Since the approximation is scale dependent, the criterion $(Ex, Ex)/(x, x)$ may be replaced with $(Ex, Ex)/(x, S_2 x)$ where S_2 is a relevant positive definite matrix, e.g. an error covariance matrix in a multivariate regression situation or a covariance matrix to which the matrix of interest has to be compared. This modified criterion is equivalent to $z'L'E'ELz/z'z$ where $L'L = S_2^{-1}$ and L an upper triangular matrix. Now the approximation C of Y will be AB' with $A = YL'U_k\Lambda_k^{-1}$ as before, and $B = M'U_k\Lambda_k$ where $L'M = I_p$ and M is a lower triangular matrix. Then $Y'Y$ will be approximated simultaneously by BB' again.

When a relevant matrix S_2 is not available one may find, given the covariance matrix $\Sigma = n^{-1}Y'Y$, a diagonal scaling matrix K such that an approximation of $K\Sigma K-I$ induced by the approximation $U_k\Lambda_k^2U_k'$ of $K\Sigma K$, namely $U_k(\Lambda_k^2-I)U_k'$, will be perfect in the diagonal elements. This idea borrowed from factor analysis is directed towards equalizing by a suitable rescaling, the variance approximation errors, and so the rescaled specific variances are set equal to one beforehand. Finding such a K is exactly what happens in maximum likelihood factor analysis, where K^{-2} is the required matrix of specific variances.

Now with $\Lambda_k^2-I = \tilde{\Lambda}_k^2$ one may choose $C = AB'$ with $A = n^{-1/2}YKU_k\tilde{\Lambda}_k^{-1}$ and $B = K^{-1}U_k\tilde{\Lambda}_k$, where A contains factor scores in agreement with Bartlett's recommendation.

In the case where Y is a contingency table N, it is preferable to rescale N to $R^{-1/2}NK^{-1/2}$ where R is a diagonal matrix of row totals of N, and K similarly of column totals. In the canonical decomposition $\sum_{j=1}^{r} \lambda_j v_{*j}u_{*j}'$ all λ_j are at most 1, while λ_1 equals one, $\lambda_1 v_{*1}u_{*1}'$ representing square roots of expected frequencies under independence. The statistic $n\lambda_2^2$ may be used for testing independence, its asymptotic null distribution being known. The rank k approximation of $N-R^{1/2}v_{*1}u_{*1}'K^{1/2}$, i.e. $R^{1/2}V_k\Lambda_kU_kK^{1/2}$ may serve the study of dependence. The present paper has been published recently:

Reference

[1] L. C. A. Corsten, *Matrix approximation, a key to application of multivariate methods*, Proceedings of the 9th Biometric Conf. Boston 1976, Vol. 1, pp. 61–77.

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 BASIC CONCEPTS OF MATHEMATICAL STATISTICS

N. N. ČENCOV

Institute of Applied Mathematics, Moscow, U.S.S.R.

The basic essential concepts and structures have been formalized rather intensively for the last ten years. In this paper[1] the geometric approaches, which naturally arise in the analysis of statistical concepts, are considered.

Let (Ω, S) be a measurable space of elementary events, let $\{P_\theta\}$ be a family of a probability distribution, a priori possible, over (Ω, S), and let $(\mathscr{E}, B)$ be a measurable space of decisions.

Any of Wald's statistical decision rules [9], both determinated and randomized, can be written as a transition probability distribution $\Pi(\omega; d\varepsilon)$ from Ω onto $(\mathscr{E}, B)$. Thus if we use the rule Π, our decision will be distributed according to the law

$$(1) \qquad Q_\theta = P_\theta\Pi: \quad Q_\theta(\cdot) = \int_\Omega P_\theta(d\omega)\Pi(\omega; \cdot).$$

The value of the parameter θ at which the observations occur is unknown to the observer; he only knows that the observed P belongs to $\{P_\theta\}$. Therefore, all a priori conclusions about the quality of the decision rule Π are based on the properties of the families $\{P_\theta\Pi\}$.

It is natural to say that the families $\{P_\theta^{(i)}\}$ on $(\Omega^{(i)}, S^{(i)})$, $i = 1, 2$, parametrized by the same parameter $\theta \in \Theta$ are *equivalent in the theory of statistical inference* if, for any space of decision $(\mathscr{E}, B)$ and for any rule $\Pi^{(i)}(\omega^{(i)}; d\varepsilon)$, $i = 1, 2$, which leads to the family of laws $P_\theta^{(i)}\Pi^{(i)} = Q_\theta$ there exists a rule $\Pi^{(j)}(\omega^{(j)}; d\varepsilon)$, $j = 2, 1$, which leads to the same family $\{Q_\theta\}$:

$$(2) \qquad P_\theta^{(j)}\Pi^{(j)} = Q_\theta = P_\theta^{(i)}\Pi^{(i)}, \quad \forall\theta \in \Theta.$$

THEOREM 1. *The families $\{P_\theta^{(1)}\}$ and $\{P_\theta^{(2)}\}$ are equivalent in the theory of statistical inference iff there exist decision rules $\mathrm{III}^{(21)}$ and $\mathrm{III}^{(12)}$ such that*

$$(3) \qquad P_\theta^{(1)} = P_\theta^{(2)}\mathrm{III}^{(21)}, \quad P_\theta^{(2)} = P_\theta^{(1)}\mathrm{III}^{(12)}, \quad \forall\theta \in \Theta.$$

[1] The text following below combines two lectures of the author: "On basic concepts of mathematical statistics" and "On testing hypotheses".

To prove it let us note that the statistical decision rules form a category. Specifically the composition

$$\int_{\Omega''} \Pi^{(12)}(\omega'; d\omega'')\Pi^{(23)}(\omega''; A) = \Pi^{(13)}(\omega', A),$$

corresponding to the sequential use of two decision rules will again be a statistical decision rule. So (2) follows from (3) if we take $\Pi^{(j)} = \text{III}^{(ji)}\Pi^{(i)}$:

$$P_\theta^{(j)}\text{III}^{(ji)}\Pi^{(i)} = P_\theta^{(i)}\Pi^{(i)} = Q_\theta, \quad \forall\theta\in\Theta.$$

Conversely, let us assume $(\mathscr{E}, B) = (\Omega^{(i)}, S^{(i)})$ and as the decision rule $\Pi^{(i)}$ take the identical one Π_0: $\omega \to \omega$, which is assigned by the transition distribution $\Pi_0(\omega; A) = \chi_A(\omega)$. The corresponding one for (2) is $\Pi^{(j)} = \text{III}^{(ji)}$.

Thus we have come to the concept of statistically equivalent families of probability laws [1].

THEOREM 2. *If two families $\{P_\theta^{(1)}\}$ and $\{P_\theta^{(2)}\}$ are statistically equivalent and if some regularity conditions are satisfied, then there exists a common sufficient statistics for them; to be more exact there exist measurable mappings f_i: $\Omega^{(i)} \to \Omega$, $i = 1, 2$, such that*

$$P_\theta^{(1)}f^{-1}(\cdot) = P_\theta^{(2)}f^{-1}(\cdot) = R_\theta(\cdot), \quad \forall\theta, \tag{4}$$

and $\{P_\theta^{(1)}\} \sim \{R_\theta\} \sim \{P_\theta^{2}\}$.

PROBLEM 1. *Under what natural regularity conditions does the theorem hold?*

This question is closely bound with the question when there exists a true conditional distribution relative to statistics. So it makes sense to consider the families of concordant Lebesgue probability measures, i.e. the measures concentrated on the mutually measurable one-to-one image $g(F)$ of the $\boldsymbol{B}^*$-measurable subset F of the unit interval. Such families tolerate a short characterization in non-traditional terms: the commutative algebra of random variables has just one generator, for example, $g^{-1}(\omega)$. Perhaps the regularity conditions should also be found in non-traditional form.

Let us consider probability laws P on (Ω, S) as "points" of the collection $\text{Cap}(\Omega, S)$ of all probability measures on (Ω, S). Then the decision rule Π assigns by (1) the Markov mapping from $\text{Cap}(\Omega, S)$ to $\text{Cap}(\mathscr{E}, B)$. Thus the statistical equivalence of two families is interpreted as the geometrical congruence of two parametrized punctiform sets with respect to the category of Markov mappings.

It is naturally desirable to study invariants of this categorical geometry. In particular we shall call a numerical function of a pair of points of one object an invariant when

$$\{P_i\Pi' = Q_i,\ Q_i\Pi'' = P_i;\ i = 1, 2\} \Rightarrow \{\varphi(P_1, P_2) = \varphi(Q_1, Q_2)\}. \tag{5}$$

In this paper we shall be interested in those invariants of the pair of points which have the meaning of distance.

THEOREM 3. *If the norm on the linear space of measures of bounded variation is such that the function ϱ: $\varrho(P_1, P_2) = \|P_1 - P_2\|$ is an invariant in Markov geometry, then $\varrho(P_1, P_2) = c|P_1 - P_2|$ where $|\mu|$ is the symbol of a total variation and c is a non-negative constant.*

Thus (see [5]) on the collection $\text{Cap}(\Omega, S)$ of all probability laws on (Ω, S) there exists a unique "reasonable" (i.e. invariant) norm. This is well known but, however strange, it does not have any specific statistical sense.

Measuring the "distance" between points may be carried out by means of Riemann geometry methods.

THEOREM 4. *In Markovian geometry there exists a unique (up to a multiplicative constant) invariant Riemann metric.*

In the simplex of all probability distributions on the finite set $\Omega = (\omega_1, \dots, \omega_n\}$ it is assigned by the invariant quadratic form

$$ds^2 = \sum_{j=1}^{n} \frac{(dp_j)^2}{p_j}, \qquad p_j = P(\omega_j),\ \forall j; \tag{6}$$

and by the Fisher form on smooth probability law families

$$ds^2 = \sum_{\alpha,\beta=1}^{k} d\theta_\alpha d\theta_\beta M_P \frac{\partial \ln p(\omega;\theta)}{\partial\theta_\alpha} \cdot \frac{\partial \ln p(\omega;\theta)}{\partial\theta_\beta}. \tag{6'}$$

This metric is also well known. Through the Fisher tensor the famous information inequality is written down. But an interesting fact is that the metric can only appear in local problems (among them angle measuring). It cannot be met in the global ones.

Meanwhile measuring the distance between laws along the shortest path in the Fisher metric would be sufficiently simple. While changing the coordinates

$$p_j = z_j^2, \quad 0 \leqslant z_j \leqslant 1, \forall j; \quad \sum_j z_j^2 = 1,$$

we map the unit simplex on the positive orthant on the sphere of radius 1 and as the length we take the double Euclidean length. Thus here the arcs of great circles are the shortest and the double lengths of these arcs are the distances:

$$ds^2 = 4\sum_{j=1}^{n} dz_j^2.$$

Another approach is possible. The geodesic lines can be introduced not as the shortest lines but as the lines of null curvature with respect to any linear connection, not necessarily generated by the Riemann metric.

It should be noted that the length scales are assigned to each geodesic line separately and no consistency rule for the scales along different lines is introduced.

In my monograph [5] it is shown that the whole family of linear connections, invariant in Markovian categorical geometry, is possible. Every such connection is completely defined by the rule of finding the "mean" $R = \psi(P, Q) = \psi(Q, P)$

of the "geodesic line segment" PQ. The simplest rule is

$$R_V(\omega_j) = r_j = \tfrac{1}{2}p_j+\tfrac{1}{2}q_j, \ \forall j; \quad p_j = P(\omega_j), \quad q_j = Q(\omega_j).$$

Apparently it corresponds to measuring the distance along the variation

$$|P-R_V| = |Q-R_V| = \tfrac{1}{2}|P-Q|.$$

For the Fisher metric the mean R_F is given by the rule:

$$\sqrt{r_j} = [h(P,Q)]^{-1/2}(\sqrt{p_j}+\sqrt{q_j}), \quad \forall j, \qquad h(P,Q) = \sum_k (\sqrt{p_k}+\sqrt{q_k})^2; \tag{7}$$

"the length" PQ is given by

$$s_F(P,Q) = 2\arccos\sum_k \sqrt{p_k q_k}. \tag{8}$$

But in the most remarkable connection the mean R_E is defined through the geometric mean

$$r_j = [H(P,Q)]^{-1}\sqrt{p_j q_j}, \ \forall j; \qquad H(P,Q) = \sum_k \sqrt{p_k q_k}. \tag{9}$$

Such a mean I have proposed to call the *geodesic mean*. In this connection the geodesic lines prove to be one-parameter exponential families. The exponential families of several parameters are completely geodesic manifolds, i. e. plane analogies and so forth.

PROBLEM 2. *Define the geodesic lines of the remaining possible (according to* [5]) *invariant linear connections. Check whether convenient approaches of measuring unlikeness of probability distribution arise in them.*

Let us consider the exponential families in more detail. Let $\mu(d\omega)$ be a certain measure on (Ω, S); let $q_j(\omega)$ (where $j = 1, \ldots, m$) be measurable functions (direction statistics of the family) and let $p_0(\omega)$ be a probability density. The family $P_{\vec{s}}$ of probability laws, given by family of densities $p(\omega, \vec{s})$ with respect to the measure μ

$$p(\omega;\vec{s}) = p_0(\omega)\exp[s^j q_j(\omega)-\Psi(\vec{s})], \tag{10}$$

where $\exp[\Psi(\vec{s})]$ is a normalizing divisor

$$\Psi(\vec{s}) = \ln\int_\Omega p_0(\omega)\exp[s^j q_j(\omega)]\mu(d\omega), \tag{11}$$

is called an *exponential family in canonical parametrization* $\vec{s}$.

The probability distributions $P_{\vec{s}}$ are for all $\vec{s}$ for which the normalizing divisor is finite.

Such $\vec{s}$ form a convex set of an m-dimensional linear space of parameters with a possibly smaller dimension.

The canonical parametrization is defined by the family itself up to a nonsingular affine transformation. But there exists another remarkable parametrization (a natural one) in exponential families:

$$t_j(\vec{s}) = \int_\Omega q_j(\omega)P_{\vec{s}}(d\omega) = M_{\vec{s}}q_j(\omega), \quad \forall j. \tag{12}$$

It is connected with the corresponding canonical one by the Legendre transformation

$$M_{\vec{s}}q_j(\omega) = t_j = \frac{\partial\Psi(\vec{s})}{\partial s^j}, \qquad \frac{\partial t_j}{\partial s^k} = M_{\vec{s}}[q_j(\omega)-t_j][q_k(\omega)-t_k] = \frac{\partial^2\Psi(\vec{s})}{\partial s^j\partial s^k}, \tag{13}$$

which can be got by the differentiation of the identity

$$1 = \int_\Omega p_0(\omega)[\exp s^j q_j(\omega)-\Psi(\vec{s})]\mu(d\omega).$$

The density of the appropriate probability distribution can easily be obtained through the canonical parameters. On the other hand, the natural parameters have an exact mathematical meaning [3].

THEOREM 5. *If the smooth probability distribution family $\mathscr{F}$ admits a (jointly) efficient estimate of a vector parameter then $\mathscr{F}$ is an exponential in a natural parametrization, and vice versa.*

A natural parametrization differs from a canonical one as a rule. The family of normal laws with a constant covariance matrix is the only exception.

We have described families which admit a parameter estimate efficient with respect to the Fréchet–Darmois–Cramér–Rao information inequality. But there exists other inequalities of the same type for a somewhat different loss function.

PROBLEM 3. *Do there exist families which admit an efficient estimate of a parameter with respect to other information inequalities? What is their characterization?*

Now let us discuss what the properties of the distance between probability distributions which would specify the unlikeness of appropriate random phenomena should be. First of all, note that the likeness relation of two objects is asymmetric. One object may have its own additional distinctive features which the other object lacks. Now we want to tell a false coin with heads on both sides from a real one by tossing results. The real coin has an additional quality. Therefore, tossing it at random, we shall sometimes see a tail and correctly conclude that the coin is real. In tossing a false coin, only heads will show. And no matter how many times we set the test, we cannot draw a faultless conclusion since the possibility would always remain of a chance run of heads. Thus the false coin resembles the real one. On the contrary, the real coin does not resemble the false one too much.

We have taken an extreme case, which is particularly obvious. Now let us apply to the classical theory of testing two random phenomena. Let two simple hypotheses $P_0(d\omega)$ and $P_1(d\omega)$ compete.

If one requires the probability of the second kind error not to exceed a fixed value b, where $0 < b < 1$, then for the optimal decision rule the probability of the first kind error would decrease according to $\exp\{-N\cdot I(P_1 : P_0)\}$. Here N is the number of independent observations of a phenomenon, and $I(P_1 : P_0)$ is the Kulback–Leibler–Sanov information deviation:

$$(14)\qquad I(P_1 : P_0) = \int_\Omega \left[\frac{dP_1}{dP_0}(\omega)\ln\frac{dP_1}{dP_0}(\omega)\right] P_0(d\omega) = \int_\Omega \left[\ln\frac{dP_1}{dP_0}(\omega)\right] P_1(d\omega).$$

In general, $I(P_1 : P_0) \neq I(P_0 : P_1)$.

This non-symmetric invariant of the pair of laws possesses so many remarkable properties that it can be taken naturally for an asymmetric likeness characteristic.

Here the most interesting thing perhaps is that the information deviation has been implicitly used in that capacity for a long time. The maximum likelihood method is well known. We shall assume that independent observations $\omega^{(1)}, \ldots, \omega^{(N)}$ have been made, $R_N(\cdot) = \frac{1}{N}\sum_{i=1}^{N} \delta_{\omega^{(i)}}(\cdot)$ is an empirical distribution and an a priori family of hypotheses P_θ is given by the smooth family of densities $p(\omega;\theta)$. Then

$$\begin{aligned} I(R : P_\theta) &= \int_\Omega \left[\ln\frac{dR}{dP_\theta}(\omega)\right] R(d\omega) \\ &= I(R : P_0) + \int_\Omega [\ln p(\omega; 0) - \ln p(\omega;\theta)] R(d\omega) \\ &= I(R : P_0) + \int_\Omega \ln p(\omega; 0) R(d\omega) - \frac{1}{N}\sum_{i=1}^{N} \ln p(\omega^{(i)};\theta). \end{aligned}$$

Thus the minimum of the information deviation takes place at the likelihood maximum. Certainly the computation carried out above is only correct for the discrete space Ω of events.

The information deviation is an asymmetric analogy of the distance square (not a distance), or, it is better to say, an asymmetrical analogy of half the square of the distance between probability measures. With small deviations this value concurs, for example, with half of the Fisher distance square:

$$I = -\sum_j p_j \ln\left(1 + \frac{\Delta_j}{p_j}\right) \approx \sum_j \frac{(\Delta_j)^2}{p_j}.$$

The most remarkable thing is that the following version of Pythagorean theorem is correct for the information deviation: the square of the slant height equals the square of the projection length plus the square of the perpendicular length. Two formulations of the theorem are possible (see Problems A and B, § 22 in [5]). We shall need that one in which the rôle of a plane (or a line) will be played by an exponential family.

THEOREM 6. *Let $\{P_{\vec{s}}\}$ be an exponential family given by* (10), *let a probability measure R be absolutely continuous w.r. $P_{\vec{s}}$ ($R \ll P_{\vec{s}}$) and let the distribution $P_{\vec{\sigma}}$ be such that*

$$(15)\qquad \int_\Omega q_j(\omega) R(d\omega) = t_j(\vec{\sigma}) = \int_\Omega q_j(\omega) P_{\vec{\sigma}}(d\omega), \qquad \forall j.$$

Then for all $P_{\vec{s}}$

$$(16)\qquad I(R : P_{\vec{s}}) = I(R : P_{\vec{\sigma}}) + I(P_{\vec{\sigma}} : P_{\vec{s}}).$$

COROLLARY. *The measure $P_{\vec{\sigma}}$ is the I-closest to point R of the family.*

From formula (10) and definition (15) of the projection it follows that

$$\begin{aligned} I(R : P_{\vec{s}}) &= \int_\Omega \left[\ln\frac{dR}{dP_{\vec{\sigma}}}(\omega) + \ln\frac{dP_{\vec{\sigma}}}{dP_{\vec{s}}}(\omega)\right] R(d\omega) \\ &= I(R : P_{\vec{\sigma}}) + \int_\Omega [(\sigma^j - s^j) q_j(\omega) - \Psi(\vec{\sigma}) + \Psi(\vec{s})] R(d\omega) \\ &= I(R : P_{\vec{\sigma}}) + \int [\;] P_{\vec{\sigma}}(d\omega) = I(R : P_{\vec{\sigma}}) + I(P_{\vec{\sigma}} : P_{\vec{s}}). \end{aligned}$$

The asymmetric information Pythagorean geometry was first developed by the author in [6] (see also [5]). Recently the same considerations but in somewhat different way have been presented by Csiszar [4].

PROBLEM 4. *Is an axiomatic description of the asymmetrical Pythagorean geometry possible?*

From Theorem 6 a curious geometrical corollary follows. Let P_0 and P_1 be two mutually absolutely continuous probability measures. The exponential family $\{P_u\}$ passing through them is specified with the likelihood function

$$(17)\qquad \ln p(\omega; u) = \ln p_0(\omega) + u[\ln p_1(\omega) - \ln p_0(\omega)] - \Psi(u).$$

Then the function $\Psi(u)$, which is convex by (13), has the only maximum at a certain value of $u = v$, $0 < v < 1$, and also

$$(18)\qquad \left.\frac{d\Psi}{du}\right|_v = 0 = \int_\Omega [\ln p_1(\omega) - \ln p_0(\omega)] P_v(d\omega).$$

From (17) and (18) it can be found that for $u = v$

$$(19)\qquad I(P_v : P_1) = I(P_v : P_0) = -\Psi(v) = J(P_0, P_1).$$

The point P_v divides the family $\{P_u\}$ into two parts:

$$\begin{aligned} I(P_u : P_0) &> I(P_u : P_1) \quad \text{for} \quad u < v, \\ I(P_u : P_0) &< I(P_u : P_1) \quad \text{for} \quad u > v. \end{aligned}$$

From this and Theorem 6, the point P_v defines the bound between the domains of "attraction" of the points P_0 and P_1. The measure R is "closer" to P_0 or to P_1 depending on whether its projection on $\{P_u\}$ lies to the left or to the right of P_v, i.e. whether the appropriate natural coordinate is negative or positive:

$$\int_{\Omega} [\ln p_1(\omega) - \ln p_0(\omega)] R(d\omega).$$

Hence, according to (16),

$$J(P_0, P_1) = \min_{R \ll P} \max_{k=0,1} I(R : P_k). \tag{20}$$

This formula permits us to give a geometric interpretation to Chernoff's result [2] about the minimax testing of two simple hypotheses.

THEOREM 7. *Under the optimal minimax testing of two simple hypotheses the probabilities of the first and second kind errors, equal to each other, decrease as* $\exp\{-N \cdot J(P_0, P_1)\}$ *where N is the number of independent observations and $J(P_0, P_1)$ is the minimum size of the domains of "attraction" according to* (20).

The maximum likelihood criterion will be the simplest asymptotically optimal test. It can easily be seen that the null hypothesis has more likelihood iff

$$\int_{\Omega} [\ln p_1(\omega) - \ln P_0(\omega)] R_N(d\omega) < 0,$$

i.e. iff the empirical distribution $R_N(\cdot) = \frac{1}{N} \sum_{l=1}^{N} \delta_{\omega^{(l)}}(\cdot)$ belongs (generally speaking in a generalized sense) to the domain of "attraction" of the distribution P_0.

As follows from [7] analogous formulations appear in testing three or more simple hypotheses. Depending on the statement of the problem, the maximum rate of decreasing the logarithm of an error probability is defined by the set of sizes of the Dirichlet domains of testing laws and by their deviations from each other.

PROBLEM 5. *Extend Theorem 7 to the testing of two composite hypotheses, assuming that a null family and an alternative one are compact and smooth, and the minimum size of the domain of attraction is defined by*

$$J = \min_{R \ll P} \max_{k=0,1} \min_{P_k \in H_k} I(R : P_k).$$

Let us now consider the possible advantages of a geometrical approach applied to the classical problems of a parameter estimation. Let us assume that a certain smooth family $\{P_\theta, \theta \in \Theta\}$ of a priori possible laws on (Ω, S) is known. From N independent observations we must estimate the value θ for the law P of the observed phenomenon.

Connected with Cramér, Rao, Darmois, Fréchet and other names, there exists an information inequality which sets the lower bound of the estimate accuracy. However, for the statistician the problem statement mentioned above may not be fully adequate for his aims. Indeed, we mostly need to estimate the law P_θ itself and not its "name" θ. But a smooth family can be smoothly reparametrized; for example, from parametrizing the family of normal laws by a dispersion one can pass to parametrizing by a mean-root-square error. From a formal point of view this can be explained by observing that it is contravariant but not invariant. Nevertheless let us derive an invariant corollary which will be an absolute limitation of the estimate accuracy.

First of all we shall measure the deviation of the "true" law P_ϑ from the estimate Q in an invariant way. So we take

$$L(P_\vartheta, Q) = 2I(Q : P_\vartheta)$$

as an appropriate loss function. We shall further let the estimate Q take a value in the whole totality $\mathrm{Cap}(\Omega, S)$ of distributions on (Ω, S) but not only in the family $\{P_\theta\}$. Note that for some problems this extension can be justified and even useful. We shall make it evident by using an example from the theory of shooting. Suppose we must hit a pin-point target placed in some line in a plane (for example, a machine-gun placed behind an embankment). If the embankment has the form of an arc, it is more advantageous to aim at point shifted a little along the arc radius; in this case the probability of hitting the target by a shell splinter increases.

Thus we have come to the risk function

$$\mathfrak{R}_{\Pi}(P_\theta) = \int_{\Omega^N} \int_{\mathrm{Cap}(\Omega, S)} 2I(Q : P_\theta) \Pi(\omega^{(1)}, \ldots, \omega^{(N)}; d[Q(\cdot)] P_\theta(d\omega^{(1)}) \ldots P_\theta(d\omega^{(N)}). \tag{21}$$

Let dV be an invariant volume generated by the invariant Riemann–Fisher metric (6′). Now define an average risk along the open domain

$$\mathfrak{M}\mathfrak{R}_{\Pi} = \frac{1}{V(\Theta)} \int_{\Theta} R_{\Pi}(P_\theta) \, dV(\theta). \tag{22}$$

THEOREM 8. *Let $\Pi(N)$ be a decision rule constructed from N independent observations to estimate an unknown law of distribution from the smooth family $\{P_\theta, \theta \in \Theta\}$. Then*

$$\lim_{N \to \infty} N \cdot \inf_{\Pi(N)} \mathfrak{M}\mathfrak{R}_{\Pi(N)} = \dim \Theta. \tag{23}$$

This beautiful theorem is a corollary of the classical information inequality, i.e. it preserves only a part of the meaning of the inequality. Easily remembered (since it is formulated in invariant terms), the asymptotic equality (23) has its own value. It shows that superefficient estimates [8] differ little from the usual efficient ones and that super-efficiency is a property which diminishes while the number of observations increases.

The fact that in (23) it is possible to put the sign of asymptotic equality instead of that of asymptotic inequality, which follows directly from the information inequality, shows that the Cramér–Rao–Darmois–Fréchet inequality is the only essential restriction on the accuracy of estimating the smooth families of mutually absolutely continuous probability distributions. All the remaining inequalities of the same type either are weaker or coincide with that asymptotically; the Cramér–Rao–Darmois–Fréchet inequality can only be made more precise by higher order corrections.

PROBLEM 6. *Find the bound for the remainder term in* (23) *in terms of the maximum of invariant curvature of the family* $\{P_\theta\}$.

Note that under the chosen loss function $2I(Q:P_\theta)$ for exponential families the problem of estimating the law P_θ reduces (due to Theorem 6) to the problem of estimating its parameter, i.e. the estimates of $Q \in \{P_\theta\}$-type form a complete class.

PROBLEM 7. *In what natural terms family smoothness should be described in Theorem 8?*

It is completely obscure how to replace the usual sufficient conditions of smoothness in terms of majorant existence in the third derivatives since the analysis then requires to be developed in non-linear, non-topological space.

PROBLEM 8. *How does the formulation of Theorem* 8 *change when the smooth family* $\{P_\theta\}$ *has points of self-intersection?*

Acknowledgement. The author considers it his pleasant duty to thank Professor R. Bartoszyński for his attention.

References

[1] D. Blackwell, Ann. Math. Statist., 24 (1953), pp. 265–272.
[2] H. Chernoff, ibid. 23 (1952), pp. 493–507.
[3] H. Cramér, *Mathematical methods of statistics*, Princeton Univ. Press. Princeton 1946.
[4] I. Csiszar, Ann. Probab. 3 (1975), pp. 146–158.
[5] N. N. Čencov, *Statistical decision rules and optimal inference*, Nauka, Moscow 1972.
[6] —, Math. Notes 4 (1968), pp. 323–332.
[7] N. P. Salikhov, DAN USSR 209 (1972).
[8] Ch. Stein, J. Roy. Statist. Soc. 24 (1962), pp. 265–296.
[9] A. Wald, Ann. Math. Statist. 10 (1939), pp. 299–326.

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

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

HSU'S THEOREM IN VARIANCE COMPONENT MODELS*

HILMAR DRYGAS**

J. W. Goethe-Universität, Fachbereich Mathematik
6 Frankfurt am Main, Robert-Mayer-Strasse 6–10, FRG

0. Introduction

This paper deals with linear models of the kind $y = X\theta + U\varepsilon$, where ε is a random vector composed of independent random variables. Therefore $\operatorname{Cov}\varepsilon = \sum_{i=1}^{k} \sigma_i^2 V_i$, where $V_i = \operatorname{diag}(0, \ldots, I_i, \ldots, 0)$ and I_i is the unit matrix of appropriate order (or any diagonal matrix). Much work has been done to investigate the problem of existence of uniformly best (invariant) quadratic unbiassed estimators if ε is normally or quasi-normally distributed. It is the purpose of this paper to extend these results to the non-normal case. This extension is done in the case of best invariant quadratic unbiassed estimators. A complicated matrix-relation turns out to ensure optimality. But in analogy to Hsu's theorem ist can be shown that this relation can be replaced by requiring it only for the diagonal elements. The obtained results still appear very complicated but it turns out that due to the diagonality of the V_i the verification of the obtained conditions is rather straightforward. This is illustrated at two examples: the balanced one-way and the balanced two-way classification model.

1. Notation, Hsu's theorem

Let X be an $n \times s$-matrix, let θ be an $s \times 1$-vector and U an $n \times r$-matrix. Consider the linear model

$$y = X\theta + U\varepsilon, \tag{1.1}$$

where $y = (y_1, \ldots, y_n)'$, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_r)'$ are random vectors. It is assumed that the components of ε behave up to their moments of order 4 as independent random

* Paper presented at the conference on "Mathematical Statistics" Wisła, Poland, 13.–18.12. 1976.

** OE 3 der GhK, Gesamthochschule Kassel, Department of Mathematics, Heinrich-Plett. Str. 40, D 3500 Kassel, FRG.