S.Johansen N.Keiding S.Lauritzen

## Discussion of the Lectures Presented at the Sixth Nordic Conference on Mathematical Statistics Lofthus, Norway, 27-30 May 1975



## CONTENTS

| Søren Johansen: | Comments to Peter Bickels lecture on "Robust Statistical Methods". |
| :---: | :---: |
| Steffen L. Lauritzen: | Comments to Peter Bicke1s lecture on <br> "Robust Statistical Methods". |
| Sфren Johansen: | Comments on Erik Torgersens lecture on <br> "Comparison of Experiments". |
| Niels Keiding: | Discussion of Jan M. Hoem's lectures <br> "Mathematical Statistics in Demography". |

Sфren Johansen<br>May 1975

Comments to Peter Bickels lecture on
"Robust Statistical Methods".

Peter Bickel mentioned in his lecture two citations concerned with the role of the model. The first one by Bessel who wanted that one should always believe in ones model and the second by Wilks who expressed that one should never believe in a mathematical mode1. These points of view are somewhat extreme and do not illiqtrate the usefulness of a mathematical model. I would like to propose that the model should be used as a tool in the analysis of the data and should express ones understanding of the structure of the variation of the data.

As a final comment to the question of the model I should like to draw your attention to the following quotation by Marc Kac [1]:
"Models are, for the most part, caricatures of reality, but if they are good, then, like good caricatures, they portray, though perhaps in distorted manner, some of the features of the real world. The main role of models is not so much to explain and to predict - though ultimately these are the main functions of science - as to polarize thinking and to pose sharp questions."

The next thing I want to comment on is the model - supermodel formulation. It seems useful to think of the variations of the data in terms of systematic and stochastic variation, and to think of the role of the model builder to find out how the variations should be classified into these groups.

It seems that the model - supermodel formulation is an attempt to avoid expressing any point of view on the structure of the random variation and some very precise structures for the systematic variation. If this is the case then the robust estimation of the variance does not have a meaning since the variance describes some aspect of the random variation and is then qualitatively different from the systematic components of the model.

Peter Bickel mentioned the situation that a man comes up with some data and wants the statistician to estimate the "centre" of the data. Bickel took that as a motivation for starting a study of robust estimation of this "centre". My final comment will be concerned with this situation and will be a citation from Lewis Carroll: Alice in Wonderland [2] where she talks with the Cheshire Cat:
"Cheshire Puss," she began, rather timidly, as she did not at all know whether it would like the name: however, it only grinned a little wider.
"Come, it's pleased so far", thought Alice, and she went on. "Would you tell me, please, which way I ought to go from here?"
"That depends a good deal on where you want to get to," said the Cat.
"I don't much care where -" said Alice.
"Then it doesn't matter which way you go," said the Cat.
"- so long as I get somewhere", Alice added as an explanation.
"Oh, you're sure to do that," said the Cat, "if you only walk long enough."

## References:

[1] Kac, M.: Some mathematical models in Science. Science 166, 695-699 (1969).
[2] Carroll, Lewis: Alice in Wonderland and through the looking-glass. Jan Förlag Stockholm (1945) p. 79.

I am sorry to say that I believe the importance of robustness studies to be extremely overestimated. I do not find it useful. The important errors that one should be afraid of when you work as a practical statistician are those that turn black into white, truth into lie, i.e. errors of reasoning and not the "gross errors" that occur in data due to wrong typing or measurements. It is therefore very important that we, the theoreticians, try to give tools to the "working" statisticians that are simple and thus give them a chance to fight the logical disasters.

We would all be a lot better off if all the brilliant brains working in the field of "how to behave optimally" in stead would concentrate their efforts on a more modest task: to avoid a few of the disasters in applied statistics that has brought our subject in miscredit among scientists.

S申ren Johansen
May 1975
Comments on Erik Torgersens lecture on
"Comparison of Experiments".

I should like to express my thanks to Erik Torgersen for presenting this rather complicated and technically difficult topic and to congratulate him with a sizeable and important contribution to the understanding of the theory of comparison of experiments.

It is difficult to discuss mathematics but one may discuss the purpose of the mathematics and sometimes the means.

My comments will concern three aspects: design of experiments, sufficiency and deficiency, and may in places be comments on the notes [2] rather than the lecture itself.

1. Design of experiments. It is possible to split up this field into "possibłe design" and "optimal design", the first part consists of the theory of randomized blocks, factorial designs etc., the other consists of optimal allocation of regression points and similar problems.

The topic of Torgersens lecture can be thought of as coming in the second group. To fix ideas let $E$ be the experiment or the model consisting of independent normally distributed randomvariables $U_{1}, \ldots, U_{n}$ such that

$$
E U_{i}=\alpha+\beta t_{i} \quad \text { and } \quad V U_{i}=\sigma^{2}, \quad i=1, \ldots, n .
$$

Similarly $F$ consists of the same type of variables except that $v_{1}, \ldots, v_{n}$ has been used as observation points. The question naturally arises which experiment is best for inference concerning $\alpha$ and $\beta$.

If the criterion is that

$$
V(\hat{\alpha}+\hat{\beta} t)
$$

should be minimum, then one can show, see Kiefer and Wolfowitz [1], that $E$ is better than $F$ if and only if

$$
\operatorname{det} A^{\prime} A \geqq \operatorname{det} B^{\prime} B,
$$

where $A$ is the design matrix for $E$ and $B$ the design matrix for $F$.

Since

$$
V\binom{\hat{\alpha}}{\hat{\beta}}=\sigma^{2}\left(A^{\prime} A\right)^{-1}
$$

we obviously have that if we want precise estimates of $\alpha$ and $\beta$ then $E$ is better than $F$ if

$$
A^{\prime} A \geqq B^{\prime} B
$$

in the natural order of positive definite matrices.

It is a beautiful and convincing result that has been proven by Torgersen and Haarvad Hansen [4] that $E$ is better than $F$ for any decision problem if and only if the above inequality holds for the design matrices.
2. Sufficiency. It seems to the reader of the lecture notes by Torgersen and Lindquist [2] that sufficiency is used as an excuse for introducing a lot of concepts and for proving a lot of theorems. It appears to be a bad excuse and I should like to substantiate this somewhat more.

It is true that the natural mappings from $\left\{P_{\theta}, \theta \in \theta\right\}$ to $\left\{Q_{\theta}, \theta \in \theta\right\}$ are the stochastic transformations (or Markov Kerne1s) M such that

$$
P_{\theta} M=Q_{\theta}, \quad \theta \in \theta
$$

It is a surprising and beautiful result due to Blackwell that the decision theoretic definition of "being more informative than" is equivalent to the Markov

Kernel criterion under certain regularity conditions.

In the discussion of sufficiency only the order $\geqq$ and not the deficiency $\Delta$ is used and its motivation must be found in a different place. One is left with the feeling after reading the notes that one has been forced through a lot of hard work around $\Delta$ without any reward.

It would be natural that also ancillarity was discussed in this connection. If $E=\left\{P_{\theta}, \theta \in \theta\right\}$ and $T: X \rightarrow Y$ then we can define $E^{t}=\left\{P_{\theta}^{t}, \theta \in \theta\right\}$, where $P_{\theta}^{t}$ is the conditional distribution given $T(X)=t$.

The following might be true

$$
E^{t} \approx E \quad \forall t \Leftrightarrow T \text { ancillary }
$$

It would also be interesting with a discussion of partial sufficiency, M - anci1larity etc., see Barndorff-Nielsen [6].

What is missing is also a discussion of approximate sufficiency, where $\Delta$ and the notion of deficiency comes in in a natural way.

In the notes by Torgersen on local comparison of experiments [3] he considers an experiment $\left\{P_{\theta}, \theta \in \Theta\right\}$ and evaluate the error in replacing this by the dichotomy $\left\{P_{\theta_{0}}, \frac{d}{d \theta_{0}} P_{\theta_{0}}\right\}$. It seems natural to define $t(x)=\frac{d}{d \theta_{0}} \ln f(x, \theta)$, where $f(x, \theta)$ is the density of $P_{\theta}$. Then

$$
f(x, \theta)=f\left(x, \theta_{0}\right) e^{t(x)\left(\theta-\theta_{0}\right)}\left(1+o\left(\theta-\theta_{0}\right)\right)
$$

and in this way we can ask what happens, if we replace the family $\left\{\mathrm{P}_{\theta}\right\}$ by the exponential family generated by $P_{\theta_{0}}$ and $t$. Locally we can replace a differentiable family by a family allowing a sufficient data reduction. One can then by means of $\Delta$ investigate the distance locally between the two families. This fits nicely with the work of Bradley Effron [5] who considers a differentiable subfamily of an exponential family and defines a notion of curvature which measures
the distance between the families. It would be interesting to compare the curvature with $\Delta$.
3. The deficiency. The third important result is also due to Blackwell and consists in representing $\left\{\mathrm{P}_{\theta}, \theta \in \Theta\right\}$ by the standard experiment in such a way that convergence of experiments is equivalent to weak convergence of a measure on a simplex. This theorem makes it very precise which topology one is working with and the theorem ought to be generalized as far as possible in order to allow for realistic experiments.

In my opinion the results about the topology are of primary importance and the results about $\Delta$ only of secondary importance. The choice of $\Delta$ is somewhat arbitrary, other norms can be introduced in the very definition of $\varepsilon$-deficiency, thereby leading to other distance measures than $\Delta$. It would be nice to see the discussion of sufficiency and approximate sufficiency carried further and less emphasize be placed on compatations of $\Delta$ in concrete cases.
4. Other problems. I should like to suggest some further problems that may be of interest in comparing experiments. The first problem is that the parameter set should be identical. It seems natural that if one has a model with a regression $\alpha+\beta t$ and another with $\alpha+\beta t+\gamma t^{2}$, then one should still be allowed to ask which experiment is best for inference about $\alpha$ for instance. If one has independent normally distributed variables with mean $\zeta$ and variance $\sigma^{2}$, one can compare this experiment with independent repetitions from a normal distribution with mean $\eta$ and variance $\sigma^{2}$ and then ask which experiment is best. This last example is of course quite easy but is just mentioned as an example.

The second problem issthe following: Take a stochastic process $\left\{X_{t}, t \in T\right\}$ where the distribution depends on $\theta$. Let $T$ and $T^{\prime}$ be stopping rules and let $E$ be $\mathrm{X}_{\mathrm{T}}$ and $F$ be $\mathrm{X}_{\mathrm{T}}$. Which of these is more informative about $\theta$ ? Can one compare sampling plans in this way? If $E=\left\{\mathrm{X}_{\mathrm{t}}, \mathrm{t} \in \mathrm{T}_{1}\right\}$ and $F=\left\{\mathrm{X}_{\mathrm{t}}\right.$, $\left.\mathrm{t} \in \mathrm{T}_{2}\right\}$ where $\mathrm{T}_{1}$
and $\mathrm{T}_{2}$ are subsets of T one can ask the following question: Which experiment is most informative for $\theta$ and for the remainder of the process, i.e. one wants to estimate $\theta$ and predict the process on the set where one has not observed it.

As a conclusion I should like to add that the work should not be "motivated" by its practical applicability. It is a piece of mathematical statistics which developes concepts for the analysis of fundamental ideas like sufficiency and which hopefully leads to a clarification of these ideas. On the other hand I think that one can see that in places the author becomes more interested in mathematics than in mathematical statistics so, although I agree on the purpose of the study, I am sceptical about some of the means.

## References:

1 Kiefer, J. and Wolfowitz, J. (1959) : Optimum designs in regression problems. Ann. Math. Stat., 30, 271-294.

2 Torgersen, E. and Lindquist, B. (1975): Notes on comparison of statistical experiments. University of Oslo. 167 pp .

3 Torgersen, E. (1972): Local comparison of experiments when the parameterset is one-dimensional. University of Oslo. 153 pp.

4 Torgersen, E. and Hansen, O.H. (1974): Comparison of linear normal experiments. Ann. Statist. Vo1. 2, No. 2, 367-373.

5
Effron, B. (1975): Defining the curvature of a statistical problem (with applications to second order efficiency). (To appear in The Annals of Statistics).

Discussion of Jan M. Hoem's lectures<br>"Mathematical Statistics in Demography"

1. I should like to start by thanking Jan Hoem for a very clear exposition of mathematical statistics in demography. We know that Jan over the years has done considerable efforts to clear up concepts in current use in demography and actuarial science and explain these concepts in a language natural to students of mathematical statistics. In the following I should like to ask some questions concerning even further clearing up. Of course most of the questions have only been made possible by the work done already.

However, let me start by quoting a very interesting recent example of observations of survival data with concomitant information, namely the Stanford Heart Transplant Project. [I then outlined the results of Clark et al. (1971), Turnbull, Brown and Hu (1974) and Crowley and Hu (1974). Other recent analyses of these data are by Mantel and Byar (1974), Brown, Hollander and Korwar (1974) and Miller (1974).]
2. An important part of the exposition was concerned with what you term analytical graduation. Let me first ask some questions regarding the motivation for analytical graduation and the interpretation of the results.

To be specific, consider the application of the Hadwiger function for fertility graduation. This function seems to have no interpretation in terms of an underlying stochastic mechanism and will thus have to be considered purely descriptive. Aside from the suitability for fitting the empirically observed fertilities, the applicability of a graduation function should be judged from its ability to represent interesting structures in larger sets of data (such as Coale's (1971) fertility schedule based on the Hutterites, or the multiplcative age-calendar time interaction proposed e.g. by Le Bras (1974), or other "relational systems" in
the terminology of Brass (1974).)

It is not clear to me whether the Hadwiger function possesses nice properties of this sort. I should therefore like to ask what kind of structures have been revealed or what sort of problems have been answered with the help of Hadwiger, or other, graduations?
3. In one of your examples with empirical fitting of Hadwiger curves to period fertility data you obtained a reasonable fit for two small strata, but with grossly different parameters. You then went on to assert the well expected fact that for the whole of Norway, the fit was bad. I think that the audience was left a little inconclusive at this example.
4. The statistical problems discussed in the exposition belong mainly to estimation theory. A straightforward theoretical statistical attack on the simple occurrence/exposure rate situation would be to make clear exactly which observations are available and write down a likelihood function as the probability density function of these under an assumed parametric model (such as constant intensities in each age group). However, this is not usually done. Instead it is assumed that more observations are available (typically the exact time of the demographic events instead of aggregated numbers at period end-points). Under the idealized observational plan, the occurrence/exposure rates are the maximum likelihood estimators. Then in the second step the o/e rates are approximated from whatever observations are available. There seems to be a strong need for systematic studies of 1) the direct approach sketched above with a comparison of the statistical properties for the resulting estimators with those of the idealized o/e rates, and 2) the statistical properties of the various approximations to the o/e rates. (Breslow and Crowley (1974) show that one commonly used approximation leads to inconsistent estimation of a mortality function except under very special circumstances.)
5. The very formulation of a statistical model leading to o/e rates, in particular such fairly artificial and theoretically unsatisfactory hypotheses as piecewise constant intensities, seems to be motivated only by the limited avai1able observations and not by considerations about demographic structures. A systematic perspective on this would be to study estimability (or identifiability) of functionals of arbitrarily varying vital intensities. A very simple example is that of mortality observed at one-year age intervals: Then obviously only the
integrals

$$
\int_{a}^{a+1} \mu(x) d x, \quad a=0,1,2, \ldots
$$

of the death intensities over the one-year age intervals are estimable.
6. Turning next to the statistical basis of analytical graduation, the straightforward approach would again be to write down the likelihood of what is actually observed so that the statistical problem becomes a more or less standard nonlinear regression analysis of binomial (or the like) variables.

The alternative choice of the o/e rates as basic to the graduation needs some motivation, which I should like to request, in particular in view of the points raised under 4. and 5. above.
7. Another question regards the status of the optimality properties of the modified minimum $\chi^{2}$ and maximum likelihood graduation methods. Certainly these properties are very close to general results in large-sample estimation theory. It might be useful to point out this connection in somewhat greater detail, and perhaps also to specify more exactly in which way the present results differ from the conventional wisdom.

To be a little more specific, the basic log likelihood under your model is

$$
\begin{aligned}
& \sum \sum\left[M(a, d) \log g_{a, d}\left(\theta_{a}\right)-L(a, d) g_{a, d}\left(\theta_{a}\right)\right] \\
& a d
\end{aligned}
$$

where $M(a, d)$ is the number of events of type $a$ in the age interval $\left[t_{d-1}, t_{d}\right]$, $L(a, d)$ the corresponding time of exposure and $g_{a, d}\left(\theta_{a}\right)$ the value of the $a^{\prime}$ th component of the graduating function at some representative age in $\left[t_{d-1}, t_{d}\right]$. Thus the statistical model is a subfamily of the exponential family obtained by substituting freely varying parameters for $\log g_{a, d}\left(\theta_{a}\right)$ and $g_{a, d}\left(\theta_{a}\right)$. A general review of this situation has recently been compiled by Efron (1975) for the case where $\theta_{a}$ is one-dimensional. An important result here is the relation

$$
i_{\theta}-i_{\theta}^{T}=E_{\theta} \operatorname{Var}_{\theta}\left\{D \log L_{\theta}(X) \mid T\right\}
$$

where $i_{\theta}$ is the Fisher information in the sample, $i_{\theta}^{T}$ the Fisher information in a one-dimensional statistic $T$, and $L_{\theta}(X)$ the 1ikelihood.

From this relation the optimality properties of the maximum likelihood estimator are deduced. Analytical graduation will often have higher-dimensional parameters and thus seems to fit into a multivariate generalization of this approach. One might also quote the dual results by Sundberg (1974) who assumes that a full exponential family forms the statistical model but that only a function of the canonical statistic is observable. Sundberg's general results on asymptotic distribution of the maximum likelihood estimators may well prove useful for the studies outlined in 4. and 6. above.
8. Let me conclude by thanking Jan once more for the very stimulating survey. I have no doubts that this will generate a lot of new activity in an area with a long and impressive Nordic tradition.

## References

Brass, W. (1974): Perspectives in population prediction: Illustrated by the statistics of England and Wales (with discussion). J. Roy. Statist. Soc. A 137, 532-583.

Breslow, N. \& J. Crowley (1974): A large sample study of the life table and product limit estimates under random censorship. Ann. Statist. 2, 437-453.

Brown, B. W., Jr., M. Hollander \& R.M. Korwar (1974): Nonparametric tests for censored data with applications to heart transplant studies. Reliability and Biometry (ed. F. Proschan \& R.J. Serfling) SIAM, Philadelphia, pp. 327-354.

Clark, D.A., E.B. Stinson, R.B. Griepp, J.S. Schroeder, N.E. Shumway \& D.C. Harrison (1971): Cardiac transplantation in man VI. Prognosis of patients selected for cardiac transplantation. Ann. Int. Med. 75, 15-21.

Coale, A.J. (1971): Age patterns of marriage. Popu1. Studies 25, 193-214.

Crowley, J. \& M. Hu (1974): Covariance analysis of heart transplant survival data. Techn. rep., Division of Biostatistics, Stanford University.

Efron, B. (1975): Defining the curvature of a statistical problem (with applications to second order efficiency). (with discussion). Ann. Statist. 3 .

Gail, M.H. (1972): Does cardiac transplantation prolong life? A reassessment. Ann. Int. Med. 76, 815-817.

Le Bras, H. (1974): Populations stables aléatoires. Population 29, 435-464. Mantel, N. \& D.P. Byar (1974): Evaluation of response-time data involving transient states: An illustration using heart-transplant data. J. Amer. Statist. Ass. 69, 81-86.

Miller, R.G., Jr. (1974): Least squares regression with censored data. Techn. rep., Division of Biostatistics, Stanford University.

Sundberg, R. (1974): Maximum likelihood theory for incomplete data from an exponential family. Scand. J. Statist. 1, 49-58.

Turnbu11, B.W., B.W. Brown, Jr. \& M. Hu (1974): Survivorship analysis of heart transplant data. J. Amer. Statist. Assoc. 69, 74-80.

