SOME SIGNIFICANCE TESTS FOR NORMAL BIVARIATE DISTRIBUTIONS

By D. S. Villars and T. W. Anderson

United States Rubber Company, Passaic, New Jersey, and Princeton University

1. Introduction. In the theory of linear regression of y on x where y is normally distributed about a linear function of x, say $\nu + \beta x$, where x is a "fixed" variate, the t-test for the hypothesis that β is zero (that y is distributed about ν; independent of x) is well known. In this paper we apply some general statistical theory to the similar problem where x and y are jointly normally distributed. This case is commonly known as the case of "error in both variates." We derive a criterion for testing the hypothesis that the population means are the coordinates of a specified point when the ratio of the variances and the population correlation coefficient are known. When the ratio of variances is known, a criterion is derived to test whether the correlation coefficient is zero.

2. The means. Let us consider a sample of n pairs of observations $(x_1, y_1; x_2, y_2; \cdots; x_n, y_n)$ from a normal bivariate population. Let the variances of x and of y be σ_x^2 and σ_y^2, respectively; and the correlation coefficient, say ρ, be zero. Suppose the ratio of the weight of y to the weight of x, say $\gamma = w_y/w_x = \sigma_x^2/\sigma_y^2$, is known although the variances are not known. It is clear then, that $\sqrt{\gamma}\, y$ has variance σ_x^2. Since the observations y_i $(i = 1, 2, \cdots, n)$ can be transformed into revised observations $\sqrt{\gamma}\, y_i = y_i'$, we lose no generality by assuming that x and y are both distributed with variance σ^2.

Under the assumption of equality of variances and independence of variates we shall derive a criterion for testing the null hypothesis that each observation x_i is of a variate distributed about the same population mean μ and each observation y_i is of a variate distributed about the same population mean ν. The hypothesis may be stated symbolically as:

$$H_0: E(x) = \mu, \qquad E(y) = \nu,$$

given $\sigma_x^2 = \sigma_y^2 = \sigma^2$ and $\rho = 0$. We can write

$$\sum_{i=1}^{n} (x_i - \mu)^2 = n(\bar{x} - \mu)^2 + S_x,$$

$$\sum_{i=1}^{n} (y_i - \nu)^2 = n(\bar{y} - \nu)^2 + S_y,$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

$$S_x = \sum_{i=1}^{n} (x_i - \bar{x})^2, \qquad S_y = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

Then $n(\bar{x} - \mu)^2/\sigma^2$ and $n(\bar{y} - \nu)^2/\sigma^2$ are each distributed independently as χ^2 with one degree of freedom and each of S_x/σ^2 and S_y/σ^2 follow the χ^2-law with $n - 1$ degrees of freedom. If we define

$$(1) \qquad r = \sqrt{(\bar{x} - \mu)^2 + (\bar{y} - \nu)^2}, \qquad S_r = S_x + S_y,$$

then nr^2/σ^2 and S_r/σ^2 have independent χ^2-distributions with 2 and $2n - 2$ degrees of freedom, respectively.

It follows from this that

$$(2) \quad R = \frac{nr^2}{2\sigma^2} \bigg/ \frac{S_r}{(2n-2)\sigma^2} = n(n-1)\frac{r^2}{S_r} = n(n-1)\frac{(\bar{x} - \mu)^2 + (\bar{y} - \nu)^2}{S_x + S_y},$$

has the F-distribution with 2 and $2n - 2$ degrees of freedom.

Let us define F_α so

$$(3) \qquad \int_{F_\alpha}^{\infty} h_{2,2n-2}(F)\, dF = \alpha,$$

where $h_{2,2n-2}(F)$ is the F-distribution with 2 and $2n - 2$ degrees of freedom and $0 \leq \alpha \leq 1$. Then the probability is α that the sample statistic R is greater than or equal to F_α, i.e.,

$$(4) \qquad P\{R \geq F_\alpha\} = \alpha.$$

In considering a sample value of R, at significance level α, one rejects the hypothesis of the means being μ and ν, respectively, if R is larger than F_α, i.e., larger than 1 and larger than the α significance point in Snedecor's tables [1].

This F-test is a straightforward generalization to the bivariate case of the usual t-test as applied to the univariate case. In each case the sum of squares of distances of the observations from the population mean is broken up into the sum of squares of distances from the sample mean plus n times the square of the distance from the sample mean to the population mean. The t-test for the univariate case depends on the ratio of the distance of the sample mean from the population mean to the square root of the sum of squares of distances from the observations to the sample mean. The proposed F-test depends upon the ratio of the square of the distance of the sample mean from the population mean to the sum of squares of distances from the observations to the sample mean.

It can easily be shown that the likelihood ratio criterion for this hypothesis is

$$(5) \qquad \lambda = \left[\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 + \sum_{i=1}^{n} (y_i - \bar{y})^2}{\sum_{i=1}^{n} (x_i - \mu)^2 + \sum_{i=1}^{n} (y_i - \nu)^2}\right]^n = \left[1 + \frac{R}{n-1}\right]^n.$$

The hypothesis considered here is one of a class of hypotheses treated by Kolodziejczyk [2] in a paper in which he considers the likelihood ratio criterion for a set of general linear hypotheses.

Equation (4) may be written

$$P\{(\bar{x} - \mu)^2 + (\bar{y} - \nu)^2 \geq r_\alpha^2\} = \alpha, \tag{6}$$

where $r_\alpha^2 = F_\alpha\ (S_x + S_y)/[n(n - 1)]$. The probability is α that the distance from the sample means $\bar{x}$, $\bar{y}$ to the population means μ, ν is greater than or equal to r_α. We may call r_α the *fiducial radius* [3], and the equation $(\bar{x} - \mu)^2 + (\bar{y} - \nu)^2 = r_\alpha^2$ defines the *confidence region* for the population means.

Suppose we have two samples of n_1 and n_2 pairs of observations, respectively, from normal bivariate distributions. If the population mean of each x variate is μ and the population mean of each y variate is ν, the population variance of each variate is σ^2, and the correlation coefficient is zero, then the sample means $\bar{x}_1$ and $\bar{y}_1$ of the first sample and $\bar{x}_2$ and $\bar{y}_2$ of the second sample follow normal distributions. Also $\bar{x}_1 - \bar{x}_2$ and $\bar{y}_1 - \bar{y}_2$ are normally distributed. Then $r'^2 = n_1 n_2/(n_1 + n_2)[(\bar{x}_1 - \bar{x}_2)^2 + (\bar{y}_1 - \bar{y}_2)^2]/\sigma^2$ has the χ^2-distribution with 2 degrees of freedom. Let

$$S'_{r'} = \sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)^2 + \sum_{i=1}^{n_1} (y_{1i} - \bar{y}_1)^2 + \sum_{i=1}^{n_2} (x_{2i} - \bar{x}_2)^2 + \sum_{i=1}^{n_2} (y_{2i} - \bar{y}_2)^2,$$

where x_{1i}, y_{1i} $(i = 1, 2, \cdots, n_1)$ are the pairs of observations in the first sample and x_{2i}, y_{2i} $(i = 1, 2, \cdots, n_2)$ are the pairs of observations in the second sample. $S'_{r'}/\sigma^2$ is distributed according to the χ^2-distribution with $(2n_1 + 2n_2 - 4)$ degrees of freedom because it is the sum of quantities independently distributed as χ^2. Then

$$R' = \frac{n_1 n_2 r'^2}{2(n_1 + n_2)\sigma^2} \Bigg/ \frac{S'_{r'}}{(2n_1 + 2n_2 - 4)\sigma^2} = \frac{n_1 n_2 (n_1 + n_2 - 2) r'^2}{(n_1 + n_2) S'_{r'}}$$

has the F-distribution with 2 and $(2n_1 + 2n_2 - 4)$ degrees of freedom. This fact yields us a significance test for the hypothesis that both the means of the x variates and the means of the y variates for the two populations are the same. We can also set up confidence regions for $\mu_1 - \mu_2$ and $\nu_1 - \nu_2$.

Now let us consider a sample from a normal bivariate population with means μ and ν, variances σ_x^2 and σ_y^2 and correlation coefficient ρ. Suppose $\gamma = \sigma_x^2/\sigma_y^2$ and ρ are known. The transformation

$$x = \frac{\sqrt{1 + \rho}\, x' + \sqrt{1 - \rho}\, y'}{\sqrt{2}}, \qquad y = \frac{\sqrt{1 + \rho}\, x' - \sqrt{1 - \rho}\, y'}{\sqrt{2}\, \gamma}, \tag{8}$$

gives us the variates x' and y' which are distributed independently and with variance σ_x^2. Applying the results above we see that

$$\begin{aligned} R &= n(n-1) \frac{(\bar{x}' - \mu')^2 + (\bar{y}' - \nu')^2}{\sum_{i=1}^{n} (x'_i - \bar{x}')^2 + \sum_{i=1}^{n} (y'_i - \bar{y}')^2} \\ &= n(n-1) \frac{(\bar{x} - \mu)^2 - 2\rho\sqrt{\gamma}\, (\bar{x} - \mu)(\bar{y} - \nu) + \gamma(\bar{y} - \nu)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2 - 2\rho\sqrt{\gamma} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) + \gamma \sum_{i=1}^{n} (y_i - \bar{y})^2} \end{aligned} \tag{9}$$

has the F-distribution with 2 and $2n - 2$ degrees of freedom. From this we derive significance tests, fiducial radii, and confidence regions as before.

The above distributions, significance tests, and confidence regions are easily generalized to multivariate normal distributions. Suppose we have a sample of n k-tuples of observations $\{x_{i\alpha}\}$ $(i = 1, 2, \cdots, k; \alpha = 1, 2, \cdots, n)$ from a k-variate normal distribution. Let the expected value of each variate x_i be zero $(i = 1, 2, \cdots, k)$, the variance of each variate be σ^2 and each correlation coefficient be zero. Then

$$R'' = \frac{n(n-1)\sum_{i=1}^{k} \bar{x}_i^2}{\sum_{i=1}^{k}\sum_{\alpha=1}^{n} (x_{i\alpha} - \bar{x}_i)^2} \tag{10}$$

has the F-distribution with k and $k(n - 1)$ degrees of freedom. Significance tests, confidence regions, and fiducial radii follow from this fact.

3. Linear Regression. If one has a sample of n pairs of observations $(x_1, y_1; x_2, y_2; \cdots; x_n, y_n)$ from a normal bivariate population and wishes to fit a straight line to the scatter of sample points, one fits the line in such a way that the sum of squares of distances from the sample points to the line is a minimum ("error in both variates").

It is easily shown that this line goes through the point whose coordinates are the sample means $(\bar{x}, \bar{y}.)$ If the slope of a line through $(\bar{x}, \bar{y})$ is $\tan \theta$, the distance from a sample point (x_i, y_i) to the line is $(x_i - \bar{x}) \sin \theta - (y_i - \bar{y}) \cos \theta$. The sum of squares of distances from sample points to the line is

$$\sin^2 \theta\, S_x - 2 \sin \theta \cos \theta\, S_{xy} + \cos^2 \theta\, S_y,$$

where

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

If we minimize the above expression with respect to θ we find

$$b = \tan \theta = \frac{S_y - S_x \pm \sqrt{(S_y - S_x)^2 + 4S_{xy}^2}}{2S_{xy}}. \tag{11}$$

Using the plus sign gives us S_p, the minimum sum of squared distances; using the minus sign gives us S_a, the maximum sum of squared distances. (The latter value of $\tan \theta$ is the negative reciprocal of the former.)

S_p is the sum of squared distances perpendicular to the regression line and S_a is the sum of squared distances along the regression line. The sum $S_p + S_a$ is equal to $S_x + S_y$ which is the sum of squares of distances from the sample points to the point $\bar{x}, \bar{y}$. We have thus decomposed $S_x + S_y$ into two components, one perpendicular to the regression line and the other along the regression line.

The joint distribution of S_p and S_a may be derived from the Wishart distribution of the sums of squares and cross products,[1]

$$(12) \qquad \frac{1}{4\pi\sigma^{2n-2}\Gamma(n-2)} \begin{vmatrix} S_x & S_{xy} \\ S_{xy} & S_y \end{vmatrix}^{\frac{1}{2}(n-4)} e^{-\frac{1}{2}(S_x+S_y)/\sigma^2}$$

Let us make the transformation

$$S_x = \cos^2\theta\, S_a + \sin^2\theta\, S_p\,,$$
$$S_y = \sin^2\theta\, S_a + \cos^2\theta\, S_p\,,$$
$$S_{xy} = \sin\theta\cos\theta\,(S_a - S_p).$$

The value of θ corresponds to the plus sign in (11). We find

$$S_x + S_y = S_p + S_a\,,$$
$$\begin{vmatrix} S_x & S_{xy} \\ S_{xy} & S_y \end{vmatrix} = S_p S_a\,.$$

The Jacobian of the transformation is $(S_a - S_p)$. Using these relations in (12) and integrating out θ we derive the distribution of S_a and S_p

$$(13) \qquad \frac{1}{4\sigma^6\Gamma(n-2)} \left(\frac{S_a S_p}{\sigma^4}\right)^{\frac{1}{2}(n-4)} e^{-\frac{1}{2}(S_a+S_p)/\sigma^2}\,(S_a - S_p).$$

It can be shown that S_a and S_p are the characteristic roots of the sample variance-covariance matrix. The distribution (13) of the characteristic roots of a variance-covariance matrix when the population correlation coefficient is zero and the variances are equal has been demonstrated by P. L. Hsu [4].

As a test of correlation (i.e., test of significance of the regression coefficient) we propose using the ratio

$$F' = S_a/S_p\,.$$

This ratio is the maximum ratio of the sum of squared deviations in one direction to the sum of squared deviations in the perpendicular direction. It is intuitively evident that this ratio is probably near unity if the null hypothesis is true, that is, if the variances are equal and the correlation is zero. If the correlation is not zero then the ratio is likely to be large.

From (13) we can deduce the distribution of F' by transforming variables and integrating out the extraneous one. This procedure yields us as the distribution of F'

$$(n-2)2^{n-3}F'^{\frac{1}{2}(n-4)}(F'+1)^{-(n-1)}(F'-1).$$

If we make the transformation

$$F' = e^{2z'},$$

[1] This distribution is equivalent to Fisher's distribution of the sample variances and correlation coefficient when the population correlation coefficient is zero.

we find the probability element of z' to be

$$(n - 2)(\cosh z')^{-(n-1)}\, d(\cosh z')$$

After integrating we see the cumulative distribution of z' is

$$1 - (\cosh z')^{-(n-2)}.$$

Critical values of z' for various levels of significance may be determined from a table of hyperbolic cosines. Table I gives some values of z' and the corresponding values of F'.

TABLE I

Percentage points for the z' (or F') distribution

n	z'					F'				
	$P_{.20}$	$P_{.10}$	$P_{.05}$	$P_{.01}$	$P_{.001}$	$P_{.20}$	$P_{.10}$	$P_{.05}$	$P_{.01}$	$P_{.001}$
3	2.292	2.993	3.688	5.298	7.601	98.0	398	1600	40,000	4,000,000
4	1.444	1.818	2.178	2.993	4.144	17.9	38.0	78.0	398	4,000
5	1.130	1.402	1.656	2.216	2.993	9.59	16.5	27.4	84.2	398
6	.958	1.178	1.381	1.818	2.412	6.79	10.6	15.8	38.0	124
7	.846	1.035	1.207	1.572	2.059	5.43	7.92	11.2	23.2	61.4
8	.766	.933	1.084	1.402	1.818	4.63	6.47	8.74	16.5	38.0
9	.704	.856	.992	1.276	1.643	4.09	5.55	7.28	12.7	26.8
10	.656	.796	.920	1.178	1.509	3.71	4.91	6.30	10.6	20.5
11	.616	.746	.862	1.100	1.402	3.43	4.45	5.61	9.02	16.5
12	.583	.705	.813	1.035	1.314	3.21	4.10	5.09	7.92	13.9
13	.554	.670	.772	.980	1.241	3.03	3.82	4.68	7.10	12.0
14	.530	.639	.736	.933	1.178	2.89	3.59	4.36	6.47	10.6
15	.508	.613	.705	.892	1.124	2.76	3.41	4.10	6.00	9.47
20	.429	.517	.593	.746	.993	2.36	2.81	3.27	4.45	6.47
25	.378	.455	.522	.654	.814	2.13	2.48	2.84	3.70	5.10
30	.342	.411	.471	.589	.732	1.98	2.28	2.57	3.25	4.32
40	.293	.352	.402	.502	.621	1.80	2.02	2.23	2.73	3.47
60	.237	.284	.324	.404	.498	1.61	1.76	1.91	2.24	2.71
120	.165	.198	.226	.281	.345	1.39	1.49	1.57	1.75	2.00

The use of F' has been suggested here to test the hypothesis that the population correlation coefficient is zero when it is known that the variances of the two variates are the same, or, more generally, when the ratio of the two variances is known. This gives a test of significance of the regression coefficient when there is error in both variates if the ratio of the variances is known. The test arises from intuitive considerations. F' can also be used to test the hypothesis that $\rho = 0$ *and* $\sigma_x^2 = \sigma_y^2$ (H_4 in Hsu's paper). C. T. Hsu [5] and J. W. Mauchly [6] have shown that the likelihood ratio criterion for this hypothesis is

$$\lambda = \left[\frac{2(S_x S_y - S_{xy}^2)}{(S_x + S_y)^2}\right]^{\frac{1}{2}n} = \left[\frac{2F'}{(F' + 1)^2}\right]^{\frac{1}{2}n}.$$

If we set the normal distribution function equal to a constant, we determine a contour ellipse in the x, y — plane. Since these ellipses of constant probability density are circles when $\rho = 0$ and $\sigma_x^2 = \sigma_y^2$, Mauchly calls the test a test of circularity. The same procedure as used to test whether these ellipses are circles can be used to test whether the ellipses have major axes in a certain direction and with a specified ratio of lengths of axes. Suppose we wish to test the hypothesis that the major axis is inclined to the x axis at an angle θ and that the ratio of lengths of the major axis to the minor axis is k. This is equivalent to the hypothesis that $\rho = \rho_0$ and $\sigma_x^2 = \gamma_0 \sigma_y^2$. To do this we rotate coordinate axes of the variables of the distribution (hence changing coordinates of all sample points) through θ and change the scale of one of the new variables by the factor of k. The transformation is

$$x = kx' \cos \theta - y' \sin \theta,$$

$$y = kx' \sin \theta + y' \cos \theta.$$

In terms of x', y' the null hypothesis is $\rho' = 0$, $\sigma_{x'}^2 = \sigma_{y'}^2$, and one proceeds as above. Of course, if γ_0 is known then this method can be used to test the null hypothesis that $\rho = \rho_0$.

4. Illustrative Example. An application of the formulae given above may be illustrated from the data in Table II, which gives two sets of electrical conductivity measurements at different field strengths. The assumption that the two variances are equal is thus reasonable.

Table of Pairs of Observations of Electrical Conductivity

x_i	y_i	x_i	y_i
5.0	5.1	5.5	5.1
7.4	7.0	5.3	5.0
7.0	7.7	4.7	4.4
8.8	7.7	8.6	7.1
7.8	6.8	7.5	7.3
5.1	5.5	5.6	6.3
6.6	7.4	7.4	6.5
8.8	7.7		

Is it reasonable to regard x and y as being independently distributed in the population on the basis of these data?

The sums of squares and cross products of deviations from the means and the calculated slope are:

$$S_x = 29.40, \qquad S_{xy} = 19.99,$$

$$S_y = 18.04, \qquad b = 0.7554.$$

The maximized variance ratio is:

$$F' = \frac{S_x + 2bS_{xy} + b^2S_y}{b^2S_x - 2bS_{xy} + S_y} = \frac{69.89}{4.615} = 15.15.$$

$$z' = \tfrac{1}{2}\ln F' = 1.36.$$

Comparing with Table I for $n = 15$ we find this value of z' very highly significant (probability less than 0.001), and at this probability level and on basis of our data, x and y cannot be considered to be independent in the population.

Since the regression is significant, it becomes of interest to compute the calculated points X_i and Y_i which fall on the regression line

$$Y = 1.35 + 0.7554\,X,$$

corresponding to each observed point x_i, y_i. They are obtained from these equations

$$\begin{aligned} Y_i &= \bar{y} + \frac{b}{1+b^2}(x_i - \bar{x}) + \frac{b^2}{1+b^2}(y_i - \bar{y}) \\ &= .481x_i + .363y_i + .86, \\ X_i &= \bar{x} + \frac{1}{1+b^2}(x_i - \bar{x}) + \frac{b}{1+b^2}(y_i - \bar{y}) \\ &= .637x_i + .481y_i - .65. \end{aligned}$$

The minimized sum of squared deviations from the regression line (i.e., squared distances between observed and calculated points) is the denominator of the expression for F' divided by the factor $(1 + b^2)$,

$$4.615/.5706 = 2.64.$$

It should perhaps be pointed out that the tests of the means described in the first part of this paper are no longer applicable since we do not know the population correlation coefficient.

REFERENCES

[1] G. W. Snedecor, *Statistical Methods*, Iowa State College Press (1940), pp. 184–187.

[2] S. Kolodziejczyk, "On an important class of statistical hypotheses," *Biometrika*, Vol. 27 (1935), pp. 161–190.

[3] R. A. Fisher, "The fiducial argument in statistical inference," *Annals of Eugenics*, Vol. 6 (1935), pp. 391–398.

[4] P. L. Hsu, "On the distribution of roots of certain determinantal equations," *Annals of Eugenics*, Vol. 9 (1939), pp. 250–258.

[5] C. T. Hsu, "On samples from a normal bivariate population," *Annals of Math. Stat.*, Vol. 11 (1940), pp. 410–426.

[6] J. W. Mauchly, "Significance test for sphericity of a normal n-variate distribution," *Annals of Math. Stat.*, Vol. 11 (1940), pp. 204–209.