotelling

Columbia University

An extremely elegant theorem given recently by A. T. Craig[1] and applied by him to establish a further theorem on independent χ^2 distributions may be stated as follows:

If A and B are the symmetric matrices of two homogeneous quadratic forms in n variates which are normally and independently distributed with zero means and unit variances, a necessary and sufficient condition for the independence in probability of these two forms is that $AB = 0$.

The proof given that the condition is sufficient is adequate, but Craig's treatment of its necessity consists essentially in its assertion. In view of the growing interest in such quadratic forms, for example in connection with serial correlation, the neatness of this theorem is likely to lead to a wide usefulness. It therefore seems worth while to give a complete proof of the necessity condition.

The form with matrix A is denoted by Q_1 and that with matrix B by Q_2. The characteristic functions, if defined as $Ee^{\frac{1}{2}\lambda Q_1}$ and $Ee^{\frac{1}{2}\mu Q_2}$, are respectively the reciprocals of the square roots of the determinants of the matrices $1 - \lambda A$ and $1 - \mu B$, while the characteristic function for Q_1 and Q_2 together, $Ee^{\frac{1}{2}(\lambda Q_1 + \mu Q_2)}$, is the reciprocal of the square root of the determinant of $1 - \lambda A - \mu B$. A necessary and sufficient condition for independence is therefore that

$$| 1 - \lambda A | \cdot | 1 - \mu B | \equiv | 1 - \lambda A - \mu B |$$

shall hold identically for all values of λ and μ. Since the determinant of the product of two matrices is the product of their determinants, the left member is the same as

$$| 1 - \lambda A - \mu B + \lambda\mu AB | .$$

From this it is immediately obvious that $AB = 0$ implies the independence of the two forms. The converse will now be proved.

---

[1] "Note on the independence of certain quadratic forms," *Annals of Math. Stat.*, Vol. 14 (1943), pp. 195–197.