Then the following ordinary search linear model of the form (1) *holds*:

$$\mathbf{y}^4 = \{[G(A_3)]\}\xi_1 + \{[G(A_3)]A_2\}\xi_2. \tag{19}$$

Furthermore, (19) *can be used to determine* ξ_1 *and* ξ_2 *as under ordinary search linear models.*

Proof. Equation (19) is obvious in view of (18) and (3). To show that $\mathbf{y}^4$ at (19) has the structure of a search design, we have to show that conditions corresponding to (2) in Theorem 1 hold; thus we need to show that

$$\operatorname{Rank}\{[G(A_3)]A_1 \vdots [G(A_3)]A_{20}\} = \operatorname{Rank}\{[G(A_3)]A_1\} + \operatorname{Rank}\{[G(A_3)]A_{20}\}, \tag{20}$$

for all $(N\times 2k)$ submatrices A_{20} contained in A_2. From (7b), (8b), and (7a), taking $Q = A_3$, and $P = [A_1 \vdots A_{20}]$, we find that the l.h.s. of (20) equals $\nu_1 + 2k$. Similarly, the two terms on the r.h.s. of (20) are respectively ν_1 and $2k$. This completes the proof of the theorem.

We close the paper by recalling, for the sake of completeness, one procedure for search and estimation under the ordinary search linear model (1) when conditions (2) hold. We first compute $[G(A_1)]\mathbf{y} = \mathbf{y}^1$, say. Clearly,

$$E(\mathbf{y}^1) = \{[G(A_1)]A_2\}\xi_2, \tag{21}$$

where in view of (2) and (8b), we have Rank $[G(A_1)]A_{20} = 2k$, for all $(N\times 2k)$ submatrices A_{20} of A_2. Then, we project $\mathbf{y}'$ on the sets of k columns of $[G(A_1)]A_2$, until (in the noiseless case) we obtain a set of k columns of A_2 which gives a perfect fit. In the noisy case, ordinary least squares projection may be used. Notice that the technique mentioned in this paragraph is essentially equivalent to method I in Srivastava [1].

References

[1] J. N. Srivastava, *Designs for searching non-negligible effects*, in: *A survey of statistica design and linear models*, ed. by J. N. Srivastava, North-Holland Publ. Company, Inc. New York 1975, pp. 507–519.

[2] —, *Some further theory of search linear models*, in: *Contribution to Applied statistics*, publ. by the Swiss–Australian Region of the Biometry Society, 1976, pp. 249–256.

[3] J. N. Srivastava and S. Ghosh, *Balanced* 2^m *factorial design of resolution V which allow search and estimation of one extra unknown effect* $4 \leqslant m \leqslant 8$, Comm. Statist. — Theor. Meth. A6 (1977), pp. 141–166.

[4] J. N. Srivastava, *Optimal search designs, or designs optimal under bias-free optimality criteria*, in: *Statistical decision theory and related topics, II*, ed. by S. C. Gupta and D. S. Moore, 1977, pp. 375–409.

[5] J. N. Srivastava and D. W. Mallenby, *Some studies on a new method of search in search linear models* (submitted for publication).

Presented to the semester
MATHEMATICAL STATISTICS
September 15–December 18, 1976

MATHEMATICAL STATISTICS
BANACH CENTER PUBLICATIONS, VOLUME 6
PWN—POLISH SCIENTIFIC PUBLISHERS
WARSAW 1980

DEVIATIONS FROM TOTAL INFORMATION AND FROM TOTAL IGNORANCE AS MEASURES OF INFORMATION

ERIK N. TORGERSEN

University of Oslo, Institute of Mathematics, Oslo, Norway

1. Introduction, notations and basic facts

Many interesting possibilities of quantifying the content of information in a statistical experiment have been proposed and studied in the literature. Among the most prominent are Fisher information and Kullback–Leibler information numbers. Several of the principles for comparing designs of experiments are based on ideas of measuring information. Most of the quantifications are designed for particular problems. It is, therefore, not surprising that comparison by different measures may lead to conflicting results. There is, of course, no hope to remedy this and no single real valued quantity is likely to qualify as "the information number". Any measure is bound to be useful within limited scopes only. The particular measures which I shall shortly describe are not exceptions — on the contrary they might even appear quite artificial. I find them more interesting because of their construction than because of their usefulness in concrete applications.

Before proceeding let me at once remark that limitations of time as well as on space, force me to present most of our results without proofs. Anyone interested will find proofs and other information on the subject in [14].

Our point of departure shall be the view of statistical decision theory, i.e. that the performance of a decision procedure is to be judged on the basis of the risk it incures. In order to give precise definitions, let us agree that a statistical experiment $\mathscr{E}$ with parameter set Θ is a family $(P_\theta;\theta\in\Theta)$ of probability measures on a common measurable space, say $(\chi, \mathscr{A})$. We may then write:

$$E = (\chi, A; P_\theta; \theta\in\Theta) = (P_\theta; \theta\in\Theta).$$

It is often convenient to identify experiments with the random variables defining them. Thus, if our observation X is χ-valued and $\mathscr{A}$-measurable and the distribution of X under θ is P_θ, then $\mathscr{E}$ may be considered as the experiment obtained by observing X.

If $\mathscr{E}_i = (P_{\theta,i};\theta\in\Theta)$, $i = 1,\ldots,n$, are experiments, then their product is the experiment $\left(\prod_{i=1}^{n} P_{\theta,i};\theta\in\Theta\right)$ and we shall use notations as $\mathscr{E}_1\times\ldots\times\mathscr{E}_n$ or $\prod_{i=1}^{n}\mathscr{E}_i$

for this experiment. Thus, if $\mathscr{E}_i$ is obtained by observing X_i and $X_1, \ldots, X_n$ are independent, then $\prod_i^n \mathscr{E}_i$ is the experiment obtained by observing $(X_1, \ldots, X_n)$. The experiment $(P_\theta^n; \theta \in \Theta)$ obtained by observing n independent replications of $\mathscr{E}$ will be denoted by $\mathscr{E}^n$.

In order to keep the mathematical apparatus within familiar bounds we shall also assume, unless otherwise stated, that our experiments are dominated.

We shall need a few concepts and facts from the theory of statistical experiments. Expositions, proofs and references may be found in Le Cam [5], [6], [8], Heyer [3], [4] and Torgersen [11], [15].

Various functionals on experiments may be defined by using homogeneous functions on R^Θ. For example, we may define without ambiguity the Hellinger transform of $\mathscr{E} = (P_\theta; \theta \in \Theta)$ as the map $H_\mathscr{E}$ which to each prior distribution t with finite support associates the number $\int \prod (dP_\theta)^{t_\theta}$. The Hellinger transform is particularily useful for studying independent combinations of experiments. The main reason for this is that the Hellinger transform of a product experiment is the product of the Hellinger transforms of the factor experiments. If the experiment $\mathscr{E} = (P_\theta; \theta \in \Theta)$ is more informative (see definition below) than the experiment $\mathscr{F} = (Q_\theta; \theta \in \Theta)$ then $H_\mathscr{E} \leqslant H_\mathscr{F}$.

Minimum Bayes' risk for a prior λ with finite support may often be written in the form $\int -\psi(dP_\theta; \theta \in \Theta)$ where ψ is a sublinear functional on R^Θ.

If h is a measurable and homogeneous function on $[0, \infty[^\Theta$ and $\mathscr{E} = (P_\theta; \theta \in \Theta)$, then we may define $h(\mathscr{E}) = \int h(dP_\theta; \theta \in \Theta)$ as $\int h(f_\theta; \theta \in \Theta) d\mu$ provided $\mu \gg \mathscr{E}$, $f_\theta = dP_\theta/d\mu$; $\theta \in \Theta$ and that the integral exists. It is easily checked that neither the existence nor the value of $h(\mathscr{E})$ depend on the choice of μ.

Le Cam [5] generalizing works of Blackwell and others, formulated the following notion of a deficiency.

Let $\mathscr{E}$ and $\mathscr{F}$ be experiments with the same parameter set Θ and let ε be a nonnegative function of θ. Then we shall say that $\mathscr{E}$ is *ε-deficient w.r.t.* $\mathscr{F}$ if to any finite decision problem where the loss function is bounded by 1 in absolute value and to any risk function s which is obtainable in $\mathscr{F}$ there is a risk function r which is obtainable in $\mathscr{E}$ such that $r \leqslant s+\varepsilon$.

The smallest constant $\varepsilon \geqslant 0$ such that this holds is the deficiency of $\mathscr{E}$ w.r.t. $\mathscr{F}$ and we shall denote it by $\delta(\mathscr{E}, \mathscr{F})$.

Associated with this deficiency is the distance Δ defined by

$$\Delta(\mathscr{E}, \mathscr{F}) = \max(\delta(\mathscr{E}, \mathscr{F}), \delta(\mathscr{F}, \mathscr{E})).$$

If $\delta(\mathscr{E}, \mathscr{F}) = 0$, then we shall say that $\mathscr{E}$ is *more informative than* $\mathscr{F}$ and occasionally write this $\mathscr{E} \geqslant \mathscr{F}$. The ordering "being more informative than" is certainly not a nice ordering. If, for example, we restrict our attention to linear normal experiments with known variances, then [2] this ordering reduces to the usual ordering of Fisher information matrices. If $\# \Theta = 2$, however, then the ordering [11] is at least complete.

If $\mathscr{E} \geqslant \mathscr{F}$ and $\mathscr{F} \geqslant \mathscr{E}$ then we shall say that $\mathscr{E}$ and $\mathscr{F}$ are *equivalent* and write this $\mathscr{E} \sim \mathscr{F}$. We might also consider deficiencies relative to a given class of decision problems, e.g. testing problems. Comparison by minimum Bayes' risks leads to weighted definiencies: Let λ be a prior distribution on Θ with finite support. The λ weighted deficiency, $\delta(\mathscr{E}, \mathscr{F}|\lambda)$, of $\mathscr{E}$ w.r.t. $\mathscr{F}$ is the greatest lower bound of all numbers $\sum_\theta \lambda_\theta \varepsilon_\theta$ such that $\mathscr{E}$ is ε-deficient w.r.t. $\mathscr{F}$. The deficiency $\delta(\mathscr{E}, \mathscr{F})$ may be expressed in terms of weighted deficiencies by

$$\delta(\mathscr{E}, \mathscr{F}) = \sup_\lambda \delta(\mathscr{E}, \mathscr{F}|\lambda).$$

One very interesting feature of deficiencies is that several reasonable and apparently different approaches lead to the same concept of deficiency. Time does not permit me to go further into this. Let me just mention Le Cam's fundamental randomization criterion [5] for the deficiency of an experiment $\mathscr{E} = (P_\theta; \theta \in \Theta)$ w.r.t. another experiment $\mathscr{F} = (Q_\theta; \theta \in \Theta)$:

$$\delta(\mathscr{E}, \mathscr{F}) = \inf_M \sup_\theta \|P_\theta M - Q_\theta\|$$

where M runs through all Markov operators from the L-space of $\mathscr{E}$ to the L-space of $\mathscr{F}$ and $\| \; \|$ indicates total variation.

2. Deficiencies and information numbers

The deficiency $\delta(\mathscr{E}, \mathscr{F})$ is a function of two variables $\mathscr{E}$ and $\mathscr{F}$. It provides a partial answer to the question: What do we loose by basing ourselves on $\mathscr{E}$ rather than on $\mathscr{F}$ under the least favorable conditions for this comparison?

The deficiency is monotonically increasing in $\mathscr{F}$ and monotonically decreasing in $\mathscr{E}$. It is also convex in each variable separately. Convexity is then defined in terms of mixtures of experiments [16].

Suppose we have a family $\mathscr{E}_t$: $t \in T$ of experiments and that $\mathscr{F}$ is some ideal and unattainable experiment. Let us say that $\mathscr{E}_t \leqslant \mathscr{F}$ for all t. Then we might consider the numbers $\delta(\mathscr{E}_t, \mathscr{F})$, $t \in T$, as information numbers. Small numbers will then correspond to informative experiments.

We might, instead of considering an ideal and unattainable experiment $\mathscr{F}$, consider some "bad" experiment $\mathscr{G}$ such that $\mathscr{G} \leqslant \mathscr{E}_t$ for all t. Then we might use the numbers $\delta(\mathscr{G}, \mathscr{E}_t)$, $t \in T$, as information numbers. A small number will then indicate that our experiment contains little information. As an example consider the tails $\mathscr{E}_t$: $X_t, X_{t+1}, \ldots$ of a Markov chain $(X_1, X_2, \ldots)$ with finite state space Θ. By the Markov property and by sufficiency this experiment is defined by X_t alone. Clearly, $\mathscr{E}_1 \geqslant \mathscr{E}_2 \geqslant \ldots$ and it follows from the compactness of Δ convergence for finite Θ, [6], that $\delta(\mathscr{E}_t, \mathscr{G}) \downarrow 0$ as $t \uparrow \infty$ for some experiment $\mathscr{G}$ such that $\mathscr{G} \leqslant \mathscr{E}_t$ for all t.

The behaviour of this deficiency is treated in detail by Lindqvist in [9].

Another example where it might be natural to consider deficiencies as information numbers is the following:

Suppose we have observed n independent and normally distributed random variables $X_1, \ldots, X_n$ such that $EX_i = \sum_{j=1}^{k} a_{ji}\beta_j$ and $\operatorname{Var} X_i = \sigma^2$, $i = 1, \ldots, n$, where the a are known constants, the β are unknown parameters and σ is known or unknown. Suppose also that we have the possibility of observing a X_{n+1} such that

$$EX_{n+1} = \sum_{i=1}^{k} a_{n+1,i}\beta_i \quad \text{and} \quad \operatorname{Var} X_{n+1} = 1$$

where the vector $(a_{n+1,1}, \ldots, a_{n+1,k})$ may be chosen freely within a certain subset of R^k. How should this vector be chosen? If we do not have any specific decision problem in mind then we might try to choose X_{n+1} so that the deficiency of the experiment defined by $(X_1, \ldots, X_n)$ w.r.t. the experiment defined by $(X_1, \ldots, X_{n+1})$ is large. This problem have been investigated by Swensen in [10]. He has also, in the same report, obtained, in several important cases, closed expressions for deficiencies between linear normal experiments.

A word of caution is, by the way, in order concerning the interpretation of large deficiencies since these may stem from decision problems of scant interest.

It is also possible to construct local measures of information based on deficiencies, [12], but we shall not dwell on this here.

3. Deviations from total information and from total ignorance

In order to investigate the properties of such measures it is tempting to consider, in spite of their artificiality, distances w.r.t. experiments which are either totally informative or totally uninformative.

A totally informative experiment is an experiment $(P_\theta; \theta \in \Theta)$ such that P_{θ_1} is P_{θ_2} singular when $\theta_1 \neq \theta_2$. Although the deficiency of any dominated experiment w.r.t. such an experiment is well defined, it is of no interest when Θ is uncountable. The reason is that in that case this deficiency always equals 2. We shall therefore, in the following, when we consider comparison w.r.t. totally informative experiments assume that Θ is countable. As any two totally informative experiments are equivalent we shall use the symbol $\mathscr{M}_a$ to denote any of them. To fix ideas we might, if we so prefer, let $\mathscr{M}_a$ denote the experiment $(\delta_\theta; \theta \in \Theta)$ where δ_θ is the one point distribution in θ.

A totally uninformative experiment is an experiment $(P_\theta; \theta \in \Theta)$ where P_θ does not depend on θ. Clearly, any two non-informative experiments are also equivalent and we shall reserve the notation $\mathscr{M}_i$ for any of them.

For any experiment $\mathscr{E}$ the informational inequalities $\mathscr{M}_i \leqslant \mathscr{E} \leqslant \mathscr{M}_a$ hold. The deficiency of $\mathscr{M}_i$ w.r.t. $\mathscr{E}$ will be denoted by $\delta_i(\mathscr{E})$ while the deficiency of $\mathscr{E}$ w.r.t. $\mathscr{M}_a$ will be denoted by $\delta_a(\mathscr{E})$. The λ-weighted deficiencies of $\mathscr{E}$ w.r.t. $\mathscr{M}_a$ will be denoted by $\delta_a(\mathscr{E}|\lambda)$.

Thus our first proposal for a measure of the content of information in the experiment $\mathscr{E}$ is the number $\delta_i(\mathscr{E})$. If this distance is small, then the chance mechanism governing the random outcome is almost independent of the various explaining theories in Θ. If, on the other hand, this distance is large, then there are situations where an observation of $\mathscr{E}$ is helpful.

$\mathscr{M}_a$ is the experiment of directly observing the underlying theory θ in Θ. An experiment $\mathscr{E}$ may be considered to contain much or little information according to whether $\mathscr{E}$ is close to $\mathscr{M}_a$ or far away from $\mathscr{M}_a$. Thus we arrive at the deficiency of $\mathscr{E}$ w.r.t. $\mathscr{M}_a$, i.e. $\delta_a(\mathscr{E})$, as a measure of the content of information in $\mathscr{E}$.

A small value of $\delta_a(\mathscr{E})$ tells that an observation of $\mathscr{E}$, provided it is properly used, is almost as good as knowing the unknown parameter. A large value, on the other hand, tells that there are decision problems such that any decision procedure is risky for some of the underlying theories.

The values of these deficiencies are often extremely large for all experiments $\mathscr{E}$ under consideration. This reflects the fact that it may be much to ambitious to compare with total information and much to modest to compare with no information.

$\delta_a(\mathscr{E})$ is related to the problem of guessing the true value of θ. This may, alternatively, be viewed as a problem of finding optimal confidence regions with extreme accuracy. If we relaxed the requirement on accuracy, then we might hope to find other and more realistic measures of information than $\delta_a(\mathscr{E})$. Thus one might expect that the usefulness of the measure $\delta_a(\mathscr{E})$ is limited to situations where the space of underlying theories is, in some sense small.

Let us first consider the deficiencies $\delta_a(\mathscr{E})$ and $\delta_a(\mathscr{E}|\lambda)$. It follows directly from the randomization criterion that:

$$\delta_a(\mathscr{E}) = 2 \inf_M \sup_\theta P_\theta(M \neq \theta)$$

while

$$\delta_a(\mathscr{E}|\lambda) = 2\big[1 - \big\|\bigvee_\theta \lambda_\theta P_\theta\big\|\big].$$

Here the inf are taken over all randomized estimators of θ. Thus $\delta_a(\mathscr{E})/2$ is the minimax probability of guessing wrongly the true distribution while $\delta_a(\mathscr{E}|\lambda)/2$ is the minimum Bayes probability of the same event.

If $\Theta = \{1, 2\}$ then $\delta_a(\mathscr{E})/2$ is the unique number α_0 in $[0, 1]$ such that the Neyman–Pearson test for "P_1" against "P_2" has power $1-\alpha_0$ in P_2. $\delta_a(\mathscr{E}|\lambda)$ may then be written $\|\lambda_1 P_1 \wedge \lambda_2 P_2\|$ while the probability of an error of the second kind for the Neyman–Pearson test for "P_1" against "P_2" at level α is the smallest number $\varepsilon \geqslant 0$ so that $\mathscr{E}$ is $(2\alpha, 2\varepsilon)$ deficient w.r.t. $\mathscr{M}_a$.

As time does not permit a discussion of both δ_i and δ_a we shall, from here on, restrict ourselves to δ_a.

4. Replications

How do these quantities behave under replications? It follows, as is well known, from the weak law of large numbers that $\mathscr{E}^n \to \mathscr{M}_a$ provided $P_{\theta_1} \neq P_{\theta_2}$ when $\theta_1 \neq \theta_2$ and that Θ is finite. This implies that, for any experiment $\mathscr{E}$ such that $\theta \rightsquigarrow P_\theta$ is 1-1, $\mathscr{E}^n$ converges weakly to $\mathscr{M}_a$ in the sense that the restrictions of $\mathscr{E}^n$ to finite sub parameter sets converge to the same restrictions of $\mathscr{M}_a$. This does not, however, exclude the possibility that $\delta_a(\mathscr{E}^n) = 2$ for all n. On the contrary, one is tempted, when Θ is infinite, to say that this is the usual case, and that strong conditions are needed to ensure convergence. If convergence takes place at all, then the next problem is to decide the rate of convergence. This cannot be done on the basis of $\delta_a(\mathscr{E})$ alone, since there are experiments $\mathscr{E}$ such that $\delta_a(\mathscr{E})$ has the maximal value 2 while $\delta_a(\mathscr{E}^2)$ is less than, say $1/10^{100}$. In that case any guessing procedure based on one observation is almost certain to guess wrongly the true distribution for at least one value of θ while two observations are as good as knowing the true value of θ. If Θ is finite, however, then there are inequalities which show that $\delta_a(\mathscr{E})$ has to be small when $\delta_a(\mathscr{E}^2)$ is small. [The topological fact that $\delta_a(\mathscr{E}_n) \to 0$ whenever $\delta_a(\mathscr{E}_n^2) \to 0$ follows in this case directly from the compactness of the Δ distance.]

Consider first the case of dichotomies; $\Theta = \{1, 2\}$ say. Note first that the inequality:

$$\min\{\lambda_1 f_1 g_1, \lambda_2 f_2 g_2\} \geqslant \min\{\lambda_1 f_1, \lambda_2 f_2\}\min\{g_1, g_2\}$$

holds for any non-negative numbers $\lambda_1, \lambda_2, f_1, f_2, g_1$, and g_2. Putting $\mathscr{E} = (P_1, P_2)$, $\mathscr{F} = (Q_1, Q_2)$; $f_i = dP_i/d(P_1+P_2)$; $g_i = dQ_i/d(Q_1+Q_2)$ we find, by integrating w.r.t. $(P_1+P_2)\times(Q_1+Q_2)$, that for any prior distributions (λ_1, λ_2) and (μ_1, μ_2):

$$\tfrac{1}{2}\delta(\mathscr{E}\times\mathscr{F}|\lambda) \geqslant \tfrac{1}{2}\delta(\mathscr{E}|\lambda)\cdot\delta(\mathscr{F}|\tfrac{1}{2}, \tfrac{1}{2}) \geqslant \tfrac{1}{2}\delta(\mathscr{E}|\lambda)\cdot\tfrac{1}{2}\delta(\mathscr{F}|\mu).$$

Maximizing w.r.t. λ and μ we find:

$$\tfrac{1}{2}\delta_a(\mathscr{E}\times\mathscr{F}) \geqslant \tfrac{1}{2}\delta_a(\mathscr{E})\cdot\tfrac{1}{2}\delta_a(\mathscr{F}).$$

Thus, in particular,

$$\tfrac{1}{2}\delta_a(\mathscr{E}) \leqslant \sqrt{\tfrac{1}{2}\delta_a(\mathscr{E}^2)}.$$

Considering replications of the same experiment, we find, by the same inequality, that

$$\tfrac{1}{2}\delta_a(\mathscr{E}^{m+n}) \geqslant \tfrac{1}{2}\delta_a(\mathscr{E}^m)\tfrac{1}{2}\delta_a(\mathscr{E}^n).$$

It follows that

$$\sqrt[n]{\delta_a(\mathscr{E}^n)} \to C \quad \text{as} \quad n \to \infty,$$

where $C = \sup\limits_n \sqrt[n]{\tfrac{1}{2}\delta_a(\mathscr{E}^n)}$.

What is C? That is, is there a simple and explicit expression for this quantity? Consider a non-negative loss function $L_\theta(t)$ such that for all θ there is at least one "correct" decision t satisfying $L_\theta(t) = 0$. Then the risk of any good procedure based on n observations should be at most $\frac{1}{2}\max L_\theta(t)\, C^n$. It follows that, asymptotically, the nth root of the risk should be at most C. Furthermore, it is not difficult to show that the rate of exponential convergence does not, except for the trivial decision problems where no observations are needed, depend on the particular decision problem. Now, Chernoff [1] found that

$$\lim \sqrt[n]{\|\lambda_1 P_1^n \wedge \lambda_2 P_2^n\|} = \inf_{0<t<1} \int dP_1^{1-t} dP_2^t$$

for any non-degenerate prior (λ_1, λ_2). Thus

$$C = \inf_{0<t<1} \int dP_1^{1-t} dP_2^t.$$

In order to be able to extend these results to larger parameter sets let us, for any experiment $\mathscr{E}$, put

$$C(\mathscr{E}) = \sup_{\theta_1\neq\theta_2} \inf_{0<t<1} \int dP_{\theta_1}^{1-t} dP_{\theta_2}^t.$$

As is well known, pairwise sufficiency implies sufficiency for dominated experiments.

It is therefore perhaps not too surprising that $C(\mathscr{E})$ defines the exponential rate of convergence whenever the parameter set is finite, i.e.

$$\sqrt[n]{\delta_a(\mathscr{E}^n)} \to C(\mathscr{E})$$

when $\# \Theta < \infty$. Again $C(\mathscr{E})$ defines the exponential rate of convergence for minimum Bayes' risk for a large class of decision problems, although not for all decision problems. In general, it yields only an upper bound for the exponential rate of convergence.

What about the case of an infinite and countable parameter set? Again $\sqrt[n]{\delta_a(\mathscr{E}^n)}$ converges, as $n \to \infty$, to a limit $\sigma(\mathscr{E})$ such that $\sigma(\mathscr{E}^r) = \sigma(\mathscr{E})^r$; $r = 1, 2, \ldots$ If Θ is infinite, however, then $\sigma(\mathscr{E})$ may be strictly larger than $C(\mathscr{E})$ so that the constant $C(\mathscr{E})$ does no longer determine the rate of exponential convergence. In the case of a finite Θ a value of $C(\mathscr{E})$ as, say, $C(\mathscr{E}) = \frac{1}{2}$ would indicate that $\delta_a(\mathscr{E}^n) \sim 2^{-n}$. If Θ is infinite, it may happen that $C(\mathscr{E}) = \frac{1}{2}$ while $\delta_a(\mathscr{E}^n) \underset{n}{\equiv} 2$ (so that $\sigma(\mathscr{E}) = 1$). Also if exponential convergence does not take place, i.e. $\sigma(\mathscr{E}) = 1$, then we might ask for the actual rate of convergence. It turns out, however, that there is no alternative to exponential convergence, except no convergence at all.

If $\mathscr{E}$ has an accumulation measure for set wise convergence of probability measures, then $\delta_a(\mathscr{E}^n) \underset{n}{\equiv} 2$. This is, in particular, the case when the sample space of $\mathscr{E}$ is finite while Θ is infinite.

I would like to conclude by mentioning a few problems.

The limit $\sigma(\mathscr{E}) = \lim\limits_n \sqrt[n]{\delta_a(\mathscr{E}^n)}$ exists for any experiment $\mathscr{E}$. If Θ is finite, then $\sigma(\mathscr{E})$ may be expressed directly in terms of $\mathscr{E}$. If Θ is infinite and countable, then our expressions for $\sigma(\mathscr{E})$ involves all replications. It would be interesting to have an expression for $\sigma(\mathscr{E})$ which involves $\mathscr{E}$ only, say in terms of the Hellinger transform of $\mathscr{E}$. Even in such a well structured case as the case of translation experiments on the integers this problem appears open. Using the fact that properly specified maximum likelihood estimators are optimal in this situation, it is not difficult to see that exponential convergence takes place (i.e. $\sigma(\mathscr{E}) < 1$) — but what is $\sigma(\mathscr{E})$?

One might consider the more general problem of the asymptotic behaviour of deficiencies $\delta(\mathscr{E}^n, \mathscr{F}^n)$ as $n \to \infty$. It is known that we may have $\mathscr{E}^n \geqslant \mathscr{F}^n$ when n is sufficiently large, although $\delta(\mathscr{E}, \mathscr{F}) > 0$. Then $\sqrt[n]{\delta(\mathscr{E}^n, \mathscr{F}^n)} \to 0$. Are there other situations where $\sqrt[n]{\delta(\mathscr{E}\ , \mathscr{F}^n)} \to 0$?

Clearly,

$$\limsup_n \sqrt[n]{\delta(\mathscr{E}^n, \mathscr{F}^n)} \leqslant \limsup_n \sqrt[n]{\delta_a(\mathscr{E}^n)} \leqslant \sigma(\mathscr{E})$$

while

$$\liminf_n \sqrt[n]{\delta(\mathscr{E}^n, \mathscr{F}^n)} \geqslant \liminf_n \sqrt[n]{[\delta_a(\mathscr{E}^n) - \delta_a(\mathscr{F}^n)]^+}\,.$$

It follows that $\sqrt[n]{\delta(\mathscr{E}^n, \mathscr{F}^n)} \to \sigma(\mathscr{E})$ whenever $\sigma(\mathscr{E}) > \sigma(\mathscr{F})$. We do not, however, know the limiting behaviour of sequences $\sqrt[n]{\delta(\mathscr{E}^n, \mathscr{F}^n)}$, $n = 1, 2, \ldots$, when $\sigma(\mathscr{E}) < \sigma(\mathscr{F})$.

References

[1] H. Chernoff, *A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations*, Ann. Math. Statist. 23 (1952), pp. 493–507.

[2] O. H. Hansen and E. N. Torgersen, *Comparison of linear normal experiments*, Ann. Statist. 2 (1974), pp. 367–373.

[3] H. Heyer, *Erschöpftheit und Invarianz beim Vergleich von Experimenten*, Z. Wahrscheinlichkeitstheorie. verw. Geb. 12 (1969), pp. 21–55.

[4] —, *Mathematische Theorie statistischer Experimente*, Springer Verlag, Berlin 1973.

[5] L. Le Cam, *Sufficiency and approximate sufficiency*, Ann. Math. Statist. 35 (1964), pp. 1419–1455.

[6] —, *Notes on asymptotic methods in statistic decision theory*, Centre de Recherches, Math. Univ. de Montréal, 1974.

[7] —, *On the information contained in additional observations*, Ann. Statist. 2 (1974), pp. 630–649.

[8] —, *Distances between experiments*. In: *A survey of statistical design and linear models*, (Ed. J. N. Srivastava), pp. 383–395, 1975.

[9] B. H. Lindquist, *How fast does a Markov chain forget the initial state? A decision theoretic approach*, Scand. J. Statist. 4 (1977), pp. 145–152.

[10] A. R. Swensen, *Deficiencies in linear normal experiments*, to appear in Ann. Statist.

[11] E. N. Torgersen, *Comparison of experiments when the parameter space is finite*, Z. Wahrscheinlicheitstheorie. verw. Geb. 16 (1970), pp. 219–249.

[12] —, *Local comparison of experiments*, Stat. Res. Report, Univ. Oslo, Oslo 1972.

[13] —, *Asymptotic behaviour of powers of dichotomies*, ibid., Oslo 1974.

[14] —, *Deviations from total information and from total ignorance as measures of information*. ibid., Oslo 1976.

[15] —, *Comparison of statistical experiments*, Scand. J. Statist. 3 (1976), pp. 186–200.

[16] —, *Mixtures and products of dominated experiments*, Ann. Statist. 5 (1977), pp. 44–64.

Presented to the semester
MATHEMATICAL STATISTICS
September 15–December 18, 1976

MATHEMATICAL STATISTICS
BANACH CENTER PUBLICATIONS, VOLUME 6
PWN—POLISH SCIENTIFIC PUBLISHERS
WARSAW 1980

ON THE CRAMÉR–RAO INEQUALITY AND ON A NEW VERSION OF THE CHI-SQUARE STATISTIC

I. VINCZE

Mathematical Institute of the Hungarian Academy of Sciences,
Budapest V, Hungary

In connection with the Cramér–Rao inequality many important investigations were made for the non-regular case, i.e. when the supports of the underlying densities in the sample space do not coincide. In their pioneering papers, D. G. Chapman and H. Robbins [2] (1951), J. Kiefer [4] (1952), D. A. S. Fraser and I. Guttmann [3] (1952) consider mainly the case of a real parameter (using also further restrictions) given various bounds for the variance of an unbiased estimator. The aim of the present lecture is to give a brief account of the results of the above-mentioned papers pointing out that almost no assumption is needed concerning the structure of the parameter space (see also Barankin [1] (1949)).

In the second part of the lecture the following modified form of Pearson's chi-square statistic is investigated:

$$\bar{\chi}^2 = \sum_{i=1}^{r} \frac{(\bar{X}_{(i)} - E_i)^2}{\sigma_i^2}\, \nu_i,$$

where E_i and σ_i $(i = 1, 2, \ldots, r)$ are the conditional expected value and the variance of the variable restricted to the ith interval of the partition, while ν_i is the number of sample elements falling into the ith interval and having arithmetic mean $\bar{X}_{(i)}$. This statistic utilizes besides the number of sample elements lying on the respective intervals of the partition also their *positions* within the intervals. In a joint paper with E. Csáki [7] the authors show that this statistic is asymptotically distributed — when the sample size n tends to infinity — according to the chi-square distribution with parameter r, i.e. the number of intervals chosen — contrary to the $r-1$ belonging to Pearson's statistic. When $n \to \infty$ and $r = O(n^\alpha)$, $0 < \alpha < 1$, the distribution of

$$\frac{\bar{\chi}^2 - r}{\sqrt{2r}}$$

tends to the normal law $N(0, 1)$. Whenever the relation

$$\sum_{i=1}^{r} \frac{(E_i - E_i^*)^2}{\sigma_i^2}\, p_i^* > 0$$