ON A METHOD OF TESTING THE HYPOTHESIS THAT AN OBSERVED SAMPLE OF n VARIABLES AND OF SIZE N HAS BEEN DRAWN FROM A SPECIFIED POPULATION OF THE SAME NUMBER OF VARIABLES

By John W. Fertig

With the Technical Assistance of Margaret V. Leary*

The problem of determining whether or not a given observation may be regarded as randomly drawn from a certain population completely specified with respect to its parameters is readily solved if the probability integral of that population be known. In particular if the population specified be a normal population, one may calculate the relative deviate $(x - a)/\sigma$, where a and σ are the population mean and standard deviation respectively, and refer to tables of the normal probability integral. The hypothesis that x was drawn from this population may be rejected if P is less than an arbitrarily fixed value, say $\leqq .01$. Generalizations of this problem may be made in two directions: 1) May a single observation simultaneously made on n variables be considered as randomly drawn from a specified population of n variables? 2) May a sample of one variable and of size N be regarded in its entirety as randomly drawn from a specified univariate population?

The solution to the first problem for the case of sampling from a normal population of n variables was given by Karl Pearson in 1908[1] as the "Generalized Probable Error." Let

$$\chi^2 = \frac{1}{P}\left\{ \underset{i,j=1}{\overset{n}{S}} P_{ij}\left[\frac{(x_i - a_i)(x_j - a_j)}{\sigma_i\sigma_j}\right]\right\}$$

where a_i and σ_i are the population mean and standard deviation respectively of the i^{th} variable, and P_{ij} is the usual cofactor of the element in the i^{th} row and j^{th} column of the determinant P of population correlation coefficients. That is,

$$P = |\rho_{ij}|;\ i, j = 1, 2, 3, \cdots, n.$$

The probability of an observation yielding a smaller discrepancy than that represented by the value of χ^2, i.e., lying between 0 and χ^2, may then be calculated from Tables of the Incomplete Normal Moment Functions[2]. The tables are entered in terms of $(\chi^2)^{\frac{1}{2}}$ and $(n - 1)$, and the tabled value multiplied by $(2\pi)^{\frac{1}{2}}$ or 2 depending upon whether n be even or odd respectively.

* From the Memorial Foundation for Neuro-Endocrine Research and the Research Service of the Worcester State Hospital, Worcester, Massachusetts.

The probability of an observation giving a greater discrepancy is then the complement of this value. Obviously, this latter probability may be obtained directly by entering tables of the X^2 distribution such as Elderton's[3] with n degrees of freedom, or through the use of Tables of the Incomplete Γ-Function[4].

The second problem, limited to the case of sampling from a normal population, was investigated by J. Neyman and E. S. Pearson in 1928[5]. The observed sample may be regarded as a point in N-dimensional space, where N is the sample size. Criteria for the acceptance or rejection of the hypothesis may be associated with contour surfaces in this space, so that in moving out from contour to contour the hypothesis becomes less and less reasonable. Frequently, contour surfaces on which the mean or standard deviation is constant are used for the testing of this hypothesis. Such surfaces are deficient inasmuch as they are not "closed" contours. Another contour system which appears more satisfactory is that of equiprobable pairs of m and s. The latter system in fact encloses roughly the same region as do the separate contours for the means and standard deviations. These systems are of course dependent on the particular statistics chosen to describe the sample and are further limited in that they do not take into account the probability of alternative hypotheses concerning the origin of the sample.

Using the principle of maximum likelihood Neyman and Pearson have developed a system of contours which is free of the above limitations. The system so derived is in fact quite similar to that of equiprobable pairs m and s. In a later paper[6], these same investigators have shown that this method of maximum likelihood does enable one to select the most efficient criteria for the testing of an hypothesis. The criterion selected on this basis is defined as

$$\lambda = \frac{\text{Likelihood that sample came from specified population}}{\text{Maximum likelihood that sample came from some other population}}$$

$$= (s^2/\sigma^2)^{N/2} e^{-N/2\left[\frac{s^2 + (\bar{x} - a)^2}{\sigma^2} - 1\right]}$$

where a and σ are the population mean and standard deviation respectively, and $\bar{x}$ and s the sample mean and standard deviation.

λ is constant upon certain contour surfaces in N-dimensional space, and diminishes on passing outward. The form of the surfaces is independent of N. It is evident that λ must lie between zero and unity. When it is close to unity we know that it is reasonable to assume that our hypothesis is true, when small we know that it is unreasonable. But we must know the probability of λ less than a certain value occurring when the hypothesis tested is true, so that we may control another source of error, namely, that of rejecting the hypothesis when it is true. In other words, we must know the sampling distribution of λ, so that we will reject the hypothesis only when the probability of obtaining a smaller value is negligible, say $P_\lambda \leqq .01$. Neyman and Pearson were not able to evaluate this distribution but they were able to integrate the original density function of the population appropriate to N-dimensional space outside of the various λ contours. This they were able to do by effecting a transformation of the density function and contours to the plane of m and s. These values of $\dot{P}_\lambda$ have been tabled by them[7], the tables being entered in terms of N and k, where

$$k = \log\left[\frac{s^2 + (\bar{x} - a)^2}{\sigma^2}\right] - \log(s^2/\sigma^2)$$

The generalization of either of the above problems requires a criterion to test an hypothesis which may be formulated as follows: Given a sample Σ of n variables and of size N with means $\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_n$, standard deviations $s_1, s_2, \cdots, s_n$, and correlation coefficients $r_{12}, r_{13}, \cdots, r_{1n}, r_{23}, \cdots, r_{2n}, \cdots, r_{(n-1)n}$, may we regard this sample as randomly drawn from a population π of n variables and completely specified with respect to all its parameters? We shall restrict our inquiries to the case where π is a normal population. In this case the distribution law is

$$f(x_1, x_2, \cdots, x_n) = \frac{1}{(2\pi)^{n/2}\sigma_1\sigma_2\cdots\sigma_n P^{\frac{1}{2}}} e^{-\phi}$$

where

$$\phi = +\frac{1}{2P}\left\{\underset{i,j=1}{\overset{n}{S}} P_{ij}\left[\frac{(x_i - a_i)(x_j - a_j)}{\sigma_i\sigma_j}\right]\right\}$$

where a_i and σ_i are the population mean and standard deviation respectively of the i^{th} variable, and P and P_{ij} are as previously defined.

Thus the probability that Σ has been drawn from π with its N values of $x_{i\alpha}$ $(i = 1, 2, \cdots, n)$ lying in the interval $x_{i\alpha} \pm \frac{1}{2}dx_{i\alpha}$; $(\alpha = 1, 2, \cdots, N)$ is given by

$$C = \left[\frac{1}{(2\pi)^{n/2}\sigma_1\sigma_2\cdots\sigma_n P^{\frac{1}{2}}}\right]^N e^{-\Theta}\, dX$$

where

$$\Theta = \frac{1}{2P}\left\{\underset{i,j=1}{\overset{n}{S}} P_{ij} \underset{\alpha=1}{\overset{N}{S}}\left[\frac{(x_{i\alpha} - a_i)(x_{j\alpha} - a_j)}{\sigma_i\sigma_j}\right]\right\}$$

$$= \frac{N}{2P}\left\{\underset{i,j=1}{\overset{n}{S}} P_{ij}\left[\frac{s_i s_j r_{ij} + (\bar{x}_i - a_i)(\bar{x}_j - a_j)}{\sigma_i\sigma_j}\right]\right\}$$

$$dX = \prod_{i=1}^{n}\prod_{\alpha=1}^{N} dx_{i\alpha}$$

The likelihood that Σ has been drawn from any other normal population, such as π', is given by

$$C' = \left[\frac{1}{(2\pi)^{n/2}\sigma_1'\sigma_2'\cdots\sigma_n' P'^{\frac{1}{2}}}\right]^N e^{-\Theta'}\, dX$$

where

$$\Theta' = \frac{N}{2P'}\left\{ \mathop{S}_{i,j=1}^{n} P'_{ij}\left[\frac{s_i s_j r_{ij} + (\bar{x}_i - a'_i)(\bar{x}_j - a'_j)}{\sigma'_i \sigma'_j}\right]\right\}$$

The population from which it is most likely that Σ has been drawn is that for which C' is a maximum. The values of the parameters of this population may be obtained by putting

$$\frac{\partial C'}{\partial a'_i} = 0, \qquad \frac{\partial C'}{\partial \sigma'_i} = 0; \qquad (i = 1, 2, \cdots, n)$$

$$\frac{\partial C'}{\partial \rho'_{ij}} = 0; \qquad (i, j = 1, 2, \cdots, n)$$

These conditions are fulfilled when

$$a'_i = \bar{x}_i;\ \sigma'_i = s_i; \quad (i = 1, 2, \cdots, n)$$

$$\rho'_{ij} = r_{ij}; \quad (i, j = 1, 2, \cdots, n)$$

So that

$$C'_{\text{max.}} = \left[\frac{1}{(2\pi)^{n/2} s_1 s_2 \cdots s_n R^{\frac{1}{2}}}\right]^N e^{-nN/2}$$

where

$$R = |r_{ij}|; \qquad i, j = 1, 2, \cdots, n$$

The appropriate criterion to select in order to test our hypothesis is thus

$$\lambda = \frac{C}{C'_{\text{max.}}} = \left[\frac{s_1 s_2 \cdots s_n R^{\frac{1}{2}}}{\sigma_1 \sigma_2 \cdots \sigma_n P^{\frac{1}{2}}}\right]^N e^{-w}$$

where

$$w = \frac{N}{2}\left\{ \mathop{S}_{i,j=1}^{n} \frac{P_{ij}}{P}\left[\frac{s_i s_j r_{ij} + (\bar{x}_i - a_i)(\bar{x}_j - a_j)}{\sigma_i \sigma_j}\right] - n\right\}$$

The equations λ = constant represent a series of contours in N-dimensional space. As we move outward from contour to contour our hypothesis becomes less and less acceptable. Although we may be confident that the use of this criterion will minimize the chance of accepting the hypothesis when it is false we must know the frequency with which samples occur outside of a given λ contour when the hypothesis is true. In other words, we must know the integral of C outside of various contours, or else we must know the sampling distribution of λ. The former is an exceedingly difficult method for n greater than unity. Thus for the case of $n = 2$ we should have to integrate some such expression as

$$k s_1^{N-2} s_2^{N-2} e^{-\Theta}(1 - r_{12}^2)^{\frac{N-4}{2}}\, d\bar{x}_1 d\bar{x}_2 ds_1 ds_2 dr_{12}$$

outside of the various contours. Nor have we so far been able to evaluate the sampling distribution. We can however give an expression for the moments of λ and thus reach an approximate distribution.

Wilks[8] has derived expressions for the moment coefficients about zero for the maximum likelihood criterion that k samples of n variables and of N_t observations each have been drawn from the same unspecified normal population of n variables. Thus,

$$\mu_h'(\lambda) = \prod_{t=1}^{k} \left\{ \left[\frac{\overset{k}{\underset{t=1}{S}} N_t}{N_t} \right]^{\frac{h\,n\,N_t}{2}} \prod_{i=1}^{n} \left[\frac{\Gamma\left(\frac{N_t(1+h)-i}{2}\right)}{\Gamma\left(\frac{N_t-i}{2}\right)} \right] \right\} \prod_{i=1}^{n} \left\{ \frac{\Gamma\left[\frac{\overset{k}{\underset{t=1}{S}} N_t - i}{2}\right]}{\Gamma\left[\frac{(1+h)\overset{k}{\underset{t=1}{S}} N_t - i}{2}\right]} \right\}$$

from which we can write expressions giving the moment coefficients about zero for the λ criterion for two samples

$$\mu_h'(\lambda) = \frac{(N_1+N_2)^{\frac{n\,h\,(N_1+N_2)}{2}}}{N_1^{\frac{n\,h\,N_1}{2}}\, N_2^{\frac{n\,h\,N_2}{2}}} \prod_{i=1}^{n} \left\{ \frac{\Gamma\left[\frac{N_1(1+h)-i}{2}\right] \Gamma\left[\frac{N_2(1+h)-i}{2}\right] \Gamma\left[\frac{N_1+N_2-i}{2}\right]}{\Gamma\left(\frac{N_1-i}{2}\right) \Gamma\left(\frac{N_2-i}{2}\right) \Gamma\left[\frac{(N_1+N_2)(1+h)-i}{2}\right]} \right\}$$

The limit of this latter expression as $N_2 \to \infty$ will be the moment coefficient about zero for the λ criterion that one sample has been drawn from a specified population. Thus

$$\underset{N_2\to\infty}{\text{Lim.}}\ \mu_h'(\lambda) = \prod_{i=1}^{n} \left\{ \frac{\Gamma\left[\frac{N_1(1+h)-i}{2}\right]}{\Gamma\left(\frac{N_1-i}{2}\right)} \right\} \left(\frac{2e}{N_1}\right)^{\frac{n\,h\,N_1}{2}} (1+h)^{\frac{-n\,N_1(1+h)}{2}}$$

Various roots of λ are distributed to a good degree of approximation according to a function of the form

$$f(t) = \frac{\Gamma(m_1+m_2)}{\Gamma(m_1)\,\Gamma(m_2)}\, t^{m_1-1}\,(1-t)^{m_2-1}$$

where

$$m_1 = \mu_1'(\mu_1' - \mu_2')/(\mu_2' - \mu_1'^2) \; ; \qquad m_2 = (1 - \mu_1')m_1/\mu_1'$$

and the value of μ_h' for roots of λ may be obtained by replacing h in the original expression by h times the desired root. Measures of the skewness and kurtosis of this distribution are given by

$$B_1 = 4(m_1 - m_2)^2(m_1 + m_2 + 1)/m_1m_2(m_1 + m_2 + 2)^2$$

$$B_2 = 3B_1(m_1 + m_2 + 2) + 6(m_1 + m_2 + 1)/2(m_1 + m_2 + 3)$$

A comparison with the true measures of skewness and kurtosis for various roots of λ as given by

$$B_1 = \mu_3^2/\mu_2^3 \; ; \qquad B_2 = \mu_4/\mu_2^2$$

will afford a measure of the goodness of the approximation and the range of values of N for which any particular root will be distributed as assumed.

Investigating the moments for n from one to four and N from three to fifty we note that in the case of samples of two and three variables, $\lambda^{1/N}$ follows the assumed distribution for N from 3 to 15; $\lambda^{2/N}$ from 15 to 30; $\lambda^{3/N}$ from 30 to 50. In the case of four variables, $\lambda^{1/2N}$ follows the distribution for N from 5 to 10; $\lambda^{1/N}$ from 10 to 20; $\lambda^{2/N}$ from 20 to 40; $\lambda^{3/N}$ from 40 to 50. It appears likely that for higher values of n, for N small, some such root as $\lambda^{1/2N}$ or $\lambda^{1/3N}$ will follow the assumed distribution, while as N increases smaller roots will follow it. For any value of n, the smallest permissible value of N is $(n + 1)$.

The probability that a smaller value of λ will be obtained when the sample has actually been drawn from π, i.e., P_λ, may thus be obtained by reference to Tables of the Incomplete B-Function[9] with $p = m_1$, $q = m_2$, $x =$ value of the particular root of the observed λ. We may also get the 1% and 5% levels of significance directly from Fisher's[10] tables of "z" or Snedecor's[11] tables of "F" $(= e^{2z})$, by taking

$$n_1 = 2m_2 \; ; \qquad n_2 = 2m_1 \; ; \qquad L = n_2/(n_2 + n_1F) \, ,$$

where L is the desired root of λ. Linear interpolation will generally suffice except for very small values of N.

For the case of $N \rightarrow \infty$, we have

$$\lim_{N\rightarrow\infty} \mu_h'(\lambda) = (1 + h)^{-\overset{n+1}{\underset{i=2}{S}} i/2}$$

Thus the quantity $(-2 \log \lambda)$ will be distributed in the χ^2 distribution with $\overset{n}{\underset{i=2}{S}} i$ degrees of freedom.

A table of the 1% and 5% levels of significance for n equal one to four, and values of N from five to ∞ is given below

5% and 1% Levels of Significance of "λ"

— N —

n		5	10	15	20	30	40	50	∞
1	5%	.025	.037	.041	.043	.045	.046	.047	.050
	1%	.003	.006	.008	.008	.009	.009	.009	.010
2	5% $\times 10^{-2}$	.046	.173	.234	.269	.308	.330	.343	.392
	1% $\times 10^{-3}$	.026	.168	.260	.305	.372	.409	.428	.525
3	5% $\times 10^{-3}$	.001	.036	.072	.097	.125	.143	.155	.211
	1% $\times 10^{-4}$	$.000^{+}$	.019	.047	.076	.101	.117	.128	.194
4	5% $\times 10^{-5}$		.026	.106	.174	.295	.356	.418	.710
	1% $\times 10^{-6}$		.007	.040	.075	.145	.185	.221	.466

A check on the accuracy of the method of approximation used may be obtained by comparing the values of P_λ for the case of $n = 1$ with the exact values given by Neyman and Pearson. For $n = 10$, $\lambda^{1/N}$ is distributed as assumed with $m_1 = 9.0562$, $m_2 = 0.9987$. For the case of $(\bar{x} - a)/\sigma = 0.2$, $s/\sigma = 1.2$, we find $k = 0.48439$, $\lambda^{1/N} = .94395$. From the Tables of the Incomplete B-Function we find $P_\lambda = .5936$, from Neyman and Pearson's tables, .5935.

No studies have been made on the extent of deviation from normality permissible for the application of the test. There is no reason to doubt, however, that as much deviation is permissible as in the case of the univariate λ. From theoretical considerations and from sampling studies Neyman and Pearson conclude that the univariate λ technique holds for deviation from normality to the extent of ± 0.5 for B_1 and 2.5 to 4.2 for B_2.

We are confident that this generalized λ technique will be found useful in biological research. If the n variables were uncorrelated we would be able to test whether the sample had been drawn from the population of n variables by successive applications of the univariate λ technique and then combining the resulting probabilities. In general, however, there will be some correlation between the variables, however slight. The method here proposed will take account of all possible intercorrelations, and consequently all multiple and partial correlations.

Now, if P_λ is less than some arbitrarily fixed value, say $\leqq .01$, we may decide which variable or variables contributes most to this result, by performing simpler λ tests. It may be due to one or more of the means, standard deviations, or correlation coefficients. As may often be the case, it is not due to any one factor but to contributions from all of them. That is, all possible factors tested separately might show a fairly reasonable value of P, but if all the separate values are combined somehow, as by means of this λ method, the resultant P may be too small. It is in such problems that this technique should provide valuable information.

In case k samples of n variables are available it should be possible to determine whether all of them have come from the same specified population of n variables by performing k λ tests and combining the separate values of P_λ. Such a hypothesis may best be tested, however, by a further extension of the λ theory which the writers are at present investigating.

The following problem is chosen to illustrate the computations involved in the application of the test. Many of the investigations pursued at the Worcester State Hospital attempt to differentiate between schizophrenic patients and normal controls. In one such type of investigation various blood constituents were determined, namely, Urea N_2 (mg./100 cc.), Uric Acid N_2 (mg./100 cc.), Creatine N_2 (mg./100 cc.) for a sample of twenty-five schizophrenic patients. Previous investigations on these same variables for a large series of normal controls yielded constants which for the purpose of the example may be considered as the population parameters. Past studies on these variables have not shown any marked degree of non-normality for the various distributions.

These variables are designated as

$$1 = \text{Urea } N_2\,; \qquad 2 = \text{Uric Acid } N_2\,; \qquad 3 = \text{Creatine } N_2$$

The parameters of the population are given by

$$a_1 = 16.03\,; \qquad a_2 = 1.40\,; \qquad a_3 = 1.25$$

$$\sigma_1^2 = 20.268\,; \qquad \sigma_2^2 = 0.029\,; \qquad \sigma_3^2 = 0.025$$

$$\rho_{12} = .3075\,; \qquad \rho_{13} = .1232\,; \qquad \rho_{23} = .3853$$

The statistics for the sample of twenty-five are

$$\bar{x}_1 = 15.56\,; \qquad \bar{x}_2 = 1.42\,; \qquad \bar{x}_3 = 1.25$$

$$s_1^2 = 10.486\,; \qquad s_2^2 = 0.043\,; \qquad s_3^2 = 0.025$$

$$r_{12} = -.0161\,; \qquad r_{13} = .0925\,; \qquad r_{23} = .2174$$

None of these statistics differs significantly from the corresponding parameters.

$$R = 0.9443\,; \qquad P = 0.7710\,;$$

$$P_{12}/P = -0.3373\,; \qquad P_{13}/P = -0.0061\,; \qquad P_{23}/P = -0.4506\,;$$

$$P_{11}/P = 1.1045\,; \qquad P_{22}/P = 1.2773\,; \qquad P_{33}/P = 1.1744$$

$$w = 12.5\ (0.3802) = 4.7531$$

$$(s_1^2\ s_2^2\ s_3^2\ R/\sigma_1^2\ \sigma_2^2\ \sigma_3^2\ P) = 0.9001$$

$$\log \lambda = 12.5 \log (0.9001) - 4.7531 \log e = \bar{3}.3641$$

$$\lambda = .0023$$

Since the 5% level of significance is about .0001, we thus conclude that the patients are not differentiated from the control population with respect to these variables.

REFERENCES

1. PEARSON, KARL. Biometrika, vol. 6, 1908. pp. 59–68.
2. Tables for Statisticians and Biometricians, Part I. pp. xxiv–xxviii, 22–23.
3. Ibid. pp. xxxi–xxxiii, 26–28.
4. Tables of the Incomplete Γ-Function, 1934.
5. NEYMAN, J. AND PEARSON, E. S. Biometrika, vol. 20, 1928. pp. 175–241.
6. Ibid. Phil. Trans. Roy. Soc. A, vol. 231, 1933. pp. 289–337.
7. Tables for Statisticians and Biometricians, Part II. pp. clxxx–clxxxv, 221–223.
8. WILKS, S. S. Biometrika, vol. 24, 1932. pp. 471–494.
9. Tables of the Incomplete Beta-Function, 1934.
10. FISHER, R. A. Statistical Methods for Research Workers. Fourth Edition, 1932.
11. SNEDECOR, G. W. Calculation and Interpretation of Analysis of Variance and Covariance, 1934.