ON CONFIDENCE RANGES FOR THE MEDIAN AND OTHER EXPECTATION DISTRIBUTIONS FOR POPULATIONS OF UNKNOWN DISTRIBUTION FORM

By William R. Thompson

About the commonest situation with which we are confronted in mathematical statistics is that where we have a sample of n observations, $\{x_i\}$, which is assumed to have been drawn at random from an unknown population, U, with a zero probability that any two values in the finite sample be equal; and we desire to obtain from this evidence some insight as to parameters of the parent population, U. If further assumptions are made as to some of the parameters or the form of U, there may result a gain in power in testing other given hypotheses or establishing *confidence ranges* for particular parameters, but at an obvious sacrifice of scope in application. Insistent problems involve estimation of mathematical expectation that in further sampling we shall find x lying within a given interval, or similar expectation with regard to parameters of U such as the unknown median. It might seem that, without further assumption, all we should claim is that it is possible to draw from U the sample actually observed. A mere description of the experience may well be considered the observer's first duty, but a restriction to this would leave entirely unused the quality of *randomness* which has been assumed. What additional statements as to U may be appropriate in view of this randomness are our immediate concern; and the object of the present communication is to show how we may obtain such expressions in the form of mathematical expectations, and to present some results. Widespread applications to problems of estimation of *normal ranges of variation* or specific confidence ranges and comparisons of sample reflections of possibly different populations are immediately suggested, and a new foundation is offered for the study of frequency-distribution from the point of view of Schmidt.[1]

Section 1

Accordingly, consider the following situation. Let $A = \{x\}$ denote the set of all real numbers; and U denote an unknown frequency-distribution law of draft from $\{x\}$ such that there exists an unknown function, $f(x)$, bounded and not negative in A, and that the probability of obtaining x in an arbitrary interval (α, β) is

$$(1) \qquad P(\alpha < x < \beta) = \int_{\alpha}^{\beta} f(x) \cdot dx \ ;$$

[1] Schmidt, R., *Annals of Math. Stat.*, *5*, 30, (1934).

and, for every positive $p < 1$, there exists a finite interval (α, β) such that $P(\alpha < x < \beta) > p$. Let U be called an *infinite population*; and let n drafts, independently thus governed, made from A *without replacements* be called *a random sample of n observations from U*. Let $S = \{x_k\}$, $k = 1, \cdots, n$, denote such a sample; the enumeration to be made in an arbitrarily determined manner. In any case $x_i \neq x_j$ for $i \neq j$.

Temporarily, let us consider k to indicate the order of draft of the values of $\{x_k\}$, and let $p_k = P(x < x_k)$ denote the probability that x, drawn at random from U, be less than x_k of S. The probability *à priori* (i.e., without regard to relative values of x in the sample) that in such random sampling p_k lie between p' and p'', where $0 \leqq p' < p'' \leqq 1$, is obviously independent of k, and equals $p'' - p'$; i.e., p_k is equally likely *à priori* to lie in either of any two equal intervals in its possible range, (0, 1). Furthermore, the probability that in the rest of the sample, S, there will be just r values less than x_k is

$$\binom{n-1}{r} \cdot p_k^r \cdot (1 - p_k)^{n-1-r},$$

where r is an integer and $0 \leqq r < n$. Of course, p_k is unknown; but we may calculate (for all cases in repeated sampling wherein the same value of r is encountered) the expectation, $\overline{P}_r$ $(p' < p_k < p'')$, that p_k lie in the interval (p', p''). This is given by

$$\overline{P}_r(p' < p_k < p'') = \frac{(r+s+1)!}{r!\,s!} \int_{p'}^{p''} p^r \cdot q^s \cdot dp\,, \tag{2}$$

where $s = n - 1 - r$, and $q = 1 - p$. This is a familiar result[2,3,4] in applications of the well-known principle of Bayes to estimation of *à posteriori* probability. The approach is convenient in that many relations which have been developed in this connection are made immediately available. However, that p_k is equally likely *à priori* to lie in either of any two equal intervals in its possible range, is not based in the present case upon an especially added assumption nor any plea concerning *equal distribution of ignorance*, but follows directly from the elementary assumptions of random sampling. Accordingly, we are enabled to develop for given ranges what may be called the *specific confidence* or mathematical expectation that a given variable lie therein.

Obviously, (2) does not depend on k if this index is the order of draft provided that just r values of the sample, S, are less than the one under consideration, x_k. To simplify notation, accordingly, let the index k for any given sample, $\{x_k\}$,

---

[2] Bayes, *Philosophical Transactions, 53*, 370 (1763). *Cf.* Todhunter, I., "A History of the Mathematical Theory of Probability," Macmillan and Co., London, 1865.

[3] Laplace, "Théorie Analytique des Probabilités," Paris, 1820; and other works, *Cf.* Todhunter, *l.c.*

[4] Pearson, K., *Philosophical Magazine, Series 6, Vol. 13*, 365, (1907).

be determined by the relations, $x_i < x_j$ for $i < j$, where $k = 1, \cdots, n$. Then, by (2) as $k = r + 1$, we have

$$(3) \qquad \bar{P}(p' < p_k < p'') = \frac{n!}{(k-1)!\,(n-k)!} \int_{p'}^{p''} p^{k-1} \cdot q^{n-k} \cdot dp\,,$$

where p_k is the probability that random sample values from U will be less than the k-th value in order of ascending magnitude from a given random sample, $\{x_k\}$, of n values from U; and $\bar{P}(p' < p_k < p'')$ denotes the expectation that in such sampling p_k will lie in the interval, (p', p'').

In general, let $E(w) \equiv \bar{w}$ denote the mathematical expectation of any variable, w, under the given sampling conditions. Then, from a well-known relation developed by Laplace, we obtain from (3) the mean expectation of p_k,

$$(4) \qquad \bar{p}_k = \frac{k}{n+1}\,;$$

and, further relations[4] of Karl Pearson yield

$$(5) \qquad E((p_k - \bar{p}_k)^2) = \overline{\sigma^2_{p_k}} = \frac{k(n-k+1)}{(n+1)^2 \cdot (n+2)}\,;$$

i.e., the mean squared error in systematic use of $\frac{k}{n+1}$ instead of the unknown p_k should have the value in (5). Specific confidence ranges for x are readily established; e.g., the expectation that in random draft from U we obtain x within the range (x_k, x_{n-k+1}) in view of the sample, S, is

$$(6) \qquad \bar{P}(x_k < x < x_{n-k+1}) = \frac{n+1-2k}{n+1}, \quad \text{for } 2k < n+1\,;$$

and $\bar{P}(x < x_k) = \bar{P}(x > x_{n-k+1}) = \frac{k}{n+1}$. For a given variate, w, the range (α, β) will be called *central* if $\bar{P}(w < \alpha) = \bar{P}(w > \beta)$, as in the case under (6). This is in accord with the development of the subject of confidence ranges by Neyman[5,6] and by Clopper and E. S. Pearson[7] following the introduction of the notion of fiducial interval by R. A. Fisher.[8,9] The estimates of p_k in (4) may be of value in studying frequency-distribution from the point of view developed by Schmidt,[1] by comparison of x_k with $\psi\left(\frac{k}{n+1}\right)$ rather than $\psi\left(\frac{2k-1}{2n}\right)$ where ψ is a univariant inverse of the integral of a given frequency function, taken to

---

[5] Neyman, J., *J. Roy. Stat. Soc.*, *97*, 589, (1934).
[6] Neyman, J., *Annals of Math. Stat.*, *6*, No. 3, 111, (1935).
[7] Clopper, C. J., and Pearson, E. S., *Biometrika*, *26*, 404, (1934).
[8] Fisher, R. A., *Proc. Camb. Phil. Soc.*, *26*, 528, (1930).
[9] Fisher, R. A., *Proc. Roy. Soc.*, *A 139*, 343, (1933).

replace the unknown $f(x)$. Obviously, $\overline{P}(x_k < x < x_{k+1}) = \dfrac{1}{n+1}$. A discussion of the special case, $n = 2$, has been prominent recently in a controversy between Jeffrey[10] and Fisher[9,11] and in an article by Bartlett.[12]

Now, in (3) for $p = p'$, and $p'' = 1$; we may write[13]

$$(7) \quad \overline{P}(p < p_k) = \sum_{\alpha=0}^{k-1} \binom{n}{\alpha} \cdot p^{\alpha} \cdot q^{n-\alpha} \equiv I_q(n-k+1, k) \equiv \frac{B_q(n-k+1, k)}{B_1(n-k+1, k)},$$

where $q = 1 - p$, and the incomplete B and I functions are those of K. Pearson[13] and Müller.[14] Now, let M be the unknown median of the infinite population, U. Then, by definition of p_k, if and only if $x_k > M$, then $p_k > \frac{1}{2}$. Therefore,

$$(8) \quad \overline{P}(M < x_k) = \overline{P}(0.5 < p_k) = \left(\frac{1}{2}\right)^n \cdot \sum_{\alpha=0}^{k-1} \binom{n}{\alpha} \equiv I_{0.5}(n-k+1, k).$$

Obviously, $\overline{P}(x_k < M < x_{k+1}) = \left(\dfrac{1}{2}\right)^n \cdot \binom{n}{k}$, and the expectation that M lie between the k-th observations from each end of the set, S, is given by

$$(9) \quad \overline{P}(x_k < M < x_{n-k+1}) = 1 - 2 \cdot I_{0.5}(n-k+1, k), \text{ for } 2k < n+1.$$

Obviously, this confidence range is *central.*

Section 2

Now, consider another infinite population, U'. In similar manner we may develop expressions for confidence ranges and distribution expectations. Let x' be the variate, and consider a sample, $S' = \{x'_m\}$, of n' observations drawn *without replacements* from A according to U' but after the sample, S, of U; i.e., so that no two of these sample values in S' are equal, nor any of them equal to a value in S. Furthermore, let m be the order of ascending magnitude of x' values in S'; and $p'_m \equiv P(x' < x'_m)$ for x' drawn at random from U', and let M' be the unknown median of U'. Then, by replacement of x, n, p_k, k, and M by x', n', p'_m, m, and M', respectively, in relations already developed for U and S, we obtain corresponding expressions for U' and S'; e.g.,

$$(10) \qquad \overline{P}(x'_m < x' < x'_{m+1}) = \frac{1}{n'+1}.$$

---

[10] Jeffreys, H., *Proc. Roy. Soc., A 138,* 48, (1932); *A 140,* 523, (1933); *A 146,* 9, (1934); *Proc. Camb. Phil. Soc., 29,* 83, (1933).

[11] Fisher, R. A., *Proc. Roy. Soc., A 146,* 1, (1934).

[12] Bartlett, M. S., *Proc. Roy. Soc., A 141,* 518, (1933).

[13] Pearson, K., *Biometrika, 16,* 202, (1924).

[14] Müller, J. H., *Biometrika, 22,* 284, (1930–31).

Now, let the index values, k_m, be defined as the number of values of $\{x_k\}$ that are less than x'_m, $m = 1, \cdots, n'$. Then, for all realized cases,

$$x_{k_m} < x'_m < x_{k_m+1}, \qquad m = 1, \cdots, n', \tag{11}$$

for the extreme members of (11) in S. Then, for x and x' drawn at random from U and U', respectively, we may write

$$0 < (n+1)(n'+1)\cdot\overline{P}(x < x') - \sum_{m=1}^{n'} k_m < n + n' + 1, \tag{12}$$

provided that the expectations for U and U' may be treated as independent. Similarly, for $\overline{P}(M < M')$ we have the relations,

$$\sum_{m=1}^{n'} \binom{n'}{m}\cdot I_{0.5}(n - k_m + 1, k_m) < 2^{n'}\cdot\overline{P}(M < M') < 1 + \sum_{m=1}^{n'} \binom{n'}{m-1}\cdot I_{0.5}(n - k_m, k_m + 1). \tag{13}$$

Of course, $I_{0.5}(n+1, 0) \equiv 0$, and $I_{0.5}(0, n+1) \equiv 1$. It may be verified readily that the inequality relations of (12) and (13) provide *best* upper and lower bounds for $\overline{P}(x < x')$ and $\overline{P}(M < M')$ under the circumstances given.

Obviously, any increasing function, $\phi(y)$, for y in A, may be used throughout the arguments, with $\phi(y)$ replacing $y = x, x_k, M, x', x'_m, M'$, respectively.

Section 3

Consider, now, the case of a finite population, U_N, of real numbers $\{x^{(i)}\}$, $x^{(i)} < x^{(j)}$ for $i < j$, $i = 1, \cdots, N$. Assume that N is known, and that a sample, S, of n values has been drawn at random from U_N without replacements. Let the sample values be $\{x_k\}$, $k = 1, \cdots, n$; and k be an arbitrarily determined index. As before, we might consider k the order of draft, temporarily, but the same analysis may be made if we let k be the order of ascending magnitude in the sample, S, and disregard its value in connection with *à priori* estimates of draft probability. Each $x_k = x^{(u_k)}$ for some unknown $u_k = 1, \cdots, N$; and, *à priori* (i.e., with no knowledge as to order of magnitude of other values in the sample), any two of these values are equally likely. Obviously, this is so if x_k is the first value drawn from U_N, and the rest of the sample may be regarded as a random draft without replacements of $n - 1$ elements from $[U_N - x_k]$. Let r be the number of these sample values less than x_k, and $s = n - 1 - r$. Then the probability of drawing such a sample after the given x_k, under the conditions given, is $\dfrac{\binom{u_k - 1}{r}\binom{N - u_k}{s}}{\binom{N-1}{n-1}}$, where $u_k - 1$ is the unknown number of

values in U_N that are less than x_k. To estimate the expectation, $\overline{P}(R = u_k - 1)$, that there are just a given number, R, of values in U_N less than x_k; we encounter the same situation considered by K. Pearson in a paper[15] subsequent to those applied to the infinite universe; and, by a simple conversion in notation, we have

$$\overline{P}(R = u_k - 1) = \frac{\binom{R}{r}\cdot\binom{N-1-R}{s}}{\binom{N}{n}}. \tag{14}$$

In previous communications[16,17] I have defined a function,

$$\psi(r, s, r', s') \equiv \frac{\sum_{\alpha=0}^{r'}\binom{r+r'-\alpha}{r}\cdot\binom{s+s'+1+\alpha}{s}}{\binom{r+s+r'+s'+2}{r+s+1}}, \tag{15}$$

for any four rational integers $r, s, r', s' \geqq 0$; and shown that Pearsons further result, equivalent here to evaluation of $\overline{P}(u_k \leqq R + 1)$ for a given R, may be expressed by means of this ψ-function. Thus, we have

$$\overline{P}(u_k \leqq R + 1) = \psi(r, s, R - r, N - R - s - 2). \tag{16}$$

It was demonstrated also[16,17] that

$$\psi(r, s, r', s') \equiv \psi(r, r', s, s') \equiv \psi(s', r', s, r) \equiv 1 - \psi(s, r, s', r') \tag{17}$$

with extension of the definition to include $\psi(r, s, -1, s') \equiv 0$, and that

$$\psi(r, s, r', s') \equiv \frac{\sum_{\alpha=0}^{\alpha' \leqq s, r'}\binom{r+r'+1}{r+1+\alpha}\cdot\binom{s+s'+1}{s-\alpha}}{\binom{r+s+r'+s'+2}{r+s+1}}. \tag{18}$$

As in the case of the infinite population, here also it is obvious that the order of draft of x_k is of no consequence in the analysis; and again we will let $k = r + 1$, whence $s = n - k$, and we may make these substitutions in (14) and (16). Then, we may write

$$\overline{P}(u_k \leqq R) = \psi(k - 1, n - k, R - k, k + N - R - n - 1); \tag{19}$$

---

[15] Pearson, K., *Biometrika, 20 A*, 149, (1928).
[16] Thompson, W. R., *Biometrika, 25*, 285, (1933).
[17] Thompson, W. R., *American Journal of Mathematics, 57*, 450, (1935).

and, obviously, $\bar{P}(u_{n-k+1} \geqq N - R + 1) \equiv \bar{P}(u_k \leqq R)$. Hence, if we let M be the unknown median of U_N; and $m \equiv \frac{N-a}{2}$, where $a = 0$, 1, and $N - a$ is even; then, as u_i is an integer,

$$\bar{P}(x_k \leqq M \leqq x_{n-k+1}) \equiv \bar{P}\left(u_k \leqq \frac{N}{2} \leqq u_{n-k+1}\right) \tag{20}$$

$$\equiv 1 - 2\cdot\psi(k-1, n-k, m-k, k+N-m-n-1),$$

which is the expectation that the median of U_N lie within the closed interval, (x_k, x_{n-k+1}), for $2k \leqq n + 1$. This gives the confidence range, analogous to that for the infinite universe. It may be noted that

$$\bar{P}(u_k \leqq R < u_{k+1}) = \bar{P}(u_k \leqq R) - \bar{P}(u_{k+1} \leqq R)$$

$$= \psi(r, s, r', s') - \psi(r+1, s-1, r'-1, s'+1)$$

where $r = k - 1$, $s = n - k$, $r' = R - k$, and $s' = k + N - R - n - 1$. Hence, (18) gives

$$\bar{P}(u_k \leqq R < u_{k+1}) \equiv \frac{\binom{R}{k}\cdot\binom{N-R}{n-k}}{\binom{N}{n}}. \tag{21}$$

The approach by way of Pearson's problem again makes it easy to evaluate the expected mean p_k and variance as in the case of the infinite population, where $p_k = P(x < x_k)$ for x drawn at random from U_N. Of course, $p_k = \frac{u_k - 1}{N}$, but u_k is unknown. From Pearson's result,[15] however, we obtain

$$\bar{p}_k = \frac{k(N+1) - n - 1}{N(n+1)} = \frac{k}{n+1}\left(1 - \frac{n}{N}\right) + \frac{k-1}{N}, \tag{22}$$

and the expected variance of p_k,

$$\overline{\sigma^2_{p_k}} = E((p_k - \bar{p}_k)^2) = \frac{k(n-k+1)(N+1)(N-n)}{(n+1)^2\cdot(n+2)\cdot N^2}. \tag{23}$$

YALE UNIVERSITY.