Discussiones Mathematicae
Probability and Statistics 20 (2000) 233–247

TESTS OF INDEPENDENCE OF NORMAL RANDOM VARIABLES WITH KNOWN AND UNKNOWN VARIANCE RATIO

Edward Gąsiorek, Andrzej Michalski
Department of Mathematics, Institute of Mathematics
Agriculture University of Wrocław
Grunwaldzka 53, 50–357 Wrocław, Poland
e-mail: max@ozi.ar.wroc.pl

and

Roman Zmyślony[1]
Institute of Mathematics, Technical University
Podgórna 50, 65–246 Zielona Góra, Poland
e-mail: r.zmyslony@im.pz.zgora.pl

Abstract

In the paper, a new approach to construction test for independence of two-dimensional normally distributed random vectors is given under the assumption that the ratio of the variances is known. This test is uniformly better than the t-Student test. A comparison of the power of these two tests is given. A behaviour of this test for some ϵ-contamination of the original model is also shown. In the general case when the variance ratio is unknown, an adaptive test is presented. The equivalence between this test and the classical t-test for independence of normal variables is shown. Moreover, the confidence interval for correlation coefficient is given. The results follow from the unified theory of testing hypotheses both for fixed effects and variance components presented in papers [6] and [7].

Keywords and phrases: mixed linear models; variance components; correlation; quadratic unbiased estimation; testing hypotheses; confidence intervals.

1999 Mathematics Subject Classiffication: 62F03, 62J10.

[1]The paper was partially supported by Grant KBN No. 8 T11 A00 313.

1 Introduction

The well known *t-Student* test for independence of normally distributed two random variables is the most powerful unbiased test if the covariance structure of this variables is completely unknown. However, under the assumption that the variances of random variables are equal or if the ratio of these variances is known it is possible to construct a more powerful test than the *t-Student* one. This problem is well known and was solved by Geisser [1], but the authors propose here another solution, which is based on the new approach to testing hypothesis presented in [6] and [7]. Moreover, using the simulation methods, the power functions of classical *t-Student* test and the new one were compared. Throughout this paper $F_{k,l}$ stands for *F-Snedecor* distribution with k degrees of freedom for the numerator and l degrees of freedom for the denominator. As usually $F_{\alpha,k,l}$ stands for α-critical value of this distribution.

2 Test for independence of two random variables with equal variances

Suppose that $y_1, ..., y_n$, where $y_i = (y_{1i,}\ y_{2i})'$, are two-dimensional iid random variables according to the normal distribution $N(\mu, \Sigma)$ with vector of mean $\mu = (\mu_1, \mu_2)'$ and the covariance matrix

$$\Sigma = \begin{bmatrix} \sigma & \gamma \\ \gamma & k\sigma \end{bmatrix}. \tag{1}$$

We assume that σ, γ and μ are unknown, while k is known. Note that $\rho = \frac{\gamma}{\sqrt{k}\sigma}$ is the correlation coefficient and two hypotheses

$$\mathrm{H} : \rho = 0 \quad \text{vs} \quad \mathrm{K} : \rho > 0 \tag{2}$$

or

$$\mathrm{H} : \rho = 0 \quad \text{vs} \quad \mathrm{K} : \rho \neq 0 \tag{3}$$

are equivalent to

$$\mathrm{H} : \gamma = 0 \quad \text{vs} \quad \mathrm{K} : \gamma > 0 \tag{4}$$

or

$$\text{(5)} \qquad \mathrm{H} : \gamma = 0 \quad \text{vs} \quad \mathrm{K} : \gamma \neq 0,$$

respectively. Define $z = By$, where $y = (y_{11}, ..., y_{1n}, y_{21}, ..., y_{2n})'$, while

$$B = diag\left\{1, \frac{1}{\sqrt{k}}\right\} \otimes I_n.$$

Now we can use an idea given in [6] and [7] for a construction of a test for the above hypothesis on the basis of the best quadratic unbiased estimator (BQUE) for γ. Note that the random vector $z = (z_{11}, ..., z_{1n}, z_{21}, ..., z_{2n})'$ is normally distributed $N(\mu_z, \Sigma_z)$ with the expected vector

$$\text{(6)} \qquad \mu_z = B(I_2 \otimes 1_n)\mu$$

and the covariance matrix

$$\text{(7)} \quad \Sigma_z = B(\Sigma \otimes I_n)B = \left(\frac{\gamma}{\sqrt{k}}(1_2 1_2^{'} - I_2) + \sigma I_2\right) \otimes I_n = \gamma_1 V + \sigma I_{2n},$$

where $\gamma_1 = \frac{\gamma}{\sqrt{k}}$ and $V = (1_2 1_2^{'} - I_2) \otimes I_n$.

Here 1_n, I_n and $\otimes$ stand for $n \times 1$ vector of one's, $n \times n$ identity matrix and Kronecker symbol, respectively. Since of the projection matrix $P = I_2 \otimes \frac{1}{n} 1_n 1_n^{'}$ on space of expected values $R(B(I_2 \otimes 1_n))$ commute with V and since $sp\{V, I\}$ is a Jordan algebra (quadratic subspace) we conclude that $sp\{MVM, M\}$ is also a quadratic subspace, where $M = I - P$. This in turn implies that there exists BUE $\widehat{\gamma_1}$ for γ_1 of (see Lemma 3.1 in [9] and [10], cf also [8]) and it is given by

$$\text{(8)} \qquad \widehat{\gamma_1} = \frac{1}{2(n-1)} z' A z = \frac{1}{2(n-1)}(z' A_+ z - z' A_- z),$$

where $A = \begin{bmatrix} 0 & M_n \\ M_n & 0 \end{bmatrix}$ with $M_n = I_n - \frac{1}{n} 1_n 1_n'$, while

$$A_+ = \frac{1}{2}\begin{bmatrix} M_n & M_n \\ M_n & M_n \end{bmatrix} \quad \text{and} \quad A_- = \frac{1}{2}\begin{bmatrix} M_n & -M_n \\ -M_n & M_n \end{bmatrix}.$$

Note that the quadratic forms $z'A_+z$ and $z'A_-z$ can be expressed as follows

$$z'A_+z = \frac{1}{2}\sum_{i=1}^{n}\left[y_{1i} + \frac{y_{2i}}{\sqrt{k}} - \left(\overline{y_{1\cdot}} + \frac{\overline{y_{2\cdot}}}{\sqrt{k}}\right)\right]^2 \tag{9}$$

and

$$z'A_-z = \frac{1}{2}\sum_{i=1}^{n}\left[y_{1i} - \frac{y_{2i}}{\sqrt{k}} - \left(\overline{y_{1\cdot}} - \frac{\overline{y_{2\cdot}}}{\sqrt{k}}\right)\right]^2, \tag{10}$$

where $\overline{y_{1\cdot}}$ and $\overline{y_{2\cdot}}$ are appropriate sample averages.

Remark 21. It is worthy of note that A is a tripotent matrix, i.e. $A^3 = A$ and therefore we can obtain immediately its decomposition as follows

$$A_+ = \frac{A^2 + A}{2} \quad \text{and} \quad A_- = \frac{A^2 - A}{2}.$$

Since $A_+A_- = 0$ and $tr(A_+) = tr(A_-) = tr(M_n) = n-1$, the ratio statistics

$$F = \frac{z'A_+z}{z'A_-z} \tag{11}$$

under null hypotheses (4) (or (5)) has an F-*Snedecor* distribution with $n-1$ degrees of freedom both for the numerator and the denominator, i.e. $F_{n-1,n-1}$ (see also Lemma 3.2 in [6]).

We accept null hypotheses (4) at significance level α if

$$F < F_{\alpha,n-1,n-1} \tag{12}$$

and accept null hypotheses (5) if

$$F_{1-\alpha/2,n-1,n-1} < F < F_{\alpha/2,n-1,n-1}. \tag{13}$$

Now we prove that F-tests (12) and (13) have optimal properties, i.e. they are the most powerful unbiased tests.

Theorem 21. *The F-test given by* (12) *and F-test given by* (13) *are the most powerful unbiased tests for testing hypotheses* (4) *and* (5), *respectively.*

Proof. To prove that statistics F given by (11) is F-distributed note that $u_{1i} = z_{1i} + z_{2i}$ and $u_{2i} = z_{1i} - z_{2i}$ are independent with mean $\eta_1 = \mu_1 + \frac{1}{\sqrt{k}}\mu_2$ and $\eta_2 = \mu_1 - \frac{1}{\sqrt{k}}\mu_2$, respectively. Moreover, random vectors $u_i = (u_{1i}, u_{2i})'$ are iid with the following covariance matrix

$$\Sigma_{u_i} = \begin{bmatrix} 2(\sigma + \frac{\gamma}{\sqrt{k}}) & 0 \\ 0 & 2(\sigma - \frac{\gamma}{\sqrt{k}}) \end{bmatrix}. \tag{14}$$

Thus the hypotheses H : $\gamma = 0$ vs K : $\gamma > 0$ or H : $\gamma = 0$ vs K : $\gamma \neq 0$ are equivalent to the well known testing problems of equality of variances for two samples with equal number of observation and thus the theorem follows ([5], 219). ∎

Remark 22. The result presented in Theorem 2.1 can be obtained using the invariance principle.

Note that the *t-Student* test is also unbiased. Thus, from the above theorem we have the following

Corollary 21. *Both one-sided test* (12) *and two-sided test* (13), *are more powerful tests than one-sided and two-sided t-Student tests for hypotheses* (2) *and* (3), *respectively.*

In the next section, we compare these two tests and by simulation study we show the behavior of the F-test.

3 Comparison of power of classical t-test with F-test under equal and unequal variances

In this section, we compare power functions for the *t-Student* test with the *F-Snedecor* test for different sample sizes ($n = 3, 6, 10, 20$) and

$$\rho = \frac{\gamma}{\sqrt{k}\sigma}\ , \rho \in (-1, 1).$$

Note that the power function for F-test with known k as in the original model is calculated on the basis of F-distribution for each fixed $\rho_0 \in (-1, 1)$ because the statistics

$$\frac{1-\rho_0}{1+\rho_0}F$$

is $F_{n-1,n-1}$ distributed (14), (see also [1]). Recall that the *t-Student* test is based on statistics

$$t = \frac{R}{\sqrt{1-R^2}}\sqrt{n-2}, \tag{15}$$

where R is the sample correlation coefficient, and null hypothesis (2) is rejected if $t > t_{\alpha,n-2}$, while the hypothesis (3) is rejected if $|t| > t_{\alpha/2,n-2}$.

For the small sizes of samples i.e. $n < 10$ and large $|\rho|$ (see Figure 1-2) it is clearly seen that the F-test is much better than the *t-Student* test. However, for $n \geq 10$ the power functions for both tests almost coincide. On the basis of study simulation we examine also the behavior of our test (in the original model defined by (1) with the ratio $k = 1$), in the surroundings $(k - \varepsilon,\ k + \varepsilon)$, what corresponds to an ϵ-contaminated normal distribution. By simulation studies we get the result showing that the power function of the F-test given by (13) is a decreasing function of ϵ. From Figures 5 to 8 it follows that if the ratio of variances belongs to the interval $(0.8, 1.25)$, then the *F-Snedecor* test is still uniformly better than the *t-Student* test. However, outside of this interval the F-test is even biased (see Figure 9,10). Similarly, it holds for one-sided tests. Computational step $\Delta\rho$ in simulations is equal to 0.01. The number of samples for each fixed value $\rho = \rho_0$ is equal to 10000. For random numbers of normal distribution we apply the ROU (ratio of uniforms method) generator proposed by Kinderman and Monahan [4].

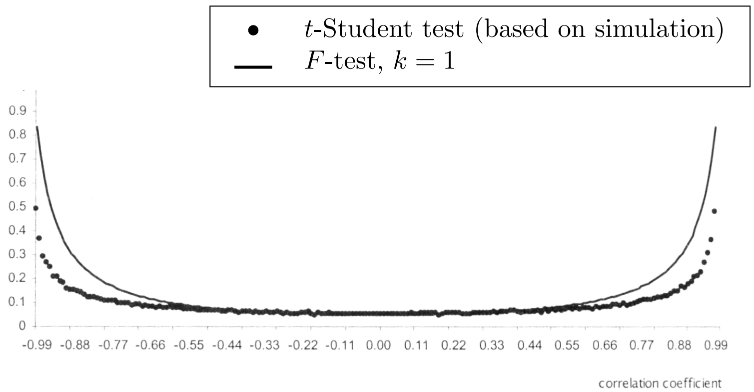

Figure 1. Comparison of power functions for $n = 3$.

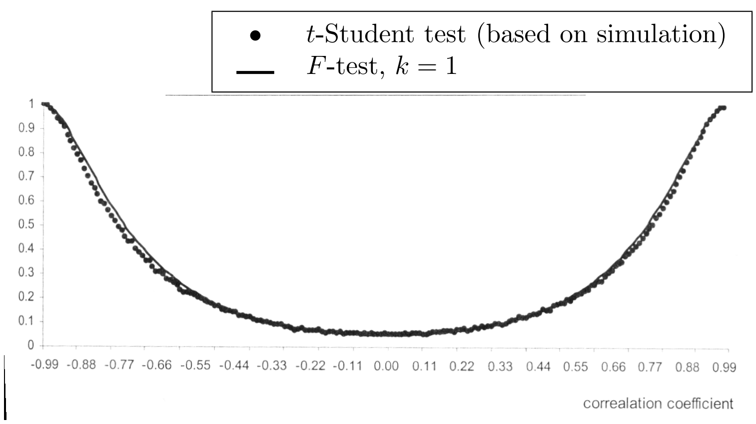

Figure 2. Comparison of power functions for $n = 6$.

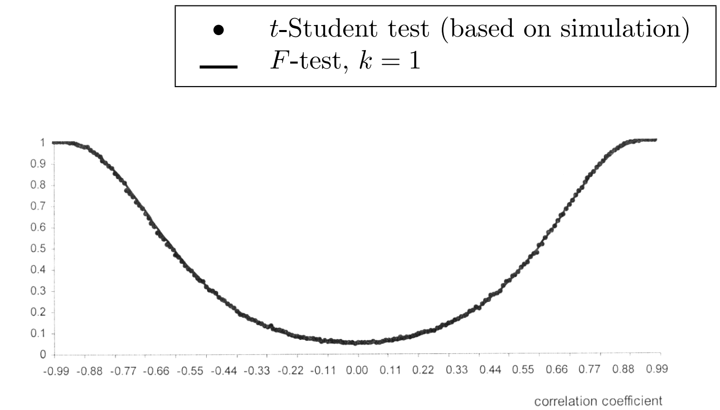

Figure 3. Comparison of power functions for $n = 10$.

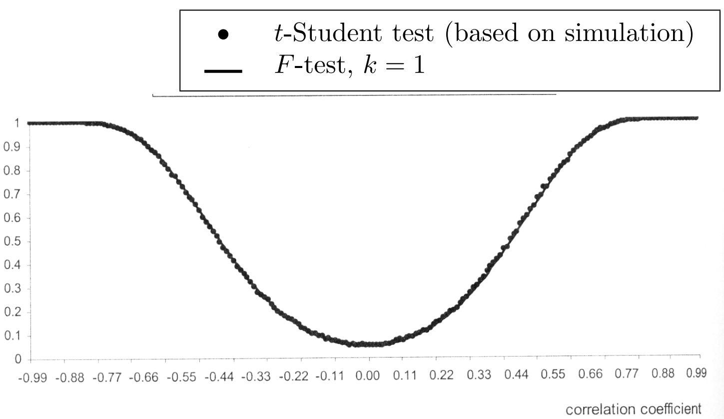

Figure 4. Comparison of power functions for $n = 20$.

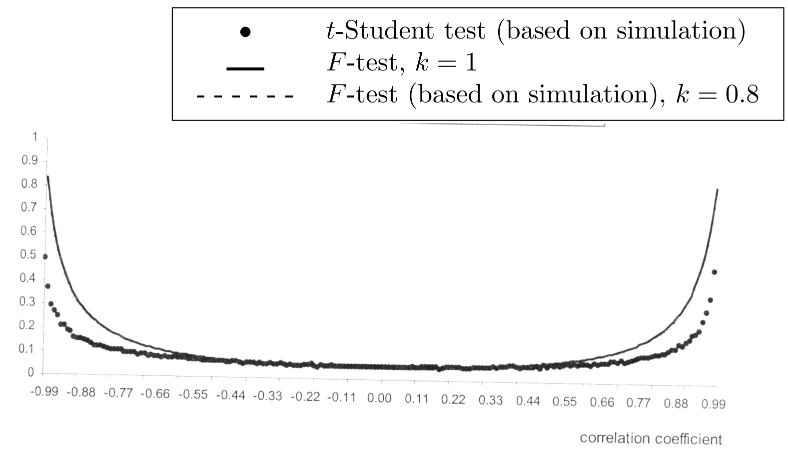

Figure 5. Comparison of power functions for $n = 3$.

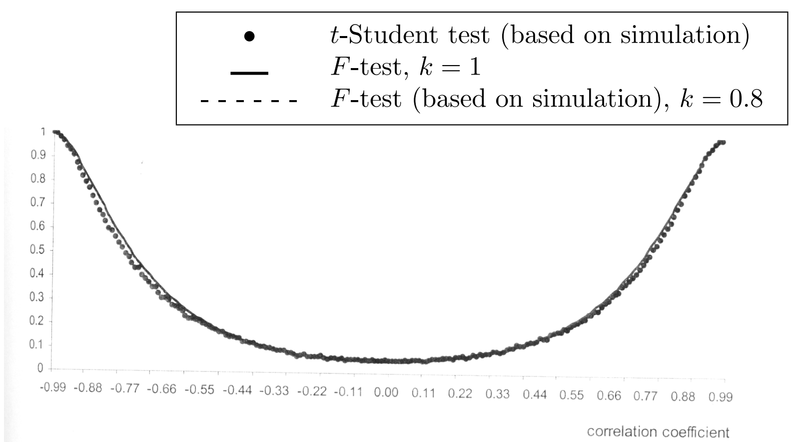

Figure 6. Comparison of power functions for $n = 6$.

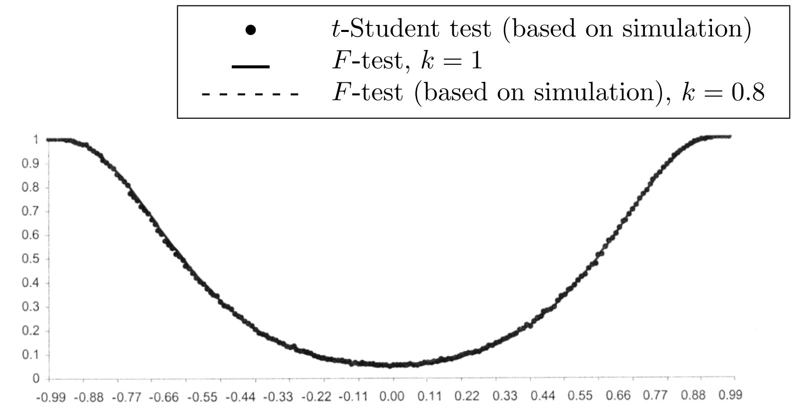

Figure 7. Comparison of power functions for $n = 10$.

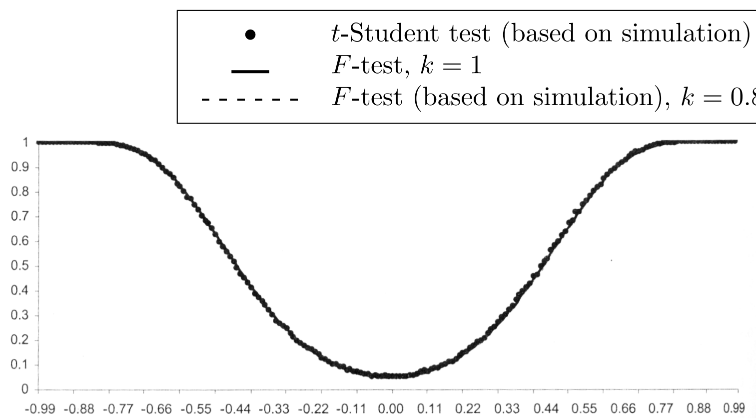

Figure 8. Comparison of powerfunctions for $n = 20$.

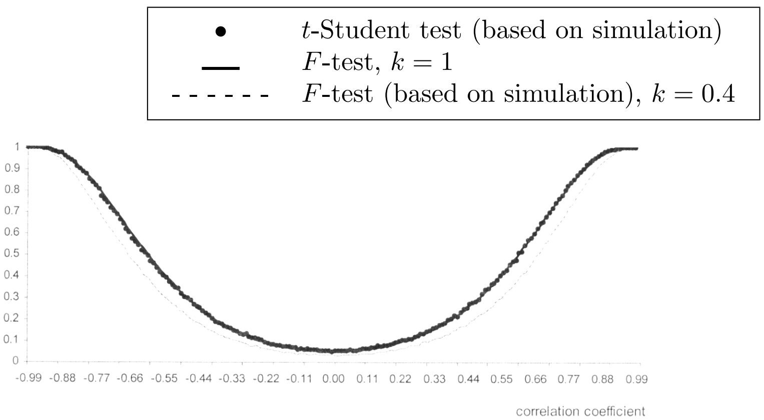

Figure 9. Comparison of power functions for $n = 10$.

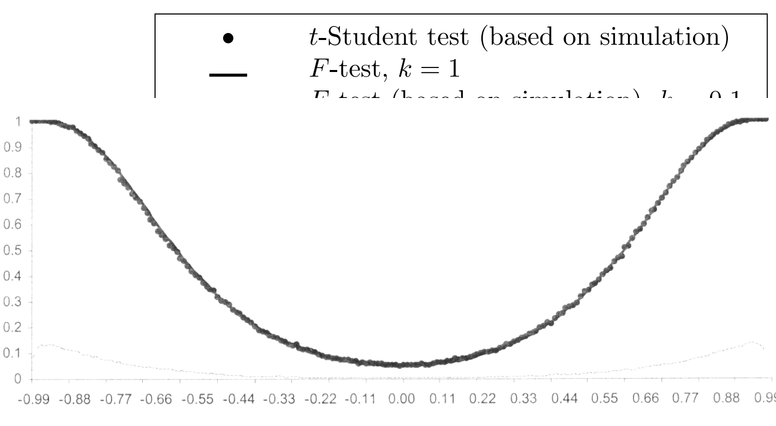

Figure 10. Comparison of power functions for $n = 10$.

4 Confidence interval for correlation coefficient ρ

In this section we construct a confidence interval for the correlation coefficient ρ. For this reason note that if the true value of covariance is γ, then $\rho = \frac{\gamma}{\sqrt{k}\sigma}$ and if $\rho = \rho_0$, then from (14) it follows that $\frac{1-\rho_0}{1+\rho_0}F$ is $F_{n-1,n-1}$ distributed. Now, similarly as for hypothesis (3) the acceptance region of a level-α test for testing

$$\text{H}(\rho_0) : \rho = \rho_0 \quad \text{vs} \quad \text{K}(\rho_0) : \rho \neq \rho_0 \tag{16}$$

has the following form

$$\frac{1+\rho_0}{1-\rho_0}F_{1-\alpha/2,n-1,n-1} < F < \frac{1+\rho_0}{1-\rho_0}F_{\alpha/2,n-1,n-1}. \tag{17}$$

Thus, we have confidence interval for ρ given by $\left[f_1^{-1}(F), f_2^{-1}(F)\right]$, where

$$f_1(\rho) = \frac{1+\rho}{1-\rho}F_{\frac{\alpha}{2},n-1,n-1}, \qquad f_2(\rho) = \frac{1+\rho}{1-\rho}F_{1-\frac{\alpha}{2},n-1,n-1}$$

and F is given by (11), while f_1^{-1} and f_2^{-1} stand for the inverse functions of f_1 and f_2, respectively. It can be easily calculated that the lower bound and upper bound are equal

$$f_1^{-1}(F) = \frac{F - F_{\frac{\alpha}{2},n-1,n-1}}{F + F_{\frac{\alpha}{2},n-1,n-1}}, \qquad f_2^{-1}(F) = \frac{F - F_{1-\frac{\alpha}{2},n-1,n-1}}{F + F_{1-\frac{\alpha}{2},n-1,n-1}},$$

respectively.

The following Figure 11 illustrates the construction of the confidence interval for ρ at confidence level $1-\alpha = 0.95$. For the sample size $n = 20$ and for the value of test statistics $F = 4.5$ we calculate that the confidence interval for ρ is [0.28035, 0.83847].

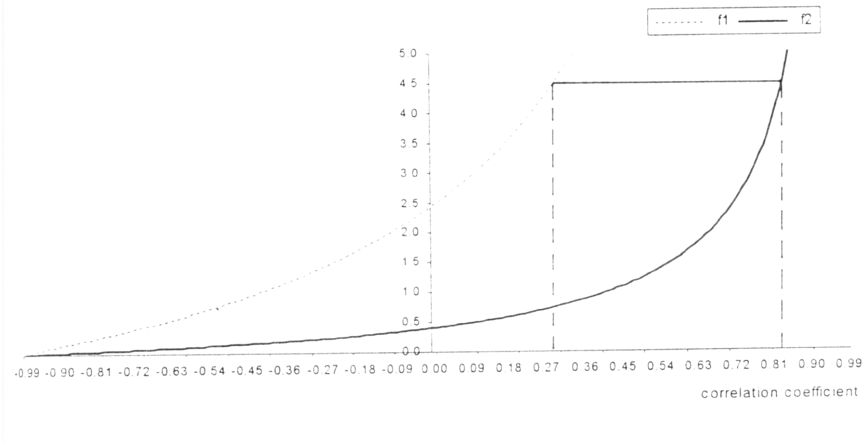

Figure 11. The 95% confidence interval for ρ.

5 Adaptive test for independence in a bivariate NORMAL DISTRIBUTION

In a general case, when the variance ratio k is unknown, we suggest to put in F-statistics given by (11) the ratio of sample variances i.e. $\frac{s_2^2}{s_1^2}$. After straightforward calculations we obtain the following statistics

$$F = \frac{1+R}{1-R}, \tag{18}$$

where R is the sample correlation coefficient.

Lemma 51. *The statistics $F = \frac{1+R}{1-R}$ under null hypothesis $H : \rho = 0$ has F-Snedecor distribution with $n-2$ degrees of freedom both for the numerator and the denominator.*

Proof. First, the distribution of R is known in the literature and among others is reviewed by Johnson and Kotz ([3], sec 32). If we suppose that $\rho = 0$, then the probability density of R has a simple form

$$g(r) = \frac{1}{\sqrt{\pi}} \frac{\Gamma(\frac{1}{2}(n-1))}{\Gamma(\frac{1}{2}(n-2))} (1-r^2)^{\frac{n}{2}-2}$$

([5], 267-270). Next we find the probability density for random variable $F = \frac{1+R}{1-R}$ at once from the formula $h(f) = g(r(f)) \cdot |r'(f)|$, where $r(f) = \frac{f-1}{f+1}$ and $r'(\cdot)$ denotes its derivative. ∎

Since the statistics F under null hypotheses (2) (or (3)) has the distribution $F_{n-2,n-2}$, we accept null hypotheses (2) at significance level α if

$$F < F_{\alpha,n-2,n-2} \tag{19}$$

and we accept null hypotheses (3) if

$$F_{1-\alpha/2,n-2,n-2} < F < F_{\alpha/2,n-2,n-2}. \tag{20}$$

Now we prove that F-tests (19) and (20) have optimal properties, i.e. they are the most powerful unbiased tests.

Theorem 51. *The adaptive tests given by* (19) *and* (20) *are the most powerful unbiased tests for testing hypotheses* (2) *and* (3), *respectively.*

Proof. From Lemma 5.1 and using the form of t-statistics given by (15) we have the following equality

$$F_{\alpha,n-2,n-2} = \frac{\sqrt{t^2_{\alpha,n-2} + n - 2} + t_{a,n-2}}{\sqrt{t^2_{\alpha,n-2} + n - 2} - t_{a,n-2}}.$$

([3], 88), where an inverse relation is presented). Since the F-statistics given by (18) is a monotonic function of the t-*Student* statistics, thus the tests based on the F-ratio are the most powerful tests and are equivalent to the classical t-*Student* test. ∎

Acknowledgments

The authors are grateful to Prof. S. Zontek for his very helpful comments and suggestions.

References

[1] S. Geisser, *Estimation in the uniform covariance case*, JASA **59** (1964), 477–483.

[2] S. Gnot and A. Michalski, *Tests based on admissible estimators in two variance components models*, Statistics **25** (1994), 213–223.

[3] N.L. Johnson and S. Kotz, *Distribution in Statistics: continuous univariate distributions* - 2, Houghton Mifflin, New York 1970.

[4] J.M. Kinderman and J.F. Monahan, *Computer generation of random variables using the ratio of uniform deviates*, ACH Trans. Math. Soft. **3** (1977), 257–260.

[5] E.L. Lehmann, *Testing Statistical Hypotheses*, Wiley, New York 1986.

[6] A. Michalski and R. Zmyślony, *Testing hypotheses for variance components in mixed linear models*, Statistics **27** (1996), 297–310.

[7] A. Michalski and R. Zmyślony, *Testing hypotheses for linear functions of parameters in mixed linear models,* Tatra Mountains Math. Publ. **17** (1999), 103–110.

[8] J. Seely, *Quadratic subspaces and completeness*, Ann. Math. Statist. **42** (1971), 710–721.

[9] R. Zmyślony, *On estimation of parameters in linear models*, Zastosowania Matematyki **15** (1976), 271–276.

[10] R. Zmyślony, *Completeness for a family of normal distributions*, Banach Center Publications **6** (1980), 355–357.

Received 5 September 2000