

IN THIS ISSUE

SPECIAL INVITED PAPER

ROBUSTNESS: WHERE ARE WE NOW?

If it is to be believed that each field of statistics has a life span ranging between ten to thirty years, the field of robustness, as a branch of statistical science, is currently in a state of fatigue. In 1964, Professor Peter Huber, following the paper on sampling from contaminated distribution by Tukey in 1960, wrote his fundamental paper in the field of robustness. Thirty years later, on May 20, 1994 during the 26th annual meeting of the Association of Statisticians and its Users (ASU) in Neuchâtel, he provided the audience with some fascinating historical developments of robustness. In this issue we are pleased to present the paper that Huber contributed to the conference (he has revised and updated his contribution for *Student*). Huber writes “Also the currently fashionable (over-)emphasis of high breakdown points transmits a wrong signal. A high breakdown point is nice to have, if it comes for free, but the potential presence of high contamination usually indicates a mixture model and calls for diagnostics. A thoughtless application of robust procedures might only hide the underlying problem”. In his paper, Professor Huber reviews certain development of selected robustness concepts from 1960 to 1995, and clarifies many disputed points such as robustness, nonparametric and distribution free methods. The small sample problems seem to him to be the most interesting and at the same time the most promising new methods for future research in statistics.

EMPIRICAL INFLUENCE FUNCTIONS FOR THE COX REGRESSION MODEL

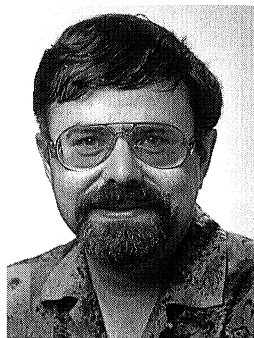
In medicine, studies are often set up to compare the survival between different groups of patients. In many situations, these patient groups differ not only in the treatment they received, but also in other variables such as age, previous treatment or past health status. The analysis of the role played by these variables requires a statistical model that relates the survival time to a series of variables thought to influence it. One frequently used approach to model this situation is the *proportional hazards model* or *Cox regression model* which originates from 1972. In her article, Françoise Dubas presents a series of empirical influence functions for the Cox regression model in order to determine its robustness. Not seemingly realistic to live beyond one hundred years after an operation at the age of 64, she shows that the Cox regression model is not robust in the presence of extremely large outliers.

Françoise Dubas was born on May 5, 1955 in Fribourg. She graduated in Biology from the University of Geneva in 1977. She obtained her Ph.D. in Physiology from the University of Aberdeen, Scotland in 1982. Having worked several years in the field of neurophysiology, she completed her postgraduate degree in Statistics from the University of Neuchâtel in 1993. She is currently working in the staff of the Swiss Science Agency as scientific advisor.



SIMPSON'S PARADOX REVISITED

The Greek word “para-doxon” means “contrary to the opinion”. The opinion should be replaced, according to Plato (4th century B.C.) by “mathema” ($\approx$ knowledge, mathematics). One of the most famous paradoxes in Statistics is Simpson's paradox. Arthur Vogt in his article shows that generalized index theory allows to treat Simpson's paradox in a new way. It turns out that the paradox stems from Drobisch, 80 years before Simpson's publication.

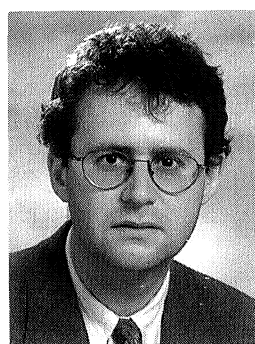


Arthur Vogt was born on December 6, 1947 in Zürich. From 1966 to 1971 he studied Mathematics at the Federal Institute of Technology. Dr. Vogt's major works are mainly influenced by Professors Markus Fierz, Eduard Stiefel and Ernst Kaiser. After working in the Swiss Statistical Office and the Swiss Private Insurance Office he came back in 1992 to the Swiss Social Insurance Office where he is in charge of Mathematics and Statistics of the Swiss pension funds.

ENVIRONMENTAL INDICES FOR THE UNITED STATES

Policy-makers, decision-takers and the general public have an increasing need for concise information about the state and development of the environment. Many countries and international organizations have recently carried out works on environmental indices and the aim is now to secure a minimum level of harmonization between them. Pascal Grosclaude's article first surveys the related literature, then describes both the conceptual framework and method and finally presents the results of a pilot study for the United States of America.

Pascal Grosclaude was born on March 2, 1966 in Bienne. He obtained his Bachelor of Science in Economics from the University of Neuchâtel in 1988 and his doctoral degree in 1995. He is currently a research fellow at the Institute for Regional and Economic Research of the University of Neuchâtel and an economist at the Department of Finance of the Canton of Neuchâtel.



RANDOM THOUGHTS ON SURVIVAL

Sir David Cox and C.R. Rao discuss some difficulties in teaching statistical materials to an audience of multiple background. They use this occasion to raise other problems associated with the current state of research in contrast to the work of the pioneers in this field.

CAPTURING HISTORY

HENRY E. DANIELS SETS THE RECORD STRAIGHT

A highlight of this issue is a letter from Henry Daniels in which he sets the record straight about the word *saddle point*. Daniels wrote up "Saddlepoint Approximations in Statistics" paper in 1952 during his one year visit at the University of Chicago. The paper was received in January 1953, was revised in March 1954 and appeared in *The Annals of Mathematical Statistics* (Vol. 25, 631-650) in 1954.

Robustness: Where Are We Now?

Peter J. Huber

University of Bayreuth, Germany

Abstract: The development of selected robustness concepts since their inception in the 1960's is sketched and their current status is reviewed.

He looked into the water and saw that it was made up of a thousand thousand thousand and one different currents, each one a different colour, weaving in and out of one another like a liquid tapestry of breathtaking complexity; and Iff explained that these were the Streams of Story, that each coloured strand represented and contained a single tale.

(Salman Rushdie, *Haroun and the Sea of Stories*, 1990)

1 The first ten years

A good case can be made that modern robustness begins in 1960, with the papers by J.W. Tukey on sampling from contaminated distributions, and by F.J. Anscombe on the rejection of outliers. Tukey's paper drew attention to the dramatic effects of seemingly negligible deviations from the model, and it made effective use of asymptotics in combination with the gross error model. Anscombe introduced a seminal insurance idea: sacrifice some performance at the model in order to insure against ill effects caused by deviations from it. Most of the basic ideas, concepts and methods of robustness were invented in quick succession during the following years and were in place by about 1970.

In 1964 I recast Tukey's general setup into an asymptotic minimax framework and was able to solve it. Important points were: the insistence on finite, but small deviations, the formal recognition that a large error

occurring with low probability is a small deviation, and the switch from the then prevalent criterion of relative efficiency to absolute performance. At the same time, I introduced the notion of maximum likelihood type or M-estimators.

Hampel (1968) added the formal definition of qualitative robustness (continuity under a suitable weak topology), infinitesimal robustness in form of the influence function (von Mises derivative of a statistical functional) and the notion of breakdown point.

A nagging early worry had been the possible irrelevancy of asymptotic approaches: conceptually, 1% gross errors in samples of 1000 are entirely different from the same error rate in samples of 5. This worry was laid to rest by Huber (1965, 1968): both for tests and for estimates, there are qualitatively identical and even quantitatively similar exact finite sample minimax robustness results.

The end of the decade saw the first steps of an extension of asymptotic robustness theory beyond location to more general parametric models, namely the introduction of shrinking neighborhoods by C. Huber-Carol (1970), as well as the first attempt at studentizing (Huber, 1970).

The basic methodology for Monte Carlo studies of robust estimators was established in Princeton 1970/71, see Andrews et al. (1972). That study more or less finished off the problem of robustly estimating a single location parameter in samples of size 10 or larger, opening the way for more general multi-parameter problems.

In my present talk I shall pick a few of the more interesting robustness ideas and follow the strands of their stories to the present. What has happened to the early strands since their invention? What important new strands have begun in the 1970's and 1980's? I shall avoid technicalities and instead give a reference to a recent survey or monograph, where feasible. Completeness is not intended; for a recent "far from complete" survey of research directions in robust statistics, with more than 500 references, see Stahel's article in Stahel and Weisberg (1991, Part II, p. 243).

2 Influence functions and pseudo-values

Among the basic robustness concepts, influence functions have become a standard tool, especially after the comprehensive treatment by Hampel et al. (1986). Also the trick to robustize classical procedures through the use of pseudo-values is becoming common knowledge, even though it has received only scant coverage in the literature. This trick is that one calculates robust fitted values $\hat{y}_i$ by iteratively applying the classical procedure to the pseudo-values $y_i^* = \hat{y}_i + r_i^*$ instead of y_i . Here, the pseudo-residual

$r_i^* = \psi(r_i)$ is obtained by cutting down the current residual $r_i = y_i - \hat{y}_i$ with the help of a function ψ proportional to the desired influence function (i.e. with the ψ -function defining an M-estimate). For examples see in particular Bickel (1976, p. 167), Huber (1979), and Kleiner, Martin and Thomson (1979). If ψ is chosen equal rather than merely proportional to the influence function, the classical formulas, when applied to the pseudo-residuals r_i^* instead of the residuals, yield asymptotically correct error estimates for ANOVA and other purposes (Huber 1981, p. 197).

There have been some very interesting extensions of influence function ideas to time series (Künsch, 1984).

3 Breakdown and outlier detection

For a long time, the breakdown point had been a step-child of the robustness literature. The paper by Donoho and Huber (1983) was specifically written to give it more visibility. Recently, I have begun to wonder whether it has given it too much, the suddenly fashionable emphasis on high breakdown point procedures has become counter-productive. One of the most striking examples of the usefulness of the concept can be found in Hampel (1985): the combined performance of outlier rejection followed by the sample mean as an estimate of location is essentially determined by the breakdown of the outlier detection procedure.

4 Studentizing

Whenever we have an estimate, we ought to provide an indication of its statistical accuracy, say by giving a 95% confidence interval. This is not particularly difficult if the number of observations is very large, so that the estimate is asymptotically normal with an accurately estimable standard error, or also in one-parameter problems without nuisance parameter, where the finite sample theory of Huber (1968) can be applied.

Otherwise, we end up with a tricky problem of studentization. To my knowledge, there has not been much progress beyond the admittedly unsatisfactory initial paper of Huber (1970). There are not only many open questions with regard to this crucially important problem, it is even open what questions one should ask! A sketch of the principal issues follows.

In the classical normal case, it follows from sufficiency of $(\bar{x}, s)$ and an invariance argument that such a confidence interval must take the form

$$(\bar{x} - k_n s / \sqrt{n}, \bar{x} + k_n s / \sqrt{n})$$

with k_n depending on the sample size, but not on the sample itself. In a

well-behaved robust version, a confidence interval might take the analogous form

$$(T - K_n S / \sqrt{n}, T + K_n S / \sqrt{n})$$

where T is an asymptotically normal robust location estimate and S is the location invariant, scale equivariant, Fisher consistent functional estimating the asymptotic standard deviation of $\sqrt{n}T$, applied to the empirical distribution. In the case of M-estimates for example, we might use

$$S(F)^2 = \frac{S_0^2 \int \psi^2 dF}{(\int \psi' dF)^2}$$

where the argument of ψ , ψ' is $y = (x - T)/S_0$, i.e. a robustly centered and scaled version of the observations, say with $S_0 = MAD$. If we are interested in 95% confidence intervals, K_n must approach $\Phi^{-1}(0.975) \approx 1.96$ for large n . But K_n might depend on the sample configuration in a non-trivial, translation and scale invariant way: since we do not have a sufficient statistic, we might want to condition on the configuration of the sample in an as yet undetermined way.

While the distribution of $\sqrt{n}T$ typically approaches the normal, it will do so much faster in the central region than in the tails, and the extreme tails will depend rather uncontrollably on details of the unknown distribution of the observations. The distribution of S suffers from similar problems, but here it is the low end which matters. The question is: what confidence levels make sense and are reasonably stable for what sample sizes? For example, given a particular level of contamination and a particular estimate, is $n = 10$ good enough to derive accurate and robust 99% confidence intervals, or do we have to be content with 95% or 90%? I anticipate that such questions can (and will) be settled with the help of small sample asymptotics, assisted perhaps by configural polysampling (below).

5 Shrinking neighborhoods

Direct theoretical attacks on finite neighborhoods work only if the problem is location or scale invariant. But for large samples, most point estimation problems begin to resemble location problems, so it is possible to derive quite general asymptotically robust tests and estimates by letting those neighborhoods shrink at a rate $n^{-1/2}$ with increasing sample size. This idea was first exploited by C. Huber-Carol (1970), followed by Rieder, Beran, Millar and others. The final word on this approach is contained in Rieder's book (1994).

The principal motivation clearly is technical: shrinking leads to a manageable asymptotic theory. But there is also a philosophical justification:

since the standard goodness-of-fit tests are just able to detect deviations of the order $O(n^{-1/2})$, it makes sense to put the border zone between small and large deviations at $O(n^{-1/2})$. Larger deviations should be taken care of by diagnostics and modelling, smaller ones are difficult to detect and should be covered (in the insurance sense!) by robustness.

This does not mean that our data samples are supposed to get cleaner if they get larger. But the shrinkage of neighborhood faces us with a dilemma, namely a choice between the alternatives:

- improve the model; or
- improve the data; or
- stop sampling.

Note that adaptive estimation is *not* among the viable alternatives. The problem is not one of reducing statistical variability, but one of avoiding bias, and the ancient Emperor-of-China paradox applies (you can get a fantastically accurate estimate of the height of the emperor by averaging the guesses of 600 million Chinese, most of whom never saw the emperor!).

The asymptotic theory of shrinking neighborhoods is, in essence, a theory of infinitesimal robustness and suffers from the same conceptual drawback as approaches based on the influence function: infinitesimal robustness (bounded influence) does not automatically imply robustness. The crucial point is that in any practical application we have a fixed, finite sample size, and we need to know whether we are inside the range of n and ε for which asymptotic theory yields a decent approximation. This range may be difficult to determine, but the breakdown point often is computable and may be a useful indicator.

6 Design

Robustness casts a shadow on the theory of optimal designs: they lose their theoretical optimality very quickly under minor violations of linearity (Huber, 1975) or independence assumptions (Bickel and Herzberg, 1979). I am not aware of much current activity in this area, but the lesson is clear: “Naive” designs usually are more robust and better than “optimal” designs.

7 Regression

Back in 1975, the discussants of Bickel (1976) raised interesting criticisms, in particular there were complaints about the multiplicity of robust procedures, and about their computational and conceptual complexity. Bickel fended them off skillfully and convincingly.

There may have been reasons for concern then, but the situation has become worse. Most of the action in the 1980's has been on the regression front. Here is an incomplete list of robust regression estimators: L1 (going back at least to Laplace), M (Huber, 1973), GM (Mallows, 1975), with variants by Hampel, Krasker and Welsch, RM (Siegel, 1982), LMS and LTS (Rousseeuw, 1984), S (Rousseeuw and Yohai, 1984), MM (Yohai, 1987), τ (Yohai and Zamar, 1988), SRC (Simpson, Ruppert and Carroll, 1991), and no end is in sight. For an up-to-date review see Davies (1993).

Bickel would not have an easy job now, much of the "Nordic" criticism, unsubstantiated in 1975, seems to be justified now. In any engineering product, an overly rapid sequence of updates sometimes is a sign of vigorous progress, but it can also be a sign of shoddy workmanship, and often it is both. In any case, it confuses the customers and hence is counter-productive...

Robustness has been defined as insensitivity to small deviations from an idealized model. What is this model in the regression case? The classical model goes back to Gauss and assumes that the carrier X (the matrix X of the "independent" variables) is error-free. X may be systematic (as in designed experiments), or haphazard (as in most undesigned experiments), but its rows only rarely can be modelled as being a random sample from a specified multivariate model distribution. Statisticians tend to forget that the elements of X often are not observed quantities, but are derived from some model (cf. the classical non-linear problems of astronomy and geodesy giving rise to the method of least squares in the first place). In essence, each individual X corresponds to a somewhat different situation and might have to be dealt with differently. Thus, multiplicity of procedures may lie in the nature of robust regression. Curiously, most of the action seems to have been focused through tunnel vision on just one aspect: safeguard at any cost against problems caused by gross errors in a random carrier.

Over the years, I too had to defend the minimax approach to distributional robustness on many occasions. The salient points of my defense were that the least favorable situation one is safeguarding against, far from being unrealistically pessimistic, is more similar to actually observed error distributions than the normal model, that the performance loss at a true normal model is relatively small, that on the other hand the classically optimal procedures may lose sorely if the normal model is just slightly violated, and that the hardest problems are not with extreme outliers (which are easy to detect and eliminate), but with what happens on the shoulders of the distributions. Moreover, the computation of robust M-estimates is easy and fast (the last paragraph of this section). Not a single one of these defense lines can be used with the modern "high breakdown point" regression

estimates.

A typical cause for breakdown in regression are gross outliers in X ; while individual such outliers are trivially easy to spot (with the help of the diagonal of the hat matrix), efficient identification of collaborative leverage groups is an open, perhaps unsolvable, diagnostic problem. However, I would advise against treating leverage groups blindly through robustness, they may hide serious design or modeling problems, and there are similar problems even with single leverage points.

The story behind an outlier among the X (“leverage point”) might for example be:

- a misplaced decimal point,
- an accurate but useless observation, outside of the range of validity of the model.

If the value at this leverage point disagrees with the evidence extrapolated from the other observations, this may be because:

- the outlying observation is affected by a gross error (in X or in y),
- the other observations are affected by small systematic errors (this is more often the case than one might think),
- the model is inaccurate, so the extrapolation fails.

The existence of several, phenomenologically indistinguishable but conceptually different situations with different consequences calls for a diagnostic approach (identification of leverage points or groups), followed by alternative “what if” analyses. This contrasts sharply with simple location estimation, where the observations are exchangeable and a minimax approach is quite adequate (although one may want to follow it up with an investigation of the causes of grosser errors).

At the root of the current confusion is that hardly anybody bothers about stating all of the issues clearly: not only the estimator and a procedure for computing it must be specified, but also the situations for which it is supposed to be appropriate or inappropriate, and criteria for judging estimators and procedures. There has been a tendency to rush into print with rash claims and procedures. In particular, what is meant by the word breakdown? For many of the newer estimates there are unqualified claims that their breakdown point approaches 0.5 in large samples. But such claims tacitly exclude designed situations: if the observations are equally partitioned among the corners of a simplex in d -space, no estimate whatsoever can achieve a breakdown point above $1/(2d + 2)$.

It is one thing to design a theoretical algorithm whose purpose is to prove, for example, that a breakdown point $1/2$ can be attained in principle,

and quite another thing to design a practical version that can be used not merely on small, but also on medium sized regression problems, with a 2000-by-50 matrix or so. This last requirement would seem to exclude all of the recently proposed robust regression estimates.

Some comments on the much maligned “plain vanilla” regression M-estimates are in order. The M-estimate approach is not a panacea (is there such a thing in statistics?), but it is easy to understand, practicable, and considerably safer than classical least squares. While I had had regression problems in mind from the very beginning (cf. Huber 1964, p.74), I had not dared to go into print until 1973, when I believed to understand the asymptotic behavior of M-estimates of regression. A necessary regularity condition for consistency and asymptotic normality of the fitted values is that the maximum diagonal element h of the hat matrix $H = X(X^T X)^{-1} X^T$ tends to 0. While watching out for large values of h does not enforce a high breakdown point, it at least may prevent a typical cause of breakdown. Moreover, with M-estimates of regression, the computing effort for large matrices typically is less than twice what is needed for calculating the ordinary least squares solution. Both calculations are dominated by the computation of a QR or SVD decomposition of the X matrix, which takes $O(np^2)$ operations for an (n, p) -matrix. Since the result of that decomposition can be re-used, the iterative computation of the M-estimate, using pseudo-values, takes $O(np)$ per iteration with fewer than 10 iterations on average.

8 Multivariate problems

Classically, “regression” problems with errors in the independent variables are solved by fitting a least squares hyperplane, that is, by solving a principal component problem and picking the plane belonging to the smallest eigenvalue. It can be argued by analogy that regression problems with potential gross errors in the carrier should be attacked through some version of robust principal components. Thus, multivariate problems, in particular of the principal component type, deserve added attention.

In 1976, Maronna showed that all M-estimates of location and scatter in d dimensions have a breakdown point $\varepsilon^* \leq 1/(d+1)$. In higher dimensions this is shockingly low, but then Stahel (1981) and Donoho (1982) independently found estimates based on projection pursuit ideas with a breakdown point approaching $1/2$ in large samples. The bad news is that with all currently known algorithms the effort for computing those estimates increases exponentially with d . We might say they break down by failing to give a timely answer!

9 Some persistent misunderstandings

Robust methods have persistently been misclassified and pooled with non-parametric and distribution-free methods. They are part of traditional parametric statistics, the only difference is that they do not assume that the parametric model is exactly true.

In part the robustness community itself is to blame for this misunderstanding. In the mid-1970's adaptive estimates - attempting to achieve asymptotic efficiency at all well-behaved error distributions - were thought by many to be the ultimate robust estimates. Then Klaassen (1980) proved a disturbing result on the lack of stability of adaptive estimates. My current conjecture is that an estimate cannot be simultaneously adaptive in a neighborhood of the model and qualitatively robust at the model.

Also the currently fashionable (over-)emphasis of high breakdown points transmits a wrong signal. A high breakdown point is nice to have, if it comes for free, but the potential presence of high contamination usually indicates a mixture model and calls for diagnostics. A thoughtless application of robust procedures might only hide the underlying problem.

There seems to be some confusion between the respective roles of diagnostics and robustness. The purpose of robustness is to safeguard against deviations from the assumptions, in particular against those that are near or below the limits of detectability. The purpose of diagnostics is to find and identify deviations from the assumptions.

The term "robust" was coined by Bayesians (Box and Andersen, 1955). It is puzzling that Bayesian statistics never managed to assimilate the modern robustness concept, but remained stuck with antiquated parametric supermodels. *Ad hoc* supermodels and priors chosen by intuition (personal beliefs) or convenience (conjugate priors) do not guarantee robustness, for that one needs some theory. Compare already Hampel (1973).

10 Future directions: small sample problems?

At present, the most interesting and at the same time most promising new methods and open problems have to do with small sample situations.

The Princeton robustness study (Andrews et al., 1972) remains one of the best designed comparative Monte Carlo studies in statistics. Among other things we learned from it how difficult it is to compare the behavior of estimates across entire families of distributions, since small systematic differences of performance easily are swamped by larger random sampling errors. A very sophisticated sampling method thereafter proposed by Tukey is based on the remark that a given sample can occur under any strictly

positive density, but it will do so with different probabilities. Thus it must be possible to compare those performances very efficiently by re-using the same sample configurations with different weights when forming Monte Carlo averages. A comprehensive account of this approach is given by Morgenthaler and Tukey (1991).

On the other hand, it was noted by Hampel that a variant of the saddle point method can give fantastically accurate asymptotic approximations down to very small sample sizes (occasionally down to $n = 1!$). A comprehensive account of this approach has recently been given by Field and Ronchetti (1990).

I hope that these approaches in the near future will permit to close several open problems in the area of small samples: studentization, confidence intervals, testing. Embarrassingly, the robustification of the statistics of two-way tables still is wide open. Typically there are so few degrees of freedom per cell that ordinary asymptotic approaches are out of the question, maybe some version of small sample asymptotics may help here too.

Résumé: L'évolution des concepts liés à la robustesse est étudiée depuis leur introduction dans les années 1960 jusqu'à nos jours.

References

- [1] Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., and Tukey, J.W. (1972), *Robust Estimates of Location: Survey and Advances*, Princeton Univ. Press, Princeton, NJ.
- [2] Anscombe, F.J. (1960), "Rejection of Outliers", *Technometrics*, **2**, 123-147.
- [3] Bickel, P.J. (1976), "Another Look at Robustness: A Review of Reviews and Some New Developments", *Scand. J. Statist.*, **3**, 145-168.
- [4] Bickel, P.J., and Herzberg, A.M. (1979), "Robustness of Design Against Autocorrelation in Time I", *The Annals of Statistics*, **7**, 77-95.
- [5] Box, G.E.P., and Andersen, S.L. (1955), "Permutation Theory in the Derivation of Robust Criteria and the Study of Departure from Assumption", *J. Roy. Statist. Soc. Ser. B*, **17**, 1-34.
- [6] Davies, P.L. (1993), "Aspects of Robust Linear Regression", *The Annals of Statistics*, **21**, 1843-1899.
- [7] Donoho, D.L. (1982), "Breakdown Properties of Multivariate Location Estimators", Ph.D. Qualifying Paper, Harvard University, Cambridge, MA.
- [8] Donoho, D.L., and Huber, P.J. (1973). "The Notion of Breakdown Point", In *A Festschrift for Erich L. Lehmann*, P. J. Bickel, K.A. Dok-

- sum, J.L. Hodges (Eds.), Wadsworth, Belmont, CA.
- [9] Field, C., and Ronchetti, E. (1990), *Small Sample Asymptotics*, IMS Lecture Notes, Monograph Series, Hayward, CA.
 - [10] Hampel, F.R. (1968), "Contributions to the Theory of Robust Estimation", Ph.D. Thesis, Univ. of Calif., Berkeley.
 - [11] Hampel, F.R. (1973), "Robust Estimation: A Condensed Partial Survey", *Z. Wahrsch. verw. Geb.*, **27**, 87-104.
 - [12] Hampel, F.R. (1985), "The Breakdown Point of the Mean Combined with Some Rejection Rules", *Technometrics*, **27**, 95-107.
 - [13] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986), *Robust Statistics. The Approach Based on Influence*, John Wiley, New York.
 - [14] Huber, P.J. (1964), "Robust Estimation of a Location Parameter", *Ann. Math. Statist.*, **35**, 73-101.
 - [15] Huber, P.J. (1965), "A Robust Version of the Probability Ratio Test", *Ann. Math. Statist.*, **36**, 1753-1758.
 - [16] Huber, P.J. (1968), "Robust Confidence Limits", *Z. Wahrsch. verw. Geb.*, **10**, 269-278.
 - [17] Huber, P.J. (1970), "Studentizing Robust Estimates", In *Nonparametric Techniques in Statistical Inference*, M. L. Puri (Ed.), Cambridge University Press, Cambridge, England.
 - [18] Huber, P.J. (1973), "Robust Regression: Asymptotics, Conjectures and Monte Carlo", *The Annals of Statistics*, **1**, 799-821.
 - [19] Huber, P.J. (1975), "Robustness and Designs". In: *A Survey of Statistical Design and Linear Models*, J. N. Srivastava (Ed.), North Holland, Amsterdam.
 - [20] Huber, P.J. (1979), "Robust Smoothing", *Proc. ARO Workshop on Robustness in Statistics*, R. N. Launer and G. N. Wilkinson (Eds.), Academic Press, New York.
 - [21] Huber, P.J. (1981), *Robust Statistics*, John Wiley, New York.
 - [22] Huber-Carol, C. (1970), "Etude asymptotique de tests robustes", Ph.D. thesis, ETH, Zürich.
 - [23] Klaassen, C. (1980), "Statistical Performance of Location Estimators", Ph.D. thesis, Mathematisch Centrum, Amsterdam.
 - [24] Kleiner, B., Martin, R.D., and Thomson, D.J. (1979), "Robust Estimation of Power Spectra", *J. Royal Statist. Soc. Ser. B*, **41**, 313-351.
 - [25] Künsch, H. (1984), "Infinitesimal Robustness for Autoregressive Processes", *The Annals of Statistics*, **12**, 843-863.
 - [26] Mallows, C.L. (1975), "On Some Topics in Robustness", Tech. Memorandum, Bell Telephone Laboratories, Murray Hill, NJ.
 - [27] Maronna, R.A. (1976), "Robust M-Estimators of Multivariate Location

- and Scatter", *The Annals of Statistics*, **4**, 51-67.
- [28] Morgenthaler, S. and Tukey, J.W. (1991), *Configural Polysampling*, John Wiley, New York.
 - [29] Rieder, H. (1994), *Robust Asymptotic Statistics*, Springer, Berlin.
 - [30] Rousseeuw, P.J. (1984), "Least Median of Squares Regression", *J. Amer. Statist. Assoc.*, **79**, 871-880.
 - [31] Rousseeuw, P.J., and Yohai, V.J. (1984), "Robust Regression by Means of S-Estimators", In *Robust and Nonlinear Time Series Analysis*, J. Franke, W.Härdle and R.D. Martin (Eds.), Lecture Notes in Statistics 26, Springer, New York.
 - [32] Siegel, A.F. (1982), "Robust Regression Using Repeated Medians", *Biometrika*, **69**, 242-244.
 - [33] Simpson, D.G., Ruppert, D., and Carroll, R.J. (1992), "On One-Step GM-Estimates and Stability of Inferences in Linear Regression", *J. Amer. Statist. Assoc.*, **87**, 439-450.
 - [34] Stahel, W.A. (1981), "Breakdown of Covariance Estimators", Research Report 31, Fachgruppe für Statistik, ETH Zürich.
 - [35] Stahel, W., and Weisberg, S. (Eds.) (1991), *Directions in Robust Statistics and Diagnostics*, Part I and II, the IMA Volumes in Mathematics and its Applications, Vols. 33-34, Springer, New York.
 - [36] Student (1927), "Errors of Routine Analysis", *Biometrika*, **19**, 151-164.
 - [37] Tukey, J.W. (1960), "A Survey of Sampling from Contaminated Distributions", In *Contributions to Probability and Statistics*, I. Olkin (Ed.), Stanford Univ. Press, Stanford, CA.
 - [38] Yohai, V.J. (1987), "High Breakdown-Point and High Efficiency Robust Estimates for Regression", *The Annals of Statistics*, **15**, 642-656.
 - [39] Yohai, V.J., and Zamar, R.H. (1988), "High Breakdown Point Estimates of Regression by Means of the Minimization of an Efficient Scale", *J. Amer. Statist. Assoc.*, **83**, 406-413.

Empirical Influence Functions for the Cox Regression Model

Françoise Dubas

Federal Department of Internal Affairs, Berne, Switzerland

Abstract: The aim of the present work was to determine the robustness of the Cox regression model (proportional hazards model) by constructing empirical influence functions. This was achieved by adding a moving outlier to a known data set and estimating the regression parameters for each position of the moving outlier. The results show that if the outlier is situated far from the original data set, its influence on the estimates of the regression parameters is maximal but does not increase after a certain limit. The influence functions for the regression parameters are hence bounded.

Key Words: Proportional hazards model, robustness, multivariate failure time data.

1 Introduction

Many medical studies are designed to determine whether a new medication, a new form of treatment, or a new therapeutic procedure is better than either no treatment or the best treatment currently available. Often, two groups of patients are compared where one group receives the established treatment or a properly designed placebo control treatment, and the second group receives the treatment under investigation.

An important subset of these clinical trials are survival studies, where the variable of interest is the amount of time that elapses between the onset of treatment and the occurrence of an event, such as the death of the patient, the rejection of a transplanted organ, or the point at which the patient is cured of the disease.

2 Proportional hazards model

In some situations, patient groups differ not only in the treatment which they receive, but also in other variables such as age, previous treatment, past health status, severity of disease, employment history, race and so on. The analysis of the role played by these variables requires a statistical model that relates the time until death (or more generally: *failure time*) to a series of measurements thought to influence the failure time (see Dawson-Saunders and Trapp, 1990).

The analysis of the relationships between explanatory variables and failure time is similar to a multivariate regression in which the relationship between a series of explanatory variables and a single response variable is established. In both cases, a model must be found and parameters for each variable must be estimated which explain the data best.

One frequently used approach to model multivariate failure time data is the *proportional hazards model*, or *Cox regression model* (see Cox, 1972). Given a set of failure time data where the value of k explanatory variables have been recorded in addition to time to failure, the aim of the analysis using the proportional hazards model is to find estimates of k regression parameters. The value of each parameter will express the risk due to one of the variables, adjusted for the values of the other variables. For example, the parameter for the variable “age” will express the influence of age on survival adjusted for “profession”, “drinking habits” and “health status” for instance. A parameter value larger than zero indicates an increased risk of failure, i.e. the larger the variable x , the shorter the survival time. A negative parameter value indicates a “protective” effect, i.e. the larger the variable x the longer the survival time.

The Cox regression model postulates in its simplest form that the hazard function associated with the survival in group A (λ_A) is proportional to the hazard function in group B (λ_B) or that the survival of group A is related to the hazard function in group B by a multiplication constant, whatever the shape or the value of the hazard function. That is

$$\lambda_B(t) = C\lambda_A(t)$$

where C is a constant. In terms of survival curves S_A and S_B , the property of proportionality of hazard functions translates as

$$S_B(t) = (S_A(t))^C.$$

In the more general case of the proportional hazards model, there are k explanatory variables and there are n groups or sets of variable values to

be compared. In the general case, the model becomes

$$\lambda_j(t; x_{1j}, x_{2j}, \dots, x_{kj}) = \lambda_0(t) \cdot C_j = \lambda_0(t) \cdot e^{\sum_{i=1}^k b_i \cdot x_{ij}}.$$

The value $\lambda_j(t; x_{1j}, x_{2j}, \dots, x_{kj})$ represents the *hazard rate* for a specific person with the j th set of variable values at time t . The function $\lambda_0(t)$ is an arbitrary, unspecified base line hazard rate at time t . The variables x_{ij} represent the values of the i th explanatory variable in the j th set of variable values. The coefficient b_i determines the influence of that variable on the survival time. In the case where there are k variables, the multiplicative constant can be decomposed as

$$C = (e^{b_1 \cdot x_{1j}}) \cdot (e^{b_2 \cdot x_{2j}}) \cdot \dots \cdot (e^{b_k \cdot x_{kj}}).$$

The value e^{b_i} is called the *relative hazard* for the i th explanatory variable. It is clear that the different explanatory variables act multiplicatively on the hazard function.

The ratio of two hazard functions (relative hazard) predicts the difference in survival between two individuals or two groups of individuals. It is expressed as

$$\frac{\lambda_1(t; x_{11}, x_{21}, \dots, x_{k1})}{\lambda_2(t; x_{12}, x_{22}, \dots, x_{k2})} = e^{\sum_{i=1}^k b_i \cdot (x_{i1} - x_{i2})}$$

where x_{i1} and x_{i2} represent different values of the variable x_i .

In terms of survival curves the general case of the proportional hazards model translates as

$$S_j(t) = (S_0(t))^{e^{\sum_{i=1}^k b_i \cdot x_{ij}}}.$$

When the hazard rate is constant and when the baseline survival curve can be approximated by an exponential, the model becomes

$$S_j(t) = (e^{-\lambda_0 \cdot t})^{e^{\sum_{i=1}^k b_i \cdot x_{ij}}}.$$

An interesting property of the proportional hazard model is that the form of the hazard function needs to be neither specified nor even known. Also, numerical and nominal or categorical variables can all be considered.

3 Objectives of the work

Survival studies are usually based on trials involving a large number of patients observed by several people, and where the data of interest is often

handled by different actors. Although drastic practical measures are taken to avoid the inclusion of errors in medical trial, it is important that the effect of possible erroneous records is minimized since their presence may drastically influence the statistical analysis of the data set. On the other hand, in many trials patients can be found with atypical features which may not have been evident at first. It is important that their records can be identified in order to find out why they are different. The use of *robust methods* (see Hampel, Ronchetti, Rousseeuw and Stahel, 1986) applicable to survival analysis, in particular to the Cox regression model, may help accomplish these operations.

It must be noted that the Cox model, in the real life situation, is robust enough. The usefulness of this paper lies only in the fact that it shows how the proportional hazards model behaves in the case that a totally aberrant value is included.

As a first step toward developing robust methods for the proportional hazards model, the general aim of this work has been to construct *empirical influence functions* for the Cox regression model to determine its robustness to the presence of an abnormal data point in the data set. This was achieved by adding a moving data point (test outlier) to a known data set and to estimate the regression parameters b_i for each position of the moving outlier. Although the additional data point is a true outlier only at certain values, it will always be called the outlier.

4 Methods

4.1 Data set

The 1980 update of the Stanford heart transplant data set was used (as presented in Miller and Halpern, 1982). This data set gives the survival times in days for 157 transplant patients who had entered the study by February 1980; whether the survival time was calculated from entry in the study or from transplant is not known. This information, however, is not relevant for the analysis shown below. In addition, 1) an indication of whether or not the survival time was censored, 2) the patient's age (in years) at the time of the transplant and, 3) a measure of the incompatibility between the donor's heart and the tissues of the patient recipient of the heart (T5 mismatch score: no units) were also given. A previous version of this data set has been used by Kalbfleisch and Prentice (1980) to discuss the proportional hazards model. The data set used in this work has been used by Miller and Halpern (1982) to study the performance of different regression methods with censored data and by Reid and Crépeau (1985) to

identify influential data points in the proportional hazards model.

4.2 Computer programs

Basic data analysis and description was conducted using Minitab. Data tables and 2-D plots were produced using Minitab or Excel. The 606 Version of SAS was used to fit the Cox proportional hazards model to the data and to estimate the values of the parameters.

For the estimation of the regression parameters b_i , a macro, called CMOD, was prepared according to the SAS Manual. This macro performs the proportional hazards regression using the non-linear least squares procedure (NLIN) and the discrete hazard method for ties (see SAS Manual). This approximates the maximization of the partial likelihood method normally used to estimate the parameters (for discussions of the different methods of estimation see Kalbfleisch and Prentice, 1980; Reid and Crépeau, 1985; Bednarski, 1989). The NLIN procedure estimates the parameters in a series of iterative steps. If the convergence criterion is not met by the 10th step, the regression parameters are not estimated. Estimation of the regression parameters is thus missing for the positions of the outlier for which convergence was not met.

Kalbfleisch and Prentice (1980) discuss several methods to estimate the regression parameters b_i , including the marginal likelihood method and the partial likelihood method proposed by Cox (1972). According to these authors, the usual approach is to maximize the partial likelihood (for a complete development of the method see Kalbfleisch and Prentice, 1980).

5 Statistical analysis and experimental results

5.1 Cox regression analysis of the original data set

An analysis of the original Stanford heart transplant data set was first performed in order to determine the regression parameters (b_i) for age and the T5 mismatch score. For the complete model in which age and T5 are considered, Table 1 below gives the estimates of the age and the T5 coefficients:

These estimates are very similar to those given by Miller and Halpern (1982) and Reid and Crépeau (1985) who estimated the T5 parameter at 0.167 and 0.170 respectively and the age parameter at 0.030.

The value of the age parameter given in Table 1 indicates that younger ages at the time of transplantation are associated with longer survival times. Table 1 also shows that the age coefficient is significant in the original data

Parameter	Estimate	Asymptotic Standard Error	Asymptotic 95% Confidence Interval
Age b_i	0.02960	0.01017	0.0094-0.0498
T5 b_i	0.17003	0.17358	-0.1749-0.5150

Table 1: Estimation of the age and T5 regression parameters for the original Stanford heart transplant data set using the SAS NLIN procedure

set while the T5 coefficient is not. If it was solely of interest to identify the best model fitting the original data set, the T5 coefficient could be dropped. However, since it seemed interesting to see what the behaviour of the T5 coefficient would be in the presence of different types of outliers, the complete model was always used in the following statistical experiments. Analysis of the original data set also shows that age and T5 values are not correlated (correlation coefficient: -0.043).

5.2 Outliers in one dimension only

5.2.1 Outliers in survival time

Procedure: The age and T5 values of the outlier were kept constant at a value close to the mode values of age and T5 in the original sample (i.e.: age: 50 years; T5: 0.9). The survival value of the additional data point was changed from 0 to 10000 days over 11 steps. Although survival times in the original data were no longer than 3600 days, the outlier was given unusually long survival times so that the behaviour of the model in the presence of extreme outliers could be analyzed. The experiment was made for censored and uncensored outliers. The regression parameters b_i and their standard error were estimated (as in Table 1 above) for each position of the test outlier. These estimates were then plotted versus the survival values of the outlier, independently for the censored and uncensored outlier (shown in Figure 1 and 2).

Observations: From this first series of influence functions, it can be concluded that the regression parameters for age and T5 are both influenced by the position of the outlier, as long as it is within the limits of the values found in the original data set. When the survival of the test outlier is longer than the longest survival time found in the original data set, the estimates of both b_i regression parameters are independent from the position of the outlier. This can be predicted from the method used to estimate the regression parameters (observations are ordered according to their survival

values).

5.2.2 Outliers in age

Procedure: The T5 value of the outlier was kept constant at 0.9. The age was varied from 0 to 600 years in 19 steps. Here also, unlikely age values were given to the outlier so that the behaviour of the model in the presence of extreme outliers could be analyzed. These age values were repeated for five different survival times.

Observations: When the additional data point is an outlier in the age dimension only, both the age and the T5 parameters are influenced by the position of the outlier. When the outlier moves within the range of the original data set, its position has a strong effect on the value of the age regression parameter and a limited effect on the value of the T5 parameter. When the outlier moves towards higher age values, the age parameter tends towards 0, first sharply, then more slowly. In contrast, the value of the T5 coefficient increases slightly, and then decreases slowly towards a constant.

5.2.3 Outliers in T5

Procedure: Here, the age of the outlier was set at 50 years and the T5 value was changed from 0 to 20 in 14 steps. The experiment was repeated for the same five survival times as were used for the age outliers and for the censored and the uncensored situation.

Observations: When the additional data point is an outlier in T5 only, both the age and the T5 parameters are influenced by the position of the outlier. The value of the T5 parameter is strongly influenced by the position of the outlier in the T5 direction, while the age parameter is much less influenced. When the additional data point moves towards higher T5 values and is no longer contained within the range of the original data set, the T5 regression parameter tends first sharply, and then slowly towards zero. In contrast, the age parameter tends towards a constant.

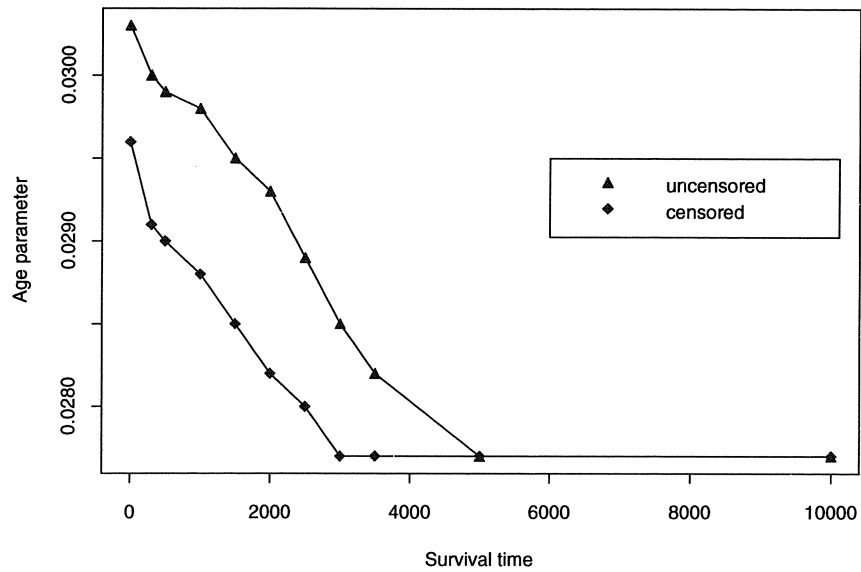


Figure 1: Empirical influence functions of an outlier in survival time (in days) on the age parameter

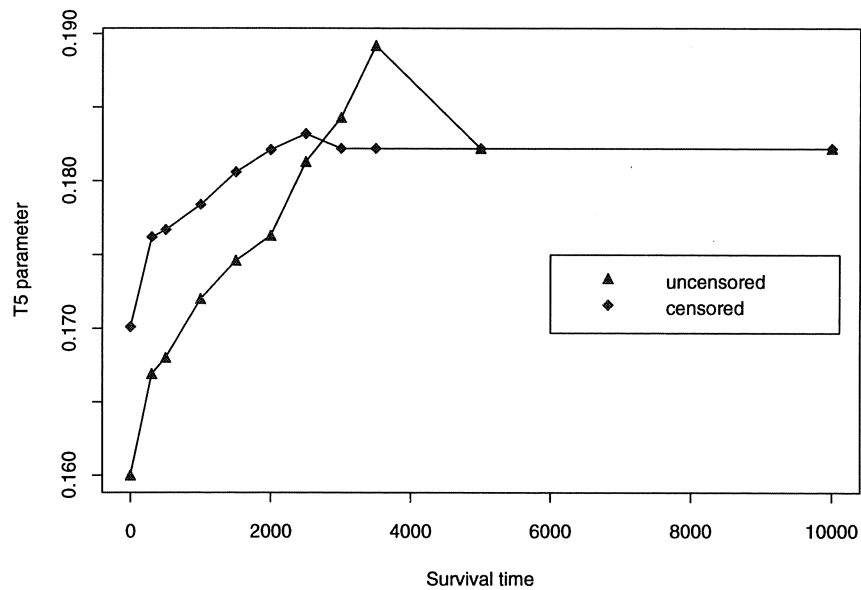


Figure 2: Empirical influence functions of an outlier in survival time (in days) on the T5 parameter

5.3 Outliers in three dimensions

Procedure: For the last series of experiments, the cross-effects between the three dimensions were studied with test outliers located where outliers are the most influential. To do this, the test outlier was given 9 different values of survival times between 0 and 3500 days, 5 different ages between 0 and 300 years, and 5 different T5 values between 0 and 10. Plots are shown in Figure 3.

Observations: The results can be summarized as follows:

a) Most influential positions of the outlier

The survival values for which the positions of the outlier are the most influential are when:

Survival is between 0 and 300 days and

- $0 < \text{age in years} < 200$
- $T5 < 0.9$ or $T5 > 3$

b) Least influential positions of the outlier

The survival values for which the positions of the outlier are the least influential are when:

Survival is less than 3600 days for the uncensored situation (or less than 5000 days in the censored situation) and

- $10 < \text{age in years} < 70$
- $T5 = 0.9$

c) Positions where the effect of the outlier is maximal but where the values of the b_i parameters are independent of outlier position

Survival is greater than 3600 days and

- $\text{age in years} > 200$
- $T5 > 5$

6 General discussion

The objective of this work was to examine the robustness of the Cox regression model in the presence of a test outlier in the sample. On the basis of the empirical functions presented above, three different cases must be distinguished: 1) where the additional data point has variable values close to the mode of each variable; 2) where the additional data point has values which situates it in the periphery of the sample and 3) where the outlier has values far outside of the original sample.

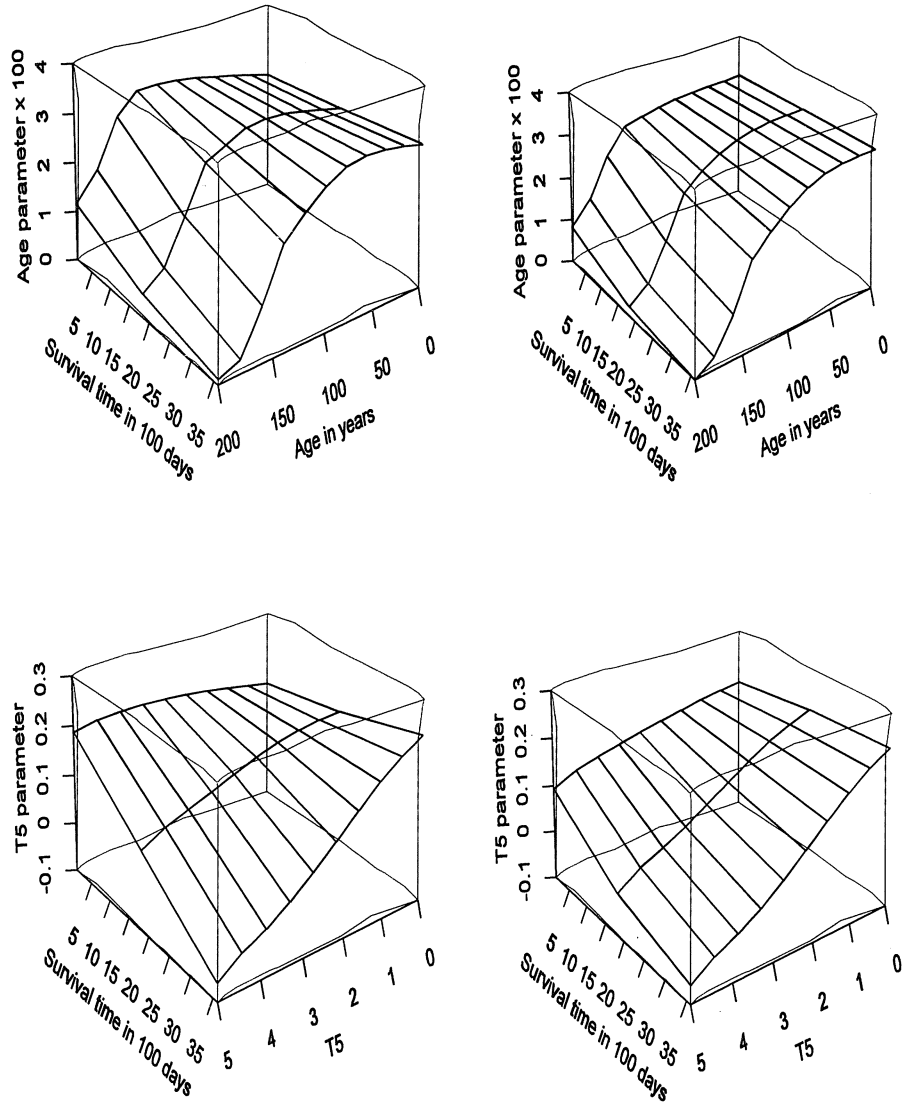


Figure 3: Empirical influence functions for censored and uncensored outliers in age, T5 and survival time

1) *Outlier close to the mode of each variable in the original data set*

In this situation the additional data point is not an outlier. When the additional data point is situated within the original sample, its location influences the estimates of the regression parameters b_i . However, this influence is limited.

2) Outlier in the periphery of the original sample

When the test outliers were positioned in the periphery of the original data set (where their age and T5 values were no more than approximately 2 to 3 times the largest values found in the original sample), their position exerted a maximum influence on the estimates of the regression parameters b_i , and the proportional hazards model was non-robust to the position of the test outlier. An important exception was that the model was totally robust to a test outlier situated outside of the range of survival values found in the original data set.

3) Outlier values much larger (or smaller) than the values found in the original data set

In this case, the effect of the outlier position on the regression parameters was maximal. However, the value of the regression parameters was not dependent on the exact position of the outlier. Thus, the influence functions for the regression parameters are bounded in this area. It must be noted, however, that, even if one single data point is localized in this area, and even if this data point is an outlier in one dimension only, the regression parameter for this variable becomes non-significant.

Acknowledgments: I would like to thank Professor A. Marazzi and his research group for their help with the scientific part of this work. In particular, my thanks go to Mr. G. Braoudakis who always found a solution to my computing problems, to Dr. G. van Melle who helped me with the graphs and to Prof. A. Marazzi himself who corrected the manuscript. A very special thank you goes to Mrs E. Haller who spent many hours teaching me some of the mis(t)ery of SAS, and who wrote the macros.

Résumé: Le but de cet article est l'étude de la robustesse de la régression de Cox (ou modèle de hasards proportionnels) à l'aide de fonctions empiriques. On a ainsi rajouté à un ensemble de données connu une valeur aberrante que l'on a fait varier pour étudier son influence sur les coefficients de la régression. Lorsque la valeur aberrante en question est plus petite que 3 fois la plus grande valeur de l'ensemble de données, sa position influence fortement les estimateurs des paramètres de la régression. Lorsque cette valeur aberrante est plus grande, son influence est maximale mais indépendante de sa position. Ainsi, le modèle n'est pas robuste en présence de valeurs aberrantes extrêmes. Cependant, les fonctions d'influence pour les paramètres de la régression sont bornées.

Mots clés: Données censurées multivariées, modèle de hasards proportionnels, robustesse.

References

- [1] Cox, D.R. (1972), "Regression Models and Life-Tables (with discussion)", *J.R. Stat. Soc. B*, **34**, 187-220.
- [2] Dawson-Saunders, B. and Trapp, R. (1990), *Basic and Clinical Biostatistics*, Prentice-Hall International, London.
- [3] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986), *Robust Statistics: the Infinitesimal Approach*, John Wiley, New York.
- [4] Kalbfleisch, J.D. and Prentice, R.L. (1980), *The Statistical Analysis of Failure Time Data*, John Wiley, New York.
- [5] Miller, R. and Halpern, J. (1982), "Regression with Censored Data", *Biometrika*, **69**, 521-531.
- [6] Reid, N. and Crépeau, H. (1985), "Influence Function for Proportional Hazards Regression", *Biometrika*, **72**, 1-9.

Simpson's Paradox Revisited

Arthur Vogt

Federal Social Insurance Office, Berne, Switzerland

Abstract: Simpson's paradox is originally formulated for probabilities or quotas. It is shown that for one event of the three in the 2x2x2-contingency table, the probabilities can be replaced by two set functions. The ratio of these two set functions is then considered, i.e. a density. One arrives to the framework of generalized index theory and to "Simpson's paradox of the second kind" with densities instead of quotas. Each price index is then a solution of Simpson's paradox. The original "solution" of the paradox corresponds to the price index by Drobisch dates back to 1871. The paradox should be called Drobisch-Simpson's paradox.

Key words: Densities, Drobisch's price index, generalized price indices, quotas.

1 Introduction

The so-called Simpson's paradox till now was mainly treated in medical statistics. Its origin is attributed to Simpson (1951). In section 3.1 the paradox is illustrated with the medical data from Hand (1994). Later in the present article other (fictitious) interpretations of the same data in different application fields are given: in Swiss pension fund statistics and in price statistics.

The easiest application field to understand the paradox is demographic fecundity measurement: let the number of births per woman in country T be higher for every age group than in country S . In spite of that, the total number of births per woman may be higher in country S if there are more elderly women in country T .

Stated in price index terms, the paradox reads: Drobisch II index (8) does not necessarily yield a value between the smallest and the biggest price relative, i.e. it does not fulfil the mean value test (Vogt, 1979, p.76).

As Drobisch in 1871 not only introduced the index (8) which we will call the “pseudosolution” of the paradox but also the solution (11), the paradox should be called *Drobisch-Simpson’s paradox*.

2 Statistical ratios: quotas, densities and indices

In the statistical literature it is a tradition to treat ratios which can be subdivided into quotas, densities and indices. The first chapter of Vogt (1979) investigates them by introducing *statistical masses* S and *submasses* $S_i \subset S$ and *set functions* m and n on them. In the context of price indices n corresponds to the value and m to the quantities. The set function n is a content, i.e. a special set function i on a sample space (Vogt, 1979, p.3). The definition of a *content* reads (Meschkowski, 1965): A content i is a set function on S and its subsets $S_i \subset S$ which has the following properties:

1. $i(S_i) \geq 0$ and $i(\emptyset) = 0$.
2. $S_1 \subset S$ and $S_2 \subset S$ and $S_1 \cap S_2 = \emptyset$ implies $i(S_1 \cup S_2) = i(S_1) + i(S_2)$.

Each statistical mass has the structure of a sample space (Vogt, 1979, p.3). Thus a content can be defined on a statistical mass.

Let S be the set of all commodities in a given economy. Two statistical submasses can be considered, e.g. the set S_1 of the umbrellas and the set S_2 of the watches. A content on S is the value n of the commodities and a set function m on S are the quantities. Contrary to value n , the quantities m are normally not a content: you should not add the numbers of umbrellas and watches, whereas their values in \$ can be added. In other contexts, m is also a content. For instance by treating demographic fecundity indices, the numbers of women in different age classes can be added.

2.1 Quotas

We distinguish two kinds of quotas. Let i be a content on S and S_i . Then a *quota of the first kind of S_i in S related to i* is

$$g_i = \frac{i(S_i)}{i(S)}. \quad (1)$$

For instance let S be the set of all Swiss pension funds and S_i the registered ones. In 1987 the quota of the registered Swiss pension funds in all Swiss pension funds related to their numbers was 27.9% and related to the number of insured persons 85.7%. In 1992 the quota related to the number fell to 25.3% and the one related to the number of insured persons rose to 89.1%.

A quota of the second kind is meaningful for different contents $i_1, i_2, \dots, i_k, i_{k+1}$ with

$$i_{k+1}(T) = \sum_{s=1}^k i_s(T) \quad (2)$$

where $T \subset S$. Then a *quota of the second kind in T of the s^{th} content relative to the $(k+1)^{\text{th}}$ content* is

$$g_s = \frac{i_s(T)}{i_{k+1}(T)}. \quad (3)$$

Quotas of the second kind are used e.g. to express capital placement prescriptions for Swiss pension funds. The quota of real estate is not allowed to be higher than 50% of the balance sheet sum and the quota of stocks not higher than 30%.

2.2 Densities

Quotas can be called *relative weights* whereas densities are *specific weights*, i.e. weights per volume unit. Hereby both “weights” and “volumes” are understood in a generalized way: they can be any contents on the statistical masses.

Let n and m be two contents on a statistical mass S . Then the ratio

$$b(S) = b(S; m, n) = \frac{n(S)}{m(S)} \quad (4)$$

is the *density of n in S relative to m* . The most common density is the population density, i.e. the population content divided by the area content of a country.

Prices are normally not densities. Prices may be called “local densities” because they are defined as “values divided by quantities” for single commodities, i.e. submasses S_i of the statistical mass S . For instance the price of a certain watch equals the value of all the watches of this type divided by the number (the quantity) of these watches. Only in cases where the set function m is a content (see section 3.4) prices are a density, the quantities measured e.g. in kilograms.

The difference between quotas and densities may be slight: shall we speak of the “density of medical doctors in the population” or of their quota? In the first case we have the two contents “number of medical doctors” and “number of inhabitants” whereas in the second case we have the statistical mass “population” and the submass “medical doctors”.

3 Indices

In Vogt (1979, p.11) it is shown that the three kinds of statistical ratios can be described by the given things: for *quotas* of the first kind two statistical masses and a content must be given. For *densities* one statistical mass and two contents are given. For *indices* two statistical masses and two set functions on them must be given. Olt (1995, p.64) shows that most price indices can be represented as quasilinear means of price relatives.

Up to today the author only treated “*indices of densities*”. Now he thinks that “*indices of quotas*” are meaningful too. In a generalized form which relates to both, Laspeyres index can be written as

$$P_{Laspeyres} = \frac{\sum_{i=1}^n \frac{m(S_i)}{m(S)} r(T_i)}{\sum_{i=1}^n \frac{m(S_i)}{m(S)} r(S_i)} \quad (5)$$

with S as the “base mass” and T as the “observed mass”. For an index of densities

$$r(S_i) = \frac{n(S_i)}{m(S_i)} \text{ and } r(T_i) = \frac{n(T_i)}{m(T_i)} \quad (6)$$

and for an index of quotas of the first kind

$$r(S_i) = \frac{m(S_{0,i})}{m(S_i)} \text{ and } r(T_i) = \frac{m(T_{0,i})}{m(T_i)}. \quad (7)$$

In the last formula, $S_{0,i}$ is a submass of S_i and $T_{0,i}$ is a submass of T_i . These submasses may be the unemployed people and (5) compares the unemployment quotas of two countries S and T , respecting the age structure of the populations.

4 Simpson’s paradox and indices

4.1 Simpson’s paradox

The following paradox is attributed to Simpson (1951). In the language of probability it reads (Agresti, 1983):

Even if

$$P(x|y \cap z) < P(x|y \cap z^c)$$

and

$$P(x|y^c \cap z) < P(x|y^c \cap z^c)$$

it is possible that

$$P(x|z) > P(x|z^c).$$

Hand (1994) illustrates the paradox with data of Table 1.

	z			z^c		
	x	x^c	Sums	x	x^c	Sums
y	255	174	429	156	102	258
y^c	88	222	310	82	175	257
Sums	343	396	739	238	277	515

Table 1: 2x2x2-contingency table given by Hand (1994)

Inserting the numbers of the table in the formulas yields (all the numbers should be divided by “the sum of the sums” $739+515=1254$ to get probabilities which sum up to 1):

Although

$$P(x|y \cap z) = \frac{p(x \cap y \cap z)}{p(y \cap z)} = \frac{255}{429} = 0.594 <$$

$$P(x|y \cap z^c) = \frac{P(x \cap y \cap z^c)}{P(y \cap z^c)} = \frac{156}{258} = 0.605$$

and

$$P(x|y^c \cap z) = \frac{P(x \cap y^c \cap z)}{P(y^c \cap z)} = \frac{88}{310} = 0.284 <$$

$$P(x|y^c \cap z^c) = \frac{P(x \cap y^c \cap z^c)}{P(y^c \cap z^c)} = \frac{82}{257} = 0.319$$

the following inequality holds

$$P(x|z) = \frac{P(x \cap z)}{P(z)} = \frac{343}{739} = 0.464 > P(x|z^c) = \frac{P(x \cap z^c)}{P(z^c)} = \frac{238}{515} = 0.462.$$

We now want to give interpretations of the events x, y, z and of their complements, namely the one by Hand (1994) and the other from the Swiss pension fund statistics for a fictitious canton (see Table 2).

Simpson's paradox, expressed by the example of Hand (1994), reads: did the quota of men rise or fall and by how much? In 1975 there were 238 men out of 515 patients, thus the quota was 0.462. In 1970 there were 343 men out of 739 patients altogether, thus a quota of 0.464. Accordingly the quota fell by 0.4%. The analogous reckoning yields for men under 65 and above 65 a rise of the quota by 1.7% and 12.4% respectively. In the same

Event	Hand (1994)	Pension fund statistics
z	Year 1970	Year 1987
z^c	Year 1975	Year 1992
y	Age under 65	Private
y^c	Age above 65	Public
x	Male	Defined benefit plan
x^c	Female	Defined contribution plan

Table 2: Two interpretations for the data of Table 1

way, the paradox reads with the pension fund example: did the quota of benefit plans rise or fall and by how much? In 1992 there were 238 benefit plans of 515 pension funds altogether, thus the quota was 0.462. In 1987 there were 343 benefit plans of 739 pension funds altogether, thus a quota of 0.464. Accordingly the quota fell by 0.4%. The analogous reckoning yields for private and public pension funds a *rise* of the quota by 1.7% and 12.4% respectively.

4.2 Simpson's paradox of the second kind: densities instead of quotas

Simpson's paradox was expressed by probabilities. In this framework quotas as numbers between zero and one fit directly. Now we want to re-interpret the events x and x^c . Instead of probabilities of events we consider a content n for x and a content m for $x \cup x^c$ in a pure formal way. The ratio between these two contents is investigated, i.e. a density according to (4). As the reformulated paradox which we want to call the paradox "*of the second kind*" fits in the price index theory we have replaced the application field "Hand (1994)" by the former. The demographic fecundity problem is given as a third illustration (see Table 3). There it is obvious that we may investigate s instead of only two age groups, i.e. a $2 \times 2 \times s$ -contingency table.

"Event"	Pension fund statistics	Price index problem	Demographic fecundity problem
z	Year 1987	Base situation 0	Base country 0
z^c	Year 1992	Observed situation 1	Observed country 1
y	Private	Commodity 1	Age group 1
y^c	Public	Commodity 2	Age group 2,3,...,s
Content n	Number of insured people in 1000	Value in \$	Number of births
Content m	Number of pension funds	Quantity in kg	Number of women

Table 3: Other interpretations for the data of Table 1

Using the same data as Table 1 the reformulated paradox reads: did the average number of insured people per pension fund rise or fall and by how much? In 1992 there were 238'000 people in 515 pension funds, thus the average was 462. In 1987 there were 343'000 people in 739 pension funds, thus the average was 464. Accordingly the average fell by 0.4%. The analogous reckoning yields for private and public pension funds a *rise* of the average by 1.7% and 12.4% respectively. In the same way, the paradox reads with the price index example: did the prices between the base and the observed situation rise or fall and by how much? In the observed situation, 515 kilograms were sold of both commodities for \$238, thus the price was \$0.462 per kilogram. In the base situation, 739 kilograms were sold of both commodities for \$343, thus the price was \$0.464 per kilogram. Accordingly the price fell by 0.4%. The analogous reckoning yields for the first and the second commodity a *rise* of the price by 1,7% and 12,4% respectively.

The first commodity's price rose by 1.7%, the second's by 12.4%. The question "how much did prices go up altogether?" has to be answered by a price index. The price index that fits Simpson's paradox stems from Drobisch (1871), namely, in the price index notation and the generalized notation

$$P_{DrobischII} = \frac{\sum_{i=1}^s v_i^1 / \sum_{i=1}^s q_i^1}{\sum_{i=1}^s v_i^0 / \sum_{i=1}^s q_i^0} = \frac{\sum_{i=1}^s n(T_i) / \sum_{i=1}^s m(T_i)}{\sum_{i=1}^s n(S_i) / \sum_{i=1}^s m(S_i)} = \frac{n(T)/m(T)}{n(S)/m(S)} \quad (8)$$

with v for values, q for quantities, above index 1 for the observed situation and above index 0 for the base situation, down index $i=1$ for the first commodity and down index $i=2$ for the second commodity. The value is

$$P_{DrobischII} = \frac{(156 + 82)/(258 + 257)}{(255 + 88)/(429 + 310)} = \frac{238/515}{343/739} = 0.996.$$

Thus a price fall by 0.4% although both commodities' prices went up!

4.3 Solutions to the paradox

By solution to Simpson's paradox we understand - illustrated by the example of Hand (1994) - an answer to the question: how much did the quota of men rise? Or with the price index example: how much did prices go up?

In price index theory, the paradox was solved exactly 80 years before Simpson (1951). In 1871, Drobisch proposed not only his price index (8) but also the arithmetic mean of Laspeyres' and Paasche's index (11) (Fisher,

1922, formulas no 52 and 8053). Every price index is a solution to Simpson's paradox.

The statement of Simpson's paradox in price index theory reads: Drobisch's index (8) does not necessarily yield a value between the smallest and the biggest price relative, i.e. it does *not fulfil the mean value test* (Vogt, 1979, p.76). Drobisch's index does not fulfil the proportionality axiom either, which states that if all goods' prices rise by the same factor, the value of the price index should be equal to this factor.

Practically there are as many solutions to Simpson's paradox as there are solutions to the index problem, i.e. an infinity, see e.g. Fisher (1922). Let us just mention the solutions (with densities) of

Laspeyres

$$P_{Laspeyres} = \frac{\sum_{i=1}^s \frac{m(S_i)}{m(S)} \cdot \frac{n(T_i)}{m(T_i)}}{\sum_{i=1}^s \frac{m(S_i)}{m(S)} \cdot \frac{n(S_i)}{m(S_i)}} = \frac{\sum_{i=1}^s m(S_i) \cdot \frac{n(T_i)}{m(T_i)}}{n(S)}, \quad (9)$$

Paasche

$$P_{Paasche} = \frac{\sum_{i=1}^s \frac{m(T_i)}{m(T)} \cdot \frac{n(T_i)}{m(T_i)}}{\sum_{i=1}^s \frac{m(T_i)}{m(T)} \cdot \frac{n(S_i)}{m(S_i)}} = \frac{n(T)}{\sum_{i=1}^s m(T_i) \cdot \frac{n(S_i)}{m(S_i)}}, \quad (10)$$

Drobisch I

$$P_{Drobisch I} = \frac{1}{2}(P_{Laspeyres} + P_{Paasche}) \quad (11)$$

and Fisher

$$P_{Fisher} = \sqrt{P_{Laspeyres} \cdot P_{Paasche}}. \quad (12)$$

The indices Drobisch I and Drobisch II were published in the same year. As (11) is better known, we call it "Drobisch I". We now want to reckon the values of these four indices for the example of Table 1.

$$P_{Laspeyres} = \frac{\frac{429}{739} \cdot \frac{156}{258} + \frac{310}{739} \cdot \frac{82}{257}}{\frac{429}{739} \cdot \frac{255}{429} + \frac{310}{739} \cdot \frac{88}{310}} = \frac{429 \cdot \frac{156}{258} + 310 \cdot \frac{82}{257}}{343} = 1.0446,$$

$$P_{Paasche} = \frac{\frac{258}{515} \cdot \frac{156}{258} + \frac{257}{515} \cdot \frac{82}{257}}{\frac{258}{515} \cdot \frac{255}{429} + \frac{257}{515} \cdot \frac{88}{310}} = \frac{238}{258 \cdot \frac{255}{429} + 257 \cdot \frac{88}{310}} = 1.0516,$$

$$P_{DrobischI} = \frac{1}{2}(1.0446 + 1.0516) = 1.0481,$$

$$P_{Fisher} = \sqrt{1.0446 \cdot 1.0516} = 1.0481.$$

Thus these four price indices register a rise of 4.5%, 5.2% and 4.8% respectively. These rates are - according to the mean value theorem - between 1.7% and 12.4%, the price rises of the two commodities.

4.4 Cases where the set function m is also a content

The (false) opinion is often expressed that the price index problem has no unique solution because the set function “quantity” is not a content. If this would be the case, the demographic fecundity problem with the number of women as content m had an unique solution. Hersch (1934, p.32) wrote according to Vogt (1979, p.10) that one should not standardize the population when comparing two countries' fecundity. Thus he strongly favoured formula Drobisch II (8). But more recent demographs standardize the population, i.e. they use solutions of the paradox and not its “pseudosolution” (8) (see e.g. Vogt, 1979, formula 1.38; Shryock and Siegel, 1973, p.423). In terms of price index theory this means that these demographs are searching fecundity indices which satisfy certain axioms.

The new Swiss wage index (“BFS-Lohnindex”) first published in April 1995 is calculated according to Laspeyres' index (9). Thus it fulfils the mean-value-property and the proportionality axiom.

5 Conclusion: from Simpson's paradox to Drobisch-Simpson's paradox

We want to call Drobisch's index (8) the “*pseudosolution of the paradox*” because it does not fulfil the proportionality axiom. As not only the pseudosolution but also the solution (11) stem from Drobisch, we want to call it “*Drobisch-Simpson's paradox*”. It has to be understood that every solution of the price index problem - not only the four mentioned above - is valid for both the original Drobisch-Simpson's paradox and that of the second kind.

To answer the question: *which is the “correct” solution of Drobisch-Simpson's paradox?*, one has to precise the problem, e.g. by mentioning the axioms its solution should fulfil. There are still questions to which the pseudosolution (8) is the correct answer! For instance, the Swiss social insurance wage index (“AHV-Lohnindex”) is calculated according to this formula (Vogt, 1977, p.289).

Résumé: On rencontre généralement le paradoxe de Simpson dans le cadre du calcul des probabilités ou des quotas; on l'illustre souvent avec une 2x2x2-table de contingence. Dans cet article, on montre que l'on peut remplacer l'un des trois événements de cette table de contingence par deux fonctions d'un ensemble, le ratio de celles-ci étant une densité. On voit alors apparaître un "paradoxe de Simpson de seconde espèce" avec des densités à la place des quotas. En abordant la théorie générale des indices, on s'aperçoit que chaque indice de prix est une solution du paradoxe de Simpson. La solution originale correspond à l'indice de prix de Drobisch, et date de 1871; on pourrait ainsi parler du paradoxe de Drobisch-Simpson.

Mots clés: Densités, indice de Drobisch, indices de prix, quotas.

References

- [1] Agresti, A. (1983), "Fallacies, Statistical", In *Encyclopaedia of Statistical Sciences*, John Wiley and Sons, New York.
- [2] Drobisch, M.W. (1871), "Ueber Mittelgrößen und die Anwendbarkeit derselben auf die Berechnung des Steigens und Sinkens des Geldwertes", *Berichte über die Verhandlungen der königlich sächsischen Gesellschaft der Wissenschaften zu Leipzig, Mathematisch-Physikalische Classe*, **23**, 25-48.
- [3] Fisher, I. (1922), *The Making of Index Numbers*, reprint of the third edition of 1927, Augustus M. Kelley, New York, 1967.
- [4] Hersch, L. (1934), "Taux globaux et taux composites en démographie", *Revue de l'Institut International de Statistique*, **2**, 5-25.
- [5] Meschkowski, H. (1965), *Mathematisches Begriffswörterbuch*, Bibliographisches Institut, Mannheim, Wien, Zürich.
- [6] Olt, B. (1995), "Axiom und Struktur in der statistischen Preisindextheorie", Ph.D. thesis at the University of Karlsruhe.
- [7] Simpson, E. H. (1951), "The interpretation of Interaction in Contingency Tables", *J.R. Statist. Soc. B*, **13**, 238-241.
- [8] Shryock, H. and Siegel, J. et al (1973), *The Methods and Materials of Demography*, US Department of Commerce.
- [9] Vogt, A. (1977), "Eine Verallgemeinerung des Preis- und Mengenindexproblems", Prof. E. Kaiser zum 70. Geburtstag, *Statistische Hefte*, 287-297.
- [10] Vogt, A. (1979), "Das statistische Indexproblem in Zwei-Situationen-Fall", Dissertation at the Federal Institute of Technology, Zürich, written under the direction of Prof. H. Bühlmann and Prof. W.Eichhorn.

Environmental Indices for the United States

Pascal Grosclaude

*Institute for Regional and Economic Research, University of
Neuchâtel, Switzerland*

Abstract: During the past 25 years, the need for a simple and general overview of the relationship between economic activities and the environment has led several countries and international organizations to work on environmental statistics. Two main approaches have been proposed: the first is to link environmental statistics to the system of national accounts (SNA), the second to calculate a limited number of environmental indicators and indices and to relate them to the SNA aggregates. The focus of this article is on the second approach, in particular on the construction of a set of three environmental indices for the United States. The function of these indices is to enable the comparison of environmental performances between geographic areas (i.e., state, region) according to environmental pressure, state and response. The article first surveys the related literature, then describes both the conceptual framework and method and finally presents the results of a pilot study for the United States.

Key words: Green GNP, sustainable development, economic growth, public opinion.

1 Introduction

Since the beginning of the 1970s, the relationship between economic growth and the environment has become a major political issue in both developed and developing countries. While economic growth has been considered essential for the reduction of environmental damage, it has also been regarded as one of the main causes of environmental degradation and depletion. In order to resolve this dilemma and to give policy-makers and decision-takers an effective statistical tool for protecting and managing the environment,

two alternative methods have been proposed.

The first of these methods links environmental statistics to the system of national accounts (SNA) and, in some cases, adjusts the SNA aggregates for environmental change. According to Scherp (1993, p. 19), the function of such a system of integrated economic and environmental accounts would be to "... provide a statistical tool that allows for the assessment as to what impact environmental policy measures would have on society, business sectors and households, and how certain macroeconomic aggregates are affected by resource depletion and environmental degradation". The second method calculates a limited number of environmental indicators or indices and relates them to the SNA aggregates. In this case, the purpose is to "...[monitor] trends of environmental problems, [identify] the relative importance of environmental problems and ... [value] the efficiency of environmental policy measures" (Scherp, 1993, p. 33). For more details see also Den Butter and Verbruggen (1994).

The two methods differ on at least three points. First, both methods have different theoretical foundations. The environmental integration of the SNA may affect the definition of the concepts of production and income and, as a result, require a modification of the theory of national accounts. On the other hand, the calculation of environmental indicators and indices does not require such amendments; in fact, the theoretical basis of environmental indicators and indices is more statistical than economic. Second, the two methods are based on different valuation techniques. Correction of national accounts aggregates requires a monetary valuation of environmental degradation and depletion. Within the SNA, the flows of good and services are usually valued at market price; but in the case of environmental goods and services, such market prices usually do not exist. It may in this case be necessary to use non-market valuation methods, which may finally alter the consistency of the system (United Nations, 1993). In the case of environmental indices, even though some valuation cannot be avoided, the ultimate valuation is left to the users of the data. As a result, the valuation may become more political than economic. Finally, both methods have different purposes. For instance, the environmental adjustment of national accounts seems to be a particularly effective statistical framework for monitoring the economic effects of environmental policy measures and for integrating economic and environmental objectives in policy making. On the other hand, the environmental indices approach is probably more appropriate for assessing environmental trends and identifying environmental problems. Therefore, these two methods should be considered more complementary than mutually exclusive.

This paper focuses on the second approach, in particular on the con-

struction of a set of three environmental indices for the United States: an Environmental Pressure Index (EPI), an Environmental State Index (ESI) and an Environmental Response Index (ERI). The function of these indices is to enable the comparison of environmental performances between two or more geographic areas according to environmental pressure, state and response. The next section surveys the related literature, Section 3 describes both the conceptual framework and method and Section 4 presents the results of a pilot study for the United States.

2 Survey of the literature

The statistical literature usually describes an index as "... a number adapted by its variations to indicate the increase or decrease of a magnitude non susceptible of accurate measurements" (Edgeworth 1925, p. 379). The concept, first used to measure the variations of price level, is now generally considered suitable for the descriptions of both economic and non-economic phenomena, whether these phenomena can be observed directly (e.g. for employment level) or not (e.g. for environmental quality) (Allen, 1975).

In the case of the environment, two types of indices can be differentiated: descriptive and normative indices. According to Adriaanse (1993b, p. 3), descriptive indices are used "... to present a certain state of the environment or related process. Normative [indices] enable an evaluation by comparing a descriptive [index] with a standard". In both cases, the index has three main functions: simplification, quantification and communication. More precisely, indices generally "...simplify in order to make complex phenomena quantifiable in such a way that communication is either enabled or promoted" (Scherp 1993, p. 10).

In practice, the main function of a *descriptive index* is to enable the comparison of at least two environmental situations, either in space or in time. In the first case, the function of the index is to compare the environmental performances of at least two geographical areas (e.g. countries, metropolitan areas). For example, the Green Metro Index developed by the World Resources Institute (1994) ranks the 75 largest metropolitans areas in the US according to eight environmental components. These eight components are first individually ranked, then aggregated using an implicitly weighted sum. In the second case, the function of the descriptive index is to describe the evolution of a given environmental situation through time. An example of such an index is given by Hope, Parker and Peake (1992) for the UK. The pilot index proposed contains nine components and covers the period 1980-1988. Each component takes a value of 100 in 1980. The index corresponds to the explicitly weighted sum - using the results of pub-

lic opinion surveys - of the nine transformed components (see also BAK, 1993; Jesinghaus, 1993).

On the other hand, the function of a *normative index* is to show the distance between a given environmental situation and a subjective reference value. This value can be either a sustainability criterion or simply a desirable future environmental state; in all cases, it might reflect either a social preference (e.g. a legal norm) or a scientific standard. Examples of such indices can be found in Adriaanse (1993a and b); using target values for the year 2000, this author developed a set of seven normative indices for the Netherlands. In each case, the index is calculated by dividing the theme component value by the corresponding policy target.

3 Conceptual framework and method

As stated at the outset of this paper, the focus of this study is on the construction of a set of three descriptive environmental indices for the United States: an Environmental Pressure Index (EPI), an Environmental State Index (ESI) and an Environmental Response Index (ERI). Generally, the function of indices is to provide policy-makers and decision-takers with a limited number of highly aggregated environmental statistics; the function of the proposed indices is specifically to enable the comparison of environmental performances between geographical areas according to environmental pressure, state and response (see Grosclaude, 1995).

The Pressure/State/Response framework has been developed by the OECD as part of its programme on environmental indicators; it is based on the idea that the links between human activities and environment form a causal chain (i.e. a chain of cause/effect relationships) between environmental pressure, state and response. First, *pressure* corresponds to the stresses caused by human activities on the environment or, in other words, to the causes of environmental problems. Second, *state* refers either to the quality of the environment or to the effects of human activities on the environment. Finally, *response* describes the social measures taken to protect or manage environment (see Figure 1).

The construction process of these three environmental indices is divided into five steps:

- 1) selection of environmental themes and components;
- 2) transformation and aggregation of the selected environmental components;
- 3) transformation of the environmental themes;
- 4) aggregation of the environmental themes; and
- 5) calculation of the indices.

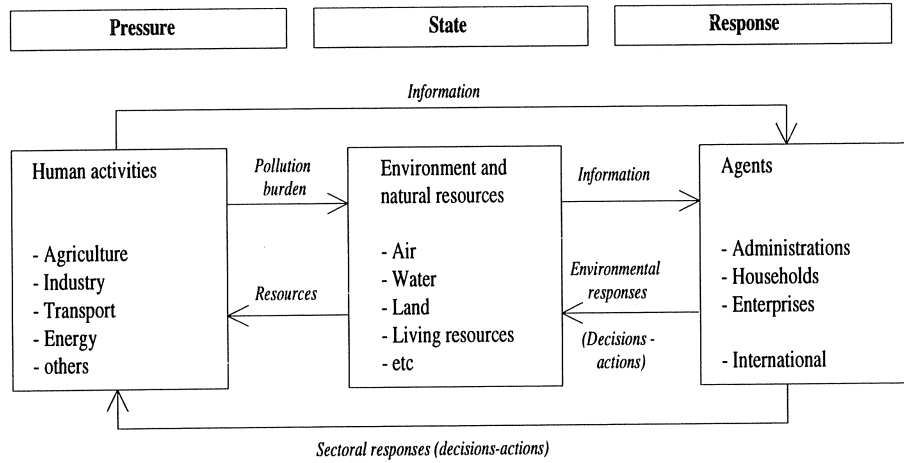


Figure 1: The Pressure-State-Response Framework (Adriaanse, 1993a)

3.1 Selection of environmental themes and components

The selection of environmental themes is based on two different sources: studies on environmental indicators and indices, and environmental surveys. These two sources reflect the three types of legitimacy defined by Eurostat, the Statistical Office of the European Community (Eurostat, 1993; Jesinghaus, 1993): competence, authority and democracy. According to Eurostat, any mixture of these three legitimization types may greatly influence the willingness of the user to accept a particular environmental theme selection.

Based on these environmental studies and surveys, ten environmental themes arise; five of these concern the pressure index, four the state index and one the response index. First, the Environmental Pressure Index distinguishes global from non-global pressures. While global pressures include the emissions of both greenhouse and ozone depleting gases, non global pressures incorporate the emissions of air pollutants, toxic substances and acid rain precursors. Second, the Environmental State Index includes the quality of drinking water, rivers, soil and air. Lastly, the Environmental Response Index is based on state environmental and natural resources expenditures.

The components of each environmental theme are then selected according to representativity, availability and aggregability. Based on these criteria, each environmental component should first give a large or general overview of the corresponding theme, second be available at a reasonable cost, and finally be supported by a relevant aggregation method (see Table 1).

Index	Theme	Component
EPI	<i>Global pressure:</i>	<i>Emissions of:</i>
	1. Ozone depleting gases	CFC
	2. Greenhouse gases	CO ₂ , CH ₄ , CFC
	<i>Non global pressure:</i>	
	3. Air pollutants	CO, SO ₂ , NO _x , VOC
ESI	4. Toxic substances	Toxic substances
	5. Acid rain precursors	SO ₂ , NO _x
	6. Drinking water quality	Population affected
	7. River quality	Rivers and streams impaired
ERI	8. Air quality	Population affected
	9. Soil quality	Hazardous waste sites
	10. Environmental expenditures	State environmental and natural resources expenditures

Table 1: Environmental Themes and Components

3.2 Transformation and aggregation of the selected environmental components

Transformation and aggregation of the selected environmental components are necessary as soon as there are more than one component per theme; affected by this process are the emissions of greenhouse gases, air pollutants and acid rain precursors.

The emissions of greenhouse gases (i.e. the emissions of CO₂, CH₄ and CFC) are transformed according to their respective global warming potentials and expressed in CO₂ equivalents; one CO₂ equivalent corresponds to one ton of CO₂, 0.047619 ton of CH₄ or 0.00017 ton of CFC (IPCC, 1990; WRI, 1992). Then the transformation of air pollutant emissions (i.e. emissions of CO, SO₂, NO_x and VOC) is based on the individual toxicity of each pollutant and expressed in CO equivalent; one CO equivalent is equal to 1 ton of CO, 0.01 ton of SO₂, 0.005 ton of NO_x or 0.005 ton of VOC (PLANCO Consulting, 1993). Finally, the emissions of acid rain precursors (i.e. the emissions of SO₂ and NO_x) are converted according to their respective acidification potentials; they are expressed in acid equivalent. One acid equivalent corresponds to 1 ton of SO₂ or 0.6957 ton of NO_x (Adriaanse, 1993 a and b). Based on these transformations, the components of each environmental theme are then aggregated into a single value, equal to the sum of the transformed components.

3.3 Transformation of the environmental themes

Once each theme is expressed as a single (aggregated) value, the next step in the indices construction process is the transformation of the environmental

Theme	Unit	Year	Source	Weight
1.	kilos per capita	1988	US EPA, 1990	0.19
2.	kilos per capita	1988	US EPA, 1990	0.13
3.	kilos per capita	1992	US EPA, 1993	0.28
4.	kilos per capita	1991	US EPA, 1991b	0.25
5.	kilos per capita	1992	US EPA, 1993	0.15
6.	% of total population	1992	NRDC, 1993	0.26
7.	% of assessed km	1992	US EPA, 1994	0.26
8.	% of total population	1991	US EPA, 1991a	0.25
9.	# by km ²	1992	US DoC, 1993	0.23
10.	% of total state expenditures	1991	CSG, 1993	1.00

Table 2: Environmental Themes: weights

themes. The function of this transformation is to standardize the units of measurement of the various environmental themes. In this case, the conventional method standardization is the Z transformation (Liu, 1976).

The Z transformation is based on a linear transformation of the original data, such that the mean of the Z score is equal to zero and its standard deviation is equal to one. Formally, for any US State i and environmental theme j , the standardized score Z_{ij} is measured by:

$$Z_{ij} = \frac{X_{ij} - \bar{X}_j}{S_j} \quad (1)$$

where X_{ij} is the original value taken by theme j for State i , $\bar{X}_j$ is the mean of theme j and S_j is its standard deviation.

3.4 Aggregation of the environmental themes

Since all environmental themes now have the same normalized unit of measurement, the standardised aggregation method "... is simply to add ... the weighted Z scores with weights being assigned to each of the variables separately" (Liu, 1976, p. 84). To be more specific, the aggregated value A_i , for any US State i , is given by:

$$A_i = \sum_{j=1}^J w_j \cdot Z_{ij} \quad (2)$$

where w_j is the weight assigned to environmental theme j . The weight assigned to each environmental theme is equal to the proportion of people who are the most concerned about that theme. According to the Gallup Poll Service (1990), 65 % of the American public personally worries a great deal about pollution of drinking water, 64 % about pollution of rivers, lakes, and reservoirs, 63 % about contamination of soil and water by toxic waste,

58 % about air pollution, 43 % about damage to the earth's ozone layer, 34 % about acid rain and 30 % about greenhouse effect or global warming. Based on these results, weights have been calculated for each theme (see last column of Table 2).

3.5 Calculation of the indices

Like the transformation of the environmental themes, the calculation of the indices is based on the Z transformation method. Therefore, every index I_i is the result of a linear transformation of the aggregated values A_i , such that the mean of the indices I_i is equal to zero and their standard deviation is equal to one. Index I_i is given by:

$$I_i = \frac{A_i - \bar{A}}{S} \quad (3)$$

where $\bar{A}$ is the mean of the aggregated values and S their standard deviation.

4 Results

Based on the conceptual framework and method presented above, a pilot set of three environmental indices has been developed for the United States; it is composed of an Environmental Pressure Index (EPI), an Environmental State Index (ESI) and Environmental Response Index (ERI) (Table 3 by Region; Table 4 by State).

The *Environmental Pressure Index* (EPI) - which takes into account the emissions of ozone depleting gases, greenhouse gases, toxic substances, air pollutants and acid rain precursors - is calculated so that the value of the index is minimal if the environmental pressure is also minimal; in other words, the smaller the index is, the smaller the environmental pressures are. The index for the United States, which corresponds to the mean, is equal to zero. The State which has the smallest index score ranks first (Figure 2 by Region; Figure 3 by State). The *Environmental State Index* (ESI) groups four themes: quality of drinking water, rivers and streams, soil, and air. Note that, according to the method used here, the smaller the index, the better the state of the environment (Figure 4 by Region; Figure 5 by State). Finally, the *Environmental Response Index* (ERI) is based on the State environmental and natural resources expenditures. The State with maximal ERI (i.e. with maximal ratio between environmental and natural resources expenditures and total expenditures) ranks first (Figure 6 by Region; Figure 7 by State).

	EPI		ESI		ERI	
	value	rank	value	rank	value	rank
East North Central	0.23	(7)	-0.13	(2)	-0.40	(8)
East South Central	0.97	(9)	-0.06	(4)	-0.42	(9)
Middle Atlantic	-0.70	(1)	0.17	(7)	-0.34	(7)
Mountain	0.08	(6)	0.07	(6)	0.59	(2)
New England	-0.46	(2)	0.40	(8)	-0.28	(6)
Pacific	-0.06	(3)	0.87	(9)	0.91	(1)
South Atlantic	-0.04	(4)	-0.49	(1)	-0.25	(5)
West North Central	0.03	(5)	-0.12	(3)	0.30	(3)
West South Central	0.70	(8)	0.05	(5)	-0.23	(4)
United States	0.00		0.00		0.00	

Table 3: EPI, ESI and ERI by Region (early 1990s)

	EPI		ESI		ERI	
	value	rank	value	rank	value	rank
Alabama (AL)	0.77	(39)	-1.05	(3)	-0.93	(50)
Alaska (AK)	-0.24	(21)	2.66	(50)	3.99	(1)
Arizona (AZ)	-0.04	(28)	0.79	(37)	-0.69	(46)
Arkansas (AR)	0.00	(30)	-0.62	(12)	-0.17	(27)
California (CA)	-0.87	(4)	1.68	(46)	0.84	(10)
Colorado (CO)	-0.48	(14)	-0.35	(20)	1.62	(4)
Connecticut (CT)	-0.36	(17)	0.65	(35)	-0.79	(49)
Delaware (DE)	-0.16	(22)	1.15	(42)	0.33	(16)
Florida (FL)	-0.71	(9)	-0.49	(16)	0.03	(22)
Georgia (GA)	-0.04	(28)	-0.49	(16)	-0.39	(37)
Hawaii (HI)	-0.83	(5)	-0.48	(18)	-0.59	(42)
Idaho (ID)	0.05	(32)	2.51	(49)	1.92	(3)
Illinois (IL)	-0.16	(22)	0.92	(39)	0.14	(20)
Indiana (IN)	1.67	(47)	-1.04	(4)	-0.75	(47)
Iowa (IA)	0.02	(31)	0.12	(26)	-0.47	(39)
Kansas (KS)	0.31	(35)	0.03	(25)	-0.33	(35)
Kentucky (KY)	0.89	(44)	-0.82	(8)	-0.30	(33)
Louisiana (LA)	2.53	(48)	1.20	(43)	0.46	(14)
Maine (ME)	-0.47	(15)	-0.61	(13)	0.16	(19)
Maryland (MD)	-0.66	(11)	-0.68	(9)	-0.24	(29)
Massachusetts (MA)	-0.45	(16)	1.40	(44)	-0.30	(33)
Michigan (MI)	-0.15	(24)	-1.38	(1)	-0.52	(40)
Minnesota (MN)	-0.30	(18)	-0.19	(22)	0.24	(17)
Mississippi (MS)	0.85	(43)	1.02	(40)	0.03	(22)
Missouri (MO)	0.17	(33)	-0.16	(24)	1.21	(7)
Montana (MT)	0.84	(40)	-0.25	(21)	0.88	(9)
Nebraska (NE)	-0.09	(27)	0.43	(28)	-0.12	(25)
Nevada (NV)	-1.39	(1)	1.14	(41)	0.13	(21)
New Hampshire (NH)	-0.52	(13)	-0.19	(22)	0.53	(13)
New Jersey (NJ)	-0.72	(8)	2.28	(48)	0.23	(18)
New Mexico (NM)	0.30	(34)	0.77	(36)	-0.26	(30)
New York (NY)	-0.97	(2)	-0.88	(6)	-0.54	(41)
North Carolina (NC)	0.84	(40)	-0.89	(5)	-0.68	(45)
North Dakota (ND)	1.36	(46)	0.50	(31)	1.26	(6)
Ohio (OH)	0.42	(37)	0.17	(27)	-0.78	(48)
Oklahoma (OK)	-0.25	(20)	0.80	(38)	-0.60	(43)
Oregon (OR)	-0.93	(3)	0.56	(33)	0.74	(11)
Pennsylvania (PA)	-0.28	(19)	-0.67	(10)	-0.37	(36)
Rhode Island (RI)	-0.60	(12)	0.55	(32)	-0.29	(31)
South Carolina (SC)	0.41	(36)	-1.06	(2)	-0.18	(28)
South Dakota (SD)	-0.67	(10)	0.43	(28)	0.68	(12)
Tennessee (TN)	1.27	(45)	-0.57	(15)	-0.29	(31)
Texas (TX)	0.52	(38)	-0.65	(11)	-0.39	(37)
Utah (UT)	0.84	(40)	0.62	(34)	0.34	(15)
Vermont (VT)	-0.76	(7)	0.46	(30)	1.32	(5)
Virginia (VA)	-0.10	(26)	-0.84	(7)	-0.15	(26)
Washington (WA)	-0.81	(6)	1.41	(45)	0.89	(8)
West Virginia (WV)	2.58	(49)	-0.40	(19)	-0.63	(44)
Wisconsin (WI)	-0.15	(24)	-0.59	(14)	-0.03	(24)
Wyoming (WY)	4.05	(50)	2.23	(47)	3.76	(2)
United States	0.00		0.00		0.00	

Table 4: EPI, ESI and ERI by State (early 1990s)

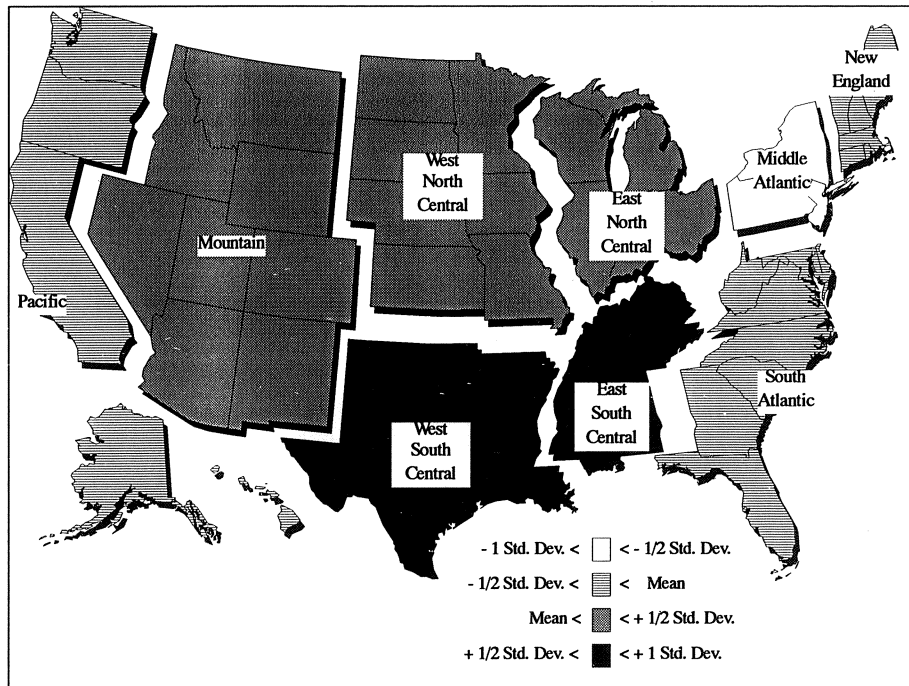


Figure 2: EPI by Region

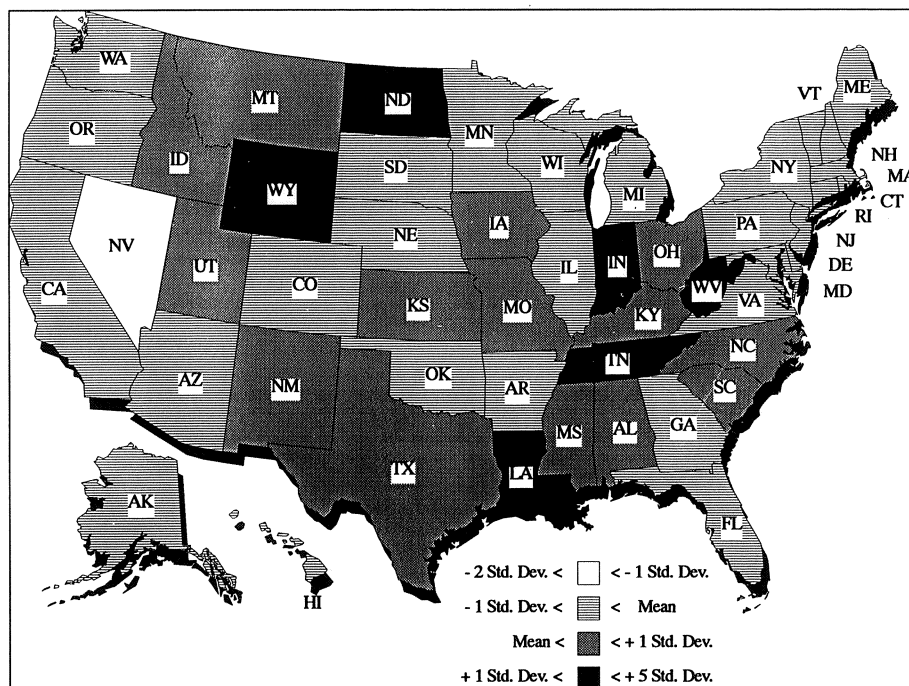


Figure 3: EPI by State

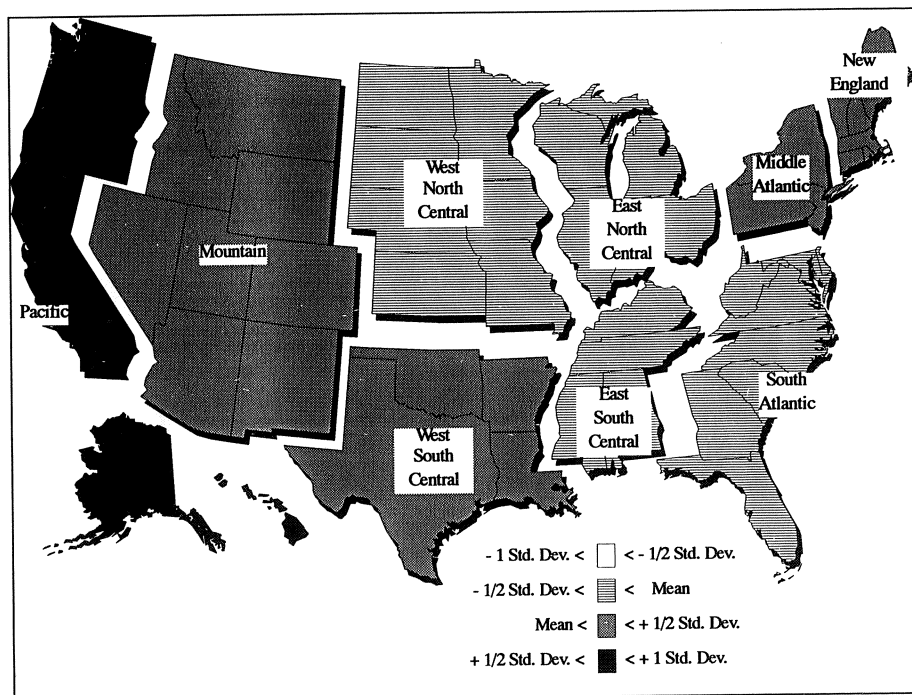


Figure 4: ESI by Region

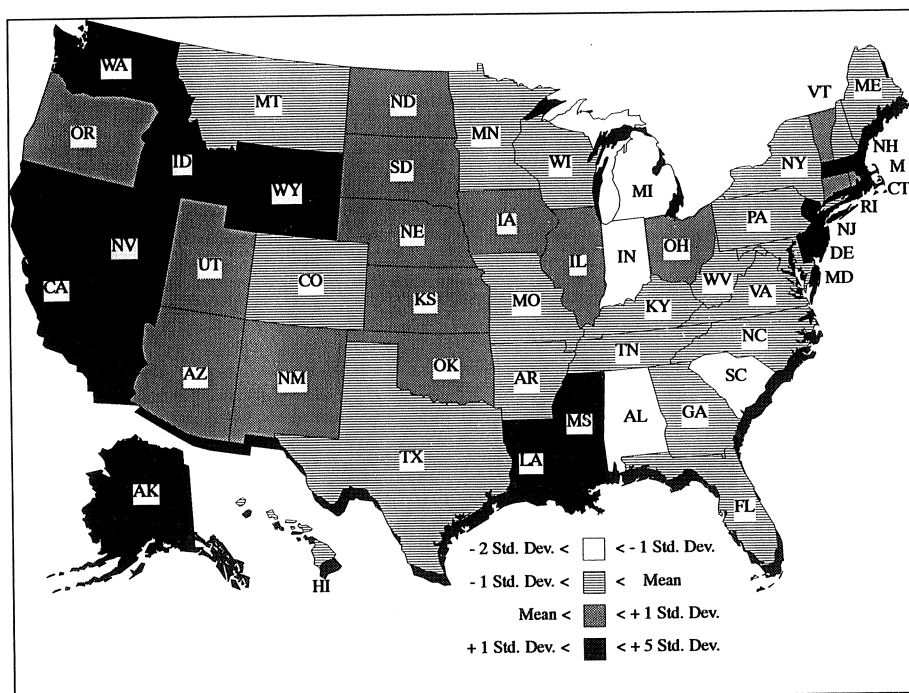


Figure 5: ESI by State

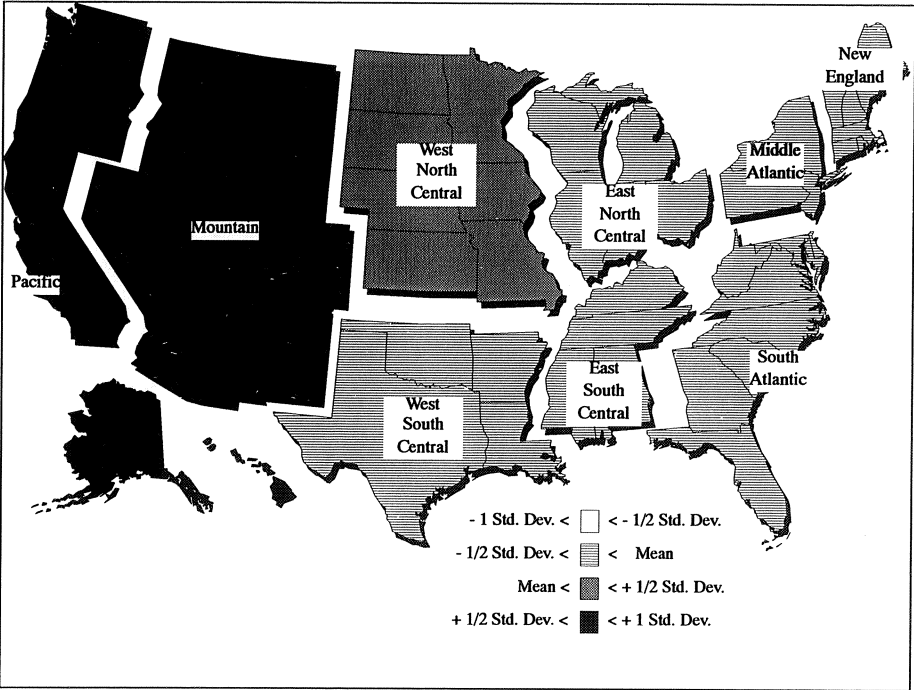


Figure 6: ERI by Region

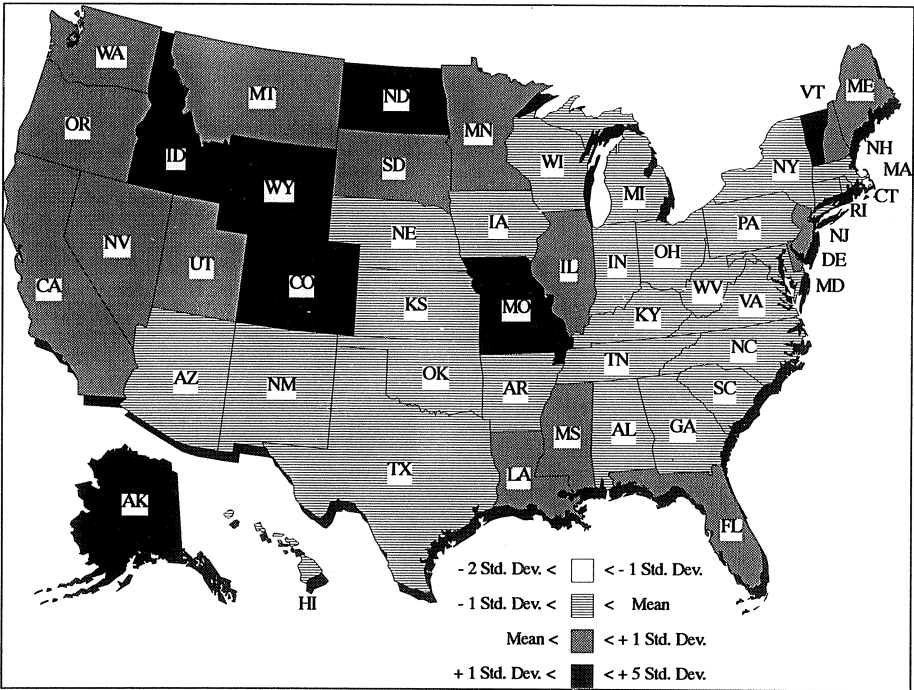


Figure 7: ERI by State

5 Conclusion

The focus of this paper has been on the construction of a set of three environmental indices for the United States: an Environmental Pressure Index (EPI), an Environmental State Index (ESI) and an Environmental Response Index (ERI). The aim of these indices is to give policy-makers, decision-takers and the general public concise information about environmental pressure, state and response. In order to facilitate the use of this information in the policy-making process, special attention has been given to the transparency of the indices and to their theoretical and empirical foundation.

Compared with economic and social indicators, environmental indices are less developed and their direct impact on decision-taking remain uncertain. However, the need for a simple and general overview of the environmental pressure, state and response has recently led several countries and international organisations to work on environmental indices. Consequently, the next step will be to secure a minimum level of harmonisation between those works in order to enable the integration of economic, social and environmental statistics.

Acknowledgments: At the time of writing, I was a visiting research assistant at the Department of Agricultural and Resource Economics, University of Maryland at College Park. I acknowledge useful comments by W. Brown, M. Nerlove, V. Taxdal and two anonymous referees.

Résumé: Durant les vingt-cinq dernières années, la nécessité d'une meilleure description des liens qu'entretiennent l'économie et l'environnement a conduit de nombreux pays et organisations internationales à s'intéresser aux statistiques d'environnement. Deux approches principales ont été proposées; la première repose sur la description comptable des phénomènes d'environnement, la seconde sur la construction d'un nombre limité d'indicateurs ou d'indices d'environnement. Cet article s'intéresse à la seconde de ces approches, en particulier à l'élaboration d'un ensemble de trois indices d'environnement pour les Etats-Unis. Ces indices ont pour objet de permettre la comparaison des performances environnementales du point de vue de la pression, de l'état ou de la réponse.

Mots clés: Croissance économique, développement soutenable, PNB vert, sondages d'opinion.

References

- [1] Adriaanse A. (1993a), *Environmental Policy Performance Indicators*, Ministry of Housing, Physical Planning and Environment (VROM), The Hague.
- [2] Adriaanse A. (1993b), "The Development of Environmental Policy Performance Indicators (EPPI's) in the Netherlands", mimeo, Ministry of Housing, Physical Planning and Environment (VROM), The Hague.
- [3] Allen R. G. D. (1975), *Index Numbers in Theory and Practice*, Aldine Publishing Company, Chicago.
- [4] BAK (1993), "Wirtschaft und Umwelt - Nordwestschweiz", BAK, Basel.
- [5] CSG (1993), *Resource Guide To State Environmental Management*, third Edition, The Council of State Governments (CSG), Lexington.
- [6] Den Butter F. A. G. and Verbruggen H. (1994), "Measuring the Trade-Off Between Economic Growth and a Clean Environment", *Environmental and Resource Economics*, 4, 187-208.
- [7] Edgeworth F. Y. (1925), "The Plurality of Index Numbers", *Economic Journal*, 35, 379-388.
- [8] Eurostat (1993), "Environmental Indicators and Accounting", Internal working papers F3, Eurostat, Luxembourg, October 21, 1993.
- [9] Gallup Poll Service (1990), "Americans Strongly in Tune with the Purpose of Earth Day", The Gallup Organization, Princeton, April 11, 1990.
- [10] Grosclaude P. (1995), *Comptabilité nationale et environnement*, Institut de Recherches Economiques et Régionales, Editions EDES, Neuchâtel.
- [11] Hope C, Parker J. and Peake S. (1992), "A Pilot Environmental Index for the UK in the 1980s", *Energy Policy*, April 1992, 335-343.
- [12] IPCC (1990), *Climate Change: the IPCC Scientific Assessment*, Houghton J.T., Jenkins G.J. and Ephraums J.J. eds, Cambridge University Press, Cambridge.
- [13] Jesinghaus J. (1993), "Environmental Problem Lists: A Methodology Proposal", Internal working papers F3, Eurostat, Luxembourg, April 27, 1993.
- [14] Liu B.-C. (1976), *Quality of Life Indicators in U.S. Metropolitan Areas - A Statistical Analysis*, Praeger Publishers, New York.
- [15] NRDC (1993), *Think Before You Drink - The Failure of the Nation's Drinking Water System to Protect Public Health*, Natural Resources Defense Council (NRDC), Washington D.C.

- [16] PLANCO Consulting (1993), "Environmental Impacts of Transport Infrastructure Projects: Evaluation Methods Used for the German Federal Transport Investment Plan", paper presented at the International Workshop on Quantitative Methods for the Environmental Impact Assessment of Land Transport, Paris, December 1993.
- [17] Scherp J. (1993), "Accounting for the Environment in Statistical Information Systems", paper presented at the fifth National Accounting Colloquium of the Association de Comptabilité Nationale, Paris.
- [18] United Nations (1993), "Integrated Environmental and Economic Accounting", United Nations, New York.
- [19] US DoC (1993), "Statistical Abstract of the United States 1993", US Department of Commerce, Washington D.C.
- [20] US EPA (1990), "National- and State-Level Emissions Estimates of Radiatively Important Trace Gases (RITGS) from Anthropogenic Sources", US Environmental Protection Agency, Washington D.C., Office of Research and Development, Washington D.C.
- [21] US EPA (1991a), "Numbers of People Living in Counties Not Meeting National Ambient Air Quality Standards (NAAQS) in 1991", In *Resource Guide To State Environmental Management*, third Edition, The Council of State Governments (CSG), Lexington.
- [22] US EPA (1991b), "1991 Toxics Release Inventory", In *Resource Guide To State Environmental Management*, third Edition, The Council of State Governments (CSG), Lexington.
- [23] US EPA (1993), "National Air Pollutant Emission Trends, 1900-1992", US Environmental Protection Agency, Office of Air Quality, Research Triangle Park.
- [24] US EPA (1994), "National Water Quality Inventory", 1992 Report to Congress, US Environmental Protection Agency, Office of Water, Washington D.C.
- [25] World Resources Institute, WRI (1992), *World Resources 1992-1993*, World Resources Institute, Washington D.C.
- [26] World Resources Institute, WRI (1994), *Environmental Almanac*, World Resources Institute, Washington D.C.

Random Thoughts on Survival

Sir David R. Cox

Nuffield College, Oxford, United Kingdom

These notes are written immediately after teaching a short course on Survival Analysis to Masters students at Neuchâtel.

It is always a particular pleasure and challenge to teach away from one's own institution, the challenge arising in part because uncertainty about the background knowledge that can reasonably be assumed, a key element in the planning of any course. The experience at Neuchâtel was notable especially for the impressive variety of subject field of specialization, experience and interests of the students. This meant that a key issue was not, as so often, that of providing motivation, but rather that of finding a level of treatment that would enable topics close to the forefront of contemporary research to be addressed without making unreasonable demands on the students' theoretical and mathematical knowledge. Of course, it is hard to judge how successfully this was accomplished.

There are two extreme strategies that may be used in selecting material for such a course. One is to choose, perhaps rather arbitrarily, two or three topics and to develop these carefully in some depth so that hopefully some understanding of principles is achieved. The other, which I find personally more appealing both when I am a student at other people's lectures, and as a teacher, is to range more broadly and to omit detail.

This raises the following general issue that arises in teaching any topic that is the focus of much research. Our subject has reached the state where in many fields the literature seems overwhelming. In survival analysis not only are there a considerable number of books, some technically very demanding, but also a very large number of research papers, some wholly theoretical, others concerned at least in part with specific applications, but often also containing new methodological points. Further new papers appear virtually daily.

Yet somehow the implicit message, which I believe to be correct, must

be conveyed that this volume of work is not to be thought overwhelming, that it should be regarded as evidence of the openness of the field and that by examining carefully new applications, novel work of value can be done. Put differently, there are a limited number of key principles and the rest is detail, important in its proper context, but not usually needed. Thus a key issue in each of the specialized areas of statistics is to separate essential principles from detail.

Much current research in statistics is highly technical; look for example at any issue of *The Annals of Statistics*. It is sometimes argued: how much easier research must have been all those years ago when simple and important results could be obtained by elementary methods.

For myself, I very much doubt the correctness of this view, to some extent from personal experience, but more importantly from thinking about the history of key developments in our subject.

Take, for example, the Cramér-Rao inequality. A simple elegant very brief proof is available requiring only the Cauchy-Schwarz inequality, itself provable by totally elementary methods. Does this mean that the pioneers had an easy task? It would be interesting to know Professor Rao's view about this. Certainly the earlier work of Aitken and Silverstone used the mathematically sophisticated ideas of the calculus of variations and was much more difficult.

The tentative conclusion I draw is that statistics was and remains a field in which new contributions can be made from all sorts of backgrounds ranging from those with powerful mathematical expertise to those who have new ideas and formulations arising fairly directly out of specific applications. The intellectual vigour of the subject relies on there being a balanced mixture and thus on our subject attracting people with a wide range of interests. One of the strengths of the Neuchâtel course is precisely that.

Comments on Sir David's Random Thoughts on Survival

C.R. Rao

Department of Statistics, The Pennsylvania State University, United States

Sir David's random thoughts on research in statistics are, indeed, thought provoking. Statistics is not a separate discipline like Mathematics, Physics, Chemistry or Biology. It has no problems of its own unlike the other disciplines. Its status may be described as the art, logic or technology of solving problems in other disciplines.

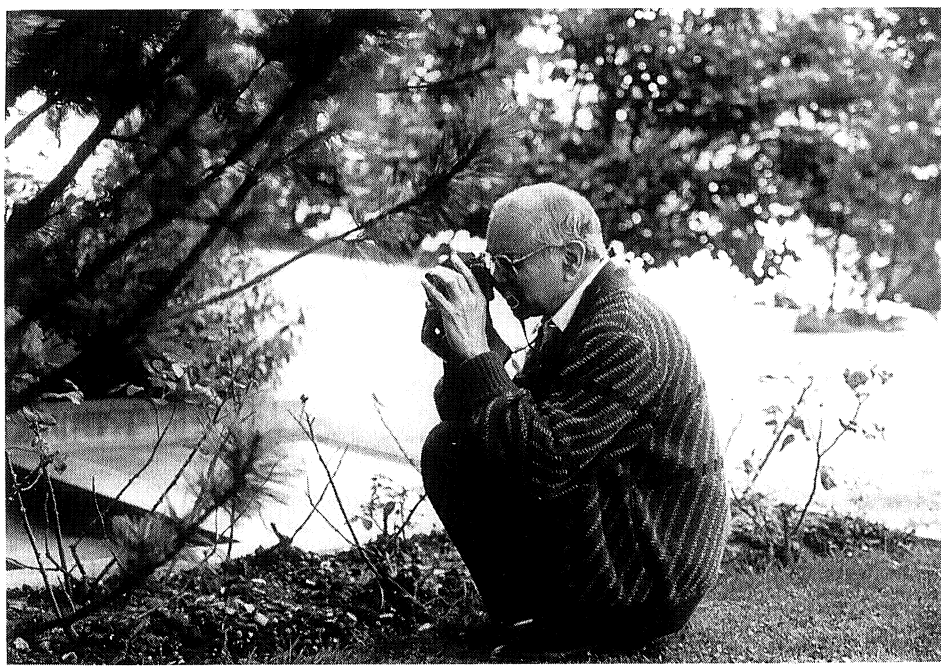
Once a problem in any area is formulated, it is necessary to acquire data to provide information to solve the problem. This is where the role of statistics come in. How do we generate data? What is the best way of extracting the relevant information from the data? In what form do we present the solution to the problem reflecting the amount of uncertainty in the available information? This statistical approach involves designing the collection of data, stochastic modeling of the data and developing efficient methods for the estimation of unknown parameters, testing of hypothesis and making decisions. The appropriate methodology for solving the problem on hand may already be available in statistical literature. Otherwise, new methods have to be developed, where mathematics come in. Sometimes the mathematics involved in forging new statistical tools may be simple and at times it can be complicated. I agree with Sir David that own subject exists and expands by interaction with other disciplines and collaboration with competent mathematicians.

It is amazing that some problems which appear complicated can be resolved using simple and elegant mathematics. A typical example is the disproof of Euler's conjecture on the nonexistence of orthogonal Latin squares of order $4n + 2$ by using combinatorial arrangements of orthogonal arrays. In my own experience, it was a great surprise for me that useful results like Cramér-Rao inequality and Rao-Blackwell theorem could be proved with

simple mathematics.

Sometimes, our initial feeling that a problem is complicated may be a deterrent in seeking for a simple solution. In the days of R.A. Fisher, the concentration of research was on small sample theory. Attempts were made to find elegant and closed form solutions to problems. Now the emphasis seems to be more on asymptotic theory and methods which Tukey describes as “asymptotitis”. Asymptotic theory is usually highly technical as it depends on advanced concepts of probability, which is not easy to understand without the knowledge of advanced probability.

The character of research in statistics is also constantly changing with advances in computing power. It is quite possible that computer intensive sample reuse methods that are being currently advocated will de-emphasize to some extent research in statistical theory.



C.R. Rao in Search of Truth, 1989

From the Editor's Notebook

On the bias of s and s^2

1. Let $\mathcal{E} = \{x_1, \dots, x_n\}$ be a sample of observations randomly drawn from a population $\mathcal{P}$. From the sample we estimate the mean μ and the variance σ^2 of the population. While the usual estimator for μ is $\bar{x} := \sum x_i/n$, the most common estimator for σ^2 is surprisingly

$$s^2 := \frac{\sum (x_i - \bar{x})^2}{n - 1}.$$

This result is certainly the first surprise that meets a student in Statistics. He or she would certainly prefer n to $(n - 1)$ as the denominator of the estimator. The reason for this curious choice is that s^2 is an unbiased estimator of σ^2 . We effectively have

$$E(s^2) = \sigma^2.$$

2. While s^2 is an unbiased estimator of σ^2 , it is not the case for the estimator $s := \sqrt{s^2}$ of the standard deviation σ of the population. We have indeed (as long as the variance of the population is not 0)

$$E(s) = \sqrt{\sigma^2 - \text{Var}(s^2)} < \sigma.$$

Since one could have more interest in the standard deviation than in the variance, it is perhaps useful to mention that the estimator s of σ is biased: σ is underestimated!

3. The estimator s^2 is biased too if the size N of the population is finite (and if the sample is drawn without replacement)! We then have

$$E(s^2) = \sigma^2 \cdot \frac{N}{N - 1}. \tag{1}$$

If we want an unbiased estimator of σ^2 , we therefore have to define

$$s_*^2 := s^2 \cdot \frac{N - 1}{N}.$$

4. Let us now examine how these estimators are implicated in the calculation of the variance of $\bar{x}$. The formula in the case of a finite population is

$$Var(\bar{x}) = \sigma_x^2 = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}. \quad (2)$$

As we usually don't know σ^2 we have to replace it in this formula by the unbiased estimator s_*^2 . Thus we obtain an unbiased estimator of σ_x^2

$$s_x^2 := \frac{s_*^2}{n} \cdot \frac{N-n}{N-1} = \frac{s^2}{n} \cdot \frac{N-n}{N}.$$

We effectively have

$$E(s_x^2) = \sigma_x^2. \quad (3)$$

It is perhaps useful to recall this point too since one could hesitate in the calculation of σ_x^2 between the two "corrector's factors" $(N-n)/(N-1)$ and $(N-n)/N$. We precise here that one has to use the first one when σ^2 is known, and the second one when σ^2 is unknown and replaced by the usual estimator s^2 .

5. We finally illustrate these points with the following example: let $\mathcal{P} = \{1, 2, 3\}$ be a population of size $N = 3$, of mean $\mu = 2$ and of variance $\sigma^2 = 2/3$. From this population it is possible to draw 3 different samples of size $n = 2$. They are $\mathcal{E}_1 = \{1, 2\}$, $\mathcal{E}_2 = \{1, 3\}$ and $\mathcal{E}_3 = \{2, 3\}$. From these samples we obtain the following estimators:

$$\begin{aligned} \bar{x}_1 &= 1.5, & \bar{x}_2 &= 2, & \bar{x}_3 &= 2.5 \\ s_1^2 &= 0.5, & s_2^2 &= 2, & s_3^2 &= 0.5 \end{aligned}$$

To illustrate (1), we calculate the mean of the s_i^2 :

$$E(s^2) = \frac{0.5 + 2 + 0.5}{3} = 1 = (2/3) \cdot \frac{3}{2} = \sigma^2 \cdot \frac{N}{N-1}.$$

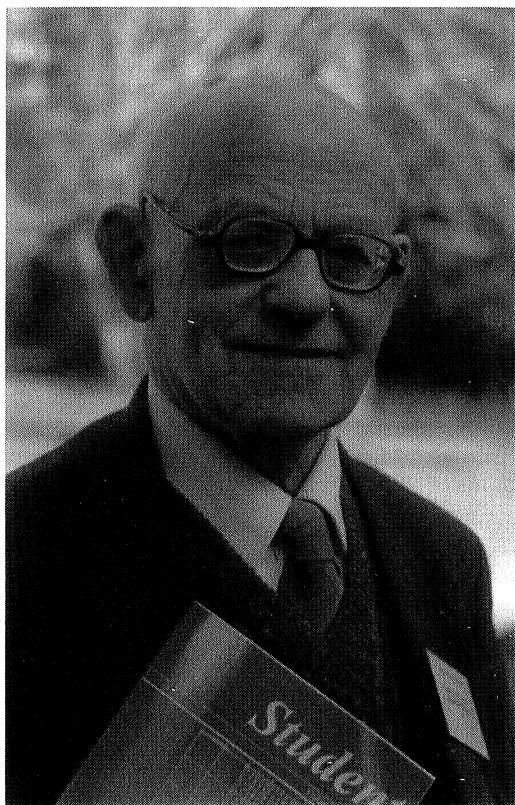
To illustrate (2), we calculate the variance of the $\bar{x}_i$:

$$Var(\bar{x}) = \frac{(1.5 - 2)^2 + (2 - 2)^2 + (2.5 - 2)^2}{3} = \frac{1}{6} = \frac{(2/3)}{2} \cdot \frac{1}{2} = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}.$$

To illustrate (3), we calculate the mean of the $s_i^2 \cdot (N-n)/(n \cdot N)$:

$$E(s_x^2) = \frac{0.5 + 2 + 0.5}{3} \cdot \frac{3-2}{2 \cdot 3} = \frac{1}{6} = \sigma_x^2.$$

Henry E. Daniels Sets the Record Straight



Henry Ellis Daniels was born in London on the second of October, 1912. He graduated from the University of Edinburgh, then went to Cambridge to continue his studies. After receiving his Ph.D. from the Edinburgh University, he returned to Cambridge as a lecturer in Mathematics in 1947. In 1957, Henry Daniels became the first professor in Mathematical Statistics at the University of Birmingham and continued in this post until his retirement in 1978. Among his awards he received the Guy Medal of the Royal Statistical Society, in silver in 1947 and in gold in 1984. In 1980, Professor Daniels was elected a fellow of the Royal Society of London.

On March 20, 1995, Henry E. Daniels was invited to deliver the opening lecture at the workshop on New Approximation Techniques for Statistical Inference, in Ascona, Switzerland. Daniels provided the audience with some remarks on the history of the saddle point approximations and the influence of his 1954 paper on Linnik's work on asymptotic expansions for stable distributions. Forty-one years after, Daniels informs us why he wrote *saddle point* as *saddlepoint*. To preserve the precious hand-writing of great statisticians, we took the opportunity to furnish the readers of *Student* with the manuscript of his Ascona lecture with his kind permission.

SADDLEPOINT APPROXIMATIONS IN STATISTICS¹

By H. E. DANIELS

University of Cambridge and University of Chicago

1. Introduction and summary. It is often required to approximate to the distribution of some statistic whose exact distribution cannot be conveniently obtained. When the first few moments are known, a common procedure is to fit a law of the Pearson or Edgeworth type having the same moments as far as they are given. Both these methods are often satisfactory in practice, but have the drawback that errors in the "tail" regions of the distribution are sometimes comparable with the frequencies themselves. The Edgeworth approximation in particular can assume negative values in such regions.

Perhaps I should start with a little personal history. I have often been asked why I spell "saddlepoint" as one word. The reason is that in the original manuscript of my 1954 paper "saddlepoint" appeared as a typist's error, but the editor of the *Annals of Mathematical Statistics* either didn't notice or didn't mind. So I decided to leave it as it matched up rather nicely with the German form *Sattelpunkt* which also appears as one word. Little did I realise the confusion and imitation it would subsequently cause to authors and editors, and I now take the opportunity to apologize to all concerned.

HED

Forty-one years after his 1954 paper, Henry E. Daniels explains why he wrote saddle point as saddlepoint.

Saddlepoint approximations - some
historical remarks and a conjecture
about "nearly exact" cases

H.E. Daniels

Ascona 1995

Abstract: After some general remarks on SP approximations, the influence of my 1954 paper on the work of Russian probabilists is indicated - in particular Richter's local limit theorems and Linnik's work on stable distributions; Demouchal's critique of the use of the asymmetrical stable distribution of index α for a test of heavy tails is related to its critical asymptotic behaviour as $\alpha \rightarrow 2$.

A class of densities with "nearly exact" SP approximation is conjectured from a number of special cases.

Saddlepoint approximation

Integral with large parameter

$$f(x) = \int e^{n\psi(x,T)} g(T) dT, \quad n \text{ large.}$$

T complex. Contour integral.

Simpler case: $T = t$, real.

Laplace approximation.

$$f(x) = \int_a^b e^{n\psi(x,t)} g(t) dt. \quad \text{Real } t.$$

max $\psi(x,t)$ at $t = \hat{t}$ where $\psi'(\hat{t}) = 0$

If $a < \hat{t} < b$, expand $\psi''(\hat{t}) < 0$

$$\psi(x,t) \sim \psi(x, \hat{t}) - \frac{1}{2} [-\psi''(x, \hat{t})] (t - \hat{t})^2$$

$$f(x) \sim \sqrt{\frac{n}{2\pi(-\psi''(x, \hat{t}))}} g(\hat{t}) e^{n\psi(x, \hat{t})} \left\{ 1 + O\left(\frac{1}{n}\right) \right\}$$

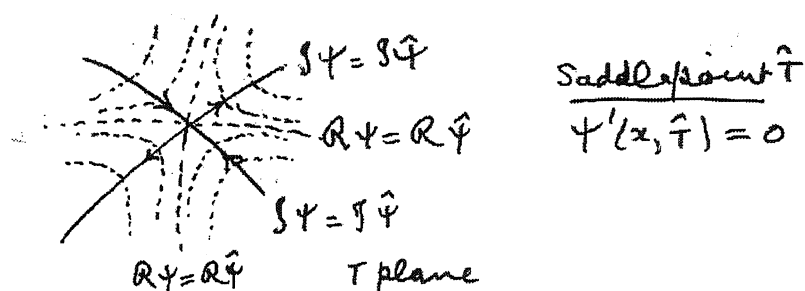
General case T complex

Method of steepest descent.

- (1) Locate suitable saddlepoint of $\psi(x, T)$ in complex plane
- (2) choose contour as steepest descent curve through SP (locally)

Riemann (1863)

Debye (1909) 2.



→ $R\psi(x, \tau)$ descends.

— Curves of constant $S\psi(x, \tau)$ through $\hat{T}$
 are curves of steepest descent of $R\psi(x, \tau)$
 (ie of $|e^{\psi(x, \tau)}|$)

.... level curves $R\psi(x, \tau) = \text{constant}$
 choose branch of $S\psi = S\hat{\psi}$ descending
 into the valley for the contour
 (locally). Then use Laplace
 approximation.

H.E.D. (1954) S.P. approx: for the mean.

Density $f(x)$.

$$\text{MGF } M(T) = e^{K(T)} = \int_{-\infty}^{\infty} e^{Tx} f(x) dx$$

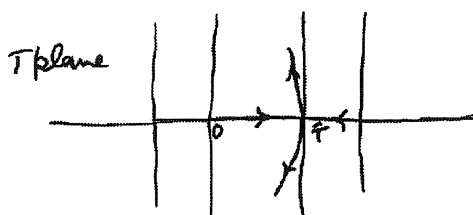
converges for $R(T)$ in open interval containing 0. (if $f(x)$ decreases at least exponentially as $|x| \rightarrow \infty$)

$$X_1, X_2, \dots, X_n \text{ i.i.d. } \bar{X} = \frac{1}{n}(X_1 + \dots + X_n)$$

$$\sum_{j=1}^n X_j \text{ has MGF } M^n(T) = e^{nK(T)}$$

Fourier inversion for density of $\bar{X}$:

$$f_n(\bar{x}) = \frac{n}{2\pi i} \int_{c-i\infty}^{c+i\infty} e^{n[K(T) - T\bar{x}]} dT$$



On real axis $K(T) - T\bar{x}$ is convex

($K''(T) > 0$). So has a minimum

at saddlepoint $\hat{T}$ where

$$K'(\hat{T}) = \bar{x}$$

choose c to be $\hat{T}$.

4.

Then $K(\tau) - \tau \bar{x}$ has a maximum
on the steepest descent curve orthogonal
to real axis.

Write $\tau = \hat{\tau} + iu$

$$K(\tau) - \tau \bar{x} = K(\hat{\tau}) - \hat{\tau} \bar{x} - \frac{1}{2} u^2 K''(\hat{\tau}) + \dots$$

$$f_n(\bar{x}) = \sqrt{\frac{n}{2\pi K''(\hat{\tau})}} e^{n[K(\hat{\tau}) - \hat{\tau} \bar{x}]} \\ \cdot \left\{ 1 + \frac{1}{n} \left(\frac{1}{8} \lambda_4(\hat{\tau}) - \frac{5}{24} \lambda_3^2(\hat{\tau}) \right) + O\left(\frac{1}{n^2}\right) \right\}$$

where $\lambda_r(\hat{\tau}) = K^{(r)}(\hat{\tau}) / [K''(\hat{\tau})]^{r/2}$

Dominant term is saddlepoint approx.

Properties:

- ① 3 exact cases. Normal, Gamma, Inverse normal
- ② Often, relative error uniformly bounded over whole range of $\bar{x}$.

Alternative approach, ("Tilting")

Escher (1932), Gamér (1938),
Khinchin (1945), Barndorff-Nielsen
and Cox (1979)

5.

use of conjugate density (Khintin)

ie embed $f(x)$ in an exponential family

$$\begin{aligned} f(x, \lambda) &= C e^{\lambda x} f(x) \\ &= e^{\lambda x - K(\lambda)} f(x) \end{aligned}$$

MGF is

$$M(T, \lambda) = e^{K(T+\lambda) - K(\lambda)}$$

$$K(T, \lambda) = K(T+\lambda) - K(\lambda)$$

$$\text{mean } K'(0, \lambda) = K'(\lambda)$$

Expand $f(x, \lambda)$ in Edgeworth series.

choose $K'(\lambda)$ to be observed $\bar{x}$, ie $\lambda = \hat{\lambda}$

Then odd terms in Edgeworth series vanish
we get SP. expansion.

Large deviations.

Cramér used this to get results
on $1 - F_n(\bar{x})$ in the far tail.

(These are a disguised form of the
SP results for tail probabilities)

6.

Russian work following H.E.D. (1954)

(1) Richter (1957), Linnik (1961).

Used SP approach as most natural way of deriving local limit theorems for densities corresponding to Cramér's large deviation results for $1 - F(x)$.

See remark by Gnedenko (1960). B.K. Symp.

(2) Stable distributions.

Linnik (1961): SP. method fills a gap in asymptotics of stable dist.ⁿs

Resumé. (book by Zolotarev)

Limiting dist.ⁿ of

$$Z_n = \frac{x_1 + x_2 + \dots + x_n}{n^{1/\alpha}} - A_n$$

$x_1, \dots, x_n$ i.i.d. $0 < \alpha \leq 2$

Moments. when $\alpha < 2$,
 $E(Z^r) = \infty$ for $r > \alpha$ ("fat tails")

Has "domains of attraction".

Exploited by Mandelbrot (1960) ... for econometric data.

Characteristic function (in standard form)

$$\phi(t; \alpha, \beta) = e^{-|t|^\alpha} e^{-\frac{1}{2} i \pi \beta (1 - |1 - \alpha|) \operatorname{sgn} t} \quad (\alpha \neq 1)$$

$$= e^{-|t|^\alpha \left\{ 1 - i \beta \frac{2}{\pi} \log |t| \operatorname{sgn} t \right\}} \quad (\alpha = 1)$$

$$0 < \alpha \leq 2, \quad -1 \leq \beta \leq 1.$$

Special cases:

$$\alpha = 2 \quad \text{Normal}$$

$$\alpha = 1 \quad \beta = 0 \quad \text{Cauchy}$$

$$\alpha = \frac{1}{2} \quad \beta = -1 \quad \text{Inverse Normal.}$$

$$\beta = \pm 1 : \text{One sided } x \geq 0 \text{ or } x \leq 0$$

Symmetric stable distributions $\beta = 0$

$$\phi(t, \alpha) = e^{-|t|^\alpha} \quad 1 < \alpha \leq 2$$

Density

$$p(x, \alpha) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{itx - |t|^\alpha} dt$$

Asymptotic expansion for large x

8.

Bergstrom (1952)

$$p(x, \alpha) \sim \frac{1}{\pi} \sum_{r=1}^{\infty} \frac{(-1)^{r-1}}{x^{r\alpha+1}} \frac{\Gamma(r\alpha+1)}{r!} \sin \frac{1}{2} r \pi \alpha$$

$$1-F(x, \alpha) \sim \frac{1}{\pi} \sum_{r=1}^{\infty} \frac{(-1)^{r-1}}{x^{r\alpha}} \frac{\Gamma(r\alpha)}{r!} \sin \frac{1}{2} r \pi \alpha$$

Dominant term

$$p(x, \alpha) \sim \frac{1}{\pi} \frac{\Gamma(\alpha+1)}{x^{\alpha+1}} \sin \frac{1}{2} \pi \alpha$$

$$1-F(x, \alpha) \sim \frac{1}{\pi} \frac{\Gamma(\alpha)}{x^{\alpha}} \sin \frac{1}{2} \pi \alpha$$

"Pareto-like" tail

But when $\alpha = 2$ all terms vanish

ie $p(x, 2) \rightarrow 0$ as $x \rightarrow \infty$ faster than any power of $1/x$

Neglected exponentially small term takes over. Evaluated

as in HSD (1954) (Exact)

(Same thing happens in one-sided cases when $\beta = -1$, $1 < \alpha < 2$, $x \geq 0$
or $\beta = +1$, $0 < \alpha < 1$, $x \leq 0$.)

9.

Anomalous behaviourNear $\alpha = 2$ put $2 - \alpha = \epsilon$

$$1 - F(x, 2 - \epsilon) \sim \frac{1}{\pi} \sum_{r=1}^{\infty} \frac{\Gamma(2r - 2\epsilon)}{r!} \cdot \frac{\sin \frac{1}{2} r \pi \epsilon}{x^{2r - 2\epsilon}}$$

When ϵ is small, first few terms are small, but when r approaches $1/\epsilon$ the terms are not small.

Statistical application.

DuMouchel (1983) . Test for $\alpha < 2$ against $\alpha = 2$
is test for "fat tail"

Sample of n , $x_1, x_2, \dots, x_n$

$$\text{Score } \sum_{j=1}^n \frac{1}{p_j} \frac{\partial p_j}{\partial \alpha} = \frac{\partial L}{\partial \alpha}$$

calculated from Bergstrom series,

Distⁿ of ML estimator $\hat{\alpha}$ becomes
non regular as $\alpha \rightarrow 2$

Woodruffe (appendix to DuMouchel): As $n \rightarrow \infty$,

$$P\left(\frac{\partial L}{\partial \alpha} < 0 / \alpha = 2\right) \rightarrow 1 \text{ ie } P(\hat{\alpha} < 2 / \alpha = 2) \rightarrow 0$$

So expect good discrimination when n is large.

10.

But

$$P(\hat{\alpha} < 2 | \alpha = 2) \sim \frac{\kappa}{\log n} \text{ (conjecture)}$$

So need very large n to discriminate

Required:

A transitional asymptotic form
for β as $\alpha \rightarrow 2$

For references, see

Ibragimov and Linnik. "Independent
and stationary sequences of random variables"

Zolotarev. "One dimensional stable
distributions".