Listy Biometryczne - Biometrical Letters
Vol. 33 (1996), No. 2, 59-69

Sampling a multi-trait representative sample

Anna Bartkowiak

Institute of Computer Science, University of Wrocław,
Przesmyckiego 20, Wrocław 51-151,
e-mail: aba@ii.uni.wroc.pl

SUMMARY

The paper is concerned with graphical comparisons of the distributional resemblance of a subsample and an entire sample – when $p > 2$ traits taking values on the continuous scale are considered. Two methods are proposed: (1) The group referenced Q-Q plot of Mahalanobis distances, and: (2) The group referenced polygon plot. Algorithms for the proposed methods are given. The methods are illustrated with a real example dealing with medical data.

KEY WORDS: sampling, representativeness, Q-Q plot, Mahalanobis distance, polygon plot

1. The problem

Sometimes the researcher faces the following situation: After collecting a large sample (called in the following: *entire sample* or *parental sample*) he is forced to choose from it a representative subsample. For instance, this might happen, when a subgroup from all the patients - included into the data base – should be selected for a more detailed examination or a follow up study. But then, a subsample obtained purely "at random", might bear - just for sampling reasons - quite a different outlook as the whole sample does – and this would be really undesirable. We want the chosen subsample to be representative – in a broad meaning – for all the gathered data contained in the entire sample.

In such a case the experimenter might proceed as follows: He samples from the entire data several subsamples satisfying the rigidity of random sampling. For each subsample he makes comparison of several distributional properties of the given subsample and the parental entire sample. He chooses for his purpose (retains finally) that subsample which bears the greatest resemblance with the parental sample.

This procedure contradicts in fact to the principles of a *purely random* sample. None the less, in a long run, it may happen that the strict random sample bears some unusual features and further experimenting with such atypical sample would provide some atypical results not contingent to the original (parent) population from which the considered sample was obtained. Therefore investigators have serious objections to deal with strongly atypical samples – of course, if they got the information, that the obtained sample differs much from the parental population.

The idea that sampling methods should adapt to local covariance structure of the target density was recently considered by Givens and Raftery (1996). They did it in the context of choosing the shape of the kernel density estimator when performing local adaptive importance sampling (LAIS) and global adaptive importance sampling (GAIS). My paper addresses a different aspect of sampling and has an entirely different scope as that considered by Givens and Raftery. I quote their paper to emphasize that in choosing a reasonable sample we may use an adaptive approach – and there are precedences of proceeding in such manner.

In last days the topic "how to defend non random samples" has been rosed in the electronic discussion group EDSTAT-L. A.o. in EDSTAT-L, digest 1211, September 1996, Richard F. Ulrich has quoted a paper by B.P. Croft (1994) with a provocative subtitle: *Chance is not such a fine thing.*

The methods proposed in the paper, along with serving for comparison among the distribution of some traits in the parental population and in a subsample, might be applied also in the following situation. Suppose, we have gathered two groups of data, say of size N and n respectively, and we would like to arrive at an statement, whether the two groups are resembling each other. The resemblance could be considered in terms of univariate or multivariate distributions.

Looking at univariate resemblance of the parental population and the subsample, or of two subsamples, is quite easy. We could look at some univariate representations like box plots, textured dot plots or mountain plots.

However, very rarely we are concerned with univariate distributions only. In most statistical tasks we have to deal with multivariate distributions, and the comparisons have to be made with respect to all traits (variables) characterizing the data. My paper is concerned just with that situation.

In the following I propose two routines for making comparisons between multivariate characteristics encountered in the parental sample and the obtained subsample. The proposed techniques are: (1) *The group referenced Q-Q plot of Mahalanobis distances* and (2) *The group referenced polygon plot.*

The proposed methods will be illustrated with a real example.

2. The group referenced Q-Q plot of Mahalanobis distances

2.1. Principles of drawing a Q-Q plot

The principles of constructing a *quantile - quantile* plot (called *Q-Q plot*) are as follows.

For a given univariate data vector $\mathbf{x} = (x_1, \ldots, x_n)$, with x_k $(k = 1, \ldots, n)$ following the usual assumption on the *iid* property of the sample values (*i*ndependent and *i*dentically *d*istributed) we consider the ordered values (order statistics) of the form:

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}.$$

In the sequence above each value $x_{(k)}$ $(k = 1, \ldots, n)$ can be considered as a kind of approximation of the $\frac{k}{n+1}$-th quantile of the cumulative distribution function $F_1(x)$ from which the sample was drawn.

In the following we will use the abbreviation *c.d.f.* for the expression *cumulative distribution function.*

For testing that the *c.d.f.* of the considered data vector follows an assumed *c.d.f.* $F_0(x)$ we evaluate for the sequence $\left\{\frac{k}{n+1}\right\}$, $(k = 1, \ldots, n)$ the respective quantiles $\left\{q_{(k)} = F_0^{-1}(\frac{k}{n+1})\right\}$.

Putting the two quantities $q_{(k)}, x_{(k)}$ together in a scatter diagram we obtain the so called *Q-Q plot* (see, e.g. Bury 1975, Hoaglin 1985, Hamilton 1992).

In a Q-Q plot usually the empirical quantiles $x_{(k)}$ – marked in the vertical axis – are put against the theoretical quantiles $q_{(k)}$ – marked in the horizontal axis. The assigning to the vertical and horizontal axis is not obligatory and the respective quantiles could be graphed in reverse order.

The Q-Q plot technique is usually applied in the situation, when one distribution (say F_1), is an empirical distribution, while the other, the referenced distribution (say F_0), is a theoretical one, e.g. the Gaussian distribution, or the χ^2 distribution.

If the two distributions $F_0(x)$ and $F_1(x)$ were the same, then also their respective quantiles should be the same – hence the quantile points in the scatterplot should be located along a straight line, speaking more precisely, the points should lie in the diagonal of the plot. Any discrepancies indicate that the two distributions differ.

Q-Q plots convey details about how the two distributions differ. After Hamilton (1992) we may say that:

1. If the quantile points lie along the diagonal line, the two distributions are *similar* in *center*, *spread* and *shape.*
2. If the points follow a straight line not parallel to the diagonal, the distributions are *similar* in *shape* but have *different spreads.*

3. If the points do not follow a straight line, the distributions have *different shapes.* They may also *differ* in *center* and *spread.*

A special case of the Q-Q plot is given by the normal probability plot, which compares an empirical distribution with the normal (Gaussian) distribution.

From the shape of the Q-Q plot we may infer also whether the tails of the considered distribution are more heavy–tailed or more light–tailed as the referenced (theoretical) distribution – see Hamilton (1992) for further details.

2.2. Constructing Q-Q plot in the entire-data and sub-sample context

Suppose that we are in the following situation. We have a big sample (entire data), say a data table of size $N \times p$ containing values of p traits (variables) gathered for N individuals. Suppose for the moment that we are concerned with one trait only (one column of the data table) denoted by Z, and its values contained in the data table are $Z_1, \ldots, Z_N$.

These values represent N realizations of a univariate random variable ζ. Let us notice that we do not know the theoretical distribution of ζ and there is little hope that we will know it ever.

However, the situation is not so completely hopeless. We got a relatively large data set (the "entire sample"), which – in our circumstances – may serve as the population.

Thus, after ordering the data values $Z_1, \ldots, Z_N$ into a nondecreasing sequence $Z_{(1)} \leq \ldots \leq Z_{(N)}$, we can construct from them an empirical *c.d.f.* $F_0(z)$ approximating the "true" parental *c.d.f.* of the trait ζ. Eventually, the obtained *c.d.f.* could be smoothed using e.g. the *lowess* or *kernel smooth* method. This will yield us the referenced (or: parental) *c.d.f.* denoted here as $F_0(z)$.

Our main goal is to generate a subsample of size n ($n < N$). The generated subsample should be representative for the entire data vector $(Z_1, \ldots, Z_N)$.

What is meant by representativeness? We are ready to agree that from several generated subsamples the most representative will be that one whose (empirical) distribution function resembles mostly the parental distribution function $F_0(z)$.

The resemblance (or discrepancy) of two distributions (an empirical one and a theoretical one) can be easily established by inspecting the Q-Q plot constructed from the two distributions.

Therefore our procedure will be as follows:

Firstly we will establish from the entire data the parental *c.d.f.* $F_0(z)$ which will serve as the referenced distribution. From this distribution we will find the quantiles q_k of order $\frac{k}{n+1}$ $(k = 1, \ldots, n)$, which will serve as the referenced quantiles. These quantiles will be computed as $q_k = F_0^{-1}(\frac{k}{n+1})$.

Next we will generate a subsample of size n and find from it the empirical quantiles $z_{(1)}, \ldots, z_{(n)}$. These quantiles will be taken simply as the order statistics established from the subsample. They will serve as the quantiles of the tested distribution.

Putting the values $(z_{(k)}, q_k)$ together into the Q-Q plot we will see at once from the obtained plot, whether the two distributions (i.e. the referenced and the tested one) resemble each other.

We may repeat the procedure of generating subsamples and constructing the Q-Q plots several times. We could then retain as the representative subsample that one which yields the Q-Q plot exhibiting a *mostly linear pattern* coinciding with the diagonal of the plot.

2.3. The group referenced Q-Q plot of Mahalanobis distances

In the previous subsection we have established the principles of choosing a representative subsample. This was done when considering one trait only, i.e. one column of the data table in which values for several (exactly: p) variables are stored. However, there is no impediment to consider instead of one observational variable ζ a general summary statistics ξ calculated on the basis of all the p variables.

One such statistics which receives much attention in data analysis is Mahalanobis distance. It measures the distance of an individual data vector from the center of the data cloud. Usually we consider the squared Mahalanobis distance which for the kth data point is defined as:

$$D_k^2 = D^2(\mathbf{x}_k, \bar{\mathbf{x}}_\cdot, \mathbf{S}) = (\mathbf{x}_k - \bar{\mathbf{x}}_\cdot)\mathbf{S}^{-1}(\mathbf{x}_k - \bar{\mathbf{x}}_\cdot)^T. \tag{1}$$

In the formula above $\bar{\mathbf{x}}_\cdot$ and $\mathbf{S}$ denote the mean and the covariance matrix of the entire data, and $\mathbf{x}_k = (x_{k1}, \ldots, x_{kp})$ is the data vector containing values of p variables (traits) observed for the k-th individual.

After defining the statistics D^2 we are - as a matter of fact - put back to the one-dimensional case. Thus, we are able to construct the Q-Q plot taking as one coordinate (say, the x coordinate) the quantiles evaluated from the entire data distribution (the referenced distribution), and as the second coordinate (say, the y coordinate) the respective quantiles of the squared Mahalanobis distances obtained from the subsample.

The algorithm for the procedure will be as follows:

1. For the entire data, of size N, calculate for all individuals their squared Mahalanobis distances according to formula (1), obtaining the statistics $D_1^2, \ldots, D_N^2$. These, after ordering, yield the nondecreasing sequence $D_{(1)}^2, \ldots, D_{(N)}^2$.
2. From the values obtained in point 1 construct the empirical cumulative distribution function $F_0(z_k)$ with $z_k = D_{(k)}^2$, $(k = 1, \ldots, N)$. Eventually perform a kind of

smoothing of the sequence $\{F_0(z_k)\}$.

3. Take from the entire data a subsample of size n (without replacement). Evaluate by formula (1) for each individual data vector $\mathbf{x}_h$ $(h = 1, \ldots, n)$ of the obtained subsample its squared Mahalanobis distance from the center of the entire data set – using as $\bar{\mathbf{x}}_.$ and $\mathbf{S}$ the values from the entire data, thus obtaining $\tilde{D}_1^2, \ldots, \tilde{D}_n^2$.
4. Order the values of the squared Mahalanobis distances (from the subsample obtained in point 3) into a nondecreasing sequence, thus obtaining: $\tilde{D}_{(1)}^2 \leq \ldots \leq \tilde{D}_{(n)}^2$.
5. Construct a scatterdiagram by putting the values $\tilde{D}_{(h)}^2$ against $F_0^{-1}(\frac{h}{n+1})$ $(h = 1, \ldots, n)$, this will result in the Q-Q plot for the considered subsample.
6. Repeat points 3, 4 and 5 several times. Retain that subsample which yields the Q-Q plot with quantile points located most closely to the diagonal (i.e. line $y = x$).

3. The group referenced polygon plot

Instead of being concerned with one summary statistics we could consider at the same time some univariate statistics evaluated for each of the p considered variables.

Say, we are interested in the statistic W evaluated for each of the p considered variables. As W, the statistics of interest, one might take: the arithmetic mean, the standard deviation, the median, the upper and lower quartiles, etc.

A convenient way for displaying simultaneously the considered statistics W for all the considered traits (variables) is to do it by constructing a polygon plot. We propose to draw the polygon plot in the following manner: Firstly we evaluate the values of W both in the referenced and the tested group obtaining the sequences:

$$w_1^{(0)}, \ \ldots \ , w_p^{(0)} \quad - \text{ for the referenced group,}$$

$$w_1^{(t)}, \ \ldots \ , w_p^{(t)} \quad - \text{ for the tested group.}$$

Our next goal is to make a comparison of the statistics of interest in both groups of data. Since the referenced group is considered as a standard or base group, we evaluate the quotients

$$r_1 = \frac{w_1^{(t)}}{w_1^{(0)}}, \ \ldots, \ r_p = \frac{w_p^{(t)}}{w_p^{(0)}}.$$

The evaluated quotients tell us, how much – relatively – the subsequent statistic $w_k^{(t)}$ exceeds (or falls below) the respective value $w_k^{(0)}$ stated in the referenced group.

Taking the values of $r_1, ..., r_p$ as the *radii* emanating from a common center O, with angular distance $\frac{2\pi}{p}$ between each pair of the radii, we construct a polygon with p vertices and p edges. Eventually, in the same plot, we might draw a circle with a

radius of unit length indicating the norm values from the referenced distribution F_0.

The plot, obtained in such a simple way, visualizes very clearly discrepancies between the statistic W evaluated for all the p variables considered in both groups of data. Moreover, on can see at once, for which variables the evaluated statistics are similar, and for which they differ.

4. An example of application

4.1. The data

The proposed methods were applied for sampling a representative subsample of size $n = 28$ from a larger data set containing $N = 125$ patients with ventilatory disorders of obstructive type. The data are described more fully in Liebhart *et.al.* (1988), however notice that for the present analysis we took only 125 patients as opposed to 128 patients analyzed in the mentioned paper – this happened because vectors for 3 patients have been showing some internal inconsistencies and we have dropped them from the present analysis.

For some reasons (not explained here) we have been considering 8 traits corresponding to the variables no. 3, 4, 5, 6, 7, 9, 10, 11 in the original data. The variables no. 3 and 4 denote age and height of the patient, the remaining ones denote some characteristics of the patient´s spirogramme.

4.2. Analysing 6 subsamples by Q-Q plots exhibiting squared Mahalanobis distances

We have drawn 6 subsamples from the entire data set. Proceeding according to the algorithm outlined in Subsection 2.3 we constructed for each subsample the respective Q-Q plot for the squared Mahalanobis distances evaluated from the subsamples. The obtained Q-Q plots are shown in Fig. 1.

All six Q-Q plots are referenced to the same parental distribution established on the base of the entire data set counting $N = 125$ patients. The referenced quantiles of that distribution are marked in the x-axis. Each of the Q-Q plots in Fig.1 has the same scale both in the x-axis and in the y-axis. The small circles in each scatterdiagram represent the points $(q_k, \tilde{D}^2_{(k)})$ in the notation of subsection 2.3 of the paper. The Q-Q plots numbered #2 and #5 have some points denoted by a "+" sign. That means that the corresponding test quantiles are beyond the maximum range $maxY$ of the scatterdiagram, which was fixed for the value $maxY = 30$. Encountering in a subsample values exceeding the prescribed range means that that subsample bears some unusual pattern – and therefore is bad for taking it as a representative sample.

Looking at the diagrams shown in Fig.1 we see that the subsamples #2 and #4 have values of the referenced quantiles exceeding the prescribed value $maxY = 30$;

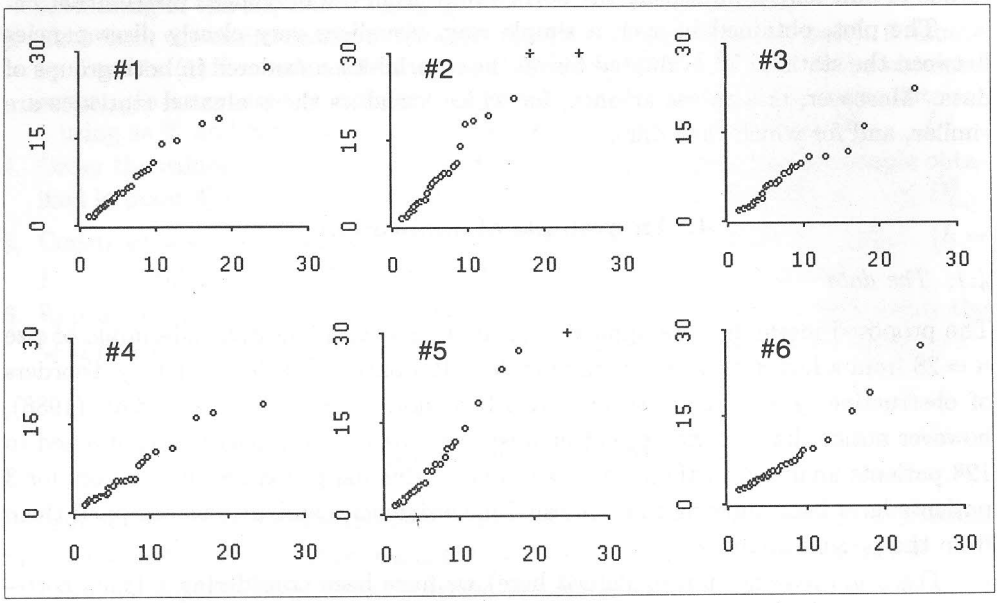

Fig. 1. Q-Q plots of squared Mahalanobis distances constructed for 6 subsamples numbered #1 through #6. Horizontally (x-axis): quantiles from the referenced distribution. Vertically (y-axis): quantiles from the tested (i.e. subsample) distribution.

thus these subsamples are somehow not good representatives of the entire data set.

The remaining four subsamples yielded values of $\tilde{D}^2$ within the presumed range. None the less, none of the corresponding Q-Q plots is exhibiting a truly linear pattern, which means that none of the considered subsamples is an ideal representative of the entire data. Samples #1 and #6 seem to be the "best".

4.3. Analysing the same 6 subsamples by polygon plots

To obtain an idea which characteristics for which variables in the subsamples are distorted, we have drawn polygon plots using the method outlined in Section 3. The obtained graphs are shown in Fig. 2. We have analysed 6 subsamples, thus figure 2 contains 6 framed panels, each representing results for one subsample.

In our analysis we have considered five statistics: mean, standard deviation, median, 1st quartile, 3rd quartile. For each subsample these characteristics were evaluated for all the 8 considered variables. Thus for each subsample we obtained 5 polygon plots. They are shown in Fig. 2 in a common framed panel. Each panel contains also a regular polygon based on a circle with a unit radius. This polygon serves

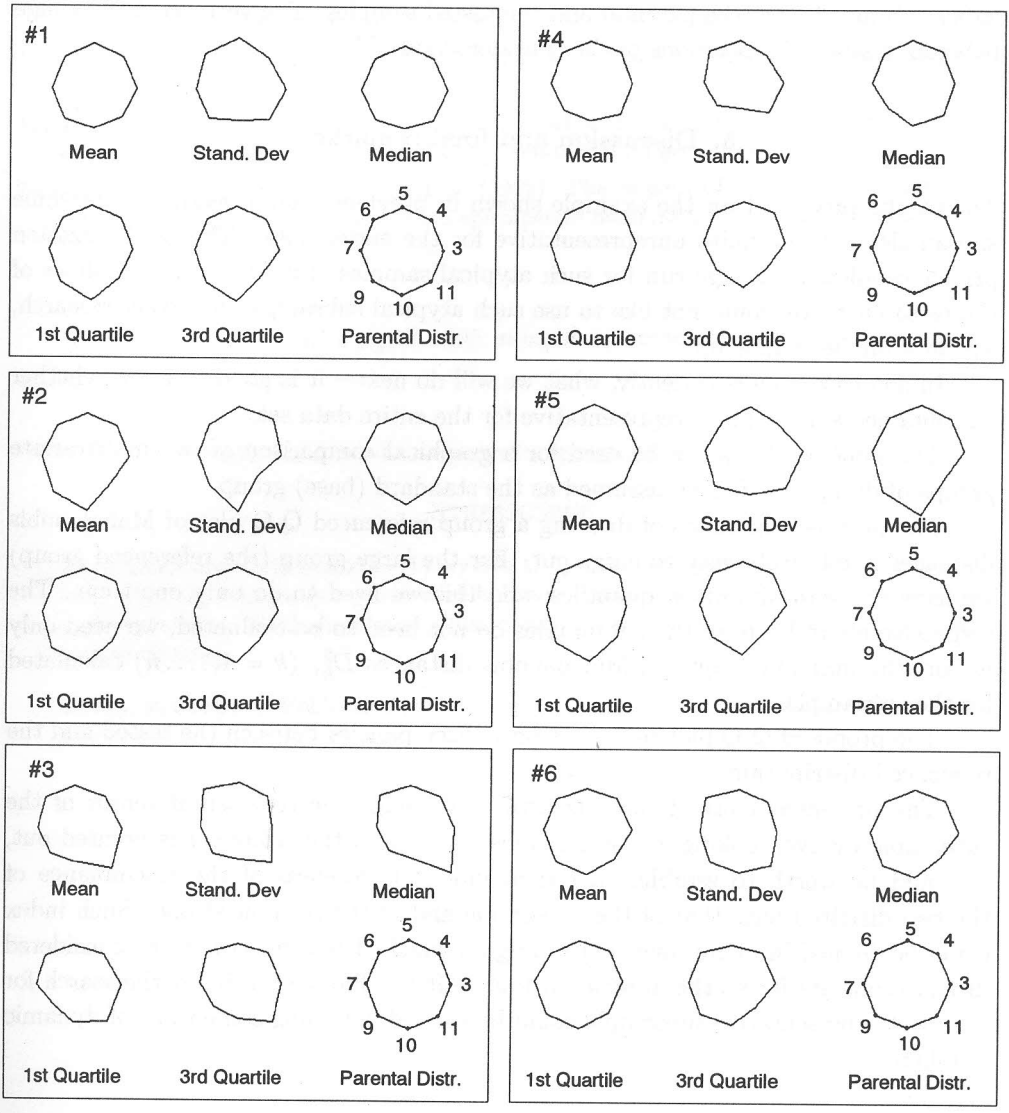

Fig. 2. Group referenced polygon plots for five statistics evaluated in 6 subsamples numbered #1 through #6. Polygons in the frames exhibit the mean, standard deviation, median, 1st quartile and 3rd quartile for each of the 8 traits of the inspected sample. A gauge polygon at the bottom right of each framed panel is added to indicate for each trait its position in the polygons

as the gauge for other polygons: it corresponds to the ideal concordance between the indices obtained from the parental and the tested samples. The vertices of the gauge polygon indicate the positions of the considered variables.

5. Discussion and final remarks

The results presented for the example shown in previous section warn us that some subsamples may be quite unrepresentative for the entire data. The randomization principles allow in a large run for such atypical samples. Putting us in the shoes of the researcher, we would not like to use such atypical subsample in further research, e.g. in a further follow up.

In any case – independently, what we will do next – it is good to know whether the obtained subsample is representative for the entire data set.

The same methods can be used for a graphical comparison of two multivariate groups of data, one of them assumed as the standard (base) group.

The proposed methods of drawing a group referenced Q-Q plot of Mahalanobis distances is relatively easy to carry out. For the large group (the referenced group) we have to establish only n quantiles, and this we need to do only one time. The respective quantiles from the subsamples do not need to be evaluated, we need only to sort the individual squared Mahalanobis distances $\tilde{D}_k^2$, $(k = 1, \ldots, n)$ calculated for the subsample.

The proposed Q-Q plot reveals at once discrepancies between the tested and the referenced distribution.

The presented methods are heuristic. We judge the representativeness of the subsample by eye, looking at some graphs. As one of the referees has pointed out, it would be worth to establish a formal index of closeness of the resemblance of the two distributions: that of the subsample and of the referenced one. Such index could be defined in many ways, depending on which characteristics of the considered distributions are for us the most important. After defining such index the search for the most representative subsample could be carried out using some tools of dynamic graphics.

References

Croft B. P. (1994). Interpreting the results of observational research: chance is not such a fine thing. *Brit. Med. Journal*, 309, 727–730. Quoted after R.F. Ulrich, EDSTAT-L, digest 1211, 23 Sept. 1996.

Bury K.V. (1975). *Statistical Models in Applied Science*. Wiley N.Y.

Givens G.H., Raftery A.E. (1996). Local adaptive importance sampling for multivariate densities with strong nonlinear relationship. *JASA* 91, Nb. 433, 132–141.

Hamilton L.C. (1992). *Regression with Graphics. A Second Course in Applied Statistics.* Brooks/Cole, Pacific Grove.

Hoaglin D.C. (1985). Using quantiles to study shape. In: Hoaglin D.C., Mosteller F., Tukey J. W. (Eds) *Exploring Data Tables, Trends and Shapes.* Wiley, New York, 417-460.

Liebhart J., Bartkowiak A., Liebhart E. (1989). The impact of outliers in the regression estimating TLC from age and some spirometric observations. *Modelling, Simulation & Control, C*, AMSE Press, 15, 1–18.

Received 2 April 1996; revised 24 September 1996

Pobieranie wielocechowej reprezentacyjnej próbki

STRESZCZENIE

Zaproponowano dwie metody: (1) wykres Q-Q na kwadratach odległości Mahalanobisa, oraz (2) relatywny wykres wielobokowy. Podano algorytmy dla sporządzania tych wykresów. Działanie metod zilustrowano na rzeczywistym przykładzie z praktyki medycznej.

SŁOWA KLUCZOWE: Pobieranie próbek, reprezentatywność, wykres Q-Q, wykres wielobokowy