AN EXTENSION OF WILKS' METHOD FOR SETTING TOLERANCE LIMITS

By Abraham Wald

Columbia University

1. Introduction. Let x be a random variable and let $f(x)$ be its probability density function. Suppose that nothing is known about $f(x)$ except that it is continuous. Let $x_1, \cdots, x_n$ be n independent observations on x. The problem of setting tolerance limits can be formulated as follows: *For some given positive values* $\beta < 1$ and $\gamma < 1$ *we have to construct two functions* $L(x_1, \cdots, x_n)$ *and* $M(x_1, \cdots, x_n)$, *called tolerance limits, such that the probability that*

$$\int_L^M f(t)\,dt \geq \gamma, \tag{1}$$

holds, is equal to β. This problem has recently been solved by S. S. Wilks[1] in a very satisfactory way when nothing is known about $f(x)$ except that it is continuous. Wilks proposes the following solution: Let $x_1, \cdots, x_n$ be the observed values of x arranged in order of increasing magnitude. Then $L = x_r$ and $M = x_{n-r+1}$ where r denotes a positive integer. The exact sampling distribution of the statistic $\int_{x_r}^{x_{n-r+1}} f(t)\,dt$ is derived by Wilks and this provides the solution for the problem of setting tolerance limits. A very important feature of Wilks' solution is the fact that the distribution of $\int_{x_r}^{x_{n-r+1}} f(t)\,dt$ is entirely independent of the unknown density function $f(x)$, i.e. the distribution of $\int_{x_r}^{x_{n-r+1}} f(t)\,dt$ is the same for any arbitrary continuous density function $f(x)$.

In this paper we shall give an extension of Wilks' method to the multivariate case. Let $x_1, \cdots, x_p$ be a set of p random variables with the joint probability density function $f(x_1, \cdots, x_p)$. Suppose that nothing is known about $f(x_1, \cdots, x_p)$ except that it is a continuous function of $x_1, \cdots, x_p$. A sample of n independent observations is drawn and the α-th observation on x_i is denoted by $x_{i\alpha}$ $(i = 1, \cdots, p;\ \alpha = 1, \cdots, n)$. The problem of setting tolerance limits for $x_1, \cdots, x_p$ can be formulated as follows: *For some given positive values* $\beta < 1$ *and* $\gamma < 1$ *we have to construct* p *pairs of functions of the observations* $L_i(x_{11}, \cdots, x_{pn})$ *and* $M_i(x_{11}, \cdots, x_{pn})$ $(i = 1, \cdots, p)$ *such that the probability that*

$$\int_{L_p}^{M_p} \cdots \int_{L_1}^{M_1} f(t_1, \cdots, t_p)\,dt_1 \cdots dt_p \geq \gamma, \tag{2}$$

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

holds, is equal to β. The functions L_i and M_i are called the lower and upper tolerance limits of x_i. A natural extension of Wilks' procedure would seem to be the following: Let $x_{i1}, \cdots, x_{in}$ be the observations on x_i arranged in order of increasing magnitude and let $L_i = x_{ir_i}$ and $M_i = x_{is_i}$ $(i = 1, \cdots, p)$ where r_i and s_i denote some integers. However, this choice of the tolerance limits does not provide a satisfactory solution of our problem, since the distribution of (2) is *not* independent of the unknown density function $f(x_1, \cdots, x_p)$. It will be shown in this paper that by a slight modification of the above procedure the distribution of (2) becomes entirely independent of the unknown density function $f(x_1, \cdots, x_p)$. In section 2 we will treat the bivariate case and in section 3 we will extend the results to multivariate distributions.

2. The bivariate case. In this section we deal with the case when $p = 2$. Let $x_{11}, \cdots, x_{1n}$ be the observations on x_1 arranged in order of increasing magnitude. We may assume that $x_{11} < x_{12} < \cdots < x_{1n}$ since the probability of an equality sign is equal to zero. We define

$$L_1 = x_{1r_1} \quad \text{and} \quad M_1 = x_{1s_1}, \tag{3}$$

where r_1 and s_1 denote some positive integers and $r_1 < s_1 \leq n$. Consider only those sample points $(x_{1\alpha}, x_{2\alpha})$ for which $x_{1r_1} \leq x_{1\alpha} < x_{1s_1}$, i.e. consider the sample points $(x_{1,r_1+1}, x_{2,r_1+1}), \cdots, (x_{1,s_1-1}, x_{2,s_1-1})$. Denote by $x'_{2,r_1+1}, \cdots, x'_{2,s_1-1}$ the values $x_{2,r_1+1}, \cdots, x_{2,s_1-1}$ arranged in order of increasing magnitude. We define

$$L_2 = x'_{2r_2} \quad \text{and} \quad M_2 = x'_{2s_2}, \tag{4}$$

where r_2 and s_2 denote some positive integers for which $r_2 < s_2 \leq s_1 - r_1 - 1$.

We will show that the distribution of the statistic

$$Q = \int_{L_2}^{M_2} \int_{L_1}^{M_1} f(t_1, t_2)\, dt_1\, dt_2, \tag{5}$$

is entirely independent of the unknown density function $f(x_1, x_2)$. Denote by $\varphi(x_1)$ the marginal distribution of x_1, i.e.

$$\varphi(x_1) = \int_{-\infty}^{+\infty} f(x_1, x_2)\, dx_2. \tag{6}$$

Furthermore denote by $\psi(x_2, L_1, M_1)$ the conditional distribution of x_2 calculated under the condition that $L_1 < x_1 < M_1$. Hence

$$\psi(x_2, L_1, M_1) = \frac{\int_{L_1}^{M_1} f(x_1, x_2)\, dx_1}{\int_{-\infty}^{+\infty} \int_{L_1}^{M_1} f(x_1, x_2)\, dx_1\, dx_2}. \tag{7}$$

Let

$$P = \int_{L_1}^{M_1} \varphi(t)\, dt \tag{8}$$

and

$$\bar{P} = \int_{L_2}^{M_2} \psi(t, L_1, M_1)\, dt \tag{9}$$

From (5), (8) and (9) it follows that

$$Q = P\bar{P}. \tag{10}$$

It is obvious that the distribution of P is given by Wilks' formula. Since Wilks derived the distribution only when $s_1 = n - r_1 + 1$, we will briefly give here the derivation for any integers r_1 and s_1.

Let $\int_{-\infty}^{x_{1r_1}} \varphi(t)\, dt = u$ and $\int_{x_{1s_1}}^{\infty} \varphi(t)\, dt = v$. Then the joint probability density function of u and v is given by

$$cu^{r_1-1}(1 - u - v)^{s_1-r_1-1} v^{n-s_1}\, du\, dv, \tag{11}$$

where c is a constant. We obviously have $P = 1 - u - v$. The joint density function of P and u is given by

$$cu^{r_1-1}P^{s_1-r_1-1}(1 - u - P)^{n-s_1}\, du\, dP, \tag{12}$$

where u is restricted to the interval $[0, 1 - P]$. Hence the distribution of P is given by

$$\begin{aligned} cP^{s_1-r_1-1} &\int_0^{1-P} u^{r_1-1}(1 - u - P)^{n-s_1}\, du \\ &= cP^{s_1-r_1-1}(1 - P)^{n-s_1+r_1-1} \int_0^{1-P} \left(\frac{u}{1-P}\right)^{r_1-1} \left(1 - \frac{u}{1-P}\right)^{n-s_1} du \\ &= cP^{s_1-r_1-1}(1 - P)^{n-s_1+r_1} \int_0^1 T^{r_1-1}(1 - T)^{n-s_1}\, dT \\ &= c'P^{s_1-r_1-1}(1 - P)^{n-s_1+r_1}. \end{aligned}$$

Since the integral of the density function of P over the range of P must be equal to 1, we find that

$$c' = \Gamma(n + 1)/\Gamma(s_1 - r_1)\Gamma(n - s_1 + r_1 + 1).$$

Hence the probability density function of P is given by

$$\frac{\Gamma(n + 1)}{\Gamma(s_1 - r_1)\Gamma(n - s_1 + r_1 + 1)} P^{s_1-r_1-1}(1 - P)^{n-s_1+r_1}\, dP. \tag{13}$$

Since $x_{2,r_1+1}, \cdots, x_{2,s_1-1}$ can be considered as $s_1 - r_1 - 1$ independent observations on a random variable t the distribution of which is given by $\psi(t, L_1, M_1)\, dt$, for any given values L_1, M_1 the conditional distribution of $\bar{P}$ is given by the expression we obtain from (13) by substituting r_2 for r_1, s_2 for s_1 and $s_1 - r_1 - 1$ for n. Hence the conditional distribution of $\bar{P}$ is given by

$$(14) \qquad \frac{\Gamma(s_1 - r_1)}{\Gamma(s_2 - r_2)\Gamma(s_1 - r_1 - s_2 + r_2)} (\bar{P})^{s_2 - r_2 - 1}(1 - \bar{P})^{s_1 - r_1 - 1 - s_2 + r_2}\, d\bar{P}.$$

Since the expression (14) does not involve the quantities L_1 and M_1, $\bar{P}$ is distributed independently of L_1 and M_1. Hence the joint density function of P and $\bar{P}$ is given by the product of (13) and (14), i.e. by

$$(15) \qquad AP^{s_1 - r_1 - 1}(1 - P)^{n - s_1 + r_1}(\bar{P})^{s_2 - r_2 - 1}(1 - \bar{P})^{s_1 - r_1 - 1 - s_2 + r_2}\, dP\, d\bar{P},$$

where A denotes the product of the constant coefficients in (13) and (14). From (15) it follows that the joint distribution of P and $Q = P\bar{P}$ is given by

$$(16) \qquad A(1 - P)^{n - s_1 + r_1} Q^{s_2 - r_2 - 1}(P - Q)^{s_1 - r_1 - 1 - s_2 + r_2}\, dP\, dQ.$$

Since the range of P is the interval $[Q, 1]$, the distribution of Q is given by

$$(17) \qquad AQ^{s_2 - r_2 - 1} \int_Q^1 (1 - P)^{n - s_1 + r_1}(P - Q)^{s_1 - r_1 - 1 - s_2 + r_2}\, dP.$$

Let $R = P - Q$. Then we have

$$\begin{aligned} &\int_Q^1 (1 - P)^{n - s_1 + r_1}(P - Q)^{s_1 - r_1 - 1 - s_2 + r_2}\, dP \\ &\quad = \int_0^{1-Q} (1 - Q - R)^{n - s_1 + r_1} R^{s_1 - r_1 - 1 - s_2 + r_2}\, dR \qquad (18) \\ &\quad = (1 - Q)^{n - 1 - s_2 + r_2}(1 - Q) \int_0^1 (1 - T)^{n - s_1 + r_1} T^{s_1 - r_1 - 1 - s_2 + r_2}\, dT. \end{aligned}$$

From (17) and (18) it follows that the probability density function of Q is given by

$$(19) \qquad \frac{\Gamma(n + 1)}{\Gamma(s_2 - r_2)\Gamma(n - s_2 + r_2 + 1)} Q^{s_2 - r_2 - 1}(1 - Q)^{n - s_2 + r_2}\, dQ.$$

3. The multivariate case. We may assume that no two elements of the matrix $\| x_{i\alpha} \|$ $(i = 1, \cdots, p;\ \alpha = 1, \cdots, n)$ are equal, since the probability of this event is equal to 1. For each α let $t_\alpha (\alpha = 1, \cdots, n)$ be the point with the coordinates $x_{1\alpha}, \cdots, x_{p\alpha}$. Let $x_{11}, \cdots, x_{1n}$ be the observations on x_1 arranged in order of increasing magnitude. Then $L_1 = x_{1r_1}$ and $M_1 = x_{1s_1}$. The quantities L_i and M_i $(i = 2, \cdots, p)$ are defined in the following manner: Let S be the set of all points t_α for which

$$L_j < x_{j\alpha} < M_j \qquad (j = 1, \cdots, i - 1).$$

Arrange the i-th coordinates of the points in S in order of increasing magnitude. Then L_i is equal to the r_i-th element and M_i is equal to the s_i-th element of this ordered sequence. We will derive the distribution of

$$(20) \qquad Q_p = \int_{L_p}^{M_p} \cdots \int_{L_1}^{M_1} f(x_1, \cdots, x_p)\, dx_1 \cdots dx_p.$$

Let

$$Q_i = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \int_{L_i}^{M_i} \cdots \int_{L_1}^{M_1} f(x_1, \cdots, x_p)\, dx_1 \cdots dx_p \tag{21}$$

$$(i = 1, \cdots, p-1).$$

Denote by $\varphi_i(x_i, L_1, M_1, \cdots, L_{i-1}, M_{i-1})$ $(i = 2, \cdots, p)$ the conditional probability density function of x_i calculated under the condition that $L_j \leq x_j \leq M_j$ $(j = 1, \cdots, i-1)$. Let

$$\bar{P}_i = \int_{L_i}^{M_i} \varphi_i(x_i, L_1, M_1, \cdots, L_{i-1}, M_{i-1})\, dx_i. \tag{22}$$

We obviously have

$$Q_{i+1} = Q_i \bar{P}_{i+1} \qquad (i = 1, \cdots, p-1). \tag{23}$$

We will prove that the probability density function of Q_i is given by

$$\frac{\Gamma(n+1)}{\Gamma(s_i - r_i)\Gamma(n - s_i + r_i + 1)} Q_i^{s_i - r_i - 1}(1 - Q_i)^{n - s_i + r_i}\, dQ_i, \quad (i = 1, \cdots, p). \tag{24}$$

This is certainly true for $i = 1, 2$. We will assume that it is true for $i = j$ and we will prove it for $i = j + 1$. It is easy to see that Q_j and $\bar{P}_{j+1}$ are independently distributed and that the probability density function of $\bar{P}_{j+1}$ is given by

$$\frac{\Gamma(s_j - r_j)}{\Gamma(s_{j+1} - r_{j+1})\Gamma(s_j - r_j - s_{j+1} + r_{j+1})} \cdot (\bar{P}_{j+1})^{s_{j+1} - r_{j+1} - 1}(1 - \bar{P}_{j+1})^{s_j - r_j - 1 - s_{j+1} + r_{j+1}}\, d\bar{P}_{j+1}. \tag{25}$$

The joint distribution of Q_j and $\bar{P}_{j+1}$ is of the same form as the joint distribution of P and $\bar{P}$ in section 2. Hence the distribution of $Q_j \bar{P}_{j+1}$ can be obtained from the distribution of $Q = P\bar{P}$ by substituting r_{j+1} for r_2 and s_{j+1} for s_2. The distribution of Q is given in (19). Making the above substitution in formula (19) we obtain formula (24) for $i = j + 1$. Hence the validity of (24) is proved for $i = 1, 2, \cdots, p$. In particular, the distribution of Q_p is given by

$$\frac{\Gamma(n+1)}{\Gamma(s_p - r_p)\Gamma(n - s_p + r_p + 1)} Q_p^{s_p - r_p - 1}(1 - Q_p)^{n - s_p + r_p}\, dQ_p. \tag{26}$$

It is interesting to note that the distribution of Q_p does not depend on the integers $r_1, s_1, \cdots, r_{p-1}, s_{p-1}$. The construction of the tolerance limits L_i, M_i $(i = 1, \cdots, p)$, as proposed here, is somewhat asymmetric, since it depends on the order of the variates $x_1, \cdots, x_p$. In practical applications the asymmetry of the construction will be very slight, since in most practical cases the integers r_p and s_p will be chosen so that $(s_p - r_p - 1)/n$ will be near to 1. If, for example, $(s_p - r_p - 1)/n \geq .95$, the tolerance limits will be affected only very slightly by a permutation of the variates $x_1, \cdots, x_p$. However, it would be desirable to find a construction which is entirely independent of the order of the variates $x_1, \cdots, x_p$.

4. Tolerance regions composed of several rectangles. For the sake of simplicity we will consider here the bivariate case. All results obtained in this section can be extended without any difficulty to the multivariate case.

In section 2 the tolerance region has been a single rectangle in the plane (x_1, x_2) determined by the four lines $x_1 = L_1$, $x_1 = M_1$; $x_2 = L_2$ and $x_2 = M_2$. If the variates x_1 and x_2 are strongly correlated, a tolerance region of rectangle shape seems to be unfavorable, since it will cover an unnecessarily large area in the (x_1, x_2) plane. The situation is illustrated in figure 1 where the scatter of a

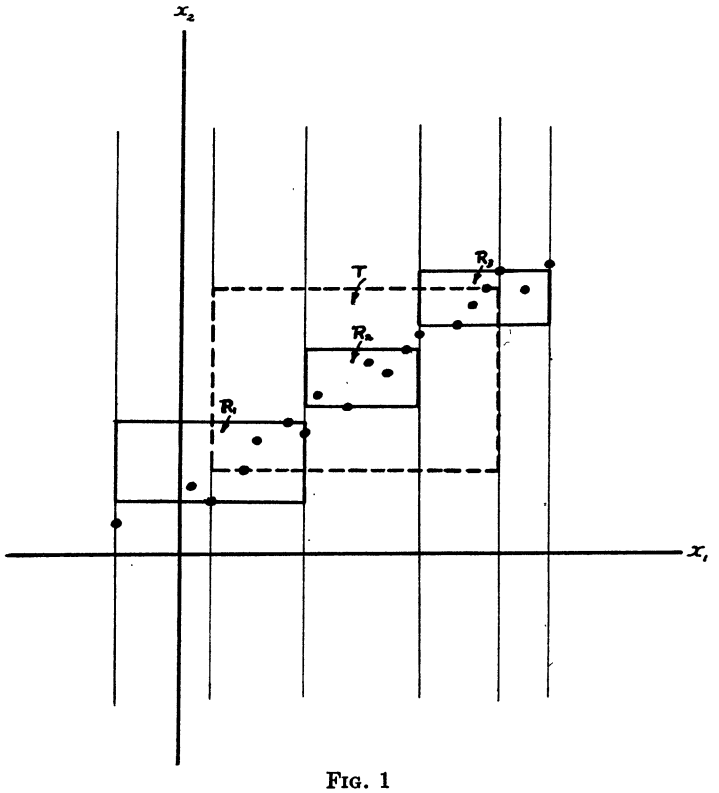

FIG. 1

bivariate sample of size $n = 19$ is shown. Suppose we choose $r_1 = 3$, $s_1 = 17$; $r_2 = 1$, $s_2 = 13$, then the tolerance region T, as defined in section 2, will be the rectangle determined by the lines $x_1 = L_1 = x_{1,3}$; $x_1 = M_1 = x_{1,17}$; $x_2 = L_2 = x'_{2,1}$; and $x_2 = M_2 = x'_{2,13}$. Now consider the tolerance region T' consisting of 3 small rectangles R_1, R_2 and R_3 defined as follows:

The rectangle R_1 is determined by the vertical lines through $x_{1,1}$ and $x_{1,7}$ and the horizontal lines through the sample points with smallest and largest ordinate, restricting ourselves to points which have abscissa values in the interior of the interval $[x_{1,1}, x_{1,7}]$. Similarly R_2 is determined by the vertical lines through

$x_{1,7}$ and $x_{1,13}$ and the horizontal lines through the sample points with largest and smallest ordinate, restricting ourselves to points with abscissa values in the interior of $[x_{1,7}, x_{1,13}]$. Finally R_3 is determined by the vertical lines through $x_{1,13}$ and $x_{1,19}$ and the horizontal lines through the sample points with largest and smallest ordinate, restricting ourselves to points whose abscissa values lie in the interior of $[x_{1,13}, x_{1,19}]$. The region T' consisting of the rectangles R_1, R_2 and R_3 has a much smaller area than the region T. As we will see later, the probability distribution of the statistic $\iint_{T'} f(x_1, x_2)dx_1\,dx_2$ is exactly the same as that of $\iint_{T} f(x_1, x_2)dx_1\,dx_2$. Thus the use of T' may be preferred to that of T.

We will consider tolerance regions T^* of the following general shape: Let $m_1, \cdots, m_k$ be k positive integers such that $1 \leq m_1, m_k \leq n$ and $m_{i+1} - m_i \geq 3$ where n is the size of the bivariate sample. Let V_i be the vertical line in the (x_1, x_2) plane given by the equation $x_1 = x_{1,m_i}$ $(i = 1, \cdots, k)$. The number of sample points which lie between the vertical lines V_i and V_{i+1} is obviously equal to $m_{i+1} - m_i - 1$. Through each point which lies between the vertical lines V_i and V_{i+1} we draw a horizontal line. In this way we obtain $m_{i+1} - m_i - 1$ horizontal lines $W_{i,1}, \cdots, W_{i,m_{i+1}-m_i-1}$ where the line $W_{i,j+1}$ is above the line $W_{i,j}$. Denote by R_{ij} $(i = 1, \cdots, k-1; j = 1, \cdots, m_{i+1} - m_i - 2)$ the rectangle determined by the lines V_i, V_{i+1}, $W_{i,j}$, $W_{i,j+1}$. Let T^* be a region composed of s different rectangles R_{ij}. The regions T and T' in the example illustrated in figure 1 are special cases of the type of regions T^* as described above. For the region T we have $k = 2$, $m_1 = 3$, $m_2 = 17$, $s = 12$, and for the region T' we have $k = 4$, $m_1 = 1$, $m_2 = 7$, $m_3 = 13$, $m_4 = 19$ and $s = 12$.

Let Q^* be given by $\iint_{T^*} f(x_1, x_2)dx_1\,dx_2$. We will prove that the probability density function of Q^* is given by

$$\frac{\Gamma(n+1)}{\Gamma(s)\Gamma(n-s+1)} Q^{*s-1}(1-Q^*)^{n-s}\,dQ^*. \tag{27}$$

Let $f_i(x_2)\,dx_2$ be the conditional distribution of x_2 under the restriction that $x_{1,m_i} < x_1 < x_{1,m_{i+1}}$. Thus, we have

$$f_i(x_2) = \frac{\int_{x_{1,m_i}}^{x_{1,m_{i+1}}} f(x_1, x_2)\,dx_1}{\int_{-\infty}^{+\infty}\int_{x_{1,m_i}}^{x_{1,m_{i+1}}} f(x_1, x_2)\,dx_1\,dx_2}. \tag{28}$$

Denote by $\varphi(x_1)\,dx_1$ the marginal distribution of x_1, i.e.

$$\varphi(x_1) = \int_{-\infty}^{+\infty} f(x_1, x_2)\,dx_2$$

Let

$$P_i^* = \int_{x_{1,m_i}}^{x_{1,m_{i+1}}} \varphi(x_1)\, dx_1 \qquad (i = 1, \cdots, k-1) \tag{29}$$

and

$$\overline{P}_i^* = \sum_j \int_{a_{ij}}^{b_{ij}} f_i(x_2)\, dx_2 \qquad (i = 1, \cdots, k-1) \tag{30}$$

where a_{ij} is the ordinate of the lower corners and b_{ij} is the ordinate of the upper corners of the rectangle R_{ij} and the summation is to be taken over all values of j for which R_{ij} is included in T^*. It is clear that

$$Q^* = P_1^* \overline{P}_1^* + \cdots + P_{k-1}^* \overline{P}_{k-1}^* . \tag{31}$$

Let y be any random variable which has a continuous probability density function, say $\psi(y)\, dy$. Furthermore let $y_1, \cdots, y_n$ be n independent observations on y. Let $\psi_i(y)\, dy$ be the conditional density function of y under the condition that y is restricted to the interval $[y_{m_i}, y_{m_{i+1}}]$. Let

$$P' = \sum_{i,j} \int_{y_{m_i+j}}^{y_{m_i+j+1}} \psi(y)\, dy \tag{32}$$

where the summation is taken over all pairs i, j for which R_{ij} is contained in T^* Let

$$P_i' = \int_{y_{m_i}}^{y_{m_{i+1}}} \psi(y)\, dy, \qquad \text{and}$$

$$\overline{P}_i' = \sum_j \int_{y_{m_i+j}}^{y_{m_i+j+1}} \psi_i(y)\, dy,$$

where the summation is to be taken over all values j for which R_{ij} is contained in T^*. We obviously have

$$P' = P_1' \overline{P}_1' + \cdots + P_{k-1}' \overline{P}_{k-1}' . \tag{33}$$

It is easy to verify that (i) the joint distribution of $P_1', \cdots, P_{k-1}'$ is the same as the joint distribution of $P_1^*, \cdots, P_{k-1}^*$; (ii) the distribution of $\overline{P}_i'$ is the same as that of $\overline{P}_i^*$ $(i = 1, \cdots, k-1)$; (iii) the variates $\overline{P}_1', \cdots, \overline{P}_{k-1}'$ are independent of each other and also of $P_1', \cdots, P_{k-1}'$; (iv) the variates $\overline{P}_1^*, \cdots, \overline{P}_{k-1}^*$ are independent of each other and also of $P_1^*, \cdots, P_{k-1}^*$. Hence it follows from (31) and (33) that the distribution of Q^* is the same as that of P'. Now we will derive the distribution of P'. The expression P' can be written in the following form:

$$P' = \sum_{i=1}^{l} \int_{y_{r_i}}^{y_{s_i}} \psi(y)\, dy, \tag{34}$$

where $r_1, s_1, \cdots, r_l, s_l$ are some positive integers for which $1 \leq r_1 < s_1 < r_2 < s_2 < \cdots < r_l < s_l \leq n$. Let

$$P'' = \sum_{i=1}^{l-1} \int_{y_{r_i}}^{y_{s_i}} \psi(y)\,dy + \int_{y_{s_{l-1}}}^{y_{s_{l-1}+s_l-r_l}} \psi(y)\,dy$$

$$= \sum_{i=1}^{l-2} \int_{y_{r_i}}^{y_{s_i}} \psi(y)\,dy + \int_{y_{r_{l-1}}}^{y_{s_{l-1}+s_l-r_l}} \psi(y)\,dy. \tag{35}$$

For any fixed value $y_{s_{l-1}}$ denote by $\psi_1(y)$ the conditional probability density of y under the restriction that $y < y_{s_{l-1}}$ and by $\psi_2(y)$ the conditional distribution of y under the restriction that $y > y_{s_{l-1}}$. Let

$$P = \int_{-\infty}^{y_{s_{l-1}}} \psi(y)\,dy \qquad P_1 = \sum_{i=1}^{l-1} \int_{y_{r_i}}^{y_{s_i}} \psi_1(y)\,dy;$$

$$P_2 = \int_{y_{r_l}}^{y_{s_l}} \psi_2(y)\,dy \quad \text{and} \quad P_3 = \int_{y_{s_{l-1}}}^{y_{s_{l-1}+s_l-r_l}} \psi_2(y)\,ay.$$

Then it follows from (34) and (35) that

$$P' = PP_1 + (1 - P)P_2,$$

$$P'' = PP_1 + (1 - P)P_3. \tag{36}$$

For calculating the distributions of P_2 and P_3 we may consider the variates $y_{s_{l-1}+1}, \cdots, y_n$ as $n - s_{l-1}$ independent observations drawn from a population which has the distribution $\psi_2(y)\,dy$. Hence, the distribution of P_2 can be derived from (13) and it is easy to verify that the distribution of P_3 is the same as that of P_2. It is clear that P_2 is independent of P and P_1. Similarly P_3 is independent of P and P_1. Hence, because of (36) the distribution of P' must be the same as that of P''.

In the same way we find that the distribution of P'' is the same as the distribution of

$$P''' = \sum_{i=1}^{l-3} \int_{y_{r_i}}^{y_{s_i}} \psi(y)\,dy + \int_{y_{r_{l-2}}}^{y_{s_{l-2}+s_{l-1}+s_l-r_l-r_{l-1}}} \psi(y)\,dy$$

Thus, by induction we see that the distribution of P' is the same as the distribution of the statistic $P_0 = \int_{y_{r_1}}^{y_{r_1+s}} \psi(y)\,dy$ where $s = \sum_{i=1}^{l} (s_i - r_i)$. From (13) it follows that the distribution of P_0 is given by

$$\frac{\Gamma(n+1)}{\Gamma(s)\Gamma(n-s+1)} P_0^{s-1}(1 - P_0)^{n-s}\,dP_0.$$

Hence, we have proved that the distribution of Q^* is given by (27).

5. Summary of the results and numerical illustrations. I shall give here a summary of the results obtained and a few illustrative examples. The multivariate case being a straightforward extension of the bivariate case, I shall discuss merely the latter. Consider a pair of random variables x and y. Denote by $f(x, y)dx\,dy$ the joint probability density function of x and y and suppose that nothing is known about $f(x, y)$ except that it is continuous. A sample of n pairs of independent observations $(x_1, y_1), \cdots, (x_n, y_n)$ is drawn from this bivariate population. The sample can be represented by n points $p_1, \cdots, p_n$ in the plane (x, y), p_i being the point with the coordinates x_i and y_i. In section 2 we have dealt with the problem of finding a rectangle T in the plane (x, y), called tolerance region, such that we can state with high probability, say with probability .98 or .99, that the proportion Q of the bivariate universe included in the rectangle T is not less than a given number b, say not less than .98 or .99. The rectangle T is constructed as follows: Suppose that the points $p_1, \cdots, p_n$ are arranged in order of increasing magnitude of their abscissa values, i.e. $x_1 < x_2 < \cdots < x_n$. We draw a vertical line V_{r_1} through the point p_{r_1} and a vertical line V_{s_1} through p_{s_1} where r_1 and s_1 are positive integers such that $1 \leq r_1$, $r_1 \leq s_1 - 3$ and $s_1 \leq n$. We consider the set S consisting of the points $p_{r_1+1}, \cdots p_{s_1-1}$ which lie between the vertical lines V_{r_1} and V_{s_1}. We draw a horizontal line H_{r_2} through the point of S which has the r_2-th smallest ordinate in S. Finally a horizontal line H_{s_2} is drawn through the point of S which has the s_2-th smallest ordinate in S. The values r_2 and s_2 are positive integers for which $r_2 < s_2$. The tolerance region T is the rectangle determined by the lines V_{r_1}, V_{s_1}, H_{r_2} and H_{s_2}. The probability p that at least the porportion $b(0 < b < 1)$ of the universe is included in T is given by

$$p = \int_b^1 \frac{\Gamma(n+1)}{\Gamma(s_2 - r_2)\Gamma(n - s_2 + r_2 + 1)} Q^{s_2 - r_2 - 1}(1 - Q)^{n - s_2 + r_2}\, dP. \tag{37}$$

It is known that if a random variable $v(0 \leq v \leq 1)$ has the distribution

$$\frac{\Gamma(c+d)}{\Gamma(c)\Gamma(d)} v^{c-1}(1 - v)^{d-1}\, dv,$$

and $2c$ and $2d$ are positive integers, then $\frac{2c}{2d}\frac{1 - v}{v}$ has the F-distribution (analysis of variance distribution) with $2d$ and $2c$ degrees of freedom. Thus,

$$\frac{2(s_2 - r_2)}{2(n - s_2 + r_2 + 1)} \frac{1 - Q}{Q} = F \tag{39}$$

has the F-distribution with $2(n - s_2 + r_2 + 1)$ and $2(s_2 - r_2)$ degrees of freedom. From (37) it follows that p is equal to the probability that

$$F \leq \frac{2(s_2 - r_2)}{2(n - s_2 + r_2 + 1)} \frac{1 - b}{b}$$

where F has the analysis of variance distribution with $2(n - s_2 + r_2 + 1)$ and $2(s_2 - r_2)$ degrees of freedom. For the case $r_1 = 1$, $s_1 = n$, $r_2 = 1$ and $s_2 =$

$n - 2$, the following table gives the value of the sample size n which is necessary for having the probability p that at least the proportion b of the universe is included in the tolerance rectangle T.

	$b = .97$	$b = .975$	$b = .98$	$b = .985$	$b = .99$
$p = .99$	332	398	499	668	1001
$p = .95$	256	309	385	515	771

Thus, if we want the probability to be .99 that the tolerance region will include at least 98 per cent of the universe, the sample size must be 499.

In section 4 tolerance regions are considered which are composed of several rectangles. Such a tolerance region may be more favorable than a single rectangle if x and y are highly correlated. As an illustration we consider tolerance regions T^* constructed as follows: Suppose that n is divisible by 4 and the sample points $p_1, \ldots, p_n$ are arranged in order of increasing magnitude of their abscissa values. We draw the vertical lines V_0, V_1, V_2, V_3 and V_4 through the points $p_1, p_{n/4}, p_{n/2}, p_{3n/4}$ and p_n. Let $R_i (i = 1, 2, 3, 4)$ be the rectangle determined by the vertical lines V_{i-1} and V_i and the horizontal lines H_i and H_i' where H_i and H_i' are defined as follows: consider only the points which lie between the two vertical lines V_{i-1} and V_i (points on the vertical lines are excluded). From these select the point with the smallest and the point with the largest ordinate. The lines H_i and H_i' are the horizontal lines which go through these two points respectively. The tolerance region T^* is composed of the four rectangles R_1, R_2, R_3 and R_4. The number of rectangles R_{ij} (defined in section 4) included in T^* is equal to $s = n - 9$. Thus, according to the results of section 4 the probability distribution of the proportion Q^* of the universe included in the region T^* is given by

$$\frac{\Gamma(n+1)}{\Gamma(n-9)\Gamma(10)} (Q^*)^{n-8} (1 - Q^*)^9 \, dQ^*.$$

Numerical calculations show that for $n = 1000$ the probability is .99 that at least 98.1 per cent of the universe will be included in the tolerance region T^*.