Listy Biometryczne - Biometrical Letters
Vol.XXVI (1989), No.1,2, 5-20

A STATISTICAL STUDY ON THE OCCURRENCE OF CONGENITAL CLEFT PALATES IN BABIES

Anna Bartkowiak, Urszula Chojnacka

Institute of Computer Science, University of Wrocław,
Przesmyckiego 20, 51-151 Wrocław, Poland

Department of Child Surgery, Medical Academy,
Skłodowskiej-Curie 66, 51-646 Wrocław, Poland

Summary

We consider the occurrence of cleft palates (a congenital malformation) in alive-born babies in Lower Silesia. The recorded data is stored in two tables: $K = \{k_{ij}\}$ and $N = \{N_{ij}\}$, $i=1,2,\ldots,27$, $j=1,2,\ldots,10$, with k_{ij} denoting the reported number of cleft palates in the i-th county during the j-th year, N_{ij} denoting the number of alive-born babies in the i-th county and j-th year. We build a mathematical model for the random variable X_{ij} denoting the number of cleft palates. Using the likelihood ratio principle we derive some statistical tests for the hypotheses that the probability of the occurrence of a cleft palate is the same in all counties and/or years. Carrying out these tests we obtain some indications that the counties of Lower Silesia might be nonhomogeneous with regard to the considered probabilities. Next we transform the pair K, N to probits $Q = \{Q_{ij}\}$ and apply some multivariate techniques (chi2-plot, principal components) allowing to visualize the mutual similarities and dissimilarities between counties.

1. INTRODUCTION

In epidemiological investigations it is important to find out, (1) whether the occurrence of a particular disease is connected with the area inhabited by the patient. Such connection would mean that in this area some factors increasing the probability of occurrence of this disease may be hidden. Another important question in epidemiology is, (2) whether the probability of the occurrence of a disease is constant over time (years).

The aim of this paper is to answer both these questions with regard

Key words: cleft palates, nonhomogeneity of counties, Poisson distribution, likelihood ratio test, chi2-plot, principal components

to a particular disease: the occurrence of cleft palates (a congenital malformation) in alive-born babies in 17 counties of Lower Silesia in the period 1961-1970.

Formally, from the statistical point of view, we deal here with counts of a rare event, which can be put together in the layout of a two-way contingency table, the one entry denoting the county and the other the year of birth of the baby. The count of each cell of this contingency table must be referred to the number of alive-born babies in the respective county during the respective year. Therefore the problem cannot be treated with classical methods of multiway contingency tables. Our task was to introduce a statistical model describing the probability of the occurrence of a cleft palate in a baby born in a given county in the given year and next to formulate two epidemiological questions mentioned above (as (1) and (2)), in terms of statistical hypotheses which can be dealt with in a known formal way.

We added to our formal statistical analysis some explorative analysis allowing to visualize graphically a nonhomogeneity among the considered units. The explorative analysis, although based on known statistical methods, seems to be quite new in the context of the considered problems.

2. THE DATA AND THE GENERAL MATHEMATICAL MODEL

A cleft palate is a rare congenital malformation. World statistics report 80-210 cleft palates among 100 000 alive born babies. A cleft palate is a very serious congenital malformation and has to be treated surgically.

We consider cleft palates observed in babies born alive in Lower Silesia in 27 counties (small administrative areas) in the years 1962-1970. Our primary data consists of records of k_{ij}, the number of cleft palates observed in the i-th county ($i = 1,2,\ldots,27$) and j-th year ($j = 1,2,\ldots,10$), and N_{ij}, the number of babies born in the i-th county and j-th year. The values of k_{ij} are small: from 0 (no baby with cleft palates) to 8 (the maximum of cleft palates reported during one year in one county). The numbers N_{ij} of babies born during one year vary from 547 to 5252. Now let us introduce the mathematical model.

Let us suppose that we have N babies with the same exposure to cleft palates. Let p ($0<p<1$) be the probability that the new-born baby has a cleft palate. We assume that this probability is the same for all babies. Further we assume that the occurrence of a cleft palate in one baby is independent of the occurrence of this malformation in another baby. In these circumstances we can imagine the occurrence or non-occurrence of a cleft palate in N babies as a series of N independent trials. For each trial the considered event (a cleft palate) can occur with the same probability p. Then X, the number of cleft palates observed in this series

of N trials, is a random variable which can take values 0,1,2,...,N. The probability Pr that the value of X equals k ($0 \le k \le N$), is given by the binomial distribution:

$$Pr(X=k) = \binom{N}{k} p^k (1-p)^{N-k} \quad . \tag{1}$$

The expected value of X is Np (see, e.g. Feller (1950)).

It is well known, see e.g. Feller (1950), that in the case when the expected number of events is small, the binomial distribution (1) can be approximated by a Poisson distribution given by the formula:

$$Pr(X=k) = \frac{(Np)^k}{k!} e^{-Np} \quad , \quad k=0,1,2,... \tag{2,}$$

Because our data are stratified into t counties and l years, we consider a set of random variables X_{ij}, each having the distribution

$$Pr(X_{ij} = k_{ij}) = \frac{(N_{ij}p_{ij})^{k_{ij}}}{k_{ij}} e^{-N_{ij}p_{ij}}, \quad i=1,...,t, \; j=1,...,l. \tag{3}$$

In the model given by (3) we assume that the probability of the occurrence of a cleft palate in one baby born in the i-th county during the j-th year is p_{ij}. We assume that the probabilities p_{ij} can be different in various counties and years.

A maximum likelihood estimator for p_{ij} is

$$\hat{p}_{ij} = k_{ij}/N_{ij} \quad , \quad i=1,...,t, \quad j=1,...,l. \tag{5}$$

Our task is to verify whether the p_{ij}'s are significantly different for various counties and years. We shall perform this task in the next chapter using likelihood ratio tests.

3. LIKELIHOOD RATIO TESTS

Let us assume the Poisson model given by (3). Generally we are interested in 3 hypotheses:

$$H_A: p_{1j} = p_{2j} = ... = p_{tj} = p_{.j}, \quad j=1,2,...,l, \tag{5a}$$

$$H_B: p_{i1} = p_{i2} = ... = p_{il} = p_{i.}, \quad i=1,2,...,t, \tag{5b}$$

$$H_{AB}: p_{11} = p_{12} = ... = p_{1t} = p_{..}. \tag{5c}$$

Assuming H_A we state that the probabilities of the occurrence of cleft palates may differ for various years, none the less, having the year j fixed, they are the same for all counties.

Assuming H_B we state that for a given county i the probabilities p_{ij}, $j=1,2,...,l$, are the same during the whole considered period.

Assuming H_{AB} we state that during the whole period the probabilities p_{ij} are the same in all counties.

To test the hypotheses H_A, H_B and H_{AB} we use test statistics derived from the likelihood ratio principle presented e.g. in Mood and Graybill (1973). The formulae for the respective test statistics are:

for testing H_A

$$X_A^2 = 2 \sum_{i=1}^{t} \sum_{j=1}^{l} [k_{ij} \ln(\hat{p}_{ij}/\hat{p}_{i.}) - N_{ij}(\hat{p}_{ij} - \hat{p}_{i.})] \qquad (6a)$$

with $t(l-1)$ degrees of freedom,

for testing H_B

$$X_B^2 = 2 \sum_{i=1}^{t} \sum_{j=1}^{l} [k_{ij} \ln(\hat{p}_{ij}/\hat{p}_{.j}) - N_{ij}(\hat{p}_{ij} - \hat{p}_{.j})] \qquad (6b)$$

with $l(t-1)$ degrees of freedom,

for testing H_{AB}

$$X_{AB}^2 = 2 \sum_{i=1}^{t} \sum_{j=1}^{l} [k_{ij} \ln(\hat{p}_{ij}/\hat{p}_{..}) - N_{ij}(\hat{p}_{ij} - \hat{p}_{..})] \qquad (6c)$$

with $tl-1$ degrees of freedom.

The maximum likelihood estimators for $p_{i.}$ and $p_{.j}$ are respectively:

$$\hat{p}_{i.} = k_{i.}/N_{i.}, \quad \hat{p}_{.j} = k_{.j}/N_{.j} \quad . \qquad (7)$$

Under H_A, H_B and H_{AB} the test statistics X_A^2, X_B^2 and X_{AB}^2 have asymptotically a χ^2 distribution with $t(l-1)$, $l(t-1)$ and $tl-1$ degrees of freedom, respectively.

The values of the test statistics for our data are given in Table 1. In this table X_{calc}^2 stands for the calculated value of the appropriate test statistic X_A^2, X_B^2 or X_{AB}^2.

None of hypotheses H_A, H_B, H_{AB} can be rejected when using the test statistics X_A^2, X_B^2 or X_{AB}^2. This means that using these statistics we can not prove any significant differences in the probabilities describing the occurrence of a cleft palate in the considered counties and/or years.

Still we can apply here another procedure. Assuming that H_B is true we can build a model with probabilities $p_{1.}, p_{2.}, \ldots, p_{t.}$ only. In this model we test the hypothesis

$$H_{A/B} : p_{1.} = p_{2.} = \ldots = p_{t.} \quad . \qquad (8)$$

The test statistic for $H_{A/B}$ is $X_{A/B}^2$, given by the formula

$$X_{A/B}^2 = 2 \sum_{i=1}^{t} [k_{i.}(\ln \hat{p}_{i.}/\hat{p}_{..}) - N_{i.}(\hat{p}_{i.} - \hat{p}_{..})] \quad , \qquad (9)$$

with $t-1$ degrees of freedom.

Table 1. Values of test statistics derived from the likelihood ratio principle

Hypothesis and its meaning		X^2_{calc}	df	$P=P(X^2>X^2_{calc}/H_o)$
H_A:	equality of p_{ij}'s in columns (counties)	245.5	243	0.44
H_B:	equality of p_{ij}'s in rows (successive years)	275.7	260	0.24
H_{AB}:	equality of all p_{ij}'s (over counties and years)	284.0	269	0.25
$H_{A/B}$:	conditional equality of the p_{ij}'s	38.5	26	0.05

For our data the calculated value of the test statistic $X^2_{A/B}$ equals $X^2_{calc} = 38.5$ with $df = 26$ degrees of freedom. Under $H_{A/B}$, the probability of obtaining equal or larger value of $X^2_{A/B}$ equals 0.0540, so it is on the border of statistical significance. At this moment we are in doubt whether $H_{A/B}$ is true. We should seek other indication on this topic.

We shall do it in next chapters of this paper using some graphical methods.

4. TRANSFORMATION TO PROBITS

Let us assume that all the p_{ij}'s are equal:

$$H_0 : p_{ij} = p_{..}, \quad i = 1,2,\ldots,t, \quad j = 1,2,\ldots,l. \tag{10}$$

The distribution of the random variable X_{ij} is given now by:

$$Pr(X_{ij} = k_{ij}) = \frac{(N_{ij}p_{..})^{k_{ij}}}{k_{ij}!} e^{-N_{ij}p_{..}} \tag{11}$$

For observed values $k_{11}, k_{12}, \ldots, k_{tl}$ the likelihood function L can be expressed as

$$L = L(k_{11}, k_{12}, \ldots, k_{tl}; p_{..}) = \prod_{i=1}^{t}\prod_{j=1}^{l} \frac{(N_{ij}p_{..})^{k_{ij}}}{k_{ij}!} e^{-N_{ij}p_{..}},$$

where from we obtain

$$\hat{p}_{..} = \frac{\sum_{i=1}^{t}\sum_{j=1}^{l} k_{ij}}{\sum_{i=1}^{t}\sum_{j=1}^{l} N_{ij}} = \frac{k_{..}}{N_{..}} \tag{12}$$

as the estimator for $p_{..}$. For our data, $\hat{p}_{..} = 0.001024$.

Assuming the Poisson model given by (11) and substituting $p_{..}$ with $\hat{p}_{..} = 0.001024$ we can for any given values k_{ij} and N_{ij} calculate the probability $Pr(k_{ij})$ that X_{ij} does not exceed the observed value k_{ij}. This probability is given by the formula

$$Pr(k_{ij}) = Pr(X_{ij} \leq k_{ij}/H_o) = \sum_{h=0}^{k_{ij}} Pr(X_{ij} = h) \quad . \tag{13}$$

For the calculated value $Pr(k_{ij})$ we find appropriate quantile of the Gauss-Laplace distribution:

$$Q_{ij} = \phi^{-1}(Pr(k_{ij})) \quad , \tag{14}$$

where

$$\phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \quad .$$

The values of Q_{ij}, $i = 1,2,\ldots,t$, $j = 1,2,\ldots,l$, are given in Table 2. They are called probits. Being quantiles of the standardized normal distribution, they practically take values from the interval (-3.0, +3.0) and are comparable each other. Moreover, they are suitable for handling with various multivariate techniques.

In the next two chapters we shall carry out an explorative analysis of the data which constitute the array $Q = \{Q_{ij}\}$. Our goal will be to find out whether the counties, which correspond to the rows of the table Q, are homogeneous.

5. INVESTIGATIONS ON THE HOMOGENEITY OF COUNTIES BY DRAWING A CHI2-PLOT

The array $Q = \{Q_{ij}\}$ comprising the probits Q_{ij} computed by formula (14) may be viewed as a set of t points situated in a l-dimensional Euclidean space R^l. For our data $t = 27$ (the number of counties) and $l = 10$ (the number of years). The center of gravity of these t points has the coordinates

$$\bar{q} = (\bar{Q}_{.1}, \bar{Q}_{.2}, \ldots, \bar{Q}_{.l}) \quad , \tag{15}$$

where

$$\bar{Q}_{.j} = \left(\sum_{i=1}^{t} Q_{ij} \right)/t \quad , \quad j = 1,\ldots,l.$$

Now for each point (which corresponds to one row of the array Q) we calculate its Mahalanobis distance from the center of gravity $\bar{q}$. In this way we obtain the Mahalanobis distances:

$$D_1^2, D_2^2, \ldots, D_t^2 \quad . \tag{16}$$

Table 2. Probits calculated for 27 counties

No. and name of the county	1961	1962	1963	1964	1965	1966	1967	1968	1969	1970	Marginal mean
1. Bolesławiec	0.52	0.03	0.74	0.78	0.15	0.69	1.00	1.04	0.26	-0.66	0.3110
2. Bystrzyca Kłodzka	1.37	-0.30	2.33	-0.12	3.06	1.82	-0.02	1.02	-0.02	1.80	1.0940
3. Dzierżoniów	-0.68	0.74	0.31	0.29	0.36	1.02	1.25	0.57	-0.15	-0.10	0.3610
4. Góra	-0.32	-0.22	-0.17	1.76	-0.09	0.06	0.07	0.06	0.01	0.03	0.1250
5. Jawor	-0.50	1.37	-0.32	-0.18	1.34	-0.29	-0.22	0.78	-0.25	0.77	0.2500
6. Jelenia Góra	-0.70	1.33	0.83	1.43	0.36	0.46	-0.12	1.06	-0.28	-0.32	0.4050
7. Kamienna Góra	-0.73	0.37	0.50	1.19	1.30	0.61	-0.25	-0.30	1.46	0.56	0.4710
8. Kłodzko	0.12	0.88	0.22	1.72	-0.50	-0.48	0.47	-0.44	0.42	1.16	0.3570
9. Legnica	0.92	1.68	-0.19	1.78	-0.92	0.71	0.03	1.85	1.25	0.49	0.7600
10. Lubań	2.03	0.24	0.26	1.09	1.23	0.54	0.59	0.56	0.49	0.48	0.7510
11. Lubin	2.24	-0.25	1.52	0.68	3.20	-0.37	0.69	0.65	1.26	-0.59	0.9030
12. Lwówek Śląski	0.43	1.32	-0.40	2.07	-0.26	-0.24	0.82	-0.12	-0.20	0.75	0.4170
13. Milicz	0.50	1.97	1.46	-0.29	0.77	-0.24	-0.16	0.74	-0.23	-0.24	0.4280
14. Nowa Ruda	0.21	1.74	1.88	1.25	2.00	-0.29	2.22	2.25	1.50	0.80	1.3560
15. Oleśnica	0.05	0.20	0.14	-0.71	-0.70	0.24	0.35	0.32	0.26	-0.69	-0.0540
16. Oława	0.43	-0.45	0.54	0.57	-0.41	0.59	-0.35	1.43	0.62	1.37	0.4340
17. Strzelin	2.04	1.43	2.71	-0.22	0.74	0.85	-0.03	-0.16	-0.12	-0.10	0.7140
18. Syców	0.08	0.11	0.13	1.23	1.16	1.21	0.13	1.18	1.10	0.08	0.6410
19. Środa Śląska	0.30	-0.50	1.25	0.55	0.56	0.61	1.49	0.69	0.68	0.64	0.6270
20. Świdnica	-0.28	0.45	1.03	0.67	0.67	0.23	-0.45	-0.45	1.79	-0.54	0.3120
21. Trzebnica	0.28	-0.48	-0.47	0.46	-0.36	2.04	2.73	1.44	-0.36	0.55	0.5830
22. Wałbrzych	-1.41	0.03	-0.42	-0.92	-0.86	-1.54	1.61	0.01	-0.01	-0.74	-0.4250
23. Wołów	0.92	1.00	1.70	0.44	1.17	-0.44	-0.41	0.50	0.55	0.54	0.5970
24. Wrocław - county	0.78	0.92	0.25	0.25	0.99	0.37	-0.56	0.30	1.71	1.64	0.6650
25. Ząbkowice Śląskie	-0.03	-0.80	0.87	0.94	0.30	1.08	0.34	0.39	1.10	0.34	0.4530
26. Zgorzelec	-0.97	1.79	1.18	-0.03	0.05	2.09	0.98	0.26	2.20	0.16	0.7710
27. Złotoryja	-0.86	1.29	0.01	0.89	1.66	0.41	-0.53	0.42	1.20	0.43	0.4920
Mean	0.250	0.586	0.663	0.651	0.630	0.384	0.432	0.594	0.602	0.319	0.5111

Under H_o given by (10) and under normality of the Q_{ij}'s each D_j^2 is distributed as a χ^2 variate with I-1 degrees of freedom.

Using the technique of probability plots we construct a chi2-plot which is shown in Fig.1. For an explanation of the method of constructing a probability plot see e.g. Bury (1975). The method of constructing a chi2-plot is explained e.g. in Gnanadesikan and Kettenring (1972) and Bartkowiak et al. (1988). For homogeneous data following a given distribution we should obtain a plot consisting of points located along a straight line.

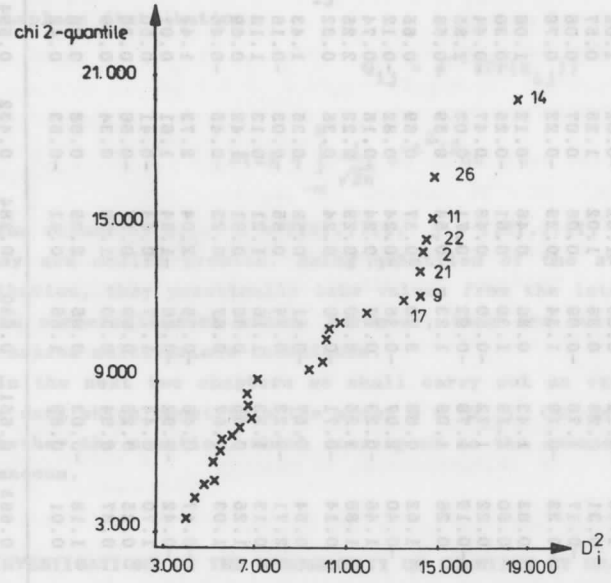

Fig. 1. Chi2-plot with Mahalanobis distances.

Looking at Fig.1 we can state that generally the points are not situated along a line. The linearity is exhibited for smaller values of D_i^2. Next we see a bend (up from the county no 17 till no 26). The county no 14 appears as an isolated point. We conclude that the considered array **Q** comprises some nonhomogeneous rows.

The points nos. 26, 11, 22, 2, 21, 9, 17 look distinct. They seem to be situated along another straight line. The point no 14 is clearly isolated from the rest of the points. We could therefore conclude that our data is a mixture of two normal distributions and one point (no 14) which can be suspected to be an outlier. Looking at the marginal probits we state that point no 14, representing the county Nowa Ruda, has the largest marginal probit (= 1.3560) indicating that the observed frequency of cleft palates is here the highest. The marginal probits for the

counties nos. 26, 11, 22, 2, 21, 9, 17 are 0.77, 0.90, -0.42, 1.09, 0.58, 0.71; they do not differ from marginal probits for other counties.

To obtain an insight into the mutual position of the counties identified with points in R^{10} we shall consider scatterdiagrams of the first two and principal components.

6. GRAPHICAL PRESENTATION OF COUNTIES USING PRINCIPAL COMPONENTS

Let us consider the array $\mathbf{Q} = \{Q_{ij}\}$, $i = 1,\ldots,27$, $j = 1,\ldots,10$. Each row $q_i = (Q_{i1}, Q_{i2},\ldots,Q_{i10})$ of $\mathbf{Q}$ is viewed as an individual point in the 10-dimensional Euclidean space R^{10}. So the 27 counties are represented as a cluster of points-individuals in this space. In the following we apply the principal component technique as described in Morrison (1967) or Jolliffe (1985).

The idea of principal components is to approximate points in R^p (in our data $p = 10$) by their projections onto a subspace of a lower dimension. Let $\mathbf{S} = (s_{ij})$, $i,j = 1,\ldots,10$ be the cross product matrix calculated from the array $\mathbf{Q}$. The matrix $\mathbf{S}$ can be reproduced by its eigenvalues $\lambda_1,\lambda_2,\ldots,\lambda_{10}$ and the eigenvectors $c_1,c_2,\ldots,c_{10}$ corresponding to these eigenvalues; that is,

$$\mathbf{S} = \sum_{h=1}^{10} \lambda_h c_h c_h^T \quad , \tag{17}$$

provided that $(c_1,\ldots,c_{10})^T(c_1,\ldots,c_{10}) = I_{10}$ (what means that the eigenvectors $c_1,\ldots,c_{10}$ are ortonormalized).

For each r ($1 \le r \le 10$) the best approximation (in the L_2 norm) of the matrix $\mathbf{S}$ may be obtained by the first r eigenvalues and eigenvectors of this matrix:

$$\mathbf{S}^{(r)} = \sum_{h=1}^{r} \lambda_h c_h c_h^T \tag{18}$$

Similarly, for each r the best approximation (in the L_2 norm) of the array $\mathbf{Q} = \{Q_{ij}\}$ by r new coordinates may be obtained when using new coordinates $y_{i1}, y_{i2},\ldots,y_{ir}$ defined as

$$y_{i1} = q_i c_1, \quad y_{i2} = q_i c_2,\ldots,y_{ir} = q_i c_r \tag{19}$$

with q_i being the i-th row of the array $\mathbf{Q}$. Denoting the old coordinates by $\mathbf{X} = (X_1,X_2,\ldots,X_{10})$, the new coordinates by $\mathbf{Y}^{(r)} = (Y_1,Y_2,\ldots,Y_r)$ and the matrix of the first r eigenvectors by $\mathbf{C}^{(r)} = (c_1,c_2,\ldots,c_r)$, we can rewrite (19) as

$$\mathbf{Y}^{(r)} = \mathbf{X}\mathbf{C}^{(r)} \tag{20}$$

For $r=10$ we have $\mathbf{Y}^{(10)} = \mathbf{X}\mathbf{C}^{(10)}$, or simply $\mathbf{Y} = \mathbf{X}\mathbf{C}$.

Now suppose that the true rank of the matrix $\mathbf{S}$ is h, $1 \le h < 10$, what

means that the cluster of points-individuals in R^{10} is h-dimensional, and among the columns of the array **Q** there are only h linearly independent columns. In this situation the matrix **S** can be approximated by $S^{(h)}$ and the variables **Y** by $Y^{(h)}$, the remaining 10-h coordinates being zero.

In practice we seldom know what is the true dimension of the considered cluster of points. Inspecting $S^{(h)}$ and $C^{(h)}$ for h = 1,2,...,p we can make some inference about the approximation based on the first h eigenvalues and eigenvectors. Izenman (1980) proposes two characteristics ΔS and ΔC defined as

$$\Delta C^{(h)} = \frac{\|\hat{C}^{(h)} - C\|}{\|C\|} , \tag{21}$$

$$\Delta S^{(h)} = \frac{\|S^{(h)} - S\|}{\|S\|} , \tag{22}$$

with $\|A\|$ being the norm of the matrix **A**.

In our consideration we assumed that $\|A\|$ is the classical Euclidean norm: $\|A\| = (tr(AA^T))^{1/2} = (\sum_{i}^{10}\sum_{j}^{10} a_{ij}^2)^{1/2}$. Then (21) and (22) reduce to:

$$\Delta C^{(h)} = (1-h/10)^{1/2} , \tag{23}$$

$$\Delta S^{(h)} = \left\{\left[\sum_{j=h+1}^{10} \lambda_j^2\right] / \left[\sum_{j=1}^{10} \lambda_j^2\right]\right\}^{1/2} . \tag{24}$$

Plotting $\Delta S^{(h)}$ against $\Delta C^{(h)}$ for h = 0,1,...,10 we obtain a rank trace plot which is shown in Fig.2. We see that, generally, $\Delta S^{(h)}$ decreases faster than $\Delta C^{(h)}$ does. This means that $\Delta S^{(h)}$ stabilizes faster than $\Delta C^{(h)}$. According to Izenman's proposal, we can assess the rank of the approximation of **S** by $S^{(h)}$ by the smallest integer value between 1 and p at which an "elbow" can be detected in the PC rank plot. Looking at Fig. 2 we do not see any clear "elbow". We guess that a kind of "elbow" appears for h=3 or h=6.

Other practical rule to determine how many principal components should be included into the model is a thumb rule called also Kaiser's rule (see, e.g. Jolliffe (1985)). This rule was constructed specially for use with correlation matrices, although it can be adapted for covariance matrices. The idea behind this rule is that if all columns of the array **Q** are independent, than the principal components are the same as the original variables and all have unit variances in the case of a correlation matrix. The variances of the subsequent principal components equal to subsequent eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$. In these circumstances any principal component with variance less than one contains less information than any of the original variables and so is not worth retaining. Taking into account the sampling variability a practical advise is to retain principal components corresponding to eigenvalues larger than $l^* = 0.7$.

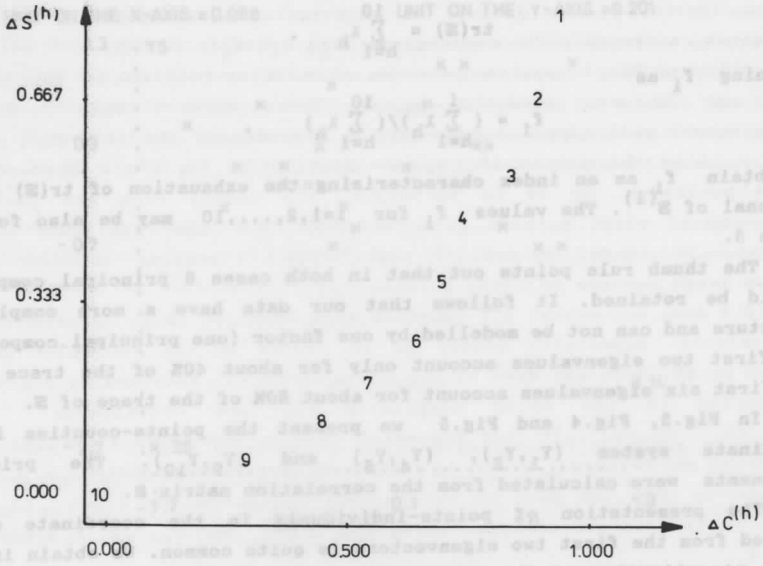

Fig. 2. Rank trace plot.

This rule can be adapted for principal components calculated from a covariance matrix. In this case the cut-off point is $l^* = 0.7\,\bar{l}$, where $\bar{l}$ is the mean of all eigenvalues calculated from the covariance matrix (Jolliffe, 1985).

The eigenvalues of the correlation matrix and of the covariance matrix for our data are given in Table 3.

Table 3. Eigenvalues λ_i and fractions f_i of exhaustion of tr(S) by $\sum_{h=1}^{i} \lambda_h$

No i	1	2	3	4	5	6	7	8	9	10
a) from the correlation matrix										
λ_i	2.22	1.52	1.44	1.06	1.03	0.87	0.61	0.50	0.45	0.30
f_i	0.22	0.37	0.52	0.62	0.73	0.81	0.87	0.92	0.97	1.00
b) from the covariance matrix										
λ_i	1.83	0.94	0.84	0.69	0.63	0.56	0.43	0.29	0.26	0.22
f_i	0.27	0.41	0.54	0.64	0.74	0.82	0.89	0.93	0.97	1.00

From (17) it follows that

$$tr(\mathbf{S}) = \sum_{h=1}^{10} \lambda_h \quad . \tag{25}$$

Defining f_i as

$$f_i = \Big(\sum_{h=1}^{i} \lambda_h\Big) / \Big(\sum_{h=1}^{10} \lambda_h\Big) \quad , \tag{26}$$

we obtain f_i as an index characterizing the exhaustion of $tr(\mathbf{S})$ by the diagonal of $\mathbf{S}^{(i)}$. The values f_i for $i=1,2,\ldots,10$ may be also found in Table 3.

The thumb rule points out that in both cases 6 principal components should be retained. It follows that our data have a more complicated structure and can not be modelled by one factor (one principal component). The first two eigenvalues account only for about 40% of the trace of $\mathbf{S}$. The first six eigenvalues account for about 80% of the trace of $\mathbf{S}$.

In Fig.3, Fig.4 and Fig.5 we present the points-counties in the coordinate system (Y_1,Y_2), (Y_4,Y_5) and (Y_9,Y_{10}). The principal components were calculated from the correlation matrix $\mathbf{R}$.

The presentation of points-individuals in the coordinate system derived from the first two eigenvectors is quite common. We obtain in this way a visualization of the mutual positions of the considered points in R^{10}. In our case the first two eigenvalues explain a fraction equal to 0.37 of the trace of $\mathbf{R}$. This is not much and therefore some relevant

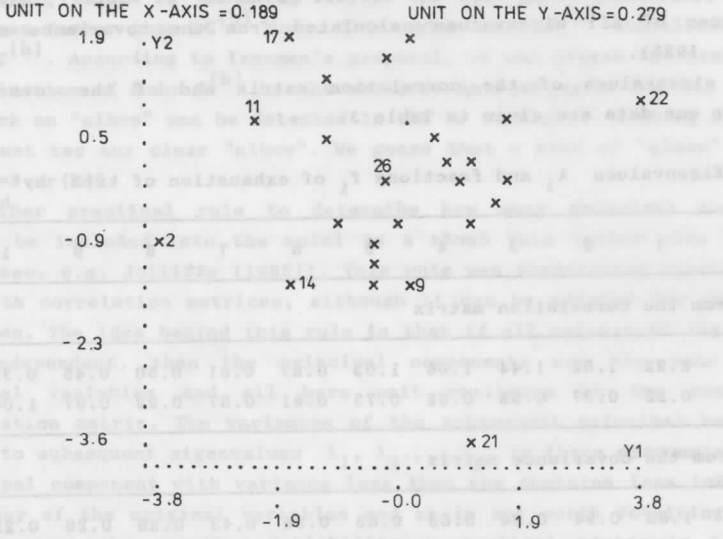

Fig. 3. Scatterdiagram of counties in the coordinate system of the first two principal components.

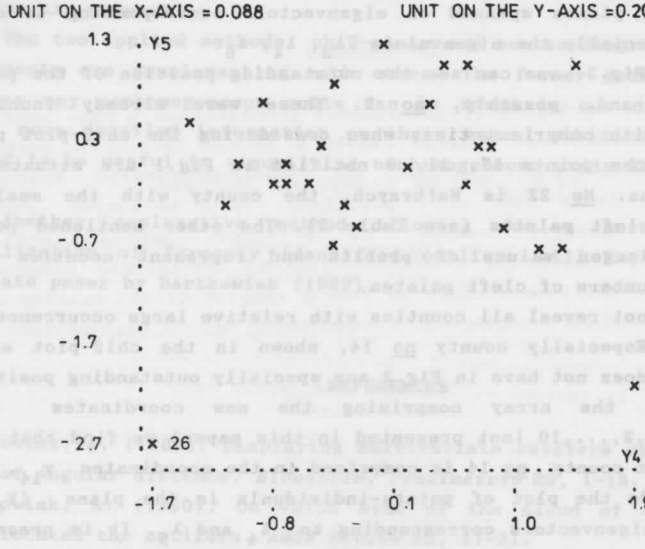

Fig. 4. Scatterdiagram of counties in the coordinate system of the 4-th and 5-th principal component.

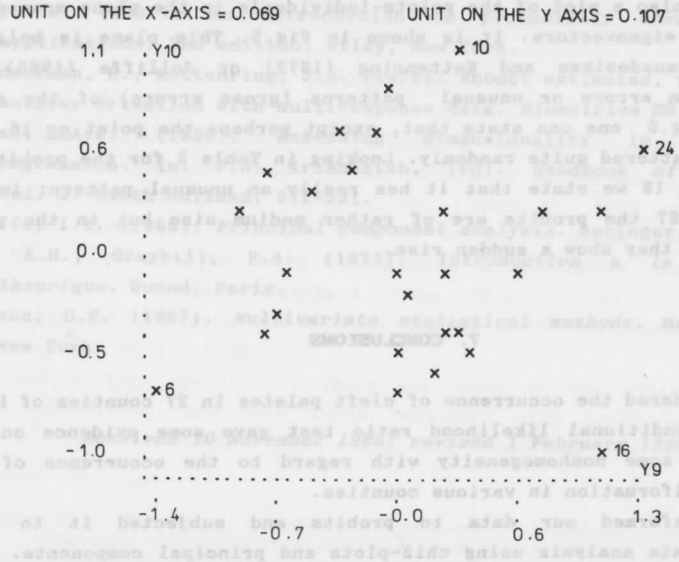

Fig. 5. Scatterdiagram of counties in the coordinate system of the last two principal components.

information concerning the mutual position of points might be obtained when considering planes spanned on eigenvectors corresponding to other eigenvalues, especially the eigenvalues λ_3, λ_4, λ_5.

Looking at Fig.3 one can see the outstanding position of the points no 21, no 22 and, possibly, no 2. These were already found as nonhomogeneous with other counties, when considering the chi2-plot given in Fig.1. Also the points 17, 11, 9 notified in Fig.1 are situated at extreme positions. No 22 is Wałbrzych, the county with the smallest occurrences of cleft palates (see Table 2). The other mentioned points correspond to larger values of probits and represent counties with relative large numbers of cleft palates.

Fig.3 does not reveal all counties with relative large occurrences of cleft palates. Especially county no 14, shown in the chi2-plot as an isolated point, does not have in Fig.3 any specially outstanding position.

Looking in the array comprising the new coordinates y_{ij}, $i=1,\ldots,27$, $j=1,2,\ldots,10$ (not presented in this paper) we find that much information about county no 14 is comprised in the coordinates y_{i4}, y_{i5}. Therefore we made the plot of points-individuals in the plane (Y_4,Y_5) spanned by the eigenvectors corresponding to λ_4 and λ_5. It is presented in Fig.4.

Looking at Fig.4 one can see the outstanding positions of points no 14 and no 26. The last county was not found as outlier in Fig.3. From Table 2 we can see that it has a considerably large marginal probit.

We made also a plot of the points-individuals in the plane spanned by the last two eigenvectors. It is shown in Fig.5. This plane is believed (see e.g. Gnanadesikan and Kettenring (1972) or Jolliffe (1985)) to reflect random errors or unusual patterns (gross errors) of the data. Looking at Fig.5 one can state that, except perhaps the point no 16, the points are scattered quite randomly. Looking in Table 2 for the probits of the county no 16 we state that it has really an unusual pattern: in the years 1961-1967 the probits are of rather medium size but in the years 1968 and 1970 they show a sudden rise..

7. CONCLUSIONS

We considered the occurrence of cleft palates in 27 counties of Lower Silesia. A conditional likelihood ratio test gave some evidence on the existence of some nonhomogeneity with regard to the occurrence of the considered malformation in various counties.

We transformed our data to probits and subjected it to some explorative data analysis using chi2-plots and principal components. This analysis confirmed our suspicions on the nonhomogeneity of counties. We found one county (no 14, Nowa Ruda) with relatively high occurrences of cleft palates. This county is an industrial area. Another county (no 22,

Wałbrzych, a mining area) has relatively small occurrences of the considered malformations. This county has mineral water spring with therapeutic properties.

The two applied methods: chi2-plots and scatterdiagrams of principal components are complementing each other. The former generally indicates whether any nonhomogeneous units can be found in our data. The latter gives more detailed information on their mutual positions. Both methods proved to be useful in visualizing nonhomogeneous units in the considered data.

Another explorative method allowing for a more comprehensive visualization of formerly identified outliers will be presented in a separate paper by Bartkowiak (1989).

REFERENCES

Bartkowiak, A. (1989). Displaying multivariate outliers using the concept of angular distance. *Biometrie, Praximetrie* **29**, 1-16.

Bartkowiak, A. (1990). On which side of the cloud of data points are located the outliers. *AMSE Review* **15**, 17-31.

Bartkowiak, A., Łukasik, S., Chwistecki, K., Mrukowicz, M., and Morgenstern, W. (1987). Location of outliers in large epidemiological data. *EDV in Medizin und Biologie* **18**, 108-114.

Bury, K.V. (1975). *Statistical models in applied science.* Wiley, New York.

Feller, W. (1961). *An introduction to probability theory and its applications.* 2nd edition, Wiley, New York.

Gnanadesikan, R., Kettenring, J.R. (1972). Robust estimates, residuals and outlier detection with multiresponse data. *Biometrics* **28**, 81-124,

Izenman, A.J. (1980). Assessing dimensionality in multivariate regression. In: P.R. Krishnaiah, (Ed). *Handbook of statistics. Vol. I.* North Holland, 571-591.

Jolliffe, I.T. (1985). *Principal component analysis.* Springer, New York.

Mood, A.M., Graybill, F.A. (1973). *Introduction a la statistique theorique.* Dunod, Paris.

Morrison, D.F. (1967). *Multivariate statistical methods.* Mc Graw Hill, New York.

Received 10 November 1988: revised 1 February 1989

BADANIA STATYSTYCZNE NAD WYSTĘPOWANIEM ROZSZCZEPÓW PODNIEBIENIA U NOWORODKÓW

Streszczenie

Rozpatrujemy występowanie rozszczepów podniebienia u żywo urodzonych noworodków w 27 powiatach Dolnego Śląska w latach 1961-1970. Pokazujemy testy na weryfikowanie hipotez, że prawdopodobieństwo wystąpienia rozszczepu jest takie samo a) w powiatach, b) w kolejnych latach. Z przeprowadzonej analizy wynika, że rozważane powiaty mogą być niejednorodne ze względu na prawdopodobieństwo wystąpienia rozszczepów. Wniosek ten może być potwierdzony za pomocą eksploratywnej analizy danych wykonanej na probitach otrzymanych z rozważanych danych statystycznych.