ON BASIC CONCEPTS OF STATISTICS

JAROSLAV HÁJEK
MATHEMATICAL INSTITUTE OF THE CZECHOSLOVAK ACADEMY OF SCIENCES,
and CHARLES UNIVERSITY, PRAGUE

1. Summary

This paper is a contribution to current discussions on fundamental concepts, principles, and postulates of statistics. In order to exhibit the basic ideas and attitudes, mathematical niceties are suppressed as much as possible. The heart of the paper lies in definitions, simple theorems, and nontrivial examples. The main issues under analysis are *sufficiency, invariance, similarity, conditionality, likelihood*, and their mutual relations.

Section 2 contains a definition of sufficiency for a subparameter (or sufficiency in the presence of a nuisance parameter), and a criticism of an alternative definition due to A. N. Kolmogorov [11]. In that section, a comparison of the principles of sufficiency in the sense of Blackwell-Girschick [2] and in the sense of A. Birnbaum [1] is added. In theorem 3.5 it is shown that for nuisance parameters introduced by a group of transformations, the sub-σ-field of invariant events is sufficient for the respective subparameter.

Section 4 deals with the notion of similarity in the x-space as well as in the (x, θ)-space, and with related notions such as ancillary and exhaustive statistics. Confidence intervals and fiducial probabilities are shown to involve a postulate of "independence under ignorance."

Sections 5 and 6 are devoted to the principles of conditionality and of likelihood, as formulated by A. Birnbaum [1]. Their equivalence is proved and their strict form is criticized. The two principles deny gains obtainable by mixing strategies, disregarding that, in non-Bayesian conditions, the expected maximum conditional risk is generally larger than the maximum overall risk. Therefore, the notion of "correct" conditioning is introduced, in a general enough way to include the examples given in the literature to support the conditionality principle. It is shown that in correct conditioning the maximum risk equals the expected maximum conditional risk, and that in invariant problems the sub-σ-field of invariant events yields the deepest correct conditioning.

A proper field of application of the likelihood principle is shown to consist of families of experiments, in which the likelihood functions, possibly after a common transformation of the parameter, have approximately normal form with constant variance. Then each observed likelihood function allows computing the risk without reference to the particular experiment.

In section 7, some forms of the Bayesian approach are touched upon, such as those based on diffuse prior densities, or on a family of prior densities.

In section 8, some instructive examples are given with comments.

References are confined to papers quoted only.

2. Sufficiency

The notion of sufficiency does not promote many disputes. Nonetheless, there are two points, namely sufficiency for a subparameter and the principle of sufficiency, which deserve a critical examination.

2.1. *Sufficiency for a subparameter.* Let us consider an experiment (X, θ), where X denotes the observations and θ denotes a parameter. The random element X takes its values in an x-space, and θ takes its values in a θ-space. A function τ of θ will be called a subparameter. If θ were replaced by a σ-field, then τ would be replaced by a sub-σ-field. Under what conditions may we say that a statistic is sufficient for τ? Sufficiency for τ may also be viewed as sufficiency in the presence of a nuisance parameter. The present author is aware of only one attempt in this direction, due to Kolmogorov [11].

DEFINITION 2.1. *A statistic $T = t(X)$ is called sufficient for a subparameter τ, if the posterior distribution of τ, given $X = x$, depends only on $T = t$ and on the prior distribution of θ.*

Unfortunately, the following theorem shows that the Kolmogorov definition is void.

THEOREM 2.1. *If τ is a nonconstant subparameter, and if T is sufficient for τ in the sense of definition 2.1, then T is sufficient for θ as well.*

PROOF. For simplicity, let us consider the discrete case only. Let T be not sufficient for θ, and let us try to show that it cannot be sufficient for τ. Since T is not sufficient for θ, there exist two pairs, (θ_1, θ_2) and (x_1, x_2), such that

$$T(x_1) = T(x_2) \tag{2.1}$$

and

$$\frac{P_{\theta_1}(X = x_1)}{P_{\theta_2}(X = x_1)} \neq \frac{P_{\theta_1}(X = x_2)}{P_{\theta_2}(X = x_2)}. \tag{2.2}$$

If $\tau(\theta_1) \neq \tau(\theta_2)$, let us consider the following prior distribution: $\nu(\theta = \theta_1) = \nu(\theta = \theta_2) = \frac{1}{2}$. Then, for $\tau_1 = \tau(\theta_1)$ and $\tau_2 = \tau(\theta_2)$,

$$P(\tau = \tau_1 | X = x_1) = P_{\theta_1}(X = x_1)/[P_{\theta_1}(X = x_1) + P_{\theta_2}(X = x_1)] \tag{2.3}$$

and

$$P(\tau = \tau_1 | X = x_2) = P_{\theta_1}(X = x_2)/[P_{\theta_1}(X = x_2) + P_{\theta_2}(X = x_2)]. \tag{2.4}$$

Obviously, (2.2) entails $P(\tau = \tau_1 | X = x_1) \neq P(\tau = \tau_1 | X = x_2)$ which implies, in view of (2.1), that T is not sufficient for τ.

If $\tau(\theta_1) = \tau(\theta_2)$ held, we would choose θ_3 such that $\tau(\theta_3) \neq \tau(\theta_1) = \tau(\theta_2)$. Note that the equations

$$\frac{P_{\theta_1}(X = x_1)}{P_{\theta_3}(X = x_1)} = \frac{P_{\theta_1}(X = x_2)}{P_{\theta_3}(X = x_2)} \tag{2.5}$$

and

$$\frac{P_{\theta_2}(X = x_1)}{P_{\theta_3}(X = x_1)} = \frac{P_{\theta_2}(X = x_2)}{P_{\theta_3}(X = x_2)} \tag{2.6}$$

are not compatible with (2.2). Thus either (2.5) or (2.6) does not hold, and the above reasoning may be accomplished either with (θ_1, θ_3) or (θ_2, θ_3), both pair satisfying the condition $\tau(\theta_1) \neq \tau(\theta_3)$, $\tau(\theta_2) \neq \tau(\theta_3)$. Q.E.D.

Thus we have to try to define sufficiency in the presence of nuisance parameters in some less stringent way.

DEFINITION 2.2. *Let $\mathcal{P}_\tau$ be the convex hull of the distributions $\{P_\theta, \tau(\theta) = \tau\}$ for all possible τ-values. We shall say that T is sufficient for τ if*

(i) *the distribution of T will depend on τ only, that is,*

$$P_\theta(dt) = P_\tau(dt), \tag{2.7}$$

and

(ii) *there exist distributions $Q_\tau \in \mathcal{P}_\tau$ such that T is sufficient for the family $\{Q_\tau\}$.*

In the same manner we define a sufficient sub-σ-field for τ.

Now we shall prove an analogue of the well-known Rao-Blackwell theorem. For this purpose, let us consider a decision problem with a *convex* set D of decisions d, and with a loss function $L(\tau, d)$, which is *convex* in d for each τ, and depends on θ only through τ, $L(\theta, d) = L(\tau(\theta), d)$. Applying the minimax principle to eliminate the nuisance parameter, we associate with each decision function $\delta(x)$ the following risk:

$$R(\tau, \delta) = \sup_{\tau(\theta) = \tau} \int L[\tau(\theta), \delta(x)]P_\theta(dx), \tag{2.8}$$

where the supremum is taken over all θ-values such that $\tau(\theta) = \tau$. Now, if T is sufficient for τ in the sense of definition 2.2, we can associate with each decision function $\delta(x)$ another decision function $\delta(t)$ defined as follows:

$$\bar{\delta}(t) = \int \delta(x)Q_\tau(dx|T = t) \tag{2.9}$$

provided that $\delta(x)$ is integrable. Note that the right side of (2.9) does not depend on τ, since T is sufficient for $\{Q_\tau\}$, according to definition 2.2, and that $\bar{\delta}(t) \in D$ for every t in view of convexity of D. Finally put

$$\delta^*(x) = \bar{\delta}(t(x)). \tag{2.10}$$

THEOREM 2.2. *Under the above assumptions,*

$$R(\tau, \delta^*) \leq R(\tau, \delta) \tag{2.11}$$

holds for all τ.

PROOF. Since the distribution of T depends on τ only, we have

$$\begin{aligned} R(\tau, \delta^*) &= \int L[\tau(\theta), \delta^*(x)]P_\theta(dx) \\ &= \int L[\tau(\theta), \delta^*(x)]Q_\tau(dx) \end{aligned} \tag{2.12}$$

for all θ such that $\tau(\theta) = \tau$. Furthermore, since $L(\tau, d)$ is convex in d,

$$(2.13)\qquad \int L[\tau(\theta), \delta^*(x)]Q_\tau(dx) \le \int L[\tau(\theta), \delta(x)]Q_\tau(dx) \le \sup_{\tau(\theta)=\tau} \int L[\tau(\theta), \delta(x)]P_\theta(dx) = R(\tau, \delta).$$

This concludes the proof.

Thus, adopting the minimax principle for dealing with nuisance parameters, and under the due convexity assumptions, one may restrict himself to decision procedures depending on the sufficient statistic only. This finds its application in points estimation and in hypothesis testing, for example.

Remark 2.1. Le Cam [13] presented three parallel ways of defining sufficiency. All of them could probably be used in giving equivalent definitions of sufficiency for a subparameter. For example, Kolmogorov's definition 2.1 could be reformulated as follows: T is called sufficient for τ if it is sufficient for each system $\{P_{\theta'}\}$ such that $\{\theta'\} \subset \{\theta\}$ and $\theta_1' \neq \theta_2' \Rightarrow \tau(\theta_1') \neq \tau(\theta_2')$.

Remark 2.2. We also could extend the notion of sufficiency by adding some "limiting points." For example, we could introduce ϵ-sufficiency, as in Le Cam [13], and then say that T is sufficient if it is ϵ-sufficient for every $\epsilon > 0$, or if it is sufficient for every compact subset of the τ-space.

Remark 2.3. (Added in proof.) Definition 2.2 does not satisfy the natural requirement that T should be sufficient for τ, if it is sufficient for some finer subparameter τ', $\tau = \tau(\tau')$. To illustrate this point, let us consider a sample from $N(\mu, \sigma^2)$ and put $T = (\bar{x}, s^2)$, $\tau = \mu$, $\tau' = (\mu, \sigma^2)$. Then T is not sufficient for τ in the sense of definition 2.2, since its distribution fails to be dependent on μ only. On the other hand, T is sufficient for τ' as is well known. The definition 2.2 should be corrected as follows.

Definition 2.2*. *A statistic $T = t(X)$ is called sufficient for a subparameter τ, if it is sufficient in the sense of definition 2.2 for some subparameter τ' such that $\tau = \tau(\tau')$.*

Remark 2.4. (Added in proof.) A more stringent (and, therefore, more consequential) definition of sufficiency for a parameter is provided by Lehmann [14] in problem 31 of chapter III: T is sufficient for τ if, first, $\theta = (\tau, \eta)$, second, $P_\theta(dt) = P_\tau(dt)$, third, $P_\theta(dx|T = t) = P_\eta(dx|T = t)$. If T is sufficient in this sense, it is also sufficient in the sense of definition 2.2 where we may take $Q_\tau = P_{\tau,\eta_1}$ for any particular η_1.

2.2. *The principle of sufficiency.* If comparing the formulations of this principle as given in Blackwell-Girshick [2] and in A. Birnbaum [1], one feels that there is an apparent discrepancy. According to Birnbaum's sufficiency principle, we are not allowed to use randomized tests, for example, while no such implication follows from the Blackwell-Girshick sufficiency principle. The difference is serious but easy to explain: Blackwell and Girshick consider only convex situations (that is, convex D and convex $L(\theta, d)$ for each θ), where the Rao-Blackwell theorem can be proved, while A. Birnbaum has in mind any possible situation. However, what may be supported in convex situations by a theorem is a rather stringent postulate in a general condition. (If, in Blackwell-Girshick, the situation is not convex, they make it convex by allowing randomized decisions.)

In estimating a real parameter with $L(\theta, d) = (\theta - d)^2$, the convexity conditions are satisfied and no randomization is useful. If, however, θ would run through a discrete subset of real numbers, randomization might bring the same gains from the minimax point of view as in testing a simple hypothesis against a simple alternative. And even in convex situations the principle may not exclude any decision procedure as inadmissible. To this end it is necessary for $L(\tau, d)$ to be strictly convex in d, and not, for example, linear, as in randomized extensions of nonconvex problems.

3. Invariance

Most frequently, the nuisance parameter is introduced by a group of transformations of the x-space on itself. Then, if we have a finite invariant measure on the group, we can easily show that the sub-σ-field of invariant events is sufficient in the presence of the corresponding nuisance parameter.

Let $G = \{g\}$ be a group of one-to-one transformations of the x-space on itself. Take a probability distribution P of X and put

$$P_g(X \in A) = P(gX \in A). \tag{3.1}$$

Then obviously, $P_h(X \in g^{-1}A) = P_{gh}(X \in A)$. Let μ be a σ-finite measure and denote by μg the measure such that $\mu g(A) = \mu(gA)$.

THEOREM 3.1. *Let $P \ll \mu$ and $\mu g \ll \mu$ for all $g \in G$. Then $P_g \ll \mu$. Denoting $p(x, g) = dP_g/d\mu$ and $p(x) = dP/d\mu$, then*

$$p(x, g) = p(g^{-1}x) \frac{d\mu g^{-1}}{d\mu}(x) \tag{3.2}$$

holds. More generally,

$$p(x, h^{-1}g) = p(h(x), g) \frac{d\mu h}{d\mu}(x). \tag{3.3}$$

PROOF. According to the definition, one can write

$$\begin{aligned} P_g(X \in A) = P(X \in g^{-1}A) &= \int_{g^{-1}A} p(x)\, d\mu = \int_A p(g^{-1}(y))\, d\mu g^{-1}(y) \\ &= \int_A p(g^{-1}(y)) \frac{d\mu g^{-1}}{d\mu}(y)\, d\mu(y). \end{aligned} \tag{3.4}$$

CONDITION 3.1. *Let $\mathcal{G}$ be a σ-field of subsets of G, and let $\mathcal{A}$ be the σ-field of subsets of the x-space. Assume that*

(i) *$\mu g \ll \mu$ for all $g \in G$,*

(ii) *$p(x, g)$ is $\mathcal{A} \times \mathcal{G}$-measurable,*

(iii) *functions $\phi_h(g) = hg$ and $\psi_h(g) = gh$ are $\mathcal{G}$-measurable,*

(iv) *there is an invariant probability measure ν on $\mathcal{G}$, that is, $\nu(Bg) = \nu(gB) = \nu(B)$ for all $g \in G$ and $B \in \mathcal{G}$.*

THEOREM 3.2. *Under condition 3.1, let us put*

$$\bar{p}(x) = \int p(x, g)\, d\nu(g). \tag{3.5}$$

Then, for each $h \in G$,

$$\bar{p}(h(x)) = \left[\frac{d\mu h}{d\mu}(x)\right]^{-1} \bar{p}(x). \tag{3.6}$$

REMARK 3.1. Note that the first factor on the right-hand side does not depend on p. Obviously $\bar{p}(x, g) = \bar{p}(x)$ for all $g \in G$.

PROOF. In view of (3.3), we have

$$\begin{aligned} \bar{p}(h(x)) &= \int p(h(x), g)\, d\nu(g) = \left[\frac{d\mu h}{d\mu}(x)\right]^{-1} \int p(x, h^{-1}g)\, d\nu \\ &= \left[\frac{d\mu h}{d\mu}(x)\right]^{-1} \int p(x, f)\, d\nu h(f) \\ &= \left[\frac{d\mu h}{d\mu}(x)\right]^{-1} \int p(x, f)\, d\nu(f) \\ &= \left[\frac{d\mu h}{d\mu}(x)\right]^{-1} \bar{p}(x), \qquad \text{Q.E.D.} \end{aligned} \tag{3.7}$$

We shall say that an event is G-invariant, if $gA = A$ for all $g \in G$. Obviously, the set of G-invariant events is a sub-σ-field $\mathcal{B}$, and a measurable function f is $\mathcal{B}$-measurable if and only if $f(g(x)) = f(x)$ for all $g \in G$. Consider now two distributions P and P_0 and seek the derivative of P relative to P_0 on $\mathcal{B}$, say $[dP/dP_0]^{\mathcal{B}}$. Assume that $P \ll \mu$ and $P_0 \ll \mu$, denote $p = dP/d\mu$, $p_0 = dP_0/d\mu$, and introduce $\bar{p}(x)$ and $\bar{p}_0(x)$ by (3.5).

THEOREM 3.3. *Under condition 3.1 and under the assumption that $\bar{p}_0(x) = 0$ entails $\bar{p}(x) = 0$ almost μ-everywhere, we have $P \ll P_0$ on $\mathcal{B}$ and*

$$[dP/dP_0]^{\mathcal{B}} = \bar{p}(x)/\bar{p}_0(x). \tag{3.8}$$

PROOF. Put $l(x) = \bar{p}(x)/\bar{p}_0(x)$. Theorem 3.2 entails $l(g(x)) = l(x)$ for all $g \in G$, namely, $l(x)$ is $\mathcal{B}$-measurable. Further, for any $B \in \mathcal{B}$, in view of (3.2) and of $[p = 0] \Rightarrow [p_0 = 0]$,

$$\begin{aligned} \int_B l(x)\, dP_0 &= \int_B l(x) p_0(x)\, d\mu = \int_B l(x) p_0(g^{-1}(x))\, d\mu g^{-1} \\ &= \int_B l(x) p_0(x, g)\, d\mu = \int_B l(x) \bar{p}_0(x)\, d\mu = \int_B \bar{p}(x)\, d\mu \end{aligned} \tag{3.9}$$

holds. On the other hand,

$$P(B) = \int_B p(x)\, d\mu = \int_B p(g^{-1}(x))\, d\mu g^{-1} = \int_B p(x, g)\, d\mu = \int_B \bar{p}(x)\, d\mu. \tag{3.10}$$

Thus $P(B) = \int_B l(x)\, dP_0$, $B \in \mathcal{B}$, which concludes the proof.

THEOREM 3.4. *Let the statistic $T = t(X)$ have an expectation under P_h, $h \in G$. Let condition 3.1 be satisfied. Then the conditional expectation of T relative to the sub-σ-field $\mathcal{B}$ of G-invariant events and under P_h equals*

$$E_h(T|\mathcal{B}, x) = \int t(g^{-1}(x)) p(x, gh)\, d\nu(g)/\bar{p}(x). \tag{3.11}$$

PROOF. The proof would follow the lines of the proofs of theorems 3.2 and 3.3.

Consider now a dominated family of probability distributions $\{P_\tau\}$ and define $P_{\tau,g}$ by (3.1) for each τ. Putting $\theta = (\tau, g)$, we can say that τ is a subparameter of θ.

THEOREM 3.5. *Under condition* 3.1 *the sub-σ-field* $\mathcal{B}$ *of G-invariant events is sufficient for τ in the sense of definition* 2.2.

PROOF. First, the G-invariant events have a probability depending on τ only, that is, $P_{\tau,g}(B) = P_\tau(B)$. Second, for

$$Q_\tau(A) = \int P_{\tau,g}(A)\,d\nu(g) \tag{3.12}$$

we have

$$\begin{aligned} Q_\tau(A) &= \int \left[\int_A p_\tau(x, g)\,d\mu\right] d\nu(g) \\ &= \int_A \int p_\tau(x, g)\,d\nu(g)\,d\mu \\ &= \int_A \bar{p}_\tau(x)\,d\mu. \end{aligned} \tag{3.13}$$

Now let P_0 be some probability measure such that $P_\tau \ll P_0 \ll \mu$, and introduce $\bar{p}_0(x)$ by (3.5). Then, according to theorem 3.3, $\bar{p}_\tau(x)/\bar{p}_0(x)$ is $\mathcal{B}$-measurable for all τ. Thus $\bar{p}(x) = [\bar{p}_\tau(x)/\bar{p}_{\tau 0}(x)]\bar{p}_0(x)$, and $\mathcal{B}$ is sufficient for $\{Q_\tau\}$, according to the factorization criterion.

REMARK 3.2. Considering right-invariant and left-invariant *probability* measures on $\mathcal{G}$ does not provide any generalization. Actually, if ν is right-invariant, then $\bar{\nu}(B) = \int_A \nu(gB)\,d\nu(g)$ is invariant, that is, both right-invariant and left-invariant.

REMARK 3.3. Theorem 3.3 sometimes remains valid even if there exists only a right-invariant countably finite measure ν, as in the case of the groups of location and/or scale shifts (see [6], chapter II).

4. Similarity

The concept of similarity plays a very important role in classical statistics, namely in contributions by J. Neyman and R. A. Fisher. It may be applied in the x-space as well as in the (x, θ)-space.

4.1. *Similarity in the x-space.* Consider a family of distributions $\{P_\theta\}$ on measurable subsets of the x-space. We say that an *event A* is *similar*, if its probability is independent of θ:

$$P_\theta(A) = P(A) \qquad \text{for all} \quad \theta. \tag{4.1}$$

Obviously the events "whole x-space" and "the empty set" are always similar.

The class of similar events is closed under complementation and formation of countable *disjoint* unions. Such classes are called λ-fields by Dynkin [3]. A λ-field is a broader concept than a σ-field. The class of similar events is usually not a σ-field, which causes ambiguity in applications of the conditionality principle, as we shall see. Generally, if both A and B are similar and are not disjoint, then $A \cap B$ and $A \cup B$ may not be similar. Consequently, the system of σ-fields contained in a λ-field may not include a largest σ-field.

Dynkin [3] calls a system of subsets a π-field, if it is closed under intersection, and shows that a λ-field containing a π-field contains the smallest σ-field over the π-field. Thus, for example, if for a vector statistic $T = t(X)$ the events $\{x : t(x) < c\}$ are similar for every vector c, then $\{x : T \in A\}$ is similar for every Borel set A.

More generally, a *statistic* $V = v(X)$ is called *similar*, if $E_\theta V = \int v(x) P_\theta(dx)$ exists and is independent of θ. We also may define similarity with respect to a nuisance parameter only. The notion of similarity forms a basis for the definition of several other important notions.

Ancillary statistics. A statistics $U = u(x)$ is called ancillary, if its distribution is independent of θ, that is, if the events generated by U are similar. (We have seen that it suffices that the events $\{V < c\}$ be similar.)

Correspondingly, a sub-σ-field will be called ancillary, if all its events are similar.

Exhaustive statistics. A statistics $T = t(x)$ is called exhaustive if (T, U), with U an ancillary statistic, is a minimal sufficient statistic.

Complete families of distributions. A family $\{P_\theta\}$ is called complete if the only similar statistics are constants.

Our definition of an exhaustive statistic follows Fisher's explanation (ii) in ([5], p. 49), and the examples given by him.

Denote by I_θ^X, I_θ^T, $I_\theta^T(u)$ Fisher's information for the families $\{P_\theta(dx)\}$, $P_\theta(dt)\}$, $\{P_\theta(dt) | U = u)\}$, respectively. Then, since (T, U) is sufficient and U is ancillary,

$$EI_\theta^T(U) = I_\theta^X. \tag{4.2}$$

If $I_\theta^T < I_\theta^X$, Fisher calls (4.2) "recovering the information lost." What is the real content of this phrase?

The first interpretation would be that $T = t$ contains all information supplied by $X = x$, provided that we know $U = u$. But knowing both $T = t$ and $U = u$, we know $(T, U) = (t, u)$, that is, we know the value of the sufficient statistics. Thus this interpretation is void, and, moreover, it holds even if U is not ancillary.

A more appropriate interpretation seems to be as follows. Knowing $\{P_\theta(dt | U = u)\}$, we may dismiss the knowledge of $P(du)$ as well as of $\{P_\theta(dt) | U = u')\}$ for $u' \neq u$. This, however, expresses nothing else as the conditionality principle formulated below.

Fisher makes a two-field use of exhaustive statistics: first, for extending the scope of fiducial distributions to cases where no appropriate sufficient statistic exists, and, second, to a (rather unconvincing) eulogy of maximum likelihood estimates.

The present author is not sure about the appropriateness of restricting ourselves to "minimal" sufficient statistics in the definition of exhaustive statistics. Without this restriction, there would rarely exist a minimal exhaustive statistic, and with this restriction we have to check, in particular cases, whether the employed sufficient statistic is really minimal.

4.2. *Similarity in the (x, θ)-space.* Consider a family of distributions $\{P_\theta\}$ on the x-space, the family of all possible prior distributions $\{\nu\}$ on the θ-space, and the family of distributions $\{R_\nu\}$ on the (x, θ)-space given by

$$R_\nu(dx\,d\theta) = \nu(d\theta)P_\theta(dx). \tag{4.3}$$

Then we may introduce the notion of similarity in the (x, θ)-space in the same manner as before in the x-space, with $\{P_\theta\}$ replaced by $\{R_\nu\}$. Consequently, the prior distribution will play the role of an unknown parameter. Thus an event Λ in the (x, θ)-space will be called similar, if

$$R_\nu(\Lambda) = R(\Lambda) \tag{4.4}$$

for all prior distributions ν.

To avoid confusion, we shall call measurable functions of (x, θ) *quantities* and not statistics. A quantity $H = h(X, \Theta)$ will be called similar if its expectation $EH = \int h(x, \theta)\nu(d\theta)P_\theta(dx)$ is independent of ν. Ancillary statistics in the (x, θ)-space are called *pivotal* quantities or *distribution-free* quantities. Since only the first component of (x, θ) is observable, the applications of similarity in the (x, θ)-space are quite different from those in the x-space.

Confidence regions. This method of region estimation dwells on the following idea. Having a similar event Λ, whose probability equals $1 - \alpha$, with α very small, and knowing that $X = x$, we can feel confidence that $(x, \theta) \in \Lambda$, namely, that the unknown θ lies within the region $S(x) = \{\theta : (x, \theta) \in \Lambda\}$. Here, our confidence that the event Λ has occurred is based on its high probability, and this confidence is assumed to be unaffected by the knowledge of $X = x$. Thus we assume a sort of independence between Λ and X, though their joint distribution is indeterminate. The fact that this intrinsic assumption may be dubious is most appropriately manifested in cases when $S(x)$ is either empty, so that we know that Λ did not occur, or equals the whole θ-space, so we are sure that Λ has occurred. Such a situation arises in the following.

EXAMPLE 1. For this example, $\theta \in [0, 1]$, x is real, $P_\theta(dx)$ is uniform over $[0, \theta + 2]$, and

$$\Lambda = \{(x, \theta) : \theta + \alpha < x < \theta + 2 - \alpha\}. \tag{4.5}$$

Then $S(x) = \emptyset$ for $3 - \alpha < x < 3$ or $0 < x < \alpha$, and $S(x) = [0, 1]$ for $1 + \alpha < x < 2 - \alpha$. Although this example is somewhat artificial, the difficulty involved seems to be real.

Fiducial distribution. Let $F(t, \theta)$ be the distribution function of T under θ, and assume that F is continuous in t for every θ. Then $H = F(T, \Theta)$ is a pivotal quantity with uniform distribution over $[0, 1]$. Thus $R(F(T, \Theta) \leq x) = x$. Now, if $F(t, \theta)$ is strictly decreasing in θ for each t, and if $F^{-1}(t, \theta)$ denotes its inverse for fixed t, then $F(T, \Theta) \leq x$ is equivalent to $\Theta \geq F^{-1}(T, x)$. Now, again, if we know that $T = t$, and if we feel that it does not affect the probabilities concerning $F(T, \Theta)$, we may write

$$(4.6)\qquad x = R(F(T, \Theta) \leq x = R(\Theta \geq F^{-1}(T, x)) = R(\Theta \geq F^{-1}(T, x)|T = t) \\ = R(\Theta \geq F^{-1}(t, x)|T = t)$$

namely, for $\theta = F^{-1}(t, x)$,

$$(4.7)\qquad R(\Theta < \theta|T = t) = 1 - F(t, \theta).$$

As we have seen, in confidence regions as well as in fiducial probabilities, a peculiar postulate of independence in involved. The postulate may be generally formulated as follows.

Postulate of independence under ignorance. Having a pair (Z, W) such that the marginal distribution of Z is known, but the joint distribution of (Z, W), as well as the marginal distribution of W, are unknown, and observing $W = w$, we assume that

$$(4.8)\qquad P(Z\in\Lambda|W = w) = P(Z\in\Lambda).$$

J. Neyman gives a different justification of the postulate than R. A. Fisher. Let us make an attempt to formulate the attitudes of both these authors.

Neyman's justification. Assume that we perform a long series of independent replications of the given experiments, and denote the results by $(Z_1, w_1), \cdots,$ (Z_N, w_N), where the w_i's are observed numbers and the Z_i's are not observable. Let us decide to accept the hypothesis $Z_i \in \Lambda$ at each replication. Then our decisions will be correct approximately in $100P\%$ of cases with $P = P(Z \in \Lambda)$. Thus, in a long series, our mistakes in determining $P(Z \in \Lambda|W = w)$ by (4.8) if any, will compensate each other.

Fisher's justification. Suppose that the only statistics $V = v(W)$, such that the probabilities $P(Z \in \Lambda|V = v)$ are well-determined, are constants. Then we are allowed to take $P(Z \in \Lambda|W = w) = P(Z \in \Lambda)$, since our absence of knowledge prevents us from doing anything better.

The above interpretation of Fisher's view is based on the following passage from his book ([5], pp. 54-55):

"The particular pair of values of θ and T appropriate to a particular experimenter certainly belongs to this enlarged set, and within this set the proportion of cases satisfying the inequality

$$(4.9)\qquad \theta > \frac{T}{2n}\chi^2_{2n}(P)$$

is certainly equal to the chosen probability P. It might, however, have been true . . . that in some recognizable subset, to which his case belongs, the proportion of cases in which the inequality was satisfied should have some value other than P. It is the stipulated absence of knowledge *a priori* of the distribution of θ, together with the exhaustive character of the statistic T, that makes the recognition of any such subset impossible, and so guarantees that in his particular case . . . the general probability is applicable."

To apply our general scheme to the case considered by R. A. Fisher, we should put $Z = 1$ if (4.9) is satisfied and $Z = 0$ otherwise, and $W = T$.

REMARK 4.1. We can see that Fisher's argumentation applies to a more specific situation than described in the above postulate. He requires that conditional probabilities are not known for any statistics $V = v(W)$ except for $V =$ const. Thus the notion of fiducial probabilities is a more special notion than the notion of confidence regions, since Neyman's justification does not need any such restrictions.

Fisher's additional requirement is in accord with his requirement that fiducial probabilities should be based on the minimal sufficient statistics, and in such a case does not lead to difficulties.

However, if the fiducial probabilities are allowed to be based on exhaustive statistics, then his requirement is contradictory, since no minimal exhaustive statistics may exist. Nonetheless, we have the following.

THEOREM 4.1. *If a minimal sufficient statistic $S = s(X)$ is complete, then it is also a minimal exhaustive statistic.*

PROOF. If there existed an exhaustive statistic $T = t(S)$ different from S, there would exist a nonconstant ancillary statistic $U = u(S)$, which contradicts the assumed completeness of S. Q.E.D.

If S is not complete, then there may exist ancillary statistics $U = u(S)$ and the family corresponding to exhaustive statistics contains a minimal member if and only if the family of U's contains a maximal member.

REMARK 4.2. Although the two above justifications are unconvincing, they correspond to habits of human thinking. For example, if one knows that an individual comes from a subpopulation of a population where the proportion of individuals with a property A equals P, and if one knows nothing else about the subpopulation, one applies P to the given individual without hesitation. This is true even if we know the "name" of the individual, that is, if the subpopulation consists of a single individual. Fisher is right, if he claims that we would not use P if the given subpopulation would be a part of a larger one, in which the proportion is known, too. He does not say, however, what to do if there is no minimal such larger subpopulation.

In confidence regions the subpopulation consists of pairs (X, Θ) such that $X = x$. It is true that we know the "proportion" of elements of that subpopulation for which the event Λ occurs, but we do not know how much of the probability mass each element carries. Thus we can utilize the knowledge of this proportion only if it equals 0 or 1, which is exemplified by example 1 but occurs rather rarely in practice. The problem becomes still more puzzling if the knowledge of the proportion is based on estimates only, and if these estimates become less reliable as the subpopulation from which they are derived becomes smaller. Is there any reasonable recommendation as to what to do if we know $P(X \in A | U_1 = u_1)$ and $P(X \in A | U_2 = u_2)$, but we do not know $P(X \in A | U_1 = u_1, U_2 = u_2)$?

REMARK 4.3. Fisher attacked violently the Bayes postulate as an adequate form of expressing mathematically our ignorance. He reinforced this attitude in ([5], p. 20), where we may read: "It is evidently easier for the practitioner of natural science to recognize the difference between knowing and not knowing than it seems to be for the more abstract mathematician." On the other hand, as we have seen, he admits that the absence of knowledge of proportions in a subpopulation allows us to act in the same manner as if we knew that the proportion is the same as in the whole population. The present author suspects that this new postulate, if not stronger than the Bayes postulate, is by no means weaker. Actually, if a proportion P in a population is interpreted as probability for an individual taken from this population, we tacitly assume that the individual has been selected according to the uniform distribution.

One cannot escape the feeling that all attempts to avoid expressing in a mathematical form the absence of our prior knowledge about the parameter have been eventually a failure. Of course, the mathematical formalization of ignorance should be understood broadly enough, including not only prior distributions, but also the minimax principle, and so on.

4.3. *Estimation.* Similarity in the (x, θ)-space is widely utilized in estimation. For example, an estimate $\hat{\theta}$ is unbiased if and only if the quantity $H = \hat{\theta} - \theta$ is similar with zero expectation. Further, similarity provides the following class of estimation methods: starting with a similar quantity $H = h(X, \Theta)$ with $EH = c$, and observing $X = x$, we take for θ the solution of equation

$$h(x, \theta) = c. \tag{4.10}$$

This class includes the method of maximum likelihood for

$$h(x, \theta) = (\partial/\partial\theta) \log p_\theta(x). \tag{4.11}$$

Another method is offered by the pivotal quantity $H = F(T, \Theta)$ considered in connection with fiducial probabilities. Since $EH = \frac{1}{2}$, we may take for $\hat{\theta}$, given $T = t$, the solution if any of

$$F(t, \theta) = \tfrac{1}{2}, \tag{4.12}$$

that is, the parameter value for which the observed value is the median.

5. Conditionality

If the prior distribution ν of θ is known, all statisticians agree that the decisions given $X = x$ should be made on the basis of the conditional distribution $R_\nu(d\theta|X = x)$. This exceptional agreement is caused by the fact that then there is only one distribution in the (x, θ)-space, so that the problems are transferred from the ground of statistics to the ground of pure probability theory.

The problems arise in situations where conditioning is applied to a family of distributions. Here two basically different situations must be distinguished according to whether the conditioning statistic is ancillary or not. Conditioning with respect to an ancillary statistic is considered in the conditionality principle as formulated by A. Birnbaum [1]. On the other hand, for example, conditioning with respect to a sufficient statistic for a nuisance parameter, successfully used in constructing most powerful similar tests (see Lehmann [14]), is a quite different problem and will not be considered here.

Our discussion will concentrate on the conditionality principle, which may be formulated as follows.

The principle of conditionality. Given an ancillary statistic U, and knowing $U = u$, statistical inference should be based on conditional probabilities $P_\theta(dx|U = u)$ only; that is, the probabilities $P_\theta(dx|U = u')$ for $u' \neq u$ and $P(du)$ should be disregarded.

This principle, if properly illustrated, looks very appealing. Moreover, all *proper* Bayesian procedures are concordant with it. We say that a Bayesian procedure is proper, if it is based on a prior distribution ν established independently of the experiment. A Bayesian procedure, which is not proper, is exemplified by taking for the prior distribution the measure ν given by

$$\nu(d\theta) = \sqrt{I_\theta}\, d\theta, \tag{5.1}$$

where I_θ denotes the Fisher information associated with the particular experiment. Obviously, such a Bayesian procedure is not compatible with the principle of conditionality. (Cf. [4].)

The term "statistical inference" used in the above definition is very vague. Birnbaum [1] makes use of an equally vague term "evidential meaning." The present author sees two possible interpretations within the framework of the decision theory.

Risk interpretation. A decision procedure δ defined on the original experiment should be associated with the same risk as its restriction associated with the partial experiment given $U = u$. Of course this transfer of risk is possible only after the experiment has been performed and u is known.

Decision rules interpretation. A decision function δ should be interpreted as a function of two arguments, $\delta = \delta(E, x)$, with E denoting an experiment from a class $\mathcal{E}$ of experiments and x denoting one of its outcomes. The principle of conditionality then restricts the class of "reasonable" decision functions to those for which $\delta(E, x) = \delta(F, x)$ as soon as F is a subexperiment of E, and x belongs to the set of outcomes of F (F being a subexperiment of E means that F equals "E given $U = u$" where U is some ancillary statistics and u some of its particular value).

In this section we shall analyze the former interpretation, returning to the latter in the next section.

The principle of conditionality, as it stands, is not acceptable in non-Bayesian conditions, since it denies possible gains obtainable by randomization (mixed strategies). On the other hand, some examples exhibited to support this principle (see [1], p. 280) seem to be very convincing. Before attempting to delimit the area of proper applications of the principle, let us observe that the principle is generally ambiguous. Actually, since generally no maximal ancillary statistic exists, it is not clear which ancillary statistic should be chosen for conditioning.

Let us denote the conditional distribution given $U = u$ by $P_\theta(dx|U = u)$ and the respective conditional risk by

$$R(\theta, U, u) = \int L(\theta, \delta(x))P_\theta(dx|U = u). \tag{5.2}$$

In terms of a sub-σ-field $\mathcal{B}$, the same will be denoted by

$$R(\theta, \mathcal{B}, x) = \int L(\theta, \delta(x))P_\theta(dx|\mathcal{B}, x), \tag{5.3}$$

where $R(\theta, \mathcal{B}, x)$, as a function x, is $\mathcal{B}$-measurable. Since the decision function δ will be fixed in our considerations, we have deleted it in the symbol for the risk.

DEFINITION 5.1. *A conditioning relative to an ancillary statistic U or an ancillary sub-σ-field $\mathcal{B}$ will be called* correct, *if*

$$R(\theta, U, u) = R(\theta)b(u) \tag{5.4}$$

or

$$R(\theta, \mathcal{B}, x) = R(\theta)b(x), \tag{5.5}$$

respectively. In (5.4) *and* (5.5), *$R(\theta)$ denotes the overall risk, and the function $b(x)$ is $\mathcal{B}$-measurable. Obviously, $Eb(U) = Eb(X) = 1$.*

The above definition is somewhat too strict, but it covers all examples exhibited in literature to support the conditionality principle.

THEOREM 5.1. *If the conditioning relative to an ancillary sub-σ-field $\mathcal{B}$ is correct, then*

$$E[\sup_\theta R(\theta, \mathcal{B}, X)] = \sup_\theta R(\theta). \tag{5.6}$$

PROOF. The proof follows immediately from (5.5). Obviously, for a noncorrect conditioning we may obtain $E[\sup_\theta R(\theta, \mathcal{B}, X)] > \sup_\theta R(\theta)$, so that, from the minimax point of view, conditional reasoning generally disregards possible gains obtained by mixing (randomization), and, therefore, is hardly acceptable under non-Bayesian conditions.

THEOREM 5.2. *As in section* 3, *consider the family $\{P_g\}$ generated by a probability distribution P and a group G of transformations g. Further, assume the loss function L to be invariant, namely, such that $L(hg, \delta(h(x)) = L(g, \delta(x))$ holds for all h, g, and x. Then the conditioning relative to the ancillary sub-σ-field $\mathcal{B}$ of G-invariant events is correct. That is, $R(g) = R$ and*

$$R(g, \mathcal{B}, x) = R(\mathcal{B}, x) = \int L[g, \delta(h^{-1}(x))]p(x, hg)\, d\nu(h)/\bar{p}(x). \tag{5.7}$$

PROOF. For $B \in \mathcal{B}$,

$$\int_B L(g, \delta(x))p(x, g)\, d\mu(x) = \int_{h^{-1}B} L[g, \delta(h(y))]p(h(y), g)\, d\mu h(y) \tag{5.8}$$
$$= \int_B L(h^{-1}g, \delta(y))p(y, h^{-1}g)\, d\mu(y).$$

The theorem easily follows from theorem 3.4 and from the above relations.

The following theorem describes an important family of cases of ineffective conditioning.

THEOREM 5.3. *If there exists a complete sufficient statistic* $T = t(X)$, *then* $R(\theta, U, u) = R(\theta)$ *holds for every ancillary statistic* U *and every* δ *which is a function of* T.

PROOF. The theorem follows from the well-known fact (see Lehmann [14], p. 162) that all ancillary statistics are independent of T, if T is complete and sufficient.

DEFINITION 5.2. *We shall say an ancillary sub-σ-field* $\mathcal{B}$ *yield the* deepest correct conditioning, *if for every nonnegative convex function* ψ *and every other ancillary sub-σ-field* $\overline{\mathcal{B}}$ *yielding a correct conditioning,*

$$E\psi[b(X)] \geq E\psi[\bar{b}(X)] \tag{5.9}$$

holds, with b *and* $\bar{b}$ *corresponding to* $\mathcal{B}$ *and* $\overline{\mathcal{B}}$ *by* (5.5), *respectively.*

THEOREM 5.4. *Under condition* 3.1 *and under the conditions of theorem* 5.2, *the ancillary sub-σ-field* $\mathcal{B}$ *of G-invariant events yields the deepest correct conditioning. Further, if any other ancillary sub-σ-field possesses this property, then* $R(\overline{\mathcal{B}}, x) = R(\mathcal{B}, x)$ *almost μ-everywhere.*

PROOF. Let ν denote the invariant measure on $\mathcal{G}$. Take a $C \in \overline{\mathcal{B}}$ and denote by $\chi_C(x)$ its indicator. Then

$$\phi_C(x) = \int \chi_C(gx)\, d\nu(g) \tag{5.10}$$

is $\mathcal{B}$-measurable and

$$P(C) = \int \phi_C(x)p(x)\, d\mu. \tag{5.11}$$

Further, since the loss function is invariant,

$$\int_C L(g, \delta(x))p(x, g)\, d\mu = \int \chi_C(g(y))L(g_1, \delta(y))p(y)\, d\mu(y) \tag{5.12}$$

where g_1 denotes the identity transformation. Consequently, denoting by $R(g, C)$ the conditional risk, given C, we have from (5.10) and (5.12),

$$P(C) \int R(g, C)\, d\nu(g) = \int \phi_C(x)L(g_1, \delta(x))p(x)\, d\mu. \tag{5.13}$$

Now, since $R(g) = R$, according to theorem 5.2, and since the correctness of $\overline{\mathcal{B}}$ entails $R(g, C) = R\bar{b}_C$, we have

$$\int R(g, C)\, d\nu(g) = R\bar{b}_C. \tag{5.14}$$

Note that

$$P(C)\bar{b}_C = \int_C \bar{b}(x)p(x)\, d\mu. \tag{5.15}$$

Further, since $\phi_C(x)$ is $\mathcal{B}$-measurable,

$$\begin{aligned} \int \phi_C(x)L(g_1, \delta(x))p(x)\, d\mu &= \int \phi_C(x)R(\mathcal{B}, x)p(x)\, d\mu \\ &= \int \phi_C(x)Rb(x)p(x)\, d\mu. \end{aligned} \tag{5.16}$$

By combining together (5.13) through (5.16), we obtain

$$P(C)\bar{b}_C = \int \phi_C(x)b(x)p(x)\,d\mu. \tag{5.17}$$

Now, let us assume $E\psi(\bar{b}(X)) < \infty$. Then, given a $\epsilon > 0$, we choose a finite partition $\{C_k\}$, $C_k \in \bar{\mathcal{B}}$, such that

$$E\psi(\bar{b}(X)) < \sum_k P(C_k)\psi(\bar{b}_{C_k}) + \epsilon, \tag{5.18}$$

and we note that, in view of (5.11), (5.17), and of convexity of ψ,

$$P(C_k)\psi(\bar{b}_{C_k}) \leq \int \phi_{C_k}(x)\psi(b(x))p(x)\,d\mu, \tag{5.19}$$

and, in view of (5.10),

$$\sum_k \phi_{C_k}(x) = 1 \tag{5.20}$$

for every x. Consequently, (5.18) through (5.20) entail

$$E\psi(\bar{b}(X)) < E\psi(b(X)) + \epsilon. \tag{5.21}$$

Since $\epsilon > 0$ is arbitrary, (5.9) follows. The case $E\psi(\bar{b}(X)) = \infty$ could be treated similarly.

The second assertion of the theorem follows from the course of proving the first assertion. Actually, we obtain $E\psi(\bar{b}(X)) < E\psi(b(X))$ for some ψ, unless $\bar{b}$ is a function of b a.e. Also conversely, b must be a function of $\bar{b}$, because the two conditionings are equally deep. Q.E.D.

A restricted conditionality principle. If all ancillary statistics (sub-σ-fields) yielding the deepest correct conditioning give the same conditional risk, and if there exists at least one such ancillary statistic (sub-σ-field), then the use of the conditional risk is obligatory.

6. Likelihood

Still more attractive than the principle of conditionality appears the principle of likelihood, which may be formulated in lines of A. Birnbaum [1], as follows.

The principle of likelihood. Statistical inferences should be based on the likelihood functions only, disregarding the other structure of the particular experiments. Here, again, various interpretations are possible. We suggest the following.

Interpretation. For a given θ-space, the particular decision procedures should be regarded as functionals on a space $\mathcal{L}$ of likelihood functions, where all functions differing in a positive multiplicator are regarded as equivalent. Given an experiment E such that for all outcomes x the likelihood functions $l_x(\theta)$ belong to $\mathcal{L}$, and a decision procedure δ in the above sense, we should put

$$\delta(x) = \delta[l_x(\cdot)] \tag{6.1}$$

where $l_x(\theta) = p_\theta(x)$.

If $\mathcal{L}$ contains only unimodal functions, an estimation procedure of the above kind is the maximum likelihood estimation method.

A. Birnbaum [1] proved that the principle of conditionality joined with the principle of sufficiency is equivalent to the principle of likelihood. He also conjectured that the principle of sufficiency may be left out in his theorem, and gave a hint, which was not quite clear, for proving it. But his conjecture is really true if we assume that *statistical inferences should be invariant under one-to-one transformations of the x-space to some other homeomorph space.* In the following theorem we shall give the "decision interpretation" to the principle of conditionality, and the class $\mathcal{E}$ of experiments will be regarded as the class of all experiments such that all their possible likelihood functions belong to some space $\mathcal{L}$.

THEOREM 6.1. *Under the above stipulations, the principle of conditionality is equivalent to the principle of likelihood.*

PROOF. If F is a subexperiment of E, and if x belongs to the space of possible outcomes of F, the likelihood function $l_x(\theta|E)$ and $l_x(\theta|F)$ differ by a positive multiplicator only, namely, they are identical. Thus the likelihood principle entails the conditionality principle.

Further, assume that the likelihood principle is violated for a decision procedure δ, that is there exist in $\mathcal{E}$ two experiments $E_1 = (\mathfrak{X}_1, P_{1\theta})$ and $E_2 = (\mathfrak{X}_2, P_{2\theta})$ such that for some points x_1 and x_2,

$$\delta(E_1, x_1) \neq \delta(E_2, x_2), \tag{6.2}$$

whereas

$$P_{1\theta}(X_1 = x_1) = cP_{2\theta}(X_2 = x_2) \qquad \text{for all} \quad \theta, \tag{6.3}$$

with $c = c(x_1, x_2)$, but independent of θ. We here assume the spaces $\mathfrak{X}_1$ and $\mathfrak{X}_2$ finite and disjoint, and the σ-fields to consist of all subsets. Then let us choose a number λ such that

$$0 < \lambda < \frac{1}{c+1} \tag{6.4}$$

and consider the experiment $E = (\mathfrak{X}_1 \cup \mathfrak{X}_2, P_\theta)$, where

$$\begin{aligned} P_\theta(X = x) &= \lambda P_{1\theta}(x) && \text{if} \quad x \in \mathfrak{X}_1, \\ &= \lambda c P_{2\theta}(x) && \text{if} \quad x \in \mathfrak{X}_2 - \{x_2\}, \\ &= 1 - \lambda(c + 1 - P_{1\theta}(x_1)) && \text{if} \quad x = x_2. \end{aligned} \tag{6.5}$$

Then E conditioned by $x \in \mathfrak{X}_1$ coincides with E_1, and E conditioned by $x \in (\mathfrak{X}_2 - \{x_2\}) \cup \{x_1\}$ coincides with $\tilde{E}_2$ which is equivalent to E_2 up to a one-to-one transformation $\phi(x) = x$, if $x \in \mathfrak{X}_2 - \{x_2\}$ and $\phi(x_2) = x_1$. Thus we should have $\delta(E_1, x_1) = \delta(E, x_1) = \delta(\tilde{E}_2, x_1) = \delta(E_2, x_2)$, according to the conditionality principle. In view of (6.2) this is not true, so that the conditionality principle is violated for δ, too. Q.E.D.

Given a fixed space $\mathcal{L}$, the likelihood principle says what decision procedures of wide scope (covering many experimental situations) are permissible. Having any two such procedures, and wanting to compare them, we must resort either to some particular experiment and compute the risk, or to assume some a priori distribution $\nu(d\theta)$ and to compute the conditional risk. In both cases we leave the proper ground of the likelihood principle.

However, for some special spaces $\mathcal{L}$, all conceivable experiments with likelihood functions in $\mathcal{L}$, give us the same risk, so that the risk may be regarded as independent of the particular experiment. The most important situation of this kind is treated in the following.

THEOREM 6.2. *Let θ be real and let $\mathcal{L}_\sigma$ consist of the following functions:*

$$l(\theta) = c \exp \left[-\tfrac{1}{2} \frac{(\theta - t)^2}{\sigma^2} \right], \qquad -\infty < t < \infty, \tag{6.6}$$

with σ fixed and positive. Then for each experiment with likelihood functions in $\mathcal{L}_\sigma$ there exists a complete sufficient statistic $T = t(X)$, which is normally distributed with expectation θ and variance σ^2.

PROOF. We have a measurable space $(\mathfrak{X}, t)$ with a σ-finite measure $\mu(dx)$ such that the densities with respect to μ allow the representation

$$p_\theta(x) = c(x) \exp \left[-\tfrac{1}{2} \frac{(\theta - t(x))^2}{\sigma^2} \right]. \tag{6.7}$$

This relation shows that $T = t(X)$ is sufficient, by the factorization criterion. Further,

$$1 = \int p_\theta(x)\mu(dx) = \int e^{-1/2\,\theta^2 + \theta t}\bar{\mu}(dt) \tag{6.8}$$

where $\bar{\mu} = \mu^* t^{-1}$ and $\mu^*(dx) = c(x) \exp [-\tfrac{1}{2}t^2(x)]\mu(dx)$. Now, assuming that there exist two different measures $\bar{\mu}$ satisfying (6.8), we easily derive a contradiction with the completeness of exponential families of distributions (see Lehmann [14], theorem 1, p. 132). Thus $\bar{\mu}(dt) = e^{-1/2t^2}\, dt$ and the theorem is proved.

The whole work by Fisher suggests that he associated the likelihood principle with the above family of likelihoods, and, asymptotically, with families, which in the limit shrink to $\mathcal{L}_\sigma$ for some $\sigma > 0$ (see example 8.7). If so, the present author cannot see any serious objections against the principle, and particularly, against the method of maximum likelihood. Outside of this area, however, the likelihood principle is misleading, because the information about the kind of experiment does not lose its value even if we know the likelihood.

7. Vaguely known prior distributions

There are several ways of utilizing a vague knowledge of the prior distribution $\nu(d\theta)$. Let us examine three of them.

7.1. *Diffuse prior distributions.* Assume that $\nu(d\theta) = \sigma^{-1}g(\theta/\sigma)\, d\theta$ where $g(x)$ is continuous and bounded. Then, under general conditions the posterior density $p(\theta|X = x, \sigma)$ will tend to a limit for $\sigma \to \infty$, and the limits will, independently of g, correspond to $\nu(d\theta) = d\theta$. This will be true especially if the likelihood function $l(\theta|X = x)$ are strongly unimodal, that is, if $-\log l(\theta|X = x)$ is convex in θ for every x, in which case also all moments of finite order will converge. There are, however, at least two difficulties connected with this approach.

Difficulty 1. However large a fixed σ is, for a nonnegligible portion of

θ-values placed at the tails of the prior distribution, the experiment will lead to such results x that the posterior distribution with $\nu(d\theta) = \sigma^{-1}g(\theta/\sigma)\, d\theta$ will differ significantly from that with $\nu(\theta) = d\theta$. Thus, independently of the rate of diffusion, our results will be biased in a nonnegligible portion of cases.

Difficulty 2. If interested in rapidly increasing functions of θ, say e^{θ}, and trying to estimate them by the posterior expectation, the expectation will diverge for $\sigma \to \infty$ under usual conditions (see example 8.5).

7.2. *A family of possible prior distributions.* Given a family of prior distributions $\{\nu_{\alpha}(d\theta)\}$, where α denotes some "metaparameter," we may either

(i) estimate α by $\hat{\alpha} = \hat{\alpha}(X)$, or

(ii) draw conclusions which are independent of α.

7.3. *Making use of non-Bayesian risks.* If we know the prior distribution $\nu(d\theta)$ exactly, we could make use of the Bayesian decision function δ_{ν}, associating with every outcome x the decision $d = d(x)$ that minimizes

$$R(\nu, x, d) = \int L(\tau(\theta), d)\nu(d\theta|X = x). \tag{7.1}$$

Simultaneously, the respective minimum

$$R(\nu, x) = \min_{d' \in D} R(\nu, x, d') \tag{7.2}$$

could characterize the risk, given $X = x$.

If the knowledge of ν is vague, we can still utilize δ_{ν} as a more or less good solution, the quality of which depends on our luck with the choice of ν. Obviously, in such a situation, the use of $R(\nu, x)$ would be too optimistic. Consequently, we had better replace it by an estimate $\hat{R}(\theta, \delta_{\nu}) = r(X)$ of the usual risk

$$R(\theta, \delta_{\nu}) = \int L(\theta, \delta_{\nu}(x))P_{\theta}(dx). \tag{7.3}$$

Then our notion of risk will remain realistic even if our assumptions concerning ν are not. Furthermore, comparing $R(\nu, x)$ with $\hat{R}(\theta, \delta_{\nu}) = r(x)$, we may obtain information about the appropriateness of the chosen prior distribution.

8. Examples

EXAMPLE 8.1. Let us assume that $\theta_i = 0$ or 1, $\theta = (\theta_1, \cdots, \theta_N)$ and $\tau = \theta_1 + \cdots + \theta_N$. Further, put

$$p_{\theta}(x_1, \cdots, x_N) = \prod_{i=1}^{N} [f(x_i)]^{\theta_i}[g(x_i)]^{1-\theta_i}, \tag{8.1}$$

where f and g are some one-dimensional densities. In other words, $X = (X_1, \cdots, X_N)$ is a random sample of size N, each member X_i of which comes from a distribution with density either f or g. We are interested in estimating the number of the X_i' associated with the density f. Let $T = (T_1, \cdots, T_N)$ be the order statistic, namely $T_1 \leq \cdots \leq T_N$ are the observations $X_1, \cdots, X_N$ rearranged in ascending magnitude.

PROPOSITION 8.1. *The vector T is a sufficient statistic for τ in the sense of definition 2.2, and*

$$p_\tau(t) = p_\tau(t_1, \cdots, t_N) = \tau!(N - \tau)! \sum_{s_\tau \in S_\tau} \prod_{i \in s_\tau} f(t_i) \prod_{j \notin s_\tau} g(t_j), \tag{8.2}$$

where S_τ denotes the system of all subsets s_τ of size τ from $\{1, \cdots, N\}$.

PROOF. A simple application of theorem 3.5 to the permutation group.

PROPOSITION 8.2. *For every t*

$$p_{\tau+1}(t)p_{\tau-1}(t) = p_\tau^2(t) \tag{8.3}$$

holds.

PROOF. See [7]. Relation (8.3) means that $-\log p_\tau(t)$ is convex in τ; that is, the likelihoods are (strongly) unimodal. Thus we could try to estimate τ by the method of maximum likelihood, or still better by the posterior expectation for the uniform prior distribution, if $X_i = 0$ or 1, the T is equivalent to $T' = X_1 + \cdots + X_N$.

This kind of problem occurs in compound decision making. See H. Robbins [17].

EXAMPLE 8.2. Let $0 < \sigma^2 < K$, μ real, $\theta = (\mu, \sigma^2)$, and

$$p_\theta(x_1, \cdots, x_n) = \sigma^{-n}(2\pi)^{-1/2n} \exp\left\{-\tfrac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^2\sigma^{-2}\right\}. \tag{8.4}$$

PROPOSITION 8.3. *The statistic*

$$s^2 = (n - 1)^{-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

is sufficient for σ^2.

PROOF. For given $\sigma^2 < K$ we take for the mixing distribution of μ the normal distribution with zero expectation and the variance $(K - \sigma^2)/n$. Then we obtain the mixed density.

$$q_\sigma(x_1, \cdots, x_N) = \sigma^{-n+1} (2\pi)^{-n/2} \exp\left\{-\tfrac{1}{2} \frac{n\bar{x}^2}{K} - \tfrac{1}{2} \sum_{i=1}^{n} (x_i - \bar{x})^2\sigma^{-2}\right\} \tag{8.5}$$

and apply definition 2.2.

REMARK 1. If $K \to \infty$, then the mixing distribution tends to the uniform distribution over the real line, which is not finite. For practical purposes the bound K means no restriction, since such a K always exists.

REMARK 2. A collection of examples appropriate for illustrating the notion of sufficiency for a subparameter could be found in J. Neyman and E. Scott [16].

EXAMPLE 8.4. If τ attains only two values, say τ_0 and τ_1, and if it admits a sufficient statistic in the sense of definition 2.2, then the respective distributions Q_{τ_0} and Q_{τ_1} are *least favorable* for testing the composite hypothesis $\tau = \tau_0$ against the composite alternative $\tau = \tau_1$ (see Lehmann [14]). In particular, if T is ancillary under $\tau(\theta) = \tau_0$, and if

$$p_\theta(x) = r_\tau(t(x))p_{\bar{b}}(x), \qquad \theta = (\tau, \bar{b}) \tag{8.6}$$

where the range of $\bar{b}$ is independent of the range of τ, then T is sufficient for τ in the sense of definition 2.2, and the respective family $\{Q_\tau\}$ may be defined so that we choose an arbitrary b, say b_0, and then put $Q_\tau(dx) = r_\tau(t(x))p_{b_0}(x)\mu(dx)$. In this way asymptotic sufficiency of the vector of ranks for a class of testing problems is proved in [6]. (See also remark 2.4.)

EXAMPLE 8.5. The standard theory of probability sampling from finite populations is linear and nonparametric. If we have any information about the *type* of distribution in the population, and if this type could make nonlinear estimates preferable, we may proceed as follows.

Consider the population values $Y_1, \cdots, Y_N$ as a sample from a distribution with two parameters. Particularly, assume that the random variables $X_i = h(Y_i)$, where h is a known strictly increasing function, are normal (μ, σ^2). For example, if $h(y) = \log y$, then the Y_i's are log-normally distributed. Now a simple random sample may be identified with the partial sequence $Y_1, \cdots, Y_n$, $2 < n < N$. Our task is to estimate $Y = Y_1 + \cdots + Y_N$, or, equivalently, as we know $Y_1 + \cdots + Y_n$, to estimate

$$Z = Y_{n+1} + \cdots + Y_N. \tag{8.7}$$

Now the minimum variance unbiased estimate of Z equals

$$\hat{Z} = (N - n)\,[B\,(\tfrac{1}{2}n - 1, \tfrac{1}{2}n - 1)]^{-1} \int_0^1 h^{-1}[\bar{x} + (2v - 1)s(n - 1)n^{-1/2}][v(1 - v)]^{1/2n-2}\,dy \tag{8.8}$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} h(y_i), \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n} [h(y_i) - \bar{x}]^2. \tag{8.9}$$

On the other hand, choosing the usual "diffuse" prior distribution $(d\mu\, d\sigma) = \sigma^{-1}\, d\mu\, d\sigma$, we obtain for the conditional expectation of Z, given $Y_i = y_i$, $i = 1, \cdots, n$, the following result:

$$\tilde{Z} = E(Z|Y_i = y_i, 1 \leq i \leq n,\nu) = (N - n)(n - 1)^{-1/2} [B\,(\tfrac{1}{2}\,n - \tfrac{1}{2}, \tfrac{1}{2})]^{-1} \int_{-\infty}^{\infty} h^{-1}(\bar{x} + vs(1 + 1/n)^{1/2}) \left(1 + \frac{v^2}{n-1}\right)^{-1/2n} dv. \tag{8.10}$$

However, for most important functions h, for example for $h(y) = \log y$, that is $h^{-1}(x) = e^x$, we obtain $\tilde{Z} = \infty$. Thus the Bayesian approach should be based on some other prior distribution. In any case, however, the Bayesian solution would be too sensitive with respect to the choice of $\nu(d\mu\, d\sigma)$.

Exactly the same unpleasant result (8.10) obtains by the method of fiducial prediction recommended by R. A. Fisher ([5], p. 116). For details, see [9].

EXAMPLE 8.6. Consider the following method of sampling from a finite population of size N: each unit is selected by an independent experiment, and the probability of its being included in the sample equals n/N. Then the sample size K is a binomial random variable with expectation n. To avoid empty samples, let us reject samples of size $<k_0$, where k_0 is a positive integer, so that K will have

truncated binomial distribution. Further, let us assume that we are estimating the population total $Y = y_1 + \cdots + y_N$ by the estimator

$$\check{Y} = \frac{N}{K} \sum_{i \in s_K} y_i, \tag{8.11}$$

where s_K denotes a sample of size K. Then

$$E((\check{Y} - Y)^2 | K = k) = \frac{N(N-k)}{k}\, \sigma^2 \tag{8.12}$$

holds where σ^2 is the population variance $\sigma^2 = (N-1)^{-1} \sum_{i=1}^{N} (y_i - \bar{Y})^2$. Thus K yields a correct conditioning. Further, the deepest correct conditioning is that relative to K.

On the other hand, simple random sampling of fixed size does not allow effective correct conditioning, unless we restrict somehow the set of possible $(y_1, \cdots, y_N)$-values.

EXAMPLE 8.7. Let $f(x)$ be a continuous one-dimensional density such that $-\log f(x)$ is convex. Assume that

$$I = \int_{-\infty}^{\infty} \left[\frac{f'(x)}{f(x)}\right]^2 f(x)\, dx < \infty. \tag{8.13}$$

Now for every integer N put

$$p_\theta(x_1, \cdots, x_N) = \prod_{i=1}^{N} f(x_i - \theta) \tag{8.14}$$

and

$$l_x(\theta) = p_\theta(x_1, \cdots, x_N), \qquad -\infty < \theta < \infty. \tag{8.15}$$

Let $t(x)$ be the mode (or the mid-mode) of the likelihood $l_x(\theta)$, that is,

$$l_x(t(x)) \geq l_x(\theta), \qquad -\infty < \theta < \infty. \tag{8.16}$$

Since $f(x)$ is strictly unimodal, $t(x)$ is uniquely defined. Then put

$$l_x^*(\theta) = c_N I^{1/2} (2\pi)^{-1/2} \exp\left[-\tfrac{1}{2}(\theta - t(x))^2 I\right] \tag{8.17}$$

where c_N is chosen so that

$$\int_{-\infty}^{\infty} l_x^*(\theta)\, dx_1 \cdots dx_N = 1. \tag{8.18}$$

Then $c_N \rightarrow 1$ and

$$\lim_{N \rightarrow \infty} \int |l_x(\theta) - l_x^*(\theta)|\, dx_1 \cdots dx_N = 0, \tag{8.19}$$

the integrals being independent of θ. Thus, for large N, the experiment with likelihoods l may be approximated by an experiment with normal likelihoods l^* (see [8]). Similar results may be found in Le Cam [12] and P. Huber [10].

EXAMPLE 8.7. Let $\theta = (\theta_1, \cdots, \theta_N)$,

$$p_\theta(x_1, \cdots, x_N) = (2\pi)^{-1/2N} \exp\left(-\tfrac{1}{2} \sum_{i=1}^{n} (x_i - \theta_i)^2\right), \tag{8.20}$$

and let $(\theta_1, \cdots, \theta_N)$ be regarded as a sample from a normal distribution (μ, σ^2). If μ and σ^2 were known, we would obtain the following joint density of (x, θ):

$$(8.21)\qquad r_\nu(x_1, \cdots, x_N, \theta_1, \cdots, \theta_N) = \sigma^{-N}(2\pi)^{-N}\exp\left(-\tfrac{1}{2}\sum_{i=1}^{N}(x_i-\theta_i)^2 - \tfrac{1}{2}\sigma^2\sum_{i=1}^{N}(\theta_i-\mu)^2\right).$$

Consequently, the best estimator of $(\theta_1, \cdots, \theta_N)$ would be $(\hat{\theta}_1, \cdots, \hat{\theta}_N)$ defined by

$$(8.22)\qquad \hat{\theta}_i = \frac{\mu + \sigma^2 x_i}{1+\sigma^2}.$$

Now, in the prior experiment (μ, σ^2) can be estimated by (θ, s_θ^2), where

$$(8.23)\qquad \bar{\theta} = \frac{1}{N}\sum_{i=1}^{N}\theta_i, \qquad s_\theta^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\theta_i-\bar{\theta})^2.$$

In the x-experiment, in turn, a sufficient pair of statistics for $(\bar{\theta}, s_\theta^2)$ is $(\bar{x}, s^2)$, where

$$(8.24)\qquad \bar{x} = \frac{1}{N}\sum_{i=1}^{N}x_i, \qquad s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})^2.$$

Now, while $\bar{\theta}$ may be estimated by $\bar{x}$, the estimation of s_θ^2 must be accomplished by some more complicated function of s^2. We know that the distribution of $(N-1)s^2$ is noncentral χ^2 with $(N-1)$ degrees of freedom and the parameter of noncentrality $(N-1)s_\theta^2$. Denoting the distribution function of that distribution by $F_{N-1}(x, \delta)$, where δ denotes the parameter of noncentrality, we could estimate s_θ^2 by $\hat{s}_\theta^2 = h(s^2)$, where $h(s^2)$ denotes the solution of

$$(8.25)\qquad F_{N-1}((N-1)s^2, (N-1)s_\theta^2) = \tfrac{1}{2}\cdot$$

(See section 4.3.) On substituting the estimate in (8.21), one obtains modified estimators

$$(8.26)\qquad \tilde{\theta}_i = \frac{\bar{x} + \hat{s}_\theta^2 x_i}{1+\hat{s}_\theta^2}.$$

The estimators (8.26) will be for large N nearly as good as the estimators (8.26) (cf. C. Stein [18]).

The estimators (8.26) could be successfully applied to estimating the averages in individual stratas in sample surveys. They represent a compromise between estimates based on observations from the same stratum only, which are unbiased but have large variance, and the overall estimates which are biased but have small variance.

The same method could be used for $\theta_i = 1$ or 0, and $(\theta_1, \cdots, \theta_N)$ regarded as a sample from an alternative distribution with unknown p. The parameter p could then be estimated in lines of example 8.1 (cf. H. Robbins [17]).

9. Concluding remarks

The genius of classical statistics is based on skillful manipulations with the notions of sufficiency, similarity and conditionality. The importance of similarity increased after introducing the notion of completeness. An adequate imbedding of similarity and conditionality into the framework of general theory of decision making is not straightforward. Classical statistics provided richness of various methods and did not care too much about criteria. The decision theory added a great deal of criteria, but stimulated very few new methods. From the point of view of decision making, risk is important only before we make a decision. After the decision has been irreversibly done, the risk is irrelevant, and for instance, its estimation or speculating about the conditional risk makes no sense. Such an attitude seems to be strange to the spirit of classical statistics. If regarding statistical problems as games against Nature, one must keep in mind that besides randomizations introduced by the statistician, there are randomizations (ancillary statistics) involved intrinsically in the structure of experiment. Such randomizations may make the transfer to conditional risks facultative.

REFERENCES

[1] ALLAN BIRNBAUM, "On the foundations of statistical inference," *J. Amer. Statist. Assoc.*, Vol. 57 (1962), pp. 269–326.

[2] D. BLACKWELL and M. A. GIRSCHICK, *Theory of Games and Statistical Decisions*, New York, Wiley, 1954.

[3] E. B. DYNKIN, *The Foundations of the Theory of Markovian Processes*, Moscow, Fizmatgiz, 1959. (In Russian.)

[4] BRUNO DE FINETTI and LEONARD J. SAVAGE, "Sul modo di scegliere le probabilita iniziali," *Sui Fondamenti della Statistica*, Biblioteca del Metron, Series C, Vol. I (1962), pp. 81–147.

[5] R. A. FISHER, *Statistical Methods and Scientific Inference*, London, Oliver and Boyd, 1956.

[6] J. HÁJEK and Z. SIDAK, *The Theory of Rank Tests*, Publishing House of the Czech Academy of Sciences, to appear.

[7] S. M. SAMUELS, "On the number of successes in independent trials," *Ann. Math. Statist.*, Vol. 36 (1965), pp. 1272–1278.

[8] J. HÁJEK, "Asymptotic normality of maximum likelihood estimates," to be published.

[9] ———, "Parametric theory of simple random sampling from finite populations," to be published.

[10] P. J. HUBER, "Robust estimation of a location parameter," *Ann. Math. Statist.*, Vol. 35 (1964), pp. 73–101.

[11] A. N. KOLMOGOROV, "Sur l'estimation statistique des parameters de la loi de Gauss," *Izv. Akad. Nauk SSSR Ser. Mat.*, Vol. 6 (1942), pp. 3–32.

[12] L. LE CAM, "Les propriétés asymptotiques des solutions de Bayes," *Publ. Inst. Statist. Univ. Paris*, Vol. 7 (1958), pp. 3–4.

[13] ———, "Sufficiency and approximate sufficiency," *Ann. Math. Statist.*, Vol. 35 (1964), pp. 1419–1455.

[14] E. L. LEHMANN, *Testing Statistical Hypotheses*, New York, Wiley, 1959.

[15] J. NEYMAN, "Two breakthroughs in the theory of decision making," *Rev. Inst. Internat. Statist.*, Vol. 30 (1962), pp. 11–27.

[16] J. NEYMAN and E. L. SCOTT, "Consistent estimates based on partially consistent observations," *Econometrica*, Vol. 16 (1948), pp. 1–32.

[17] H. ROBBINS, "Asymptotically subminimax solutions of compound decision problems," *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, 1950, pp. 131–148.

[18] CHARLES STEIN, "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution," *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, 1956, Vol. I, pp. 197–206.