Informační Bulletin

Czech Statistical Society English issue, 1995, vol. 6.

The National Museum and the development of statistical science

Prokop Závodský

On the 15th of April 1993 the Czech cultural public remembered the 175th anniversary of the foundation of the Patriotic Museum in Bohemia (now the National Museum) in 1818.It is less known that the significant articles on the developement of political arithmetics (a predecessor of contemporary statistical science) were already publishing in the first volumes of museum journals.

The situation at Prague University till the year of 1848 (teaching was done in German) did not at all encourage the developement of original scientific thinking, the scientific work of professors was not necessary but inconvenient. Prescribed textbooks, in most cases rather obsolete, were usually written by Austrian authors. G. N. Schnabel (1791–1857), a professor of statistics and at that time the most important personality at the Law School, wrote a great number of remarkable statistical publications. In spite of that he was forced to teach from old official textbooks written by Austrian professors Zizius and Bisinger.

The Bohemian Museum, which was founded by a group of aristocratic enthusiasts as a provincial scientific institution, started in 1827 publishing the German and the Czech Museum Journals (F. Palacký, a famous Czech historian and politician of Middle Europe, was a longtime editor of the journal; today his portrait forms a main motif on our banknote of one thousand crowns value). The museum partly took the role of the university as an organizer and supporter of scientific development. Both journals were also appretiated by J. W. Goethe in a critical journal issued in Berlin. At the beginning the German Monatschrift der Gesellschaft des vaterländischen Museums in Böhmen was the representative organ of the Museum Society. With time the Czech journal (Časopis Českého Museum) was increasing its standards. The German journal was published monthly and was orientated on the provincial patriotism, which transformed into modern czech and german nationalism. The journal stopped publishing in 1831.

In the first volumes of the Museum Journal one can find reviews and news from statistical literature (by prof. Schnabel), regular reports on meteorological observations in Prague elaborated by means of elementary statistical methods (by prof. Hallaschka) and contributions to economical statistics of Bohemia (gubernial councillor K.A.Neumann, well-known natural historian The Earl Kaspar von Sternberg, prof. Schnabel and others were among the authors).

F.Palacký also published in the German Museum Journal two remarkable analysises of statistical datas on the population of Bohemia (Gradation der Bevölkerung Böhmens seit den letzen 60 Jahren, Statistisch–topographische Notizen über die Bevölkerung Böhmens im J.1830). Various people can be surprised by the author's knowledge of contemporary statistical literature, as well as of simple methods of political arithmetics (Palacký writes politische Rechenkunst).

It was a young graduate from the Prague and Vienna Universities, Karl Czoernig (1804-1889), the first one in our country who was concerned with the theoretical questions of political arithmetics. In 1831 he published in the German Museum Journal a commented translation of the essay from the English journal Edinburg Quaterly Review completed by his own extensive introduction.

Already Czoernig's attempt to define political arithmetics as a scientific branch was interesting: "Political arithmetics deals with those phenomena in the life of inhabitans of the state that can be expressed by quantitative relations..." He places political arithmetics among "state sciences" next to its "older sister" – statistics. Czoernig also discerned some limitations of the developement of political arithmetics as a scientific branch. This was discovered before its integration into modern statistical science which occured in the following decades. He points out not only an entire lack of exact and reliable data (if they were found out, they have often been concealed) but also an unclear conception of this branch, developing itself largely out of universities – due to amateurs and advisers of insurance companies. Czoerning, who used remarkable methods of data analysis already in his older work on Liberec (smoothing of time series, seasonal indexes), explains further the principle of the construction of life tables and the calculation of certain derivated characteristics (probability of surviving to specific age, expectation of life). He talks about methods of the construction of life tables used by Western political arithmeticians (W. Kersseboom, A. Déparcieux, R. Price, J. Milne). Czoernig's work indicates the possibility of inspecting the dependance among numerical variables – probably for the first time in our country – (he uses the research of a French statistician L.R.Villermé on the dependance of mortality in Parisien's districts on the ratio of the poor).

In 1841 the imperial court appointed K. Czoerning a principal of the statistical service. The fact that this post was reached by a man of a low birth (in contrast to other state offices) is worth remembering. He conducted Austrian statistics for almost 25 years (from 1852 as Freiherr von Czernhausen – according to his native village Černousy near Frýdlant).As a motto to his coat of arms he chose a device of F. Bacon Scientia est potentia).

Franz Alois Stelzig (1784-1856) was another representative of political arithmetics in the 20s and 30s. He was a physician general in the Prague Old Town and a graduate from the Prague Medical School. One of his most interesting works was a two-volume publication entitled Versuch einer medizinischen Topographie von Prag (1824). It was the first work on medical topography in our country. He also wrote an extensive essay published in the Museum Journal dedicated exclusively to the analysis of demographic datas (Resultate der Geburts – und Sterbeverhältnisse – 1830).

The publishing of Bohemian life tables (1800 – 1828) was a result of the essay. The author confronts them especially with life tables of J. P. Süssmilch, a famous Prussian political arithmetician of the 18th century. He also argues with Süssmilch's convinction of constancy of discovered demographic regularities in time and space ("divine order in transfomations of human mankind"). This refers to the consequences of medical progress (vaccination, various hygienical measures) and to different reproduction situations in various regions (city versus country) and in various groups of population, etc.

Stelzig explains as well a principle of various kinds of life insurance, life annuities, tontines, widow's and orphan's institutions, and he discusses the use of some life tables functions and other statistical methods for activity of these institutions.

Biometrics: Notes of a biometrician

Stanislav Komenda

The source of this paper was the experience of the statistician having left Charles University, Faculty of Mathematics and Physics at the very end of the 1950s – to enter the world of physicians. The reason had the name of statistical applications to carry out. No complex system of knowledge is to expect in the paper – rather short commentaries at the margin of the interaction between the biomedical environment and the statistical viewpoint.

Biometrics is characterized by its openness for the principles and methods to apply in data analysis. It means, on the other side, that the majority of biometrical methods can be met to support data analysis in the very differing areas of human intellectual activities – from archeology on the one-hand side to economic a engineering disciplines on the other side.

There is evidence that the research of the problems in very distant intellectual regions led to the discovery of the same statistical principle. Just that is the best witness of the unity of science. As an example the story of the Kaplan and Maier method of analysis of the censored survival data can be given: their common paper published almost 40 years ago became the most frequently cited statistical paper at all. The motivation to study the problem was the survival of cancer disease in one case while in the 2nd case it was the survival of vacuum tubes.

Biological and medical sciences inspired the discovery of numerous principles of statistics and statistical induction. Among the statistical

"Fathers Founders" were the personalities well known as the specialists in biology, medicine and agriculture. Francis Galton, who introduced quantitative methods into biology, Ronald Aylmer Fisher, researcher in eugenics and genetics, William Sealy Gosset, researcher and manager in the Dublin brewery of Guinness – to mention only three of many. Biological "grounds" can be identified also in the invention of the famous "normal" probability distribution. It was Adolphe-Lambert Quetelet who recognized the normal curve to be an excellent model to fit the frequency of occurrence of the stature size classes among the Scotch soldiers. An outstanding member in this gallery is the founder of genetics Gregor Mendel, the priest educated at the University in Olomouc, whose experiments carried out in Brno proved the surprisingly excellent fit with the binomial model.

An extraordinarily significant proof for the power and efficiency of the statistical reasoning is the story of doctor Ignaz Filip Semmelweis, assistant professor of the obstetrical clinic in Vienna. This story dates back to the middle of the nineteenth century, about fifty years before the inductive statistics started its knocking on the door of the empirical science.

Considering over the unexplained difference in the childbed fever maternal mortality (over 10% at the 1st clinic educating medical students against 3–4% at the midwives educating 2nd clinic) Semmelweis suspected the post-mortem dissections practised by his colleagues and students as the possible reason of the childbed fever. Due to hygienic measures recommended by him maternal mortality at the 1st clinic decreased significantly – which entitled his opinion – although the final reason, bacterial agens responsible for the disease, was discovered only 30 years after that by Pasteur and Rosenbach.

The name of biometrics itself was created in the way similar to many others: psychometrics, anthropometrics, econometrics, sociometrics, educometrics – and demonstrates the same semantic move as known from the geometry. In the contemporary interpretation the stress is not so much on the primary measurement, but more on the analytical, mainly statistical methods of evaluation of what the measurement yields.

Interaction between the specialist and the statistician is now under the significant influence of the computer. While 30 years ago this interaction of the statistician and his/her client could be summarized as the three-step process

(a) choice of an adequate experimental design and the model of statistical evaluation
(b) technology of the computations needed
(c) interpretation of the results derived,

nowadays this interaction reduces itself to the 1st and 3rd steps only. The burden of the intermediate computations moved from the statistician to his/her computer.

This question is in the close correspondence with the optimization of the statistical education of non-statisticians. Of course, there exists a coincidence in the opinion that teaching principles (like data reduction, inductive reasoning) has the priority to teaching particular techniques. On the other side, no theoretical principle can be implemented in the non-statistical mind without demonstrating of its application in the particular situations suitably selected.

Due to the openness of biometrics the paper like that hardly could be able to exhaust the topic anyway. The experience evaluated is doubtlessly conditioned by the specific circumstances under which the Department of Biometrics of the Faculty of Medicine, Palacký University in Olomouc, plays its part. Among the disciplines relying on the statistical support not only those of medicine take the part. So let us mention some less standard and more specific applications of this kind.

Psychometrics

Cooperation on the long-term program of testing of the effect of psychotic drugs on learning offered an opportunity to apply some formalized, probabilistic models of the paired-association learning and, particularly, to study the problem how to measure the information loss induced by replacing the primary data by the reducing statistical characteristics of learning. The primary data is necessary to reduce – for the purpose of easier manipulation – this reduction being mainly that of the data dimension. As a typical example the replacement of the sequence of correct and incorrect responses by the length of the 1st run of incorrect and the 1st run of correct responses can be mentioned. Fisher's measure of expected information on the distribution parameter as contained in the sample space of the primary data and that contained in the sample space of the values of these reducing characteristics proved to be the valid and useful index of the information lost.

Educometrics

Knowledge assessment is among the basic problems of the general didactics and of the common educational practice, as well. Without it no feedback between the teacher and the subject taught would be possible.

Stochastic model of the situation postulates the following three theoretical concepts of the knowledge. The 1st one is the concept of the "actual knowledge", non-accessible for the direct, immediate measurement. The actual knowledge is supposed to influence significantly the behaviour of the testee, as manifested through his/her test score. The object of assessment should be the actual knowledge – while the assessment rules refer to the knowledge manifested: the space of possible responses of the testee being decomposed in the disjoint and exhaustive way into the system of subsets such that each of them corresponds to just one level of the assessment scale.

Between the level of the subject's directly accessible responses (i.e. test scores) there is an intermediate level of his actual knowledge of the test items (it means not of the whole topic) as expressed by the number of test items the testee knows (or does not know). Also at this level no direct measurement is possible – due to the mechanism of guessing which enables the testee to reach the correct solution by chance. In every decision-making scheme the so-called operation characteristics of knowledge assessment can be considered (and computed, in case of a school-achievement test): for the actual knowledge given, to each grade (assessment level) there is the probability that the subject with this knowledge will reach just this grade. Thus the operation characteristics are the functions of the actual knowledge. In the model of the school-achievement test the actual knowledge plays the role of the parameter in the conditional distribution of the test score, i.e. of the subject's behaviour. Such an access makes it possible to quantify the notion of the "assessment injustice" and also to consider the optimization and efficiency control of the knowledge assessment by means of the size of the test, number of grades applied and by the modification of the assessment rule.

Anthropometrics

In physical anthropology a series of situations is known suitable for the statistical methods to be applied.

One of them is the case of the so-called growth standards where the "normal" growth and development of children and youth is to quantify. The utilization of statistics, distribution functions and percentiles to determine and specify standards (norms) is a particular chapter in the methodology of research.

Within the last decades statistics is invited by the archeologists to cooperate when the anthropological problems are studied on the osteological remains. Among the questions to solve the following ones are of significant importance:

(a) age determination of the subject to whom the bone remains belonged
(b) sex determination of such a subject
(c) stature (body height) reconstruction
(d) identification of the subject based on the presence or absence of certain specific features.

Similar problems are presented to the statistician also from the side of forensic medicine, in the cases of juridical expertises. The respective decision-making is supported by the results of the multidimensional statistical analyses (discriminant, cluster and factor analysis, multinomial correlation and regression). An important heuristic problem is the applicability of the analyses carried out on the recent bone material also for the decision-making in case of the remains older by centuries and millenniums. In the context of more numerous findings (Old Egyptian cemeteries, Old Slavonic cemeteries) sometimes the terms of paleostatistics and historical statistics are being used.

Another important region of anthropological standardization is ergonomics, which is the discipline studying the optimum fit of the living and working conditions to the parameters of human anatomy, physiology and mind. Working place design of the operator (airplane pilot cabin, car driver seat), but also the sizing system of the products of the industrial mass production (clothes, underwear, shoes, furniture, tools) are the problems of this kind. In contradistinction to the traditional craftsman production where the fit of the item to the customer was possible to be made repeatedly, the industrial mass production is producing its items for anonymous consumers – the members of certain population. What can be used to support the production design are the statistical parameters of this population: mean values, variances and correlations of body dimensions important from the viewpoint of the construction of the particular product. The solution of the problem is a multidimensional lattice situated in the space of the basic body dimensions. Besides this lattice of the type figures the system of regression functions is derived which enable to compute the values of other dimensions corresponding to the particular type figures.

Similar problem is to solve in the area of school hygiene when an optimum size system of chairs and desks is to determine – in dependence on the body height of school children. From the known variability of this body dimension the number of size alternatives is possible to compute for each classroom.

The regression integrating time can be find also in the classical criminological problem – when the body height of the subject is to forecast by means of his/her known foot length (as derived from the trace left on the ground).

An interesting part of biometrics is the so-called adequate body mass determination. The problem has its history dating back to the French physician Paul de Broca, whose proposal was to derive the adequate body mass (in kg of body weight) W from the known body height X (in cm) by means of the equation q

$$w(x) = x - 100,$$

where (x, w) are body height and body mass of the subject. When accepted, this equation is to interpret as the regression function (x – regressor, w – regressand) with the regression coefficient equal 1 and having the physical dimension kg.cm^{-1} and the constant member of the value 100 with the dimension kg.

Within the last decades the standard evaluation of the respective overweight or obesity is supported by the so-called BMI_2 (Body Mass

Index) W/X^2. Through this index body mass is related to the body surface. Since the higher efficiency is declared in comparison with the traditional proposal by Broca, a statistical verification has its reason.

The procedure we have introduced is based on the following principles:

(1) By the body mass adequacy its adequacy to the skeletal dimensions is meant

(2) The relationship between the skeletal dimensions and the body mass should be derived by means of the analysis of the relations actually existing in the reference populations in which these relations are to be a priori considered adequate; the participants of the Czechoslovak Spartakiade 1985 (public physical training exhibitions) were taken as these populations

(3) As the prediction efficiency of the methods by means of which the adequate body mass was to forecast – the relative decrease of the body mass variance was taken, in the conditional probability distribution of it, in relation to the unconditional, initial body mass variance.

Statistical properties of the following indices were investigated:

$$Q = W/X \qquad G = (W/X^2)10^3 \qquad R = (W/X^3)10^6$$

$$D = W/Y \qquad E = (W/Y^2)4\pi 10^3 \qquad F = (W/Y^3)6\pi^2 10^3$$

$$C = (W/XY)103 \qquad K = (W/XY^2)4\pi 10^3$$

where W is the body mass in kg, X and Y are stature and chest circumference both measured in cm. Physical dimension of these indices is that of the density human body reaches in the respective virtual skeletal space: on the line, inside the square and inside the volume of the cube determined by the stature X, on the circle circumference, inside this circle and inside the sphere determined by the chest circumference Y, and on the surface and inside the volume of the cylinder of the height X and the circumference Y.

The forecasting formulas through which the expected body mass $w(x), w(y)$ and $w(x, y)$ is to derive in dependence on the respective skeletal space (x, y or both x and y given) were specified by the linear regression where always just one parameter was to estimate. The analyses carried out concluded that the index G now recommended generally by the specialists in the nutrition, hypertension, diabetes, cardiovascular disorders, arthritis and vertebrogenous disorders actually reaches somewhat higher efficiency (in the adult populations) when compared with the index Q. Nevertheless, this efficiency is remaining low. Moreover, completing of the prediction formulas by other skeletal dimensions is able to improve prediction efficiency twice to three times.

There is an objection against the application of the variable Y in the prediction mechanisms – due to that Y is not a purely skeletal dimension. As a response to this challenge the indices L and M were introduced as an analogy to the indices C and K, where the chest circumference Y was substituted by the transversal and sagittal diameters T and S of the chest,

$$L = (W/X\sqrt{TS}\pi)10^3 \qquad\qquad M = (W/XTS\pi)4.10^3$$

As we have found, an introduction of chest diameters T and S into the prediction formulas L and M led to somewhat lower efficiency in comparison with the indices C and K. Nevertheless, this efficiency is always twice over that reached by the formulas based on the known value of X only.

Author's address:

On the coefficient of determination: simple but ...

Jan Ámos Víšek

ABSTRACT An example of possible misleading role of the basic characteristics of the classical *LS* regression analysis is given. Another example using high breakdown point estimators demonstrates that in the case of contaminated data various estimators may give considerably different estimates. Consequently, a simple solution, routinely leading to the true model, need not exist. A proposal of one possibility how to cope with the situation is given and another is refered to.

Key words: Classical and robust regression analysis, studentized estimates of regression coefficients.

AMS classification: Primary 62J20; secondary 62J05;

Introduction

There are so much papers devoted to the behavior of the coefficient of determination (see e. g. [9],[10],[21] *&* and the all monographies devoted to regression analysis, e. g. [3] or [20]) that to try to write something more is not only somewhat superfluous but just an outrageous impertinence. However, I secretly hope that some readers may generously depress their sorrow for the wasted paper and they will read the paper up to the end.

The coefficient of determination is, together with studentized estimates of regression coefficients and with Fisher-Snedecor F-statistic, one of the basic characteristics of classical regression analysis. It is easy to see why. It is posssible to read nice (and relatively simple) lectures on these characteristics, it is a joy to derive their distributions and after all, the topic may be ellegantly utilized for the examination of students. That is why these characteristics have found their fix place in many monographies, PC-libraries and temporarily also in the minds of some students.

These reasons probably caused that people have constructed the tables for the corresponding distributions and they established (by the way, really tight, see [1]) approximations to them, which may be used on

PC (if, by an unexplainable coincidence of circumstancies, it happened that no error penetrated into the implementation of the corresponding algorithms). The frequency of the references on these distributions (t and F) can be beaten only by the frequency of references on the normal law. The latter, although empirically indistinguishable from t (even for rather low degrees of freedom and relatively large samples) allows to derive even more interesting results which are assumed to be a treasure of Her Majesty Statistics. Perhaps it is (also) due to the fact that the warning of Sir Ronald Aylmer Fisher (see [5]), pointing out that the efficiency of these results may considerably decrease (even down to zero) when data are governed by the t-distribution instead of by normal one, has stealthily disappeared from the textbooks.

The above claimed simplicty of some statistics may however appear to be equally betraying as peats, see [4]. In the Table 1 the data simulated according to the model

$$(1)\quad Y_i = 10 + 11 \cdot X_{i1} + 12 \cdot X_{i2} + ... + 20 \cdot X_{i10} + \varepsilon_i, \quad i = 1, 2, ..., 18$$

are gathered. The values of the regressors X_i's were generated as independent realization of uniform random variable from $[-2, 2]^{10}$ with one coordinate shifted about $+10$ or -10 and the random fluctuations ε's have the distribution of the normal random variable with mean zero and variance $\sigma^2 = 0.04$. It means that the fluctuations are of so small magnitude that data are nearly "deterministically" goverened by the linear model, so that the estimation of the regression coefficients is to be very easy. Finally two points, namely points 2 and 12 were contaminated. One may immediately verify that some coordinates of these points were changed and it was done in a such way that it represents 6 false hits among 1 000 hits which were necessary for the input of data (e. g. an error in the decimal point). Maybe that at this moment some reader may put a question why just the data of this character was used (those who are not interested in the answer on this question let go directly to the label ♡).

In 1979 R. A. Maronna, O. H. Bustos and V. J. Yohai [11] showed that there is among the solutions of the equations

$$(2)\qquad \sum_{i=1}^{n} \psi \left(Y_i - \sum_{j=1}^{p} X_{ij}\, \beta_j \right) X_{ik}\, w(X_i) = 0, \qquad k = 1, 2, \ldots, p$$

(where $X_i = (X_{i1}, X_{i2}, ..., X_{ip})^T$) at least one which has the breakdown point smaller than $\frac{1}{p}$. In the other words, if the system (2) has unique solution, then (somewhat paradoxically) corresponding M-estimate has for larger dimension the breakdown point rather small (and notice that we speak about weighted M-estimators which obtained, as a gift into the cradle from the Fates (Hampel, Krasker, Welsh etc., see e. g. [6]) a possibility to "depress" or "weight down" those data, which have in the factor space too "individualistic" positions). Recalling this result, we may ask what is a (geometric) reason for this strange fact ? Earlier than giving an explanation, let us remind that the breakdown point is the smallest number of points (divided by the number of all observations) which is necessary to change (in an arbitrary way) to cause explosion (or implosion) of the estimate.

Now let us consider a cloud of some regression data, let us say near the origine. Our counterplayer (nature) in order to damage the value of the estimate may change the first coordinate of the m_1 points, making from them leverage points. Due to large values of X_{i1}'s (for corresponding i's) the influence of these m_1 points on the solution of the first equation in the system (2), i. e. on the equation

$$\sum_{i=1}^{n} \psi \left(Y_i - \sum_{j=1}^{p} X_{ij}\, \beta_j \right) X_{i1}\, w(X_i) = 0,$$

will be larger than the influence of other points. Since we do not know whether these leverage points are good or whether they are simultaneously outliers, we give them small weights. Now the counterplaying nature will increase the (absolute) values of the second coordinate of some m_2 points, so that we will give them also small weights, etc. Finally we will be forced to give small weights to $\sum_{k=1}^{p} m_k$ points. However this sum cannot be larger than $n - p - 1$, to have at least $p + 1$ "good" points for determination of the "true" model (if it happens by unbelievable stroke of good luck that they will have large weights). It implies that at least for one k we have $\frac{m_k}{n} \leq \frac{1}{p}$.

♡ Let us return to our data, they are as follows:

Table 1

Data governed by model (1)

case	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	Y
1	9.39	-0.74	0.15	-0.58	1.39	1.99	0.28	-0.75	1.12	0.08	165.2
2	-1.22	0.55	11.83	0.37	-0.13	0.08	31.58	0.21	1.29	-11.7	9.4
3	0.62	-10.2	1.44	0.04	0.93	-0.70	0.52	0.52	0.75	0.46	-42.2
4	-1.22	0.55	1.83	0.37	-0.13	0.08	11.58	0.21	1.29	1.66	289.4
5	-1.35	-1.37	-0.71	0.55	-1.42	0.22	0.39	10.12	2.00	0.69	200.4
6	-1.21	0.78	0.83	1.90	-1.43	-0.81	-0.08	-0.56	-11.4	1.95	-179.0
7	-1.35	-1.37	-0.71	10.55	-1.42	0.22	0.39	0.12	2.00	0.69	160.4
8	0.07	11.95	0.73	-0.55	-0.54	-0.87	-1.25	-1.46	-1.52	-0.45	49.0
9	0.03	-1.36	-1.39	-1.52	-11.1	-0.36	-0.45	1.05	1.10	1.49	-156.1
10	0.07	1.95	0.73	-0.55	-0.54	9.13	-1.25	-1.46	-1.52	-0.45	88.0
11	10.03	-1.36	-1.39	-1.52	-1.15	-0.36	-0.45	1.05	1.10	1.49	103.9
12	0.62	-0.18	1.44	0.04	0.93	-10.7	0.52	0.52	10.75	10.46	181.7
13	1.23	-1.04	0.65	-11.1	1.77	1.23	-1.08	-1.57	1.61	0.61	-93.7
14	1.23	-1.04	0.65	-1.10	1.77	1.23	-11.1	-1.57	1.61	0.61	-123.7
15	-0.73	-0.16	-1.57	0.92	-0.99	-1.90	0.45	-11.6	-1.18	-0.93	-294.4
16	1.10	-0.40	-8.29	-0.65	0.92	-1.40	-1.70	-0.30	-0.96	1.44	-131.5
17	-0.98	-1.14	1.23	-0.34	-0.46	1.30	-1.12	1.01	-1.42	11.68	216.3
18	0.61	1.59	-0.19	-0.52	0.48	-1.97	0.96	0.47	1.67	-11.2	-165.7

Diagonal elements of the projection (hat) matrix are

Table 2

Diagonal elements of the projection matrix

case	1	2	3	4	5	6	7	8	9
diag.	0.49	0.85	0.51	0.27	0.55	0.83	0.60	0.64	0.94
case	10	11	12	13	14	15	16	17	18
diag.	0.57	0.55	0.80	0.59	0.45	0.65	0.74	0.44	0.54

Since the value of the ninth element is 0.938 a suspicion may appear that this point is not O. K.. Such suspicion however disappears when we find in some textbook that a recommended critical value for the diagonal elements of hat matrix is $\frac{2p}{n} = 1.22$, see e. g. [2], [3] or [20]. On the other hand, only exceptionally one can find in the textbooks an explicit statement that the diagonal element of the hat matrix may attain a value out of interval $(0, 1)$ only in the case when the evaluation of it contains an error. That is why it is larger than 1.22 only rarely. May be that someone can object at this moment that in some monographies we may find a recommendation that te diagonal elements of hat matrix should be smaller than 0.2, see e. g. [7]. This requirement implicitly includes an endeavour for a balance of number of observations and of the dimension of problem. Since we wanted to keep the size of paper in a reasonable limits we could not consider data containing too much points, we have to omit this ask. A more complete discusion devoted to an acceptable values of diagonal elements of hat matrix can be found in [3].

Suma sumarum: We have not too large data and we would not like to waste them, and so we will give a pardon to the value 0.938 . Moreover, we are in a God-like position, so that we know that the point 9 is not a contamination (an objection that we have too small number of observations will be discussed at the end of paper).

After an application of the least squares we obtain the estimates of the regression coefficients and the correponding P-values.

Table 3

LS-estimates and corresponding P-values

	β_0	β_1	β_2	β_3	β_4
estimates	-5.82	10.67	11.56	1.916	14.39
P-values	0.782	0.119	0.047	0.814	0.025

	β_5	β_6	β_7	β_8	β_9	β_{10}
estimates	14.11	20.28	9.240	18.71	16.29	21.26
P-values	0.075	0.013	0.038	0.007	0.022	0.001

Coefficient of determination attained the value .919, the parametr of scale was estimated as $\hat{\sigma} = 75.3$ and the sum of squares is 39 674.8.

We may try to delete those regressors which are indicated to be insignificant on the level of 5% (which means: intercept, X_1, X_3 and X_5) and then we recalculate the hat matrix. The diagonal elements now look like this:

Table 4

Diagonal elements of the hat matrix for the reduced data

case	1	2	3	4	5	6	7	8	9
diag.	0.05	0.80	0.41	0.19	0.41	0.68	0.52	0.55	0.02

case	10	11	12	13	14	15	16	17	18
diag.	0.50	0.02	0.76	0.52	0.16	0.57	0.03	0.39	0.40

The value of $\frac{2p}{n}$ is now .777 and it hints that the point 2 could be "leverage point". Deleting this point an recalculating once again the estimates we obtain

Table 5

Diagonal elements of the hat matrix for data after deletion of point 2

case	1	3	4	5	6	7	8	9
diag.	0.05	0.41	0.55	0.41	0.69	0.52	0.56	0.02

case	10	11	12	13	14	15	16	17	18
diag.	0.50	0.02	0.77	0.52	0.49	0.57	0.04	0.43	0.43

and $\frac{2p}{n} = 0.824$. Further

Table 6

LS-estimates of the coefficiennts and corresponding P-values

	β_2	β_4	β_6	β_7	β_8	β_9	β_{10}
estimate	10.43	10.34	22.42	16.84	16.52	20.29	18.15
P-value	0.069	0.079	0.006	0.008	0.011	0.006	0.002

From Table 6 it follows that perhaps regressors X_2 and X_4 are still insignificant. Deleting them we finally arrive to

Table 7

LS-estimates of coefficients and corresponding P-values

	β_6	β_7	β_8	β_9	β_{10}
estimate	19.93	18.09	16.15	17.00	16.97
P-value	0.026	0.012	0.027	0.029	0.007

So we have finally obtained a model in which all regressors are significant. The coefficient of determination a little bit decreased to 0.736 but nevertheless it is above traditional magic boundary of 60%. So we may be satisfied. The only shortage of the analysis is that it produced a completely false model.

On the other hand, applying least trimmed square estimator

$$\hat{\beta}^{\text{LTS}} = \underset{\beta \in R^{11}}{\operatorname{argmin}} \sum_{i=1}^{15} r^2_{(i:18)}(\beta)$$

where $r^2_{(i:18)}(\beta)$ is the i-th order statistics among $r^2_i(\beta) = [Y_i - \beta_0 - - \sum_{i=1}^{10} X_{ij}\beta_j]^2$, $i = 1, 2, ..., 18$, we obtain the estimates of coefficients together with P-values

Table 8

LTS-estimates of coefficients and corresponding P-values

	β_0	β_1	β_2	β_3	β_4
estimate	10.13	11.01	12.03	12.98	14.02
P-value	0.000	0.000	0.000	0.000	0.000

	β_5	β_6	β_7	β_8	β_9	β_{10}
estimate	15.01	15.93	17.01	18.03	18.97	19.99
P-value	0.000	0.000	0.000	0.000	0.000	0.000

The estimate of scale is $\hat{\sigma} = .0361$.

It seems that as a conclusion we could recommend:
Let us use the procedures with high breakdown point (at least for the first diagnostics), may be in the second step accompanied by some M-estimator, and let us refuse least squares! **Unfortunately, this recommendation can be misleading as well as the belief in the almighty of market economy.** It is not difficult to demonstrate that the high breakdown point procedures may give also result which may be strange, e. g. we may easy find data for which various high breakdown point estimators give rather different results. Let us give one example. To create a little more complete idea about a possible diversity of estimates we shall include into the following table results of several methods. To be sure that the paper is understandable without requiring too large a priori knowledge, let us recall the definitions of them. Let us put

(4) $$r_i(\beta) = Y_i - \sum_{j=1}^{p} X_{ij}\,\beta_j, \quad i = 1,2,...,n \qquad h = \left[\frac{n}{2}\right] + \left[\frac{p+1}{2}\right]$$

and let $r^2_{(i:n)}(\beta)$ be the i-th order statistics among $r^2_i(\beta)$, $i = 1,2,...,n$. Further, let us recall that (with h given by (4))

(14)
$$\hat{\beta}^{\mathrm{LS}} = \underset{\beta\in R^p}{\operatorname{argmin}} \sum_{i=1}^{n} r_i^2(\beta), \qquad \hat{\beta}^{\mathrm{LMS}} = \underset{\beta\in R^p}{\operatorname{argmin}}\ r^2_{(h:n)}(\beta),$$
$$\hat{\beta}^{\mathrm{LTS}} = \underset{\beta\in R^p}{\operatorname{argmin}} \sum_{i=1}^{h} r^2_{(i:n)}(\beta), \qquad \hat{\beta}^{(\rho_k)} = \underset{\beta\in R^p}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_k(r_i(\beta))$$

with ρ_1 – Huber's function (with $\psi_1(t) = \frac{\mathrm{d}\rho_1(t)}{\mathrm{d}t} = t$ for $|t| < c$ and $\psi_1(t) = c \cdot \operatorname{sign} t$ otherwise) and ρ_2 – Hampel's function (with $\psi_2(t) = \psi_1(t)$ for $|t| < 1.2c$, $\psi_2(t) = [c - \frac{5}{9}(t - 1.2c)] \cdot \operatorname{sign} t$ for $1.2c < |t| < 3c$ and zero otherwise). For both Huber's and Hampel's functions c was equal to 1.2. Finally, let

$$\hat{\beta}^{\mathrm{L}_1} = \underset{\beta\in R^p}{\operatorname{argmin}} \sum_{i=1}^{n} |r_i(\beta)| \qquad \text{and} \qquad \hat{\beta}^{\mathrm{TLS}} = \underset{\beta\in R^p}{\operatorname{argmin}} \sum_{i\in\mathcal{I}_\alpha} r_i^2(\beta)$$

where $\mathcal{I}_\alpha$ is the index-set of points obtained by the symmetric trimming according to regression α-quantiles of Koenker and Bassett [8] (value of α was 0.2 to trim away the same number of points as was trimmed by LMS and LTS, i. e. $n - h = 21$).

Example 1. : US Crime Data (47 cases, [12]).

These data are for the crime in U.S.A., and they concern 47 states. The goal of the investigation was to find how the crime rate (number of offence known to the police per 10^6 population) in 1960 depended on age distribution, on the fact whether the offence was accomplished in southern state, on the educational level, on the police expenditure, on the labour force participation rate, on ratio of males in population, on the total number of population in the state, on the ratio of whites in population, on the unemployment rate, on the median family wealth and on the income inequality. The regressors were selected in the following way: The variables which appeared in the complete LS and in the complete LTS analysis as "highly" insignificant have been deleted (P-value over 0.2). Then LS and LTS analyses were repeated and the variables which were still significant were taken into account (of course, we do not want to claim that it is the only possibility how to choose). So that the variables used in the example are: Age distribution (the number of males aged $14 - 24$ per 10^3 of total state population), Educational level (mean number of years of schooling of the population 25 years old and over), Police expenditure (the per capite expenditure on police protection by state and local government in 1960) and Income inequality (the number of families per 10^3 earnings below one half of the median income). Their P-values for LS model are $0.0^4 13$, 0.032714, $0.0^3 773$, 0.0^6 and $0.0^3 137$ (the "power" means number of zeros before the first significant digit), and for the LTS 0.0^6, $0.0^3 353$, 0.0^6, 0.0^6 and 0.0^6.

Table 9
US Crime Data

Method	LS	LMS	LTS	TLS	L_1	Huber	Hampel
intercept	-424.922	-234.216	-424.369	375.158	450.269	406.826	403.081
Age	0.760	0.472	0.633	0.541	0.426	0.476	0.477
Education	1.660	0.294	2.099	0.337	-0.018	0.241	0.281
Police	1.298	1.675	0.817	-1.875	-2.096	-2.073	-2.119
Income	0.641	0.464	0.665	-0.912	-0.795	-0.819	-0.781

First of all, we should say that when clasifying differences among the estimates we should not take into account the differences among the estimates of intercept because these differences may be large due to small differences among the estimates of slopes together with position of data in the factor space (imagine data which are far from origine). The differences between LTS- and LMS-estimates of coefficients, expressed by their ratio, are 1.34, 7.14, 0.49 and 1.43, respectively. The coefficient of determination in LS analysis is 70.0% and in LTS even 91.4%.

And it is not difficult to find an example of one dimensional data for which LMS- and LTS-estimates are orthogonal each to other (see e. g. [13]) or [18]).

So let us recapitulate seemingly terrible situation:
"Failure" of the high breakdown point estimators as well as of the classic ones and a misleading behaviour of simple characteristics, as coefficient of determination and studentized estimates of regression coefficients, can be interpreted from a little bit more general point of view (see [16]). Statisticians, but not only them, consider some basic assumptions (if you want principles or axioms) which are usually supported by convinsing heuristic arguments, and reasonable requirements (e. g. maximization of likelihoods, minimization of sum of squares, consistency, maximization of the value of breakdown point, minimization of maximal bias (maximum taken over some family of distributions, efficiency, minimization of some loss, etc.) and they are permanently inventing new ones (see e. g. [14], otherwise they cannot obtain grants). These heuristic requirements, reformulated into some mathematical criteria, are naturally directly resublimated into some plausible properties of resulting procedures (after all, the plausibility or quality of results is "measured" in fact according to a criterium which is based on the same ideas which were included into principles and axioms, so that the success is a priori ensured – if we leave aside a painful fact that we need to cope with some technical difficulties when looking for a proof). However a hope that the heuristic arguments which we gathered at the start of the research guarantee a "reasonability" of the resulting procedures, especially for finite samples of data, is dim.

So it seems that very near to some "pragmatic truth" is a statement:
We cannot blindly rely on the heuristics (which are somewhere in the

background of the methods) and apply routinely methods or characteristics steming from them (as e. g. minimal biased estimators, see [15] *and* [16]*, or the coefficient of determination). Besides the other reasons for it is the fact that the good behaviour of such methods (or reliability of the information contained in characteristics) is frequently connected with sufficiency and efficiency of some statistics. However, both these propeties are intimately associated with distributions (e. g. normal one) which we cannot empirically distinguish (at least for small and modest sample sizes) from those (e. g. student one) for which the corresponding statistics are not always sufficient and are usually considerably deficient (see* [7] *or* [6]*).*

So quality or acceptability of an estimate of regression model should be probably judged by more complex criteria, in fact by the all available ones, especially by a "global look" on residuals, e. g. by the normal plot. However, the normal plot, although being much more sensitive to some "irregularities" among data than coefficient of determination, brings information which cannot substitute information offered by coefficient of determination and hence should be used together with coefficient of determination. On the other hand, the normal plot as a diagnostic tool has (at least) two disadvantages:

- Firstly, it is applicable only for normally distributed residuals.
- Secondly, it does not offer numerically clasifiable test.

The first shortage can be removed easily by utilizations of approprite quantiles of some other distribution. To cope with the second one is somewhat more difficult because there are some test of good fit e. g. for normality of residuals but they are *exteremely* useless. E. g. anybody who sometimes tried to prepare some simulated data knows that it is not a rare case when data which passed through the test for normality, passed through the same test after having been contaminated much easier. In other words, the tests of good fit are nice topics to be read on the lecture but for practicl use they are nearly completely useless. So it seems better, however much more complicated, to apply some more complex criteria, e. g. to compare estimates of density of residuals in two disjoint parts of factor space, see [13].

Two short remarks at the end. We should admit that two objections may appear. At first, one may say that the first example contained to small number of point. Secondly, someone else may propose to use the diagnostic tool based on the formula for the change of the LS or M-estimate when one point is excluded from data. The formula for LS reads

$$(5) \qquad \hat{\beta}^{LS,n} - \hat{\beta}^{LS,n-1,\ell} = \left[(X^{\ell})^T X^{\ell}\right]^{-1} X_{\ell}(Y_{\ell} - X_{\ell}^T \hat{\beta}^{LS,n})$$

where $\hat{\beta}^{LS,n-1,\ell}$ denotes LS estimate for data after deletion of the ℓ-th observation, X_{ℓ} denotes the ℓ-th row of the design matrix X and $X^{(\ell)}$ is the design matrix for the reduced data, see e. g. [3] or [20]. For the M-estimators the formula is similar however it would require an introduction of some additional notation, so we only refer to [18]. It is true that using (5) and looking for the point, deletion of which causes the largest change of $\|\hat{\beta}^{LS,n} - \hat{\beta}^{LS,n-1,\ell}\|$ we would find successively points 2 and 12, and finally the estimate of true model.

Both objections can be "annulled" by simulating more data of the same type as above (i. e. also with the same level of contamination). It is necessary to have such number of points (in our case about 72) that we would have somewhat more than $p+1$ "contaminated" points. Since the contaminated points (contaminated in the same way as the points 2 and 12 above) are (very) modest leverage points LS estimate will take into account just these points and the result is similar as above. But in this case the formula (5) does not help. What would help? An analoguous formula but for

$$\max_{I_k \subset \{1,2,\ldots,n\}} \|\hat{\beta}^{LS,n} - \hat{\beta}^{LS,I_k}\|$$

where $I_k = \{i_1, i_2, ..., i_k\}$ with $1 \leq i_1 < i_2 < ... < i_k \leq n$ (an asymptotic representation even for M-estimators can be found in [19]). Unfortunately just described data are not appropriate for presentation in the paper (because of space necessary for them).

As a final conclusion let us say a "serious" word. All of readers probably felt that the text of the paper was somewhat but we hope that not intolerably overstated. We hope that for those of us who are more interested in applications it will lead to thinking the classical tools of regression analysis once again. For those of us who are more attracted by theoretical work it will inspire a feeling that in the flood of new, especially estimating methods it would be worthwhile to create also diversified, however relatively simply applicable diagnostic tools with the good properties working already for finite sample sizes.

Reference

[1] Abramowitz, M., Stegun, I. A. (1964): *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables.* Dover Publications.

[2] Antoch, J., Vorlíčková, D.: *Vybrané metody statistické analýzy dat. (Selected methods of statistical analysis of data.)* Academia, Praha, 1992.

[3] Chatterjee, S., Hadi, A. S. (1988): *Sensitivity Analysis in Linear Regression.* New York: J. Wiley & Sons.

[4] Doyle, A. C. (1892): *The Hound of the Baskerviles.* London, Logman 1990.

[5] Fisher, R. A. (1922): On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London Ser. A 222, pp. 309–368.*

[6] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A. (1986): *Robust Statistics – The Approach Based on Influence Functions.* New York: J.Wiley & Sons.

[7] Huber, P.J.(1981): *Robust Statistics.* New York: J.Wiley & Sons.

[8] Koenker,R., Bassett, G. (1978): Regression quantiles. *Econometrica, 46, 33-50.*

[9] Kozák, J. (1993): *Znovu ke koeficientu determinace. (On the coefficient of determination once again.)* Informační bulletin České statistické společnosti (Information bulletin of the Czech statistical society), srpen (august) 1993, 12–16.

[10] Kozák, J. (1994): *STATGRAPHICS a koeficient determinace (STATGRAPHICS and coefficient of determination).* Informační bulletin České statistické společnosti (Information bulletin of the Czech statistical society), duben (april) 1994, 16–20.

[11] Maronna, R.A., Bustos, O. H., Yohai, V. J. (1979): Bias- and efficiency-robustness of general M-estimators for regression with random carriers. In *Smoothing Techniques for Curve Estimation. Eds. T Gasser and M. Rosenblatt, New York: Springer-Verlag, 91 – 116.*

[12] Vandaele, W. (1978): Participation in illegitimate activities: Erlich revisted. In *Deterrence and incapacitation.* Eds. Blumstein,A., Cohen, J., Nagin, D., Washington. D. C.: National Academy of Sciences, pp. 270 – 335.

[13] Rubio, A. M., Víšek, J. Á. (1994): Diagnostics of regression model: Test of goodness of fit, *Transactions of the Fifth Prague Symposium on Asymptotic Statistics, eds. M. Hušková & P. Mandl, Springer Verlag, 423 – 432.*

[14] Simpson, D. G., Ruppert, D., Carroll, R. J. (1992): On one-step GM estimates and stability of inferences in linear regression. *Journal of American Statistical Association, vol. 87, 439 – 450.*

[15] Víšek, J. Á. (1994 a): High robustness and an illusion of truth. *Transactions of ROBUST'94, JČMF (Union of Czech Mathematicians), eds. J. Antoch & G. Dohnal, ISBN 80-7015-492-6, 172-185.*

[16] Víšek, J. Á. (1994 b): On the heuristics of statistical results. To appear in *Proceedings of 'PROBASTAT'94'.*

[17] Víšek, J. Á. (1995 a): Sensitivity analysis of M-estimates. To appear in *Annals of the Institute of Statistical Mathematics, Tokyo.*

[18] Víšek, J. Á. (1995 b): On the diversity of estimates. Submitted to *Computational Statistics and Data Analysis.*

[19] Víšek, J. Á. (1996): Data subsample influence in M-estimation of the non-linear regression model. *Preprint.*

[20] Zvára, K. (1989): *Regresní analýza (Regression Analysis).* Praha: Academia.

[21] Zvára, K. (1993): *Který model je ten pravý. (Which model is that true.)* Informační bulletin České statistické společnosti (Information bulletin of the Czech statistical society), květen (may) 1993, 8–11.

Author's address: Pod vodárenskou věží 4, 182 08 Prague, e-mail: visek@@utia.cas.cz

Randomized response

Martin Anděl

Abstract The aim of the randomized response methods is to decrease the percentage of untruthful respondent's replies in sample surveys. Two basic methods of randomized response model and their applications in our republic as well as abroad are mentioned in the paper.

1. Introduction

Surveys dealing with sensitive or highly personal matters disturb privacy of respondents. Results of this surveys are influenced by high number of refusing replies and also certain number of untruthful replies. Respondents answering untruthfully are ashamed of the true answer or they are aware of a persecution. This difficulties are not eliminated even by assuring the respondents about anonymity of interviewing and using data only for statistical interpretation. An approach, which guarantees anonymity of respondents also against interviewer, is mentioned in the following chapter. The approach deals with questions replying only yes or no.

2. Warner's model

Let a characteristic A dividing population of people into two mutually exclusive groups be given. The group of people with attribute A we denote $\mathcal{A}$. The characteristic A may assign for example respondent distil illegally alcohol at home or respondent with homosexual orientation. We want to estimate ratio p_A of people who belong to the group $\mathcal{A}$. The usual approach is following one. We randomly select n respondents from the population and query them:

"Are you a member of the group $\mathcal{A}$?"

The total of positive replies is denoted n_A. The estimate of ratio p_A is given by

$$\hat{p}_A = \frac{n_A}{n}. \tag{2.1}$$

The variance of this estimate is

$$var(\hat{p}_A) = \frac{p_A(1-p_A)}{n}. \tag{2.2}$$

Respondent is confronted with two statements in Warner's randomized response model (see [1]):

1. *I am a member of the group $\mathcal{A}$.*
2. *I am not a member of the group $\mathcal{A}$.*

Respondent with the aid of a randomization trial, whose result is unknown for the interviewer, choose the statement to which he answers. Thus it is not possible for the interviewer to know whether respondent's "yes" or "no" confirms or, on the contrary, contests his pertinence to group $\mathcal{A}$. We introduce following notation:

$p_A =$ real proportion of people who belong to the group $\mathcal{A}$,
$P =$ probability that the first statement is randomly chosen,
$n =$ size of sample,
$m =$ total number of positive answers for both statements,
$\lambda =$ probability of positive answer.

Probability λ is given by formula

$$\lambda = Pp_A + (1-P)(1-p_A). \tag{2.3}$$

Inserting the estimate $\hat{\lambda} = m/n$ into (2.3) for λ we receive an estimate for the proportion p_A

$$(\hat{p}_A)_W = \frac{1}{2P-1}\left(P - 1 + \frac{m}{n}\right), \qquad P \neq \frac{1}{2}. \tag{2.4}$$

Variance of this estimate is

$$var((\hat{p}_A)_W) = \frac{p_A(1-p_A)}{n} + \frac{P(1-P)}{n(2P-1)^2}. \tag{2.5}$$

The first therm in (2.5) is the common binomial variance connected with direct question. The second therm in (2.5) is the price that we pay for uncertainty connected with randomized reply. Variance in (2.5) decreases with increasing distance of P from 0.5. But if we take P too near to 0 or 1, then the respondent will be probably unwilling to reply truthfully. It seems that the choice of parameter P between 0.1 and 0.3 or between 0.7 and 0.9, is a good compromise between minimization of variance of estimate and respondent's willingness to give truthful answer.

3. Do you distil alcohol at home ?

Warner's randomized response model was successfully applied by Ing. Josef Machek CSc. and Huver Fernández Rodríguez in 80*th* years in Cuba (see [3]). The aim of the survey was to estimate the percentage of households illegally distilling alcohol. Survey was done in the three areas. A method for direct response was used also for comparison. The results are displayed in the following table:

area	direct response	randomized response	
	estimate	size of sample	estimate
1	9%	384	60%
2	13%	380	30%
3	23%	576	40%

4. Two unrelated questions

This method is a modification of Warner's approach. Following choice of statements is introduced for respondent:

1. *"I am a member of the group $\mathcal{A}$."*
2. *"I am a member of a group $\mathcal{Y}$."*

The group $\mathcal{Y}$ is a group of people characterized by an undefective social attribute. Here we shall assume that we know the probability p_Y of event that respondent belongs to the group $\mathcal{Y}$. For probability λ of positive answer we have

$$\lambda = Pp_A + (1-P)p_Y . \tag{4.1}$$

Inserting the estimate $\hat{\lambda} = m/n$ in (4.1) for λ we get estimate for ratio p_A

$$(\hat{p}_A)_Y = \frac{1}{P}\left(\frac{m}{n} - p_Y(1-P)\right) . \tag{4.2}$$

The variance of this estimate is

$$var((\hat{p}_A)_Y) = \frac{1}{nP^2}\lambda\,(1-\lambda) . \tag{4.3}$$

It is convenient to choose the values of parameters as follows. Parameter P should be near to one, for example between 0.7 a 0.9. Choice of parameter p_Y depends on the value of the unknown probability p_A. If we expect that the probability p_A is less than 0.5, we try to choose p_Y near to 0. If we expect that the probability p_A is larger than 0.5, we try to put p_Y near to 1. In case that the probability $p_A = 0.5$, we try to choose p_Y near to 0 or near to 1. When choosing p_Y "near to 0" or "near to 1" it is necessary to consider that variance in (4.3) decreases with parameter p_Y tending to 0 or to 1, but when p_Y is very near to 0 or to 1 respondent's willingness to answer truthfully is low.

5. The estimate of the ratio of homosexuals in the population of Czech Republic

In the survey of company GfK Praha 889 respondents older than 18 years replied to the following couple of questions:

1. *"Are you homosexually oriented?"*
2. *"Did you read newspaper MLADÁ FRONTA DNES yesterday (if Monday, then on the day before yesterday)?"*

Interviewers were equipped by six cards numbered 1, 2, 3, 4, 5, and 6. After shuffling the cards they let the respondent draw randomly one card. If the drawn card had number 2, than respondent answered the second question. In all other cases respondent answered the first one. The frequencies of replies are displayed in the following table:

answer	frequency
yes	45
no	730
no answer	114

Respondents who refused to answer are left out for further computing. Values of parameters are following:

$$P = \frac{5}{6}, \quad n = 775, \quad m = 45.$$

From print media research we took the value

$$p_Y = 0.14$$

(see [2]). Substituting in (4.2) we have estimate of the ratio of homosexuals in the population in Czech republic

$$(\hat{p}_A)_Y \doteq 0.042.$$

If we replace λ by $\hat{\lambda}$ in (4.3)

$$\hat{\lambda} = P(\hat{p}_A)_Y + (1 - P)p_Y,$$

we have estimate for variance

$$var((\hat{p}_A)_Y) \doteq 0.0001.$$

Published results of print media research are weighted by actual demographic values given by the Czech Statistical Institute. The value $p_Y = 0.14$ is unweighted. Weighted value p_Y is 0.16. The estimate $(\hat{p}_A)_Y$ computed with this weighted value is 0.038.

Final value agrees to common meaning, that the proportion homosexual in population is about 4%.

Company DEMA that made a survey about sexual behavior of people in Czech Republic for Prague's Sexual Institute, published in the newspaper Mladá fronta DNES of 9. 3. 1994 that the ratio of homosexually orientated men and women in our population is less than 2%.

Reference

[1] Greenberg B. G., Abul-Ela A. A., Simmons W. R., Horvitz D. G. (1969): *The unrelated question randomized response model: Theoretical Framework.* J. Amer. Statist. Assoc., **64**, 520–539.

[2] *Media projekt 94,* AISA MEDIA, GfK PRAHA, SKMO. 3. kvartál 1. 8. - 30. 9. 1994, Praha 1994 (in czech).

[3] Rodríguez H. F. (1984): *Respuesta aleatorizada. Extension de la tecnica de Warner a modelos muestrales de uso frecuente en poblaciones humanas.* PhD Thesis, Universidad de la Habana.

Author's address: GfK Praha, Ltd., Market Research Institute, Újezd 40/450 (ČOS), 118 01 Praha 1.

The mean is within one mean deviation of any median

Václav Čermák

Let F denote the distribution function of a population, let μ and σ^2 denote its mean and variance, asumed finite, and let m denote any median of F. Suppose without loss of generality that $m < \mu$; the general case follows from this one by reversing signs in the population.

The American Statistician recently published some proofs of the inequality

$$|\mu - m| \leq \sigma, \tag{1}$$

i.e. that the mean is within one standard deviation of any median (cf. the article [1] which started a spate of letters to the Editor). My goal here is to present a better (more strength) inequality by changing the standard deviation with the mean deviation.

Begin the defining the two most familiar mean deviations of the distribution F. Let δ_1 is the mean deviation about mean and let δ_2 is the mean deviation about median, i.e.

$$\delta_1 = \int |x - \mu| dF \qquad \text{and} \qquad \delta_2 = \int |x - m| dF. \tag{2}$$

Well-known are following two inequalities :

(a) $\delta_1 \leq \sigma$, where equality holds only in case of causal or two-valued symmetric distribution; otherwise, the inequality is strict.

(b) $\delta_2 \leq \delta_1$, where equality holds only for symmetric distributions; otherwise, the inequality is strict.

Less known is the inequality $|\xi^\star| \leq 1$, where $\xi^\star$ is the Bonferroni (Pearson's modified) measure of skewness

$$\xi^\star = \frac{\mu - m}{\delta_2} \tag{3}$$

(cf. [2] and [3]). Here, $\xi^\star = 1$ holds only in case of maximal (extremally) skewed distribution, i.e. for a set of n values $x_1 = x_2 = \cdots = x_{n-1} = a$, $x_n = b,\ \ a < b$.

The above then gives the following string of inequalities

$$\boxed{|\mu - m| \leq \delta_2 \leq \delta_1 \leq \sigma} \tag{4}$$

as required.

Appendix (a direct proof). Without loss of generality, we can assume that $m = 0$. Introducing inequality

$$|\int xdF| \leq \int |x|dF \tag{5}$$

which holds for any distribution F that has a finite mean, we have

$$|\mu - m| = |\mu| = |\int xdF| \leq \int |x|dF = \int |x - m|dF = \delta_2. \tag{6}$$

Reference

[1] O'Cinneide, C. A.: *The mean is within one standard deviation of any median.* The American Statistician, **4** (1990), 44, pp. 292 – 293.

[2] Bonferroni, C. E.: *Elementi di Statistica Generale.* Torino, Litografia E.Gili, 1933. (Second ed. 1941.)

[3] Frosini, B. V.: *Lezioni di Statistica, Parte prima.* Milano, Vita e Pensiero, 1987.

[4] Čermák, V.: *Discrete and Continuous Distributions – Formulae, Graphs and Tables.* Handbook (in Czech), Prague School of Economics, Prague, 1993.

Contents

Informační Bulletin České statistické společnosti vychází čtyřikrát do roka v českém vydání a jednou v roce v anglické verzi. Předseda společnosti: Ing. Zdeněk Roth, CSc, SZÚ Praha, MSP, Šrobárova 48, 100 42 Praha 10, E-mail: roth@szu.cz. ISSN 1210-8022

Redakce: Dr. Gejza Dohnal, Jeronýmova 7, 130 00 Praha 3, E-mail: dohnal@fsik.cvut.cz.