Listy Biometryczne – Biometrical Letters
Vol. 30 (1993), No. 1, 39-52

Distance measures between statistical populations – a Monte Carlo study of their robustness

Agnieszka Rossa

Department of Statistical Methods, University of Łódź
41, Rewolucji 1905 r., 90-214 Łódź

Abstract

The paper is concerned with the robustness investigation of some measures of the distance between two statistical populations. Three different distance measures are examined: the Kullback Leibler divergence (a special case of which is the Mahalanobis distance), the Hellinger distance and the Bhattacharyya distance.

The definitions and the basic notions connected with the given distances are presented. In particular, the formulae of the distances for the normal populations are stressed.

Using the simulation methods, the robustness of each of the above mentioned distances to departure from normality is studied. For this purpose, a Monte Carlo method is employed to evaluate the values of the distance measures for two populations distributed according to the exponential density or the non-central Student density with n freedom degrees. The Student distribution is treated as the approximating distribution of the normal one. The values of the distances evaluated for populations distributed according to the non-central Student densities (with decreasing freedom degrees) are compared with the theoretical values, evaluated from formulae valid for the normal populations.

1. Distance measures between statistical populations

Mahalanobis (1936) introduced the notion of a distance between two populations. Let Π_1 and Π_2 be two populations characterized by the multivarate normal distributions with a common covariance matrix, i.e. $\Pi_1 \sim N(\mathbf{m}_1, \mathbf{\Sigma})$ and $\Pi_2 \sim N(\mathbf{m}_2, \mathbf{\Sigma})$. Then the Mahalanobis distance between the populations Π_1 and Π_2 may be expressed as follows

Key words: distance measures, robustness, departure from normality, Monte Carlo investigations

$$\Delta = (\mathbf{m}_1 - \mathbf{m}_2)^T \boldsymbol{\Sigma}^{-1} (\mathbf{m}_1 - \mathbf{m}_2) \ . \tag{1.1}$$

Kullback and Leibler (1951) introduced a measure of divergence between statistical populations, defined in terms of the measure of information, called simply "divergence". Let two multivariate populations Π_1 and Π_2 have the respective probability densities $f_1(x_1, x_2, \ldots, x_p)$ and $f_2(x_1, x_2, \ldots, x_p)$ which are equivalent, i.e.

$$\int_A f_1(x_1, x_2, \ldots, x_p) = 0 \Leftrightarrow \int_A f_2(x_1, x_2, \ldots, x_p) = 0$$

for any $A \in \mathcal{B}(R^p)$. Then the divergence between Π_1 and Π_2 was defined by Kullback and Leibler in the following form

$$J(1,2) = \int_{R^p} [f_1(\mathbf{x}) - f_2(\mathbf{x})] \log \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} d\mathbf{x} \ , \tag{1.2}$$

where $\mathbf{x} = (x_1, \ldots, x_p)$.

In the case of the normal populations (when $\Pi_1 \sim N(\mathbf{m}_1, \boldsymbol{\Sigma}_1)$ and $\Pi_2 \sim N(\mathbf{m}_2, \boldsymbol{\Sigma}_2)$), the divergence formula is expressed by

$$J(1,2) = \frac{1}{2} \mathrm{Tr}[(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2)(\boldsymbol{\Sigma}_2^{-1} - \boldsymbol{\Sigma}_1^{-1})] + \frac{1}{2} (\mathbf{m}_1 - \mathbf{m}_2)^T (\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{-1})(\mathbf{m}_1 - \mathbf{m}_2) \ . \tag{1.3}$$

In the particular case, when $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, the $J(1,2)$ becomes

$$J(1,2) = (\mathbf{m}_1 - \mathbf{m}_2)^T \boldsymbol{\Sigma}^{-1} (\mathbf{m}_1 - \mathbf{m}_2) \ . \tag{1.4}$$

Comparing (1.1) and (1.4) we can notice that the Mahalanobis distance is a special case of the Kullback-Leibler divergence for normal populations with the common covariance matrix.

Besides the notion of divergence, there are some other distance measures, among them the Bhattacharyya and Hellinger distances (Bhattacharyya 1943, Kailath 1967, Kobayashi 1970, Beran 1977). Let the so-called Bhattacharyya coefficient be defined as

$$\rho(1,2) = \int_{R^p} [f_1(\mathbf{x}) \cdot f_2(\mathbf{x})]^{1/2} d\mathbf{x} \ . \tag{1.5}$$

Then the Bhattacharyya and Hellinger distances have the forms, respectively

$$B(1,2) = -\ln \rho(1,2) \ , \tag{1.6}$$

$$H(1,2) = \sqrt{1 - \rho(1,2)} \ . \tag{1.7}$$

Simple calculations for the normal case lead to the formula

$$B(1,2) = \frac{1}{2}\ln\left(\frac{\det \boldsymbol{\Sigma}}{\sqrt{\det \boldsymbol{\Sigma}_1 \cdot \det \boldsymbol{\Sigma}_2}}\right) + \frac{1}{8}(\mathbf{m}_1 - \mathbf{m}_2)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\mathbf{m}_1 - \mathbf{m}_2) \ , \qquad (1.8)$$

where $\boldsymbol{\Sigma} = \frac{1}{2}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)$. Of course, from (1.6) and (1.7) it follows that

$$\rho(1,2) = e^{-B(1,2)}$$

and

$$H(1,2) = \sqrt{1 - e^{-B(1,2)}} \ . \qquad (1.9)$$

In the particular case when $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, we obtain

$$B(1,2) = \frac{1}{8}(\mathbf{m}_1 - \mathbf{m}_2)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{m}_1 - \mathbf{m}_2) = \frac{1}{8} J(1,2)$$

For any case, there is

$$B(1,2) \le \frac{1}{8} J(1,2) \ .$$

2. Monte Carlo approaches to studying the distance measures robustness

According to Kullback (1952), the notion of divergence enables one to find the "best" linear function for discriminating between two normal populations Π_1 and Π_2 without limitation to the case of equal covariance matrices. The "best" linear function is found by maximizing the divergence between the distributions of a linear discriminant function.

Several authors, for example, Hopkins and Clay (1963), Holloway and Dunn (1967), Koichi (1969) investigated the effect of departure from normality for Hotteling's T^2 statistics which is closely related to the Mahalanobis distance. Everitt (1979) applied a Monte Carlo investigation of the effect of departures from normality upon one- and two-sample T^2 tests. The multivariate distributions he used were: multivariate normal, uniform, exponential and lognormal distribution.

Johnson et al. (1979) investigated the robustness of Fisher's linear discriminant function to departures from the normal distribution. The Johnson system of distributions was used in their study. This suggested to me that I could apply the t-Student distribution to measure the departure from normality.

Both the cited papers (Everitt 1979, Johnson et al. 1979) are the starting point for my investigations. The aim is to study the effect of departure from normality for divergence, the Bhattacharyya and Hellinger distances, by means of Monte

Carlo methods. For this purpose, the formulae of the Kullback-Leibler divergence and the Bhattacharyya coefficient have to be presented in the form giving an easy way of estimating their values for a large number of Monte Carlo trials. We have

$$J(1,2) = \int_{R^p} [f_1(\mathbf{x}) - f_2(\mathbf{x})] \ln \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} d\mathbf{x} = \int_{R^p} f_1(\mathbf{x}) \ln \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} d\mathbf{x} - \int_{R^p} f_2(\mathbf{x}) \ln \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} d\mathbf{x} =$$

$$= \mathrm{E}_1 [\ln \frac{f_1(\mathbf{X})}{f_2(\mathbf{X})}] - \mathrm{E}_2 [\ln \frac{f_1(\mathbf{Y})}{f_2(\mathbf{Y})}] ,$$

where the random variable $\mathbf{X}$ has the distribution with the p-variate density f_1 and $\mathbf{Y}$ with the p-variate density f_2. Similarly we can write

$$\rho(1,2) = \int_{R^p} [f_1(\mathbf{x}) \cdot f_2(\mathbf{x})]^{1/2} d\mathbf{x} = \int_{R^p} f_2(\mathbf{x}) [\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}]^{1/2} d\mathbf{x} = \mathrm{E}_2 \{[\frac{f_1(\mathbf{Y})}{f_2(\mathbf{Y})}]^{1/2}\} ,$$

where the random variable $\mathbf{Y}$ is distributed according to the density f_2.

Using N trials when generating p-variate random numbers distributed according to the densities f_1 and f_2, we can easily calculate the expressions

$$\hat{J} = \frac{1}{N} \sum_{i=1}^{N} \ln \frac{f_1(\mathbf{x}_i)}{f_2(\mathbf{x}_i)} - \frac{1}{N} \sum_{i=1}^{N} \ln \frac{f_1(\mathbf{y}_i)}{f_2(\mathbf{y}_i)} , \tag{2.1}$$

$$\hat{\rho} = \frac{1}{N} \sum_{i=1}^{N} [\frac{f_1(\mathbf{y}_i)}{f_2(\mathbf{y}_i)}]^{1/2} , \tag{2.2}$$

where $\mathbf{x}_i$ (i=1,2,...,N) are the random values generated from the distribution f_1 and $\mathbf{y}_i$ (i=1,2,...,N) are the random values generated from the distribution f_2.

According to (1.6) and (1.7) we obtain

$$\hat{B} = -\ln \hat{\rho} , \tag{2.3}$$

$$\hat{H} = \sqrt{1 - \hat{\rho}} . \tag{2.4}$$

We assume, similarly as in the cited articles (Everitt 1979, Johnson et al. 1979), that in multivariate cases each of the p variables is generated independently and from the same distribution. Without loss of generality we shall show, for simplicity, the effects for one variable only.

First of all, we use the normal distribution to check the adequacy of the random number generator used in the study. As an example we take some pairs of the univariate normal distributions with various standard deviations. The results are contained in the Table 1.

One can see that the values $\hat{J}$, $\hat{H}$ and $\hat{B}$ obtained by means of the Monte Carlo method are not far from the exact values J, H, B calculated from the theoretical formulae (1.3), (1.8) and (1.9) valid for normal distributions.

Table 1

The comparison between the exact and the estimated values of the Kullback-Leibler divergence, the Hellinger distance and the Bhattacharyya distance (N = 10000 trials)

m_1	m_2	σ_1	σ_2	J	$\hat{J}$	H	$\hat{H}$	B	$\hat{B}$
2	7	1.0	1	25.00	25.06	0.978	0.977	3.125	3.109
2	7	1.5	1	18.40	18.51	0.927	0.930	1.965	2.007
2	7	2.0	1	16.75	16.89	0.862	0.870	1.361	1.415
2	7	2.5	1	16.71	16.81	0.806	0.812	1.048	1.074
2	7	3.0	1	17.44	17.21	0.765	0.765	0.880	0.880
2	7	4.0	1	20.31	20.20	0.725	0.723	0.745	0.739

Legend:

J, B, H – the exact values of the Kullback-Leibler divergence, the Bhattacharyya distance and the Hellinger distance, evaluated for two normal distributions $N(m_1,\sigma_1)$ and $N(m_2,\sigma_2)$ according to the formulae (1.3), (1.8) and (1.9).

$\hat{J}$, $\hat{B}$, $\hat{H}$ – the estimated values of J, B and H obtained by means of the Monte Carlo method according to the formulae (2.1), (2.3) and (2.4).

We also use the exponential distribution with distinct values of the parameter λ. For that distribution, the exact values of divergence and the distances can easily be obtained from the respective integrals. The percentages of differences between the true values of the distances and those calculated from the formulae valid for the normal distributions are shown in the Table 2.

One can see from the Table 2 that the influence of the departures from normality for the exponential distribution is the greatest in the case of the divergence and the least – in the case of the Hellinger distance.

3. Application of *t*-Student distribution for measuring the departure from normality

There are two reasons for using Monte Carlo methods to estimate any distances. First, only for a small number of densities the integrals (1.2) and (1.5) expressing the divergence and the Bhattacharyya coefficient can be evaluated analytically. Second, we do not know very often the exact distribution of the data which we elaborate and for which we use the procedures derived and valid under normality assumption. Here is the reason why we are interested in the study of

Table 2

The percentages of differences between the true (exact and estimated) and theoretical values of the Kullback-Leibler divergence, the Hellinger distance and the Bhattacharyya distance (N = 10000 trials)

λ_1	λ_2	J	$\hat{J}$	J_{Gauss}	$\frac{\lvert J_{\text{Gauss}} - J\rvert}{J} \cdot 100\%$	$\frac{\lvert J_{\text{Gauss}} - \hat{J}\rvert}{\hat{J}} \cdot 100\%$
1	10	8.100	7.903	89.910	1010.00	1037.67
2	9	2.722	2.731	15.577	472.26	470.38
3	8	1.042	1.026	4.210	304.03	310.33
4	7	0.321	0.319	1.068	232.71	234.80
5	6	0.033	0.035	0.101	206.06	188.57
λ_1	λ_2	H	$\hat{H}$	H_{Gauss}	$\frac{\lvert H_{\text{Gauss}} - H\rvert}{H} \cdot 100\%$	$\frac{\lvert H_{\text{Gauss}} - \hat{H}\rvert}{\hat{H}} \cdot 100\%$
1	10	0.652	0.653	0.797	22.24	22.05
2	9	0.478	0.481	0.661	38.28	37.42
3	8	0.331	0.320	0.506	52.87	58.13
4	7	0.195	0.185	0.322	65.13	74.05
5	6	0.064	0.063	0.111	73.44	76.19
λ_1	λ_2	B	$\hat{B}$	B_{Gauss}	$\frac{\lvert B_{\text{Gauss}} - B\rvert}{B} \cdot 100\%$	$\frac{\lvert B_{\text{Gauss}} - \hat{B}\rvert}{\hat{B}} \cdot 100\%$
1	10	0.554	0.556	1.010	82.31	81.65
2	9	0.260	0.263	0.574	120.77	118.25
3	8	0.116	0.108	0.295	154.31	173.15
4	7	0.039	0.035	0.109	179.49	211.43
5	6	0.004	0.004	0.012	200.00	200.00

Legend:

J, B, H – the exact values of the Kullback-Leibler divergence, the Bhattacharyya distance and Hellinger distance evaluated for two exponential distributions with parameters λ_1 and λ_2 according to the formulae (1.2), (1.6) and (1.7).

$\hat{J}$, $\hat{B}$, $\hat{H}$ – the estimated values of J, B and H obtained by means of the Monte Carlo method according to the formulae (2.1),(2.3) and (2.4).

J_{Gauss}, B_{Gauss}, H_{Gauss} – the theoretical values of the Kullback-Leibler divergence, the Battacharyya distance and the Hellinger distance evaluated from the formulae (1.3), (1.8) and (1.9) under assumption that the populations are Gaussian $N(\lambda_1,\lambda_1)$ and $N(\lambda_2,\lambda_2)$.

the effect of departure from normality. The basic question arises how to approach the problem of departure from normality.

We try to approximate the normal distribution $N(m,\sigma)$ by means of the t-Student distribution with n freedom degrees. It can easily be verified that the density function of the form

$$g(x;\, m,\sigma,n) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\sigma(n-2)}\,\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{1}{2}\right)}\left[\frac{1}{n-2}\left(\frac{x-m}{\sigma}\right)^2+1\right]^{-\frac{n+1}{2}} \tag{3.1}$$

is pointwise convergent to the density function of the normal distribution $N(m,\sigma)$ when n tends to infinity. Instead of studying the distances between two normal distributions with the densities $f_1(x) \sim N(m_1,\sigma_1)$ and $f_2(x) \sim N(m_2,\sigma_2)$ we have replaced them by the approximating density functions $g(x;\, m_1,\sigma_1,n)$ and $g(x;\, m_2,\sigma_2,n)$. We have caused the departure from normality by decreasing the freedom degrees n in the approximating densities. The results of such a departure from normality induced on the studied distances are summarized in Tables 3,4 and 5.

It follows from the Table 3 that the divergence changes considerably with decreasing the freedom degrees. Thus the divergence is very sensitive to small departures from normality.

The Table 4 shows, in comparison with the Table 3, that the Hellinger distance seems to be much more robust to decreasing the freedom degrees than the divergence.

Table 3

The percentages of differences between the true (estimated) and theoretical values of the Kullback-Leibler divergence (N = 10000 trials)

m_1	m_2	σ_1	σ_2	n	$\hat{J}$	J_{Gauss}	$\frac{\lvert J_{\text{Gauss}} - \hat{J}\rvert}{\hat{J}} \cdot 100\%$
2	7	1	2	300	15.89	16.75	5.39
				200	15.49	16.75	8.16
				100	14.65	16.75	14.33
				50	13.48	16.75	24.27
				40	12.97	16.75	29.18
				30	12.14	16.75	38.02
				20	11.30	16.75	48.25
				10	9.60	16.75	74.42
				4	8.43	16.75	98.59

Legend:

$\hat{J}$ – the estimated values of the Kullback-Leibler divergence evaluated for two populations with density functions $g(x;\, m_1,\sigma_1,n)$ and $g(x;\, m_2,\sigma_2,n)$ according to the formula (2.1).

J_{Gauss} – the theoretical values of the divergence evaluated from (1.3) under assumption that the populations are Gaussian $N(m_1,\sigma_1)$ and $N(m_2,\sigma_2)$.

Table 4

The percentages of differences between the true (estimated) and theoretical values of the Hellinger distance (N = 10000 trials)

m_1	m_2	σ_1	σ_2	n	$\hat{H}$	H_{Gauss}	$\frac{\lvert H_{\text{Gauss}} - \hat{H}\rvert}{\hat{H}} \cdot 100\%$
				300	0.869	0.862	0.79
				200	0.858	0.862	0.47
				100	0.861	0.862	0.11
				50	0.864	0.862	0.12
2	7	1	2	40	0.868	0.862	0.65
				30	0.854	0.862	0.96
				20	0.860	0.862	0.33
				10	0.844	0.862	2.22
				4	0.857	0.862	0.60

Legend:

$\hat{H}$ – the estimated values of the Hellinger distance evaluated for two populations with density functions $g(x; m_1,\sigma_1,n)$ and $g(x; m_2,\sigma_2,n)$ according to the formula (2.4).

H_{Gauss} – the theoretical values of the Hellinger distance evaluated from (1.9) under assumption that the populations are Gaussian distributed: $N(m_1,\sigma_1)$ and $N(m_2,\sigma_2)$.

Table 5

The percentages of differences between the true (estimated) and theoretical values of the Bhattacharyya distance (N = 10000 trials)

m_1	m_2	σ_1	σ_2	n	$\hat{B}$	B_{Gauss}	$\frac{\lvert B_{\text{Gauss}} - \hat{B}\rvert}{\hat{B}} \cdot 100\%$
				300	1.409	1.362	3.37
				200	1.335	1.362	1.98
				100	1.355	1.362	0.47
				50	1.369	1.362	0.53
2	7	1	2	40	1.400	1.362	2.76
				30	1.308	1.362	4.09
				20	1.343	1.362	1.41
				10	1.244	1.362	9.45
				4	1.327	1.362	2.58

Legend:

$\hat{B}$ – the estimated values ot the Bhattacharyya distance evaluated for two populations with density functions $g(x; m_1,\sigma_1,n)$ and $g(x; m_2,\sigma_2,n)$ according to (2.3).

B_{Gauss} – the theoretical values of the Bhattacharyya distance evaluated from (1.8) under assumption that the populations are Gaussian $N(m_1,\sigma_1)$ and $N(m_2,\sigma_2)$.

One can easily see from the Table 5 that the Bhattacharyya distance is more robust to decreasing the freedom degrees than the divergence although not so robust as the Hellinger distance.

On the basis of the results obtained, we conclude that all the statistical procedures based on the divergence are not robust. A statistical procedure is called robust if it is insensitive to departures from assumptions on which the theoretical model is based. Statistical procedures employing the divergence could not be called robust in the above mentioned sense. Statistical procedures using the Hellinger distance seem to be the most robust.

4. Graphical illustrations of the distance measures sensitivity

The reason of the high sensitivity of the divergence to small deviations from normality can be explained graphically.

Let us denote by Δ_J the difference between the theoretical and the exact values of the divergence

$$\Delta_J = J_{\text{Gauss}} - J_S = \ ,$$

$$= \int_R [f_1(x) - f_2(x)] \cdot \ln \frac{f_1(x)}{f_2(x)} dx - \int_R [g_1(x) - g_2(x)] \cdot \ln \frac{g_1(x)}{g_2(x)} dx =$$

$$= \int_R \left\{ [f_1(x) - f_2(x)] \cdot \ln \frac{f_1(x)}{f_2(x)} - [g_1(x) - g_2(x)] \cdot \ln \frac{g_1(x)}{g_2(x)} \right\} dx =$$

$$= \int_R h(x; m_1, \sigma_1, m_2, \sigma_2, n)\, dx$$

where

$$h(x; m_1, \sigma_1, m_2, \sigma_2, n) = [f_1(x) - f_2(x)] \cdot \ln \frac{f_1(x)}{f_2(x)} - [g_1(x) - g_2(x)] \cdot \ln \frac{g_1(x)}{g_2(x)} \qquad (4.1)$$

and

$$f_1(x) \sim N(m_1, \sigma_1) \ ,$$

$$f_2(x) \sim N(m_2, \sigma_2) \ ,$$

$$g_1(x) = g(x; m_1, \sigma_1, n) \ ,$$

$$g_2(x) = g(x; m_2, \sigma_2, n) \ ,$$

with $g(x; \cdot, \cdot, \cdot)$ defined in (3.1).

The plots of the function $h(x; m_1,\sigma_1,m_2,\sigma_2,n)$ expressed in (4.1) for distinct n and for $m_1 = 2$, $\sigma_1 = 1$, $m_2 = 7$, $\sigma_2 = 2$ are shown in Figure 1. The areas below the plots express differences between exact and theoretical values of the divergence.

Analogous functions can be drawn for the Bhattacharyya coefficient (1.5) used for constructing the Hellinger and Bhattacharyya distances according to the formulae (1.7) and (1.6).

Let Δ_B be the difference between the theoretical and the exact values of the Bhattacharyya coefficient. We have

$$\Delta_B = \rho_{\text{Gauss}} - \rho_S = \ ,$$

$$= \int_R [f_1(x) \cdot f_2(x)]^{1/2} dx - \int_R [g_1(x) \cdot g_2(x)]^{1/2} dx =$$

$$= \int_R \left\{ [f_1(x) \cdot f_2(x)]^{1/2} - [g_1(x) \cdot g_2(x)]^{1/2} \right\} dx =$$

$$= \int_R k(x; m_1,\sigma_1,m_2,\sigma_2,n)\, dx$$

where

$$k(x; m_1,\sigma_1,m_2,\sigma_2,n) = [f_1(x) \cdot f_2(x)]^{1/2} - [g_1(x) \cdot g_2(x)]^{1/2} \qquad (4.2)$$

and

$$f_1(x) \sim N(m_1,\sigma_1) \ ,$$

$$f_2(x) \sim N(m_2,\sigma_2) \ ,$$

$$g_1(x) = g(x; m_1,\sigma_1,n) \ ,$$

$$g_2(x) = g(x; m_2,\sigma_2,n) \ ,$$

with $g(x; \cdot, \cdot, \cdot)$ defined in (3.1).

The plots of the function $k(x; m_1,\sigma_1,m_2,\sigma_2,n)$ defined in (4.2) are drawn for distinct n and for $m_1 = 2$, $\sigma_1 = 1$, $m_2 = 7$, $\sigma_2 = 2$ (see Figure 2). They show a much less amplitude.

Our final conclusion is that the Hellinger distance is much better for robust procedures than the divergence.

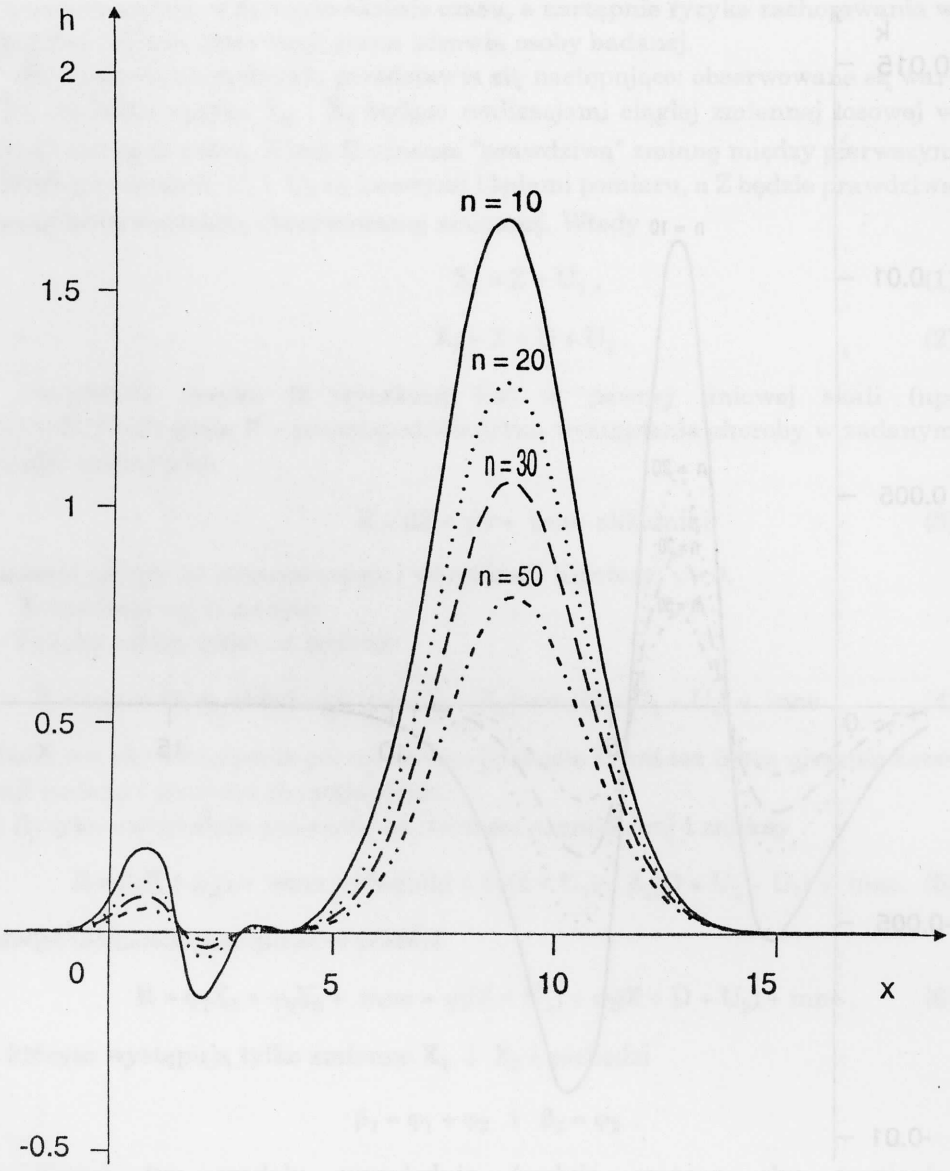

Figure 1. Plots of the function $h(x; m_1,\sigma_1,m_2,\sigma_2,n)$ for various n. The areas below the plots express differences between the theoretical and exact values of the divergence (n – freedom degrees of the t-Student distribution).

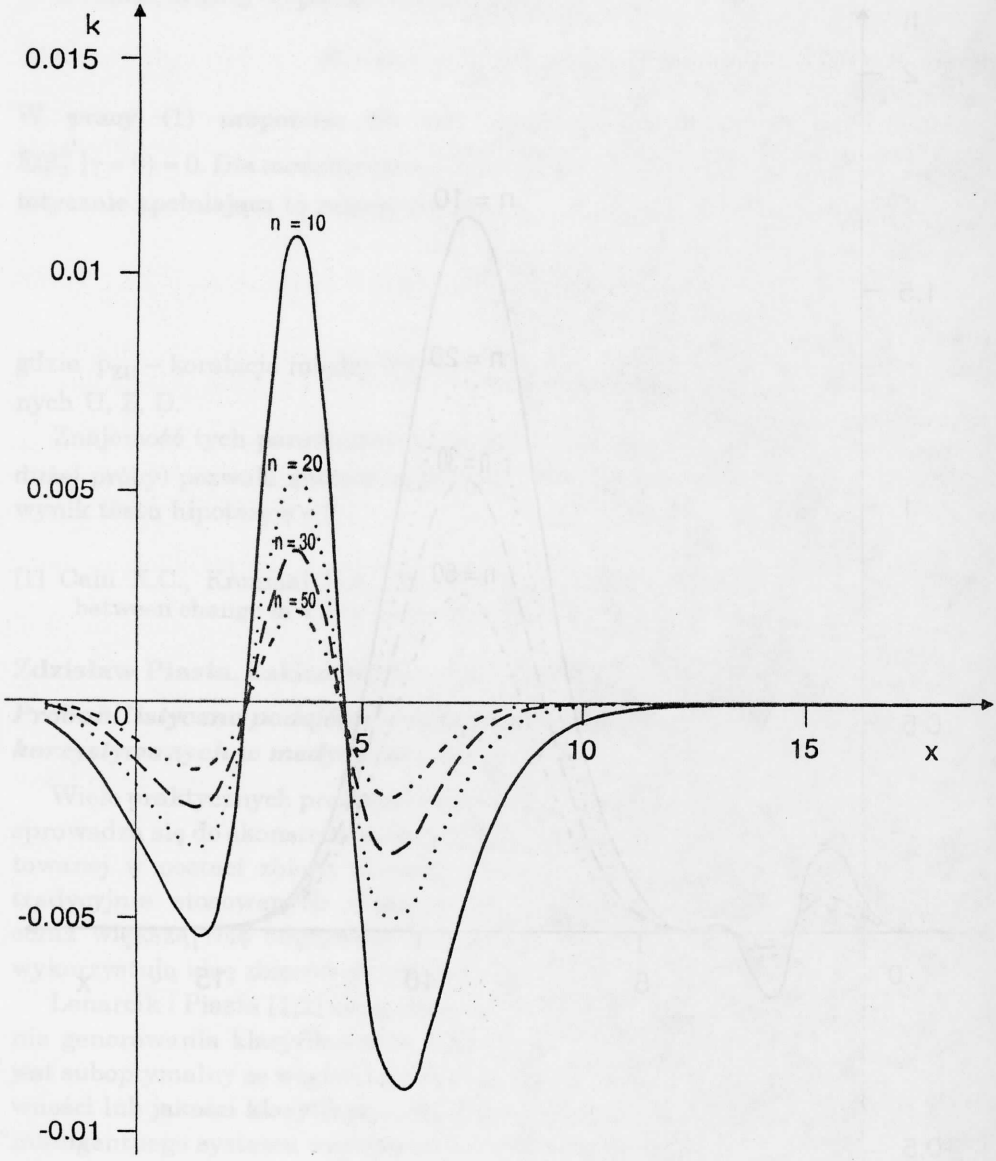

Figure 2. Plots of the function $k(x; m_1,\sigma_1,m_2,\sigma_2,n)$ for various n. The areas below the plots express differences between the theoretical and exact values of the Bhattacharyya coeficient (n – freedom degrees of the t-Student distribution).

REFERENCES

Beran R. (1977). Minimum Hellinger distance estimates for parametric models, *Annals of Stat.* **5**, 445-463.

Bhattacharyya A. (1943). On a measure of divergence between two statistical populations defined by their probability distributions. *Bull.Calcutta Math.Soc.* **35**, 99-109.

Everitt B.S. (1979). A Monte Carlo investigation of the robustness of Hotteling's one- and two-sample T^2 tests. *Journal Amer.Stat.Assoc.* **74**, 48-51.

Holloway L.S., Dunn O.J. (1967). The robustness of Hotteling's T^2. *Journal Amer.Stat.Assoc.* **62**, 124-136.

Hopkins J.W., Clay P.P.F. (1963). Some empirical distributions of bivariate T^2 and homoscedasticity criterion M under unequal variance and leptocurtosis. *Journal Amer.Stat.Assoc.* **58**, 1048-1053.

Johnson M.E., Wang C., Ramberg J.S. (1979). Robustness of Fisher's linear discriminant function to departures from normality. *Informal Report LA-8068-MS*, Los Alamos Scientific Laboratory, University of California, October 1979.

Kobayashi H.M. (1970). Distance measures and asymptotic relative efficiency. *IEEE Trans.Inform.Theory* IT-**16**, 288-291.

Kailath T. (1967). The divergence and Bhattacharyya distance measures in signal selection. *IEEE Trans.Communication Technology* COM-**15**, 52-60.

Koichi I. (1969). On the effect of heteroscedasticity and non-normality upon some multivariate test procedures. *Multivariate Analysis*, vol.2, New York: Academic Press, 87-120.

Kullback S., Leibler R.A. (1951) On information and sufficiency. *Annals of Math.Stat.* **22**, 79-86.

Kullback S. (1952). An application of information theory to multivariate analysis. *Annals of Math.Stat.* **23**, 88-102.

Mahalanobis P.C. (1936). On the generalized distance in statistics. *Proc.Nat.Inst. Sci.India* **12**, 49-55.

Received 3 February 1993; revised 22 June 1993

Miary odległości pomiędzy populacjami – symulacyjne badanie ich odporności

Praca zawiera wyniki badań symulacyjnych wybranych miar odległości pomiędzy dwiema statystycznymi populacjami. Przedmiotem badania jest odporność tych miar na odchylenia rozkładu populacji od rozkładu normalnego. Uwzględniono następujące miary: dywergencja Kullbacka-Leiblera, odległość Hellingera oraz odległość Bhattacharyya.

W pracy podano definicję powyższych miar ze szczególnym uwzględnieniem formuł stosowanych w przypadku populacji o rozkładach normalnych.

Do badania wykorzystano metodę Monte Carlo szacowania wartości poszczególnych miar odległości dla populacji o dowolnych rozkładach. W szczególności wzięto pod uwagę dwa następujące typy rozkładów: rozkład wykładniczy oraz niecentralny

rozkład Studenta. Ten ostatni wykorzystano do aproksymowania rozkładu normalnego o zadanych parametrach. Liczba stopni swobody rozkładu aproksymującego została potraktowana jako miara odejścia od rozkładu normalnego.

Badanie odporności oparto na obliczeniu względnych różnic pomiędzy wartościami teoretycznymi, wyznaczonymi przy założeniu normalności rozkładu badanych populacji, a wartościami faktycznymi oszacowanymi dla populacji o zadanych rozkładach.

Słowa kluczowe: miary odległości, odporność, odejście od rozkładu normalnego, metody Monte Carlo