A NEW TEST OF SIGNIFICANCE IN SAMPLING FROM FINITE POPULATIONS, WITH APPLICATION TO HUMAN INBREEDING

By S. D. JAYAKAR and J. B. S. HALDANE
Genetics and Biometry Laboratory, Government of Orissa, Bhubaneswar-3, Orissa, India

Introduction

Geneticists are constantly faced with the need for testing the significance of the difference between two sets of samples, and in particular between their means. When both samples are large (say each exceeding 200) we may use a test based on the normal distribution with some confidence. For even if the distribution within one or both samples is far from normal, the sampling distribution of their means will be nearly so. For smaller samples, Gossett's ("Student", 1908) or Fisher's (1925) "*t*-test" based on it, are recommended. These tests are based on the assumption that the distributions of the parent population is normal. Sometimes this is patently false. However it has been shown by Bartlett (1935), and others, that this test is fairly "robust", that is to say still nearly valid when the parent distribution is not normal. It is however desirable to produce a test in which no assumption is made about the parent distribution. Further, the logical basis of Student's test seems somewhat weaker when it is applied to the difference of two means, rather than to the significance of a single mean, for which it was originally designed. For it is assumed that variances are equal in the populations compared.

We consider the total of the two samples compared, and ask what is the probability that a sample drawn from it at random should have a mean differing from that of the total by the observed amount or more. This problem is, in theory, soluble with complete accuracy, without any assumptions beyond those made when calculating the probabilities in drawing from a pack of cards. The principles involved may be illustrated from the work of Dronamraju and Meera Khan (1963), which interested us in this problem. Their paper will be referred to as "D and M" in future. They investigated the parentage of 746 hospital patients, and determined the coefficient of inbreeding of each one. 38 of these patients suffered from pulmonary tuberculosis, and had a high average coefficient of inbreeding. Was this significantly high ? We can frame the question a little more concretely as follows. A card was made out for each patient containing data from which his or her coefficient of inbreeding could be calculated in a few seconds. It could have been written on the cards. We ask what is the probability of drawing a "hand" of 38 cards from this "pack" of 746, such that the sum of the coefficients of inbreeding on the cards of this "hand" was equal to, or greater than, that found for the patients with pulmonary tuberculosis. This can be calculated exactly, but to do so by a manually operated calculating machine would take some weeks. If the prophylaxis of pthisis were considered as important as the calculation of missile orbits, business profits, or future stellar temperatures, electronic equipment

would be available for this purpose. We can, however, obtain some of the exact moments of the distribution of this sum.

The theory of sampling from a finite population without replacement was extensively developed in the earlier years of this century, but, in the words of Kendall and Stuart (1958) "the algebra rapidly becomes unmanageable". However Tukey (1950) opened up new possibilities. He showed that the expectation of any of Fisher's (1928) k-statistics in a random sample is equal to the value in the finite population from which it is drawn at random. These statistics are often described as cumulant estimates, but they have other properties which may be more important. Wishart (1952) gave expressions for the sampling distributions of some of them; but these are not always in a form suited for computation at least from such data as those of D and M. In what follows we have definite practical applications in view, though we have gone beyond them. We may have sacrificed algebraical elegance for this reason. Some of our results have certainly been obtained before, but we do not think that the expressions (2·4) and (2·5) which are the basis of our test of significance, have been given in the form obtained here. Our proofs are not based on "polykays", but are derived by elementary methods.

We hope that our expressions may be of some general use in the comparison of sample means.

SYMBOLISM

We consider a population of n individuals $A_1, A_2, \cdots A_i, \cdots A_n$. Each has one or more numerical characteristics. x'_i, y'_i, $z'_i \cdots$. We use symbols x' etc. for the following reason. If we put $x_i = x'_i - \bar{x}'_i$, where $\bar{x}'_i$ is the population mean, we greatly simplify most of the algebra. This is an entirely legitimate procedure, since all the values of x' are supposed to be known.

Let K_r be the rth k-statistic (Fisher, 1928) of the distribution of x'. If $r>1$ it is the same statistic of the distribution of x. Let K_{rst} be the k-statistic of given subscripts for the joint distribution of $x', y', z' \cdots$. For example $K_{11}=K_{110}$ is the covariance of x' and y', or x and y. Tukey and Wishart used such symbols (polykays) with reference to univariate distributions. We do not employ them in this sense.

A sample of s members is drawn without replacement from the population of n. Other authors have used N for our n, and n for our s. Let k_r be the rth univariate cumulant and k_{rst} a multivariate cumulant of the distribution of the sample. The probability that the sample includes any specified ρ members of the population is $\frac{s^{(\rho)}}{n^{(\rho)}}$, where $n^{(\rho)}$ means $\frac{n!}{(n-\rho)!}$. For symmetrical functions from the population we use the symbol S, so that $Sx_1^2 y_1 y_2$ means the sum of all values of $x_i^2 y_i y_j$, where $j \neq i$. We use S_r or S_{rst} for power sums. Thus $S_{21}=Sx^2 y=\sum_i x_i^2 y_i$. Symmetric functions of the sample are denoted by Σ. Kendall and Stuart (1958) give expressions for K statistics in terms of power sums. These are greatly simplified if $S_1=S_{10}=0$, etc. Thus

$$K_{22}=[(n-1)(n-2)(n-3)]^{-1}[n(n+1)S_{22}-(n-1)(S_{20}S_{02}+2S_{11}^2)].$$

We require expressions for the power sums in terms of the k-statistics. For the univariate case,

$$\left.\begin{aligned} S_2&=(n-1)K_2,\\ nS_3&=(n-1)(n-2)K_3,\\ n(n+1)S_4&=(n-1)[3(n-1)^2\ K_2^2+(n-2)(n-3)K_4],\\ n^2(n+5)S_5&=(n-1)(n-2)[10(n-1)^2K_2K_3+(n-3)(n-4)K_5]. \end{aligned}\right\} \quad (1.1)$$

For the bivariate case

$$\left.\begin{aligned} S_{11}&=(n-1)K_{11},\\ nS_{21}&=(n-1)(n-2)K_{21},\\ n(n+1)S_{31}&=(n-1)[3(n-1)^2K_{20}K_{11}+(n-2)(n-3)K_{31}],\\ n(n+1)S_{22}&=(n-1)[(n-1)^2\ (K_{20}K_{02}+2K_{11}^2)+(n-2)(n-3)K_{22}]. \end{aligned}\right\} \quad (1.2)$$

We also need expressions for the symmetric functions in terms of power sums. They are derived from the published expressions, e.g. David and Kendall (1949, 1951, 1953, 1955) by putting $S_1=0$, or $S_{10}=S_{01}=0$, etc. Using non-augmented symmetric functions the univariate expressions of degree 4 are:—

$$Sx^4=S_4,\ Sx_1^3x_2=-S_4,\ 2Sx_1^2\ x_2^2=S_2^2-S_4,\ 6Sx_1^2x_2x_3=S_4-2S_2^2,\ 24Sx_1x_2x_3x_4=2S_2^2-S_4. \quad (1.3)$$

If a symmetric function of degree r has ρ suffixes, that is to say corresponds to a partition of r into ρ parts, we have

$$\mathcal{E}\left(\Sigma x_1^\alpha x_2^\beta\cdots x_\rho^\theta\right)=\frac{s^{(\rho)}}{n^{(\rho)}}\,S\,x_1^\alpha x_2^\beta x\cdots x_\rho^\theta\,. \quad (1.4)$$

For the term $x_1^\alpha\ x_2^\beta\cdots x_\rho^\theta$ will occur in Σ if A_1, A_2,$\cdots A\rho$ are present in the sample of s. The probability of this is $\dfrac{s^{(\rho)}}{n^{(\rho)}}$. For example $\mathcal{E}\ (\Sigma x_1^2\ x_2x_3)=\dfrac{s(s-1)(s-2)}{n(n-1)(n-2)}\ Sx_1^2\ x_2x_3$.

The same multiplier is clearly used for multivariate distributions. Thus

$$\mathcal{E}\ (\Sigma x_1^2\ y_1\ x_2\ y_3)=\frac{s(s-1)(s-2)}{n(n-1)(n-2)}\ S\ x_1^2\ y_1\ x_2y_3,$$

for $\rho=3$, that is to say three members of the population must be included in the sample,

Distribution of the Univariate Sum, Mean, or Difference of Means

The distribution now considered may be regarded as that of the sample mean k_1, of the sample sum $X'=sk_1=\Sigma x'$, or of the difference of two means. If we are comparing two samples of s_1 and s_2 members, we may regard them as samples from a population of $n=s_1+s_2$. Their difference $D=\dfrac{X'-nk_1}{s_2}$.

Consider $X=\Sigma x$. To each sample of s there corresponds an equiprobable sample of $n-s$ members, whose sum is $-X$. So the distributions of X in samples of s and $n-s$ differ in sign only. If n is even, the odd moments vanish when $n=2s$, and all moments vanish when $s=0$ or $n-s=0$, i.e. $s(n-s)=0$. Since $(n-2s)^2=n^2-4s\ (n-s)$, we see that s can only appear in expressions for the moments or cumulants as $n-2s$ or a power of it. Any even moment or cumulant of X is thus of the form $s(n-s)\ [f_1(n)+s(n-s)\ f_2(n)+\text{---}]$, and any odd moment or cumulant after the first of the form

$$s(n-s)\ (n-2s)\ [f_1(n)+s(n-s)\ f_2(n)+\text{----}].$$

This is a useful check. As an example of the algebra involved

$$\begin{aligned}
\kappa_3(X') &= \kappa_3(X) = \mathcal{E}(X^3) \\
&= \mathcal{E}[\Sigma x^3 + 3\ \Sigma x_1^2\ x_2 + 6\ \Sigma x_1\ x_2\ x_3] \\
&= \frac{s}{n} S x^3 + \frac{3s(s-1)}{n(n-1)}\ S x_1^2 x_2 + \frac{6s(s-1)\ (s-2)}{n(n-1)\ (n-2)}\ S x_1 x_2 x_3 \qquad \text{[from (1.4)]} \\
&= \frac{s}{n}\left[1 - \frac{3(s-1)}{n-1} + \frac{2(s-1)\ (s-2)}{(n-1)\ (n-2)}\right] S_3 \\
&= \frac{s(n-s)\ (n-2s)}{n(n-1)\ (n-2)}\ S_3 \\
&= \frac{s(n-s)\ (n-2s)}{n^2}\ K_3.
\end{aligned}$$

The first five cumulants of the distribution of X' are thus

$$\left.\begin{aligned}
&\kappa_1 = sK_1, \\
&\kappa_2 = s(n-s)n^{-1}\ K_2, \\
&\kappa_3 = s(n-s)\ (n-2s)n^{-2}\ K_3, \\
&\kappa_4 = s(n-s)n^{-2}\ (n+1)^{-1}\ [n(n+1)K_4 - 6s(n-s)K_2^2], \\
&\kappa_5 = s(n-s)\ (n-2s)n^{-3}\ (n+5)^{-1}\ [n\ (n+5)K_5 - 12s(n-s)\ (K_5+5K_2K_3)].
\end{aligned}\right\} \qquad (2.1)$$

The first three measures of departure from normality, where $G_r = K_{r+2}\ K_2^{-\frac{1}{2}r-1}$, are

$$\left.\begin{aligned}
&\gamma_1 = (n-2s)\ [ns\ (n-s)]^{-\frac{1}{2}} G_1 \\
&\gamma_2 = \frac{nG_2}{s(n-s)} - \frac{6(1+G_2)}{n+1} \\
&\gamma_3 = (n-2s)\ [ns(n-s)]^{-\frac{1}{2}} \left[\frac{nG_3}{s(n-s)} - \frac{12(5G_1+G_3)}{n+5}\right].
\end{aligned}\right\} \qquad (2.2)$$

These are the expressions needed in tests of significance. We see that the sign of G_1 is reversed if $s>\frac{1}{2}n$, and that the sample distribution is platykurtic if the population is mesokurtic ($G_2=0$), and if (for large n) $\cdot 2114\ n < s < \cdot 7886n$, it can be platykurtic even

if the population is leptokurtic. When sn^{-1} is small γ_2 will generally have the same sign as G_2. The cumulants of the distribution of k_1 are

$$\left.\begin{aligned}&\kappa_1 = K_1,\\ &\kappa_2 = (n-s)(ns)^{-1}K_2,\\ &\kappa_3 = (n-s)(n-2s)(ns)^{-2}K_3,\\ &\quad \text{etc.},\end{aligned}\right\} \tag{2.3}$$

the measures of departure from normality being as before.
For the difference of means D, the cumulants are

$$\left.\begin{aligned}&\kappa_1 = 0,\\ &\kappa_2 = n(s_1\ s_2)^{-1}K_2,\\ &\kappa_3 = (s_2-s_1)n(s_1s_2)^{-2}K_3,\\ &\quad \text{etc.}\end{aligned}\right\} \tag{2.4}$$

The measures of departure from normality are

$$\left.\begin{aligned}&\gamma_1 = (s_2-s_1)\ (ns_1s_2)^{-\frac{1}{2}}G_1,\\ &\gamma_2 = \frac{nG_2}{s_1s_2} - \frac{6(1+G_2)}{n+1},\\ &\gamma_3 = (s_2-s_1)\ (ns_1s_2)^{-\frac{1}{2}}\left[\frac{n\ G_3}{s_1s_2} - \frac{12(5G_1+G_3)}{n+5}\right].\end{aligned}\right\} \tag{2.5}$$

They can be made the basis of tests which are possibly preferable to Student's test in some cases.

The Sampling Variance of the Variance Estimate

The expectation of the sample variance estimate k_2 is of course K_2.
$k_2^2 = (s-1)^{-2}[(\Sigma x^2)^2 - 2s^{-1}\ \Sigma x^2\ (\Sigma x)^2 + s^{-2}\ (\Sigma x)^4]$
The expectation of each term can be worked out. It is convenient to subtract K_2^2 at once from the first term. The algebra is like that for (5.5) given later. We find

$$\begin{aligned}\text{Var}\ (k_2) &= \frac{n-s}{n(n+1)s(s-1)}\ [2ns\ K_2^2 + (ns-n-s-1)K_4] \qquad (3.1)\\ &= \frac{n-s}{n+1}\left[\frac{2K_2^2}{s-1} + \frac{\left\{1-\frac{s+1}{n(s-1)}\right\}}{s}\ K_4\right].\end{aligned}$$

The second form shows its relation to Fisher's expression $\frac{2K_2^2}{s-1} + \frac{K_4}{s}$ which is approached as n tends to infinity.

Some Covariances

It is easy to show that

$$\left.\begin{aligned}&\text{Cov}\ (k_1, k_2) = (n-s)\ (ns)^{-1}K_3,\\ &\text{Cov}\ (k_1, k_3) = (n-s)\ (ns)^{-1}K_4\end{aligned}\right\} \tag{4.1}$$

and from Wishart's results it follows that

$$\text{Cov}\ (k_1, k_r) = (n-s)\ (ns)^{-1} K_{r+1} \tag{4.2}$$

at least up to $r=7$.

SOME BIVARIATE RESULTS

The most important bivariate result is that

$$\mathcal{E}(k_{11}) = K_{11},$$

that is to say the expectation of the covariance estimate of the sample is equal to the population covariance estimate.

The only third order expression is

$$\text{Cov}\ (k_{20}, k_{01}) = (n-s)\ (ns)^{-1} K_{21}, \tag{5.1}$$

which is easily proved.

There are, however, three interesting results of the fourth order, of which we derive one in detail.

$$k_{11} = (s-1)^{-1}\ [\Sigma xy - s^{-1}\ \Sigma x \Sigma y]$$

$$\text{So Var}\ (k_{11}) = (s-1)^{-2}\ \mathcal{E}\ [(\Sigma xy)^2 - 2s^{-1}\ \Sigma xy \Sigma x \Sigma y + s^{-2}\ (\Sigma x)^2\ (\Sigma y)^2] - K_{11}^2$$

$$(s-1)^{-2}\ \mathcal{E}\ [(\Sigma xy)^2] - k_{11}^2 = (s-1)^{-2}\ \mathcal{E}\ [\Sigma x^2 y^2 + 2\ \Sigma\ x_1 x_2 y_1 y_2] - K_{11}^2$$

$$= (s-1)^{-2} \left[\frac{s}{n}\ S_{22} - \frac{s(s-1)}{n(n-1)}\ (S_{11}^2 - S_{22}) \right] - K_{11}^2$$

$$= \frac{(n-s)}{n(s-1)} \left[K_{11}^2 + \frac{s}{(n-1)\ (s-1)}\ S_{22} \right]. \tag{5.2}$$

$$s^{-1}\ (s-1)^{-2}\ \mathcal{E}\ [\Sigma xy\ \Sigma x\ \Sigma y] = s^{-1}(s-1)^{-2}\ \mathcal{E}\ [\Sigma x^2 y^2 + \Sigma(x_1^2 y_1 \bar{y}_2 + x_1 x_2 y_1^2 + 2x_1 x_2 y_1 y_2) + \Sigma x_1 x_2 y_2 y_3]$$

$$= n^{-1}\ (s-1)^{-2} \left[S_{22} + \frac{s-1}{n-1}\ (S_{11}^2 - 3S_{22}) + \frac{(s-1)\ (s-2)}{(n-1)\ (n-2)}\ (2S_{22} - S_{11}^2) \right]$$

$$= \frac{n-s}{n(n-2)\ (s-1)} \left[(n-1) K_{11}^2 + \frac{n-2s}{(n-1)\ (s-1)}\ S_{22} \right]. \tag{5.3}$$

$$s^{-2}\ (s-1)^{-2}\ \mathcal{E}\ [(\Sigma x)^2\ (\Sigma y)^2] = s^{-2}\ (s-1)^{-2}\ \mathcal{E}\ [\Sigma x^2 y^2 + \Sigma(2x_1^2\ y_1 y_2 + 2x_1 x_2 y_1^2 + x_1^2 y_2^2$$
$$+ 4x_1 x_2 y_1 y_2) + 2\ \Sigma\ (x_1^2 y_2 y_3 + x_2 x_3 y_1^2 + 2x_1 x_2 y_1 y_3) + 4\Sigma x_1 x_2 y_3 y_4]$$

$$= \frac{1}{ns(s-1)^2} \left[S_{22} + \frac{s-1}{n-1}\ (S_{20}\ S_{02} + 2S_{11}^2 - 7S_{22}) + \frac{2(s-1)\ (s-2)}{(n-1)\ (n-2)}\ (6S_{22} - S_{20} S_{02} - 2S_{11}^2) \right.$$
$$\left. + \frac{2(s-1)\ (s-2)\ (s-3)}{(n-1)\ (n-2)\ (n-3)}\ (S_{20}\ S_{02} + 2S_{11}^2 - 3S_{22}) \right]$$

$$= \frac{n-s}{n(n-2)\ (n-3) s(s-1)} \left[(n-1)\ (n-s-1)\ (K_{20} K_{02} + 2K_{11}^2) + \right.$$
$$\left. \{n(n+1) - 6s(n-s)\}\ (n-1)^{-1} (s-1)^{-1}\ S_{22} \right]. \tag{5.4}$$

Combining (5·2), (5·3) and (5·4) we have

$$\text{Var}\,(k_{11}) = \frac{n-s}{n(n-2)\,(n-3)\,s(s-1)}\Big[\,(n-1)\,(n-s-1)K_{20}\,K_{02} + \{2(n-1)^2-(n^2-n-2)s\}\,K_{11}^2 + (n-1)^{-1}\,(ns-n-s-1)\,S_{22}\Big].$$

But $n(n+1)\,S_{22} = (n-1)\,[(n-1)^2\,(K_{20}\,K_{02}+2K_{11}^2)+(n-2)\,(n-3)\,K_{22}]$.

$$\text{So Var}\,(k_{11}) = \frac{n-s}{(n+1)\,(s-1)}\Big[(n-2)^{-1}\,\{(n-1)\,K_{20}\,K_{02}+(n-3)\,K_{11}^2\} + (ns)^{-1}\,(ns-n-s-1)\,K_{22}\Big]$$

$$= \frac{n-s}{(n+1)\,(s-1)}\left[K_{20}\,K_{02}+K_{11}^2+\frac{K_{20}\,K_{02}-K_{11}^2}{n-2}+\frac{(ns-n-s-1)}{ns}\,K_{22}\right]. \qquad (5.5)$$

This tends to $\dfrac{K_{20}\,K_{02}+K_{11}^2}{s-1}+\dfrac{K_{22}}{s}$ as n tends to infinity.

Using quite similar algebra, we find

$$\text{Cov}\,(k_{20},k_{11}) = \frac{n-s}{(n+1)\,(s-1)}\left[2K_{20}\,K_{11}+\frac{(ns-n-s-1)K_{31}}{ns}\right]. \qquad (5.6)$$

$$\text{Cov}\,(k_{20},k_{02}) = \frac{n-s}{(n+1)\,(s-1)}\left[\frac{2(n-1)K_{11}^2-2K_{20}\,K_{02}}{n-2}+\frac{(ns-n-s-1)K_{22}}{ns}\right]. \qquad (5.7)$$

Clearly these approximate to the classical results as n tends to infinity. And with errors of order (s^{-1}, n^{-1}) they are equal to these values multiplied by $(n-s)\,n^{-1}$.

Now let $\rho = K_{11}\,(K_{20}\,K_{02})^{-\frac{1}{2}}$

$r = k_{11}\,(k_{20}\,k_{02})^{-\frac{1}{2}}$.

Clearly $\mathcal{E}\,(r)$ approximates to ρ. But we cannot obtain exact expressions for the bias and variance of r, any more than we can in the case of sampling from a probability distribution with arbitrary cumulants. However if

$k_{11}=K_{11}+\alpha$, $k_{20}=K_{20}+\beta$, $k_{02}=K_{02}+\gamma$, then

$$r=\rho\left(1+\frac{\alpha}{K_{11}}\right)\left(1+\frac{\beta}{K_{20}}\right)^{-\frac{1}{2}}\left(1+\frac{\gamma}{K_{02}}\right)^{-\frac{1}{2}}.$$

$$\text{So } \mathcal{E}\,(r)=\rho+\rho\mathcal{E}\left[\frac{-\alpha\beta}{2K_{11}\,K_{20}}-\frac{\alpha\gamma}{2K_{11}\,K_{02}}+\frac{3\beta^2}{8K_{20}^2}+\frac{3\gamma^2}{8K_{02}^2}+\frac{\beta\gamma}{4K_{20}\,K_{02}}\right]+O\left(\frac{n-s}{ns}\right)^2.$$

But $\mathcal{E}\,(\alpha\beta)=\text{Cov}\,(k_{11},k_{20})$, $\mathcal{E}\,(\beta^2)=\text{Var}\,(k_{20})$ and so on. Similarly

$$\text{Var}\,(r)=\rho^2\mathcal{E}\left[\frac{\alpha^2}{K_{11}^2}-\frac{\alpha\beta}{K_{11}\,K_{20}}-\frac{\alpha\gamma}{K_{11}\,K_{02}}+\frac{\beta^2}{4K_{20}^2}+\frac{\gamma^2}{4K_{02}^2}+\frac{\beta\gamma}{2K_{20}\,K_{02}}\right] + O\left(\frac{n-s}{ns}\right)^2. \qquad (5.8)$$

Hence the bias and variance of ρ approximate to the values already known when cumulants are given, multiplied by the factor $(n-s)\,n^{-1}$. This information is sufficient for many practical applications.

Table 1. *Inbreeding of hospital patients*

Disease	s	f=1/8	f=1/16	X	sK_1	Δ	$\Delta\sigma^{-1}$
Pulmonary tuberculosis	38	6	8	20	10·086	+9·914	+2·944
Other respiratory	58	6	11	23	15·394	+7·606	+1·864
Congenital malformations	24	2	7	11	6·370	+4·630	+1·719
Cancer	68	5	11	21	18·048	+2·952	+0·673
Other tuberculosis	39	1	10	12	10·351	+1·649	+0·487
Deficiency diseases	35	2	5	9	9·290	−0·290	−0·090
Central nervous system	62	3	8	14	16·456	−2·456	−0·584
Injuries	40	1	6	8	10·617	−2·617	−0·763
Cardio-vascular system	61	1	10	12	16·190	−4·190	−1·006
All respiratory diseases	96	12	19	43	25·480	+17·520	+3·448

PRACTICAL APPLICATION

D and M investigated relationships of the parents of 746 hospital patients. Their Table 8 shows the frequencies of various types of inbreeding among patients of various groups. Our Table 1 summarises it. Since they give good reasons for thinking that marriages more remote than those of first cousins are incompletely recorded we confine our calculations to marriages with a niece, nephew or first cousin. This only reduces the coefficient of inbreeding for inpatients from 0·01745 to 0·01671, so we only lose 4·2% of the information, which is the least reliable part of it. We also simplify the computation.

The coefficient of inbreeding of unions of degree 3 is $\frac{1}{8}$, for those of degree 4 (first cousins), $\frac{1}{16}$. Multiplying by 16, a patient is assigned a score $x=2$ if his parents were uncle and niece or aunt and nephew and of 1 if they were first cousins. Of the 746 patients 43 had $x=2$, 112 had $x=1$ and 591 had $x=0$. The k-statistics of the distribution of x are thus

$$K_1=\cdot265416,\ K_2=\cdot310668,\ K_3=\cdot346920,\ K_4=0\cdot284119,$$
$$G_1=+2\cdot0035,\ G_2=+2\cdot9161.$$

Hence from (2·2), in the absence of any effects of inbreeding, we should expect the total X from a sample of s to be distributed with standard deviation

$$=[s(n-s)\ (\cdot310668)\ n^{-1}]^{\frac{1}{2}}=\sqrt{s(746-s)}\ \ (0\cdot020407),$$

about a mean sK_1. In Table 1 we give the values of sK_1, the deviation $\Delta=X-sK_1$ and the deviation divided by the standard, $\Delta\sigma^{-1}$. Inbreeding would produce a positive

deviation if recessive genes lowered the resistance to a disease. If the sampling distribution of X were normal, we should regard the excess of inbreeding in the patients with pulmonary tuberculosis as very significant ($P=\cdot00162$) and that among the patients with other lung diseases as moderately so ($P=\cdot0312$). As there are nine samples the probability that one should have as low a P as $\cdot00162$ is $\cdot014$, but the significance of the value for the lung diseases would be doubtful. However the distribution is not normal, and we must allow for this.

$G_1=+2\cdot0035$ and $G_2=+2\cdot9161$. So from (2·2)
$\gamma_1=0\cdot29963$ $\gamma_2=0\cdot4941$.

γ_3 will be considerably smaller. It is possible to use one of the formulae for the normalisation of a non-normal distribution, but Pearson and Hartley's (1954) Table 42 suffices. With $\gamma_1=0\cdot3$ and $\gamma_2=\cdot05$, the upper 0·5% point is raised from 2·58 to 2·83. Thus the increased inbreeding in pulmonary tuberculosis is significant with $P<0\cdot005$. Cornish and Fisher's (1937) correction, including only terms of order of $s^{-\frac{1}{2}}$ and s^{-1} gives a normalised variate $\zeta=2\cdot764$ whence $P=\cdot00286$. The probability that the highest of nine deviations should reach the observed value is about 0·0254. The significance of the results for other lung diseases and for congenital malformations is lessened.

If we combine tubercular and other lung diseases, which seems justifiable, we find $\Delta\sigma^{-1}=3\cdot448$, $\gamma_1=0\cdot16268$, $\gamma_2=\cdot00341$. The sampling distribution is now fairly close to normal. The uncorrected value of P is $\cdot00028$, the corrected value about $\cdot0006$. That is to say there is less than once chance in 1500 that there is not a real excess of inbreeding among the parents of all patients of respiratory diseases. The excess is significant among the tuberculous, not so among the non-tuberculous. However on combining them, the significance is raised.

Let us now compare our test with Student's test. This would be made as follows. We compare the sample of 38 pthisical patients with the remaining 708. Defining x as above, the relevant values in the two samples are

$n_1=38$ $x_1=0\cdot526316$ $S_1=S\,(x_1-\bar{x}_1)^2=21\cdot4737$
$n_2=708$ $x_2=0\cdot251411$ $S_2=S\,(x_2-\bar{x}_2)^2=207\cdot2486$.

Hence $s^{-2}=\dfrac{n_1\,n_2}{S_1+S_2}\left(1-\dfrac{2}{n_1+n_2}\right)=123\cdot21245$,

$t=3\cdot055$, giving $P=\cdot0012$.

It is seen that t is almost the same as the deviation 2·944 of Table 1. But for a large total sample its distribution is taken to be normal. In fact this difference of means is not nearly normally distributed if either sample is small.

Discussion

The first point to be considered is whether the test here developed is likely to be needed in practice. We have no doubt that this is so where the distribution in the

total of the two samples is far from normal, and one of the samples to be compared contains less than 100 members. It may be that such cases will not be very common. But they are, we think, certain to occur if the effects of human inbreeding are studied by comparing the amounts of inbreeding in samples of patients with different diseases, rather than by comparing an inbred with an outbred group. In the case which we have considered, the probability of obtaining the observed result as the result of random sampling turned out to be about twice that calculated by Student's method. Our method is therefore at least sometimes of value.

It is clear that a whole array of formulae similar to Fisher's $\kappa(a^a\, b\beta\, ---)$ can be calculated. Perhaps this will best be done from Tukey's polykays. Several of our mathematical results are not new. But so far as we know our formulae (2·1) and (5·5) to (5·7) are new, and may prove useful. It should be possible to devise rules, as has been done for Fisher's expressions. But before a more comprehensive paradigm of formulae is set out, it may be desirable to discuss both the best symbolism, and the easiest methods of derivation.

While we agree with D and M that much more work is required along the lines which they have laid down, we think that their result as to pulmonary tuberculosis is more significant than they claim it to be. If they had had no reason to suspect that inbreeding might raise the frequency of pulmonary tuberculosis we might ask what was the probability that for at least one disease or group of diseases the inbreeding should be as high as their highest figure. This probability, as we saw, is about 2·5%. But given the results of Sutter and Tabah (1952) in France, it was reasonable to ask whether inbreeding also conduced to pulmonary tuberculosis in India. If the question is framed in this way, then results are significant at a level of 0·29%. Still greater significance is obtained if all pulmonary diseases are pooled.

SUMMARY

Expressions are given for the distributions of means, variances, co-variances, and correlation, in a sample from a finite population. It is claimed that this method is, in some cases, more accurate than Student's method for testing the significance of a difference between means. On applying our test to Dronamraju and Meera Khan's (1963) data on a hospital population, we find that their coefficient of inbreeding is very significantly raised, both in patients with pulmonary diseases, and in patients with respiratory diseases of all kinds.

REFERENCES

BARTLETT, M. S. (1935). The effect of non-normality on the *t*-distribution. *Proc. Camb. Phil. Soc.*, **31**, 223.

CORNISH, E. A. AND FISHER, R. A. (1937). Moments and cumulants in the specification of distributions. *Rev. Inst. Int. Stat.*, **5**, 307.

DAVID, F. N. AND KENDALL, M. G. (1949, 1951, 1953, 1955). Tables of symmetric functions. *Biometrika*, **36**, 431; **38**, 435; **40**, 427; **42**, 223.

DRONAMRAJU, K. R. AND MEERA KHAN, P. (1963). The frequency and effects of consanguineous marriages in Andhra Pradesh. *J. Genet.*, **58**, 387-401.

FISHER, R. A. (1925). Application of 'Student's' distribution. *Metron*, **5**, 90.

FISHER, R. A. (1928). Moments and product-moments of sampling distributions. *Proc. Lond. Math. Soc.*, **30**, 199.

KENDALL, M. G. AND STUART, A. (1958). *The Advanced Theory of Statistics. Vol. I. Distribution Theory.* Charles Griffin and Co. Ltd.: London.

PEARSON, E. S. AND HARTLEY, H. O. ed. (1954). *Biometrika Tables for Statisticians. Vol. I.* University Press: Cambridge.

"STUDENT" (1908). On the probable error of a mean. *Biometrika*, **6**, 1.

SUTTER, J. AND TABAH, J. (1952). Effets de la consanguinité et de l'endogamie. *Population*, **7**, 249.

TUKEY, J. W. (1950). Some sampling simplified. *J. Am. Stat. Ass.*, **45**, 501.

WISHART, J. (1952). Moment coefficients of the *k*-statistics in samples from a finite population. *Biometrika*, **39**, 1.