BIOLOGICAL APPLICATIONS OF NORMAL RANGE AND ASSOCIATED SIGNIFICANCE TESTS IN IGNORANCE OF ORIGINAL DISTRIBUTION FORMS*

By William R. Thompson

The word *normal* has been used in many senses—commonly by statisticians to designate a well-known distribution function. Another use familiar to biologists, particularly in experimental work and medicine, is to denote an untreated or control part of a universe, or a part whose members are free from specified characteristics such as evidence of past or present disease or malformation. Closely related to this last usage are attempts to delimit so-called normal ranges of variation for a quantitative attribute of the members of part or all of a universe in question. Interpretations are often vague, as when the interval between the least and greatest values observed in either a large or a small number of instances is taken to estimate a normal range. We shall consider the problem of using ranked data for estimating normal ranges as defined in the next paragraph.

If the instances have been drawn at random from a universe (U) of all possible observations obtainable in a prescribed manner, and are enumerated in ascending order of magnitude, $\{x_i\}$ for $i = 1, \cdots, n$; then it is proposed to show in the present communication how ranges of the type (x_k, x_{n+1-k}) may be used to estimate *normal ranges*, R_f, where the subscript f is the theoretical probability that a random value, x, drawn from U will lie within the range R_f, g that it will lie above, and g that it will lie below (where $2g = 1 - f$). Furthermore, it is proposed to show how these ranges may be used as the basis of significance tests where altered conditions appear to lead to abnormal biological variation. The form of frequency-distribution of U is supposed unknown, and is without effect upon the analysis. *Section* 1 is a development of the theory of range estimation, treated briefly in a previous paper [1] together with illustrations of its application. *Section* 2 deals with significance tests.

1. **The Method of Range Estimation.** Let x be a real variate, a random value drawn from an infinite universe or population U. Let $f(x)$ be the frequency function of x in U, supposed unknown; and $\int_{-\infty}^{\infty} f(x)\,dx = 1$. Then for any given α and β, where $\alpha < \beta$, and

$$P(\alpha < x < \beta) \equiv \int_{\alpha}^{\beta} f(x)\,dx. \tag{1}$$

* Presented at a meeting of the American Statistical Association, December 28, 1937, Atlantic City, N. J.

To facilitate development, suppose that in any finite sampling under consideration no two values of x may be exactly the same. Let $S = \{x_k\}$, $k = 1, \cdots, n$, denote a random sample from U, where the order of enumeration is arbitrary, but temporarily taken as a random order (to fix the ideas, consider this the order obtained in drawing). Let p_k be defined by

$$(2) \qquad p_k = P(x < x_k) \equiv \int_{-\infty}^{x_k} f(x)\,dx \quad \text{from which} \quad dp_k = f(x_k)\,dx_k.$$

Then p_k is the probability that a random x from U shall be less than any number x_k. Then obviously if x_k is drawn at random from U, p_k is a random variable whose distribution is the unit rectangle; i.e., $P(p' < p_k < p'') = p'' - p'$. Furthermore, the joint probability that x_k will lie in the interval x_k, $x_k + dx_k$ and that exactly r values in the sample S will be less than x_k is, to within terms of order dp_k, $\binom{n-1}{r} \cdot p_k^r \cdot (1 - p_k)^{n-1-r} dp_k$.

Then, in repeated sampling as above, for the case where just r of the n random values $\{x_i\}$ are less than the k-th drawn, let $P_{n,r}(p' < p_k < p'')$ denote the probability that p_k lies in the interval (p', p''). Then

$$(3) \qquad P_{n,r}(p' < p_k < p'') = \frac{(r + s + 1)!}{r! \cdot s!} \cdot \int_{p'}^{p''} p^r \cdot q^s \cdot dp,$$

where $s = n - 1 - r$, and $q = 1 - p$. Obviously, the expression on the right of (3) does not depend on k if this index is the order of draft or a random index, but only upon the condition that exactly r of the n random values from U be less than a value x_k drawn at random from the sample of n values. Accordingly, we obtain the same result if we enumerate the n values $\{x_i\}$ in ascending order of magnitude ($x_i < x_j$, if $i < j$). Then $k = r + 1$, in the cases considered, and (3) may be written,

$$(4) \qquad P_n(p' < p_k < p'') = \frac{n!}{(k-1)!(n-k)!} \cdot \int_{p'}^{p''} p^{k-1} \cdot q^{n-k} \cdot dp,$$

for $0 \leqq p' \leqq p'' \leqq 1$. Obviously, the result is the same if we deal instead with the k-th value (x_k) of every random sample S drawn. In passing it may be noted that for $p' = 0$ and $p'' = p$ in (4) we have

$$(5) \qquad P_n(p_k < p) = I_p(k, n - k + 1),$$

which may be evaluated for k, $n - k + 1 \leqq 50$ by means of the *Tables of the Incomplete Beta-Function* [2].

Of course, $P_n(0 < p_k < 1) = 1$, and (4) gives $\bar{p}_k$, the mean value of p_k in repeated random sampling of n values from U, as

$$(6) \qquad \bar{p}_k = \frac{n!}{(k-1)!(n-k)!} \cdot \int_0^1 p^k \cdot q^{n-k} \cdot dp = \frac{k}{n+1}.$$

Similarly, the variance, $\sigma^2_{p_k}$, of p_k is given by

$$\sigma^2_{p_k} = E[(p_k - \bar{p}_k)^2] = \frac{k(n-k+1)}{(n+1)^2 \cdot (n+2)}. \tag{7}$$

Now suppose that we want to find a range (α, β) such that, in random drafts from U, the theoretical relative frequency of drawing x less than α is g, and the same as that of drawing x greater than β. (α, β) may be called a *central confidence range* with a *confidence* $f = 1 - 2g$ that x drawn at random from U will lie within the range. For $g = k/(n+1)$ we may take the range $R_f = (x_k, x_{n-k+1})$; and likewise with $g = 5\%$ we may estimate, or approximate by interpolation where $20k > n + 1 > 20(k-1)$, a range R_f for normal biological variation of a specified character, and this may be called briefly the estimated 90% *central normal range.*

Of course the probability of drawing $x < \alpha$ is $\int_{-\infty}^{\alpha} f(x)\,dx$, and that of drawing $x > \beta$ is $\int_{\beta}^{\infty} f(x)\,dx$; and these probabilities are unknown, as the frequency function $f(x)$ is unknown; but with $\alpha = x_k$ and $\beta = x_{n-k+1}$ the theoretical relative frequency in each case is $k/(n+1)$ regardless of the universe.

It has been shown [1] also that if the sample S were drawn at random from a finite ordered population of aggregate number N, denoted by U_N, and Np_k is the number of values in U_N that are less than the k-th member of the given random sample in ascending order of magnitude; then, for S a sample of n values as before, the mean value of p_k in repeated sampling is

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

$$\sigma^2_{p_k} = \frac{k(n-k+1)}{(n+1)^2 \cdot (n+2)} \cdot \left(1 + \frac{1}{N}\right)\left(1 - \frac{n}{N}\right).$$

An example is furnished by an analysis of data reported by Wadsworth and Hyman [3] in a study of influences of antigenic treatment of horses upon their plasma concentration of esterified cholesterol, free cholesterol, and phospholipids. As in chart 1 for normal horses, a graph has been constructed for each horse studied, using time as abscissa and a logarithmic ordinate scale for observed values of plasma concentration of the constituents:

1. Esterified Cholesterol,
2. Free Cholesterol, and
3. Phospholipids *times one-tenth,*

the respective successive points for each being joined to form three polygon curves. As these are in all cases discrete and lie in the order of enumeration from top to bottom of the graph, no special label seemed needed; but estimated normal ranges for the central 90% of variation have been indicated in each

case by two horizontal lines between brackets at the right, numbered to correspond with the enumeration above. The ranges are based on observations on 62 plasma samples, each from a different presumably normal horse. The normal horses in the chart show about the same individual variations; but, of course, the ranges are not to be interpreted to indicate normal variation for an individual animal.

Chart 2 presents in like manner the data obtained for horses under immunization against tetanus and the streptococcus. The tetanus immunization treat-

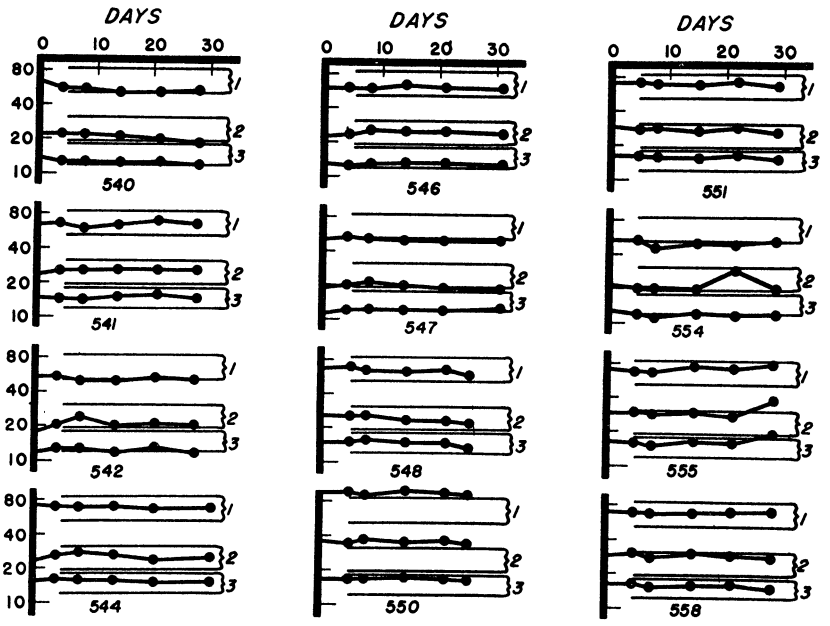

CHART 1. On each graph for a given normal horse, the number of which appears below, the curves in descending order respectively represent (1) esterified cholesterol, (2) free cholesterol, and (3) one-tenth phospholipid concentration in plasma (in mg. per 100 cc.). Corresponding 90-per-cent normal range estimates are indicated.

ment appears to produce marked and sustained depression in all three curves of at least five of the six animals observed.

That this is statistically significant seems obvious. A single observation below the 90% normal range should be expected once in twenty random trials if normal causes of variation may be assumed unaffected by the treatment in question. The expectation of obtaining 5 or more such values in six independent trials is obviously much less, and may be accurately estimated by means of relations developed in the following section.

2. **Significance Tests.** Now consider as in *section* 1 another sample S' of n' values; $\{x'_{k'}\}$, $k' = 1, \cdots, n'$ (where $x'_i < x'_j$ if $i < j$), drawn at random from an infinite universe U' as was S from U; but where U' and U are not necessarily

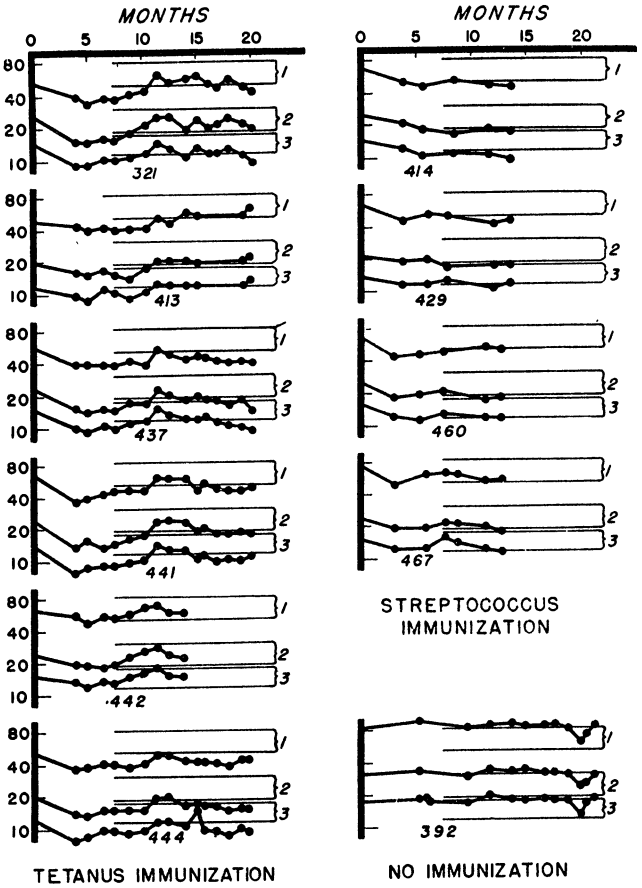

CHART 2. On each graph for horses receiving the indicated antigenic treatment and one untreated horse, the curves in descending order respectively represent (1) esterified cholesterol, (2) free cholesterol, and (3) one-tenth phospholipid concentration in plasma (in mg. per 100 cc.). Corresponding 90-per-cent normal range estimates are indicated.

the same universe. In like manner it may be shown that, if x' is drawn at random from U' and $p'_{k'}$ denotes $P(x' < x'_{k'})$, then

$$P_{n'}(\phi' < p'_{k'} < \phi'') = \frac{(v + w + 1)!}{v!w!} \cdot \int_{\phi'}^{\phi''} p^v \cdot q^w \cdot dp \tag{8}$$

where $q = 1 - p$, $v = k' - 1$, $w = n' - k'$, and $0 \leqq \phi' \leqq \phi'' \leqq 1$.

The probabilities in (4) and (8) are independent, obviously, whether U' is the same as U or not. Accordingly, these relations make possible an evaluation

of $P(p_k < p'_{k'})$ under the circumstances where repeated sampling is applied to both the case of S and to that of S'. With this understanding, then

$$(9) \qquad P(p_k < p'_{k'}) = \frac{(r+s+1)!(v+w+1)!}{r!\cdot s!\cdot v!\cdot w!}\cdot\int_0^1 p_0^r\cdot q_0^s\cdot dp_0\cdot\int_{p_0}^1 p^v\cdot q^w\cdot dp,$$

where, as before, $r = k-1$, $s = n-k$, $v = k'-1$, $w = n'-k'$, $q \equiv 1-p$, and $q_0 \equiv 1-p_0$.

In a previous paper [4] a Ψ-function was defined as

$$(10) \qquad \Psi(r, s, r', s') \equiv \frac{\sum_{\alpha=0}^{r'}\binom{r+r'-\alpha}{r}\binom{s+s'+1+\alpha}{s}}{\binom{r+s+r'+s'+2}{r+s+1}}$$

for any four rational integers $r, s, r', s' \geqq 0$; and it was shown in detail that the right member of (9) is equal to $\Psi(r, s, v, w)$; whence we may write

$$(11) \qquad P(p_k < p'_{k'}) = \Psi(k-1, n-k, k'-1, n'-k').$$

Obviously, if U and U' are the same universe, then $p_k < p'_{k'}$ if and only if $x_k < x'_{k'}$, and then we have

$$(12) \qquad P(x_k < x'_{k'}) = \Psi(k-1, n-k, k'-1, n'-k')$$

in repeated random sampling applied to both sample types, S and S', respectively of n and of n' observations. In the paper just mentioned, and in another [5] the Ψ-function was further developed by extension of definition to include $\Psi(r, s, -1, s') \equiv 0$, and it was shown that

$$(13) \quad \Psi(r, s, r', s') \equiv \Psi(r, r', s, s') \equiv \Psi(s', r', s, r) \equiv 1 - \Psi(s, r, s', r').$$

Further demonstrations [5] included the relation,

$$(14) \qquad \Psi(r, s, r', s') \equiv \frac{\sum_{\alpha=0}^{\alpha\leqq s, r'}\binom{r+r'+1}{r+1+\alpha}\binom{s+s'+1}{s-\alpha}}{\binom{r+s+r'+s'+2}{r+s+1}},$$

which offers another form for calculation. The identities in (13) are particularly useful to facilitate calculation where one of the four arguments is small. A system for forming a table has also been developed [4, 5] in an economical form, but tabulation has been given only for the arguments not exceeding 5.

Now, in applying a test based on relation (12) or on that for the complementary probability, $P(x'_{k'} < x_k)$ which obviously, by (13), equals $\Psi(n-k, k-1, n'-k', k'-1)$, we may wish to exclude from the *normal* set of observations those values obtained from animals later given the treatment in question in the statistical significance test. The purpose would be to avoid violation of the condition of independent sampling required. In the case of the tetanus antigen treatment, we have an experience wherein 5 or more of 6 horses treated yield

values for a given plasma constituent less than the third in ascending order of magnitude (namely x_3) in our independent set of *normal* values. Here $n' = 6$, and $n = 62 - 6 = 56$. In accordance with the hypothesis that the treatment in question does not affect normal causes of variation in the plasma constituent under investigation we have $P(x'_{k'} < x_3)$ is $\Psi(53, 2, 6 - k', k' - 1)$. This is approximately $1.891(10)^{-5}$ for $k' = 5$, and $4.555(10)^{-7}$ for $k' = 6$. Obviously, a rule for establishing the value of k to be used in such tests should be fixed in advance without prejudice, as in the present case where we have taken $k \geqq g(n + 1) > k - 1$ for $g = 5\%$.

In the case of streptococcus immunization treatment, the corresponding test would have $n = 58$, $n' = 4$, $k = 3$, and $k' = 4$, 3, or 2; which would yield approximately $2.689(10)^{-5}$, $1.031(10)^{-3}$, or $1.817(10)^{-2}$, respectively for $P(x'_{k'} < x_3)$. Thus it appears that where such values are found (intuitively it would appear a fortiori if we compare instead with x_3 of the entire normal set of 62 values), their low magnitude appears to discredit the hypothesis that such discrepancies are ascribable to mere chance normal variation in the quantitative attribute investigated.

The tests proposed are free from any assumption concerning the form of the original distribution $f(x)$. The illustrative material is only a part of that presented with similar statistical treatment in the paper of Wadsworth and Hyman [3], which makes it apparent that the tests suggested here may be useful and powerful in analysis of biological and other experimental data. From a similar point of view, Hotelling and Pabst [6] developed tests of bi-variate correlation, and Milton Friedman has elaborated a multi-variate rank analysis [7], the tests being likewise free from any assumption about the form of the original distributions. In a previous paper [1] confidence ranges for the median are based similarly, employing relation (5) for the special case $p = \frac{1}{2}$.

DIVISION OF LABORATORIES AND RESEARCH
NEW YORK STATE DEPARTMENT OF HEALTH
ALBANY, N. Y.

REFERENCES

[1] W. R. THOMPSON, *Annals of Mathematical Statistics*, Vol. 7 (1936), p. 122.
[2] *Tables of the Incomplete Beta-function*, edited by Karl Pearson, (Office of *Biometrika*, University College, London), 1934, p. 494.
[3] AUGUSTUS WADSWORTH AND L. W. HYMAN, *Jour. Immunol.*, Vol. 35 (1938), p. 55.
[4] W. R. THOMPSON, *Biometrika*, Vol. 25 (1933), p. 285.
[5] W. R. THOMPSON, *American Journal of Mathematics*, Vol. 57 (1935), p. 450.
[6] H. HOTELLING AND M. R. PABST, *Annals of Mathematical Statistics*, Vol. 7 (1936), p. 29.
[7] MILTON FRIEDMAN, *Jour. Amer. Stat. Assoc.*, Vol. 32 (1937), p. 675.