KYBERNETIKA — VOLUME 12 (1976), NUMBER 5

Statistics and Computability

TOMÁŠ HAVRÁNEK

In the present paper conditions under which statistical procedures with definite probabilistic meta-properties (such as unbiasedness, uniformly most powerfulness, etc.) can be identified with realizing computational procedures are investigated.

"Perhaps neither mathematicians specialized in probability theory or statistics nor experts in electronic data processing will look on computational statistics as a serious scientific subject". This is the first sentence of COMPSTAT 74 – Proceedings in Computational Statistics [3]. On the other hand, perhaps nobody can deny the practical meaning of the connection between computer applications and statistics. The above mentioned proceedings can serve as an excellent example of the practical power of computer techniques in statistics. Another field of such interactions between statistics and computers is the mechanized Hypothesis Formation or similar methods of Artificial Intelligence (see [9], [7] or a systematic approach in [10]).

But already in the first book that has the words "Computational Statistics" in its title (Freiberger and Grenader [5]), the possibility of theoretical problems of this topic is mentioned. The aim of the present paper is to give some suggestions concerning their solution.

Some of the results of this paper were published in a very sketchy form in [11]. The present paper is highly influenced by the authors cooperation with P. Hájek in writing [10]. I wish to express my gratitude to P. Hájek for many discussions, suggestions and friendly help.

The theory of computable functions (= recursive functions, cf. [18]) is well known in the mathematical foundations of computer science. The class of computable functions covers the class of functions computable in real time on a real computer. More deep investigations of such functions leads to the theory of concrete computational complexity (see, e.g. [19]). Clearly, before one considers concrete computational complexity of some functions it is necessary to know whether they are computable in principle.

Statistical theory constructs statistics (statistical decision functions) without any respect to their computational properties. In usual statistics this fact is not very dangerous from practical point of view, because the practical statistician, guided by his intuition, uses only decision functions that are computable in the informal sense. Fortunatelly, various statistics optimal w.r.t. various statistical criteria are intuitively computable. Nevertheless, from the logical point of view, the question of computability of statistics (in a formal sense) deserves serious investigation.

Note the following fact: all data obtained as a result of experiments, measurements, etc. are *rational* numbers. All our computational procedures which realize our statistical estimates, tests, etc. work with these *rational* numbers. But statistical procedures are constructed mainly with respect to all real numbers.

The theory of computability has been developed on functions whose arguments and values are natural numbers. It is rather trivial to change it as to apply to functions whose arguments and values are rational numbers; but there is no obvious way how to define computability for real-valued functions. Thus, speaking on computable procedures, we shall mean procedures dealing with rational numbers. Probabilistic properties of such procedures are not specified. But using appropriate restrictions, we shall guarantee that each computable procedure determines uniquely a statistics (on real numbers) whose probabilistic properties are well defined.

This situation is typical in the mechanized hypothesis formation.

We can illustrate our problems by two examples. It is clear that these examples are rather artificial but from the logical point of view they cannot be ignored.

(1) We investigate samples of independent and identically distributed random variables having normal distribution. As an estimate of expectation we use the arithmetical mean which has some optimal statistical properties. Suppose that our samples are of the cardinality m. Then the mean f maps $\mathbf{R}^m$ into $\mathbf{R}$, where $\mathbf{R}$ is the set of real numbers. What we really deal with, is a result of measurement, i.e. values in $\mathbf{Q}^m$, where $\mathbf{Q}$ is the set of rational numbers. Consider a statistic $f' : \mathbf{R}^m \to \mathbf{R}$ such that $f' \restriction \mathbf{Q}^m \to c$, where c is a rational constant and $f' \restriction (\mathbf{R}^m - \mathbf{Q}^m) = f \restriction \restriction (\mathbf{R}^m - \mathbf{Q}^m)$. The statistic f' has the same optimal statistical properties as f and for all results of measurement it gives the value c.

(2) We consider in the same situation an optimal test statistic t for testing one-sided hypothesis about expectation. If we put $t_1 \restriction \mathbf{Q}^m = 0$ and $t_1 \restriction (\mathbf{R}^m - \mathbf{Q}^m) = = t \restriction (\mathbf{R}^m - \mathbf{Q}^m)$, t_1 is an optimal (e.g. uniformly most powerful) test for this situation, but it never rejects the hypothesis on the base of experimental data.

One of the aims of the present paper is to find such features of statistical procedures which help to avoid such paradoxa. It means to find a good relation between probabilistic properties of statistics and computational procedures operating on rational numbers and realizing these statistics. The second aim is to establish clear logical foundations to mechanized formation of hypotheses based on statistical data (this point of view is developed in [10]). The third motivation is the following: the investigation of the relation between computability and probabilistic properties is the necessary first step towards the construction of statistical procedures optimal with respect to computational complexity — and this could be a very practical task.

Some other problems arise if we choose an exact frame for our considerations. We shall use the usual Kolmogorov probability; we use probability in theoretical and meta-level, where we need no computations.

If now $\underline{\Sigma} = \langle \Sigma, \mathscr{R}, P \rangle$ is a probability space (Σ — the set of states, $\mathscr{R}$ — a σ-field of subsets of Σ, and P — a probability measure) and if $V \subseteq \mathbf{R}$, then we define a *regular $\underline{\Sigma}$-random V-structure* of the type $\langle 1^n \rangle$ as an $(n + 1)$-tuple $\boldsymbol{U} = \langle U, q_1, \dots, q_n \rangle$ in which each q_i is a mapping of $U \times \Sigma$ into V such that, for each $u \in U$, $q_i(u, -)$ is a random variate, and each sequence of n-dimensional random variates $\{\langle q_1(u_k, -), \dots, q_n(u_k, -)\rangle\}_{k \in \mathbf{N}}$ (where $u_k \in U$) is stochastically independent. ($\mathbf{N}$ means the set of natural numbers, $\mathbf{N}^+ = \mathbf{N} - \{0\}$.) Let the type n and the probability space $\boldsymbol{\Sigma}$ be arbitrary but fixed in the sequel. Now we present some denotations: Denote by $\mathscr{M}^V$ the set of all structures of the form $\langle M, f_1, \dots, f_n \rangle$ where M is a finite set and $f_1, \dots, f_n$ are functions from M into V; $\mathscr{M}^V_M$ is the set of such structures with fixed domain M. The meaning of $\mathscr{M}^{V \cap \mathbf{Q}}$ and $\mathscr{M}^{V \cap \mathbf{Q}}_M$ is clear. Now if $\boldsymbol{U}$ is a $\boldsymbol{\Sigma}$-random V-structure, if $M \subseteq U$ is a finite sample and if $\sigma \in \Sigma$ is a random state, we have a uniquely determined sample structure $\boldsymbol{M}_\sigma \in \mathscr{M}^V_M$, corresponding to the experimental data. (Pedantically we should write $M^{\boldsymbol{U}}_\sigma$.) For each M we define a metric on $\mathscr{M}^V_M$, e.g. as

$$\varrho(\boldsymbol{M}_1, \boldsymbol{M}_2) = \max_{u \in M, i = 1, \dots, n} |f_i(u) - g_i(u)|$$

if $\boldsymbol{M}_1 = \langle M, f_1, \dots, f_n \rangle$, $\boldsymbol{M}_2 = \langle M, g_1, \dots, g_n \rangle$. In the further, we shall suppose on $\mathscr{M}^V_M$ the topology given by ϱ. If Φ is a sentence speaking about random structures we shall write $\boldsymbol{U} \gg \Phi$ if Φ is true in $\boldsymbol{U}$. (We use this unusual denotation for typografical reasons.)

One kind of statistical inference we shall consider, namely hypothesis testing, has then the following form:

We have two theoretical sentences (sentences speaking about random structures, cf. [8; 11] and/or [10]) Φ and Ψ; we have accepted Φ (and we call Φ the frame assumption) and we ask whether to accept Ψ or not. To decide this question we first fix a set $V_0 \subseteq V$ and a function f associating with each structure $\boldsymbol{M}_\sigma$ a value $f(\boldsymbol{M}_\sigma) \in V$. Then we make observations (get a particular structure $\boldsymbol{M}_\sigma$) and compute $f(\boldsymbol{M}_\sigma)$. If $f(\boldsymbol{M}_\sigma) \in V_0$ we accept Ψ. This procedure is justified, at least, by choosing f and V_0 such that the following holds: if $\boldsymbol{U} \gg \Phi$ and $\boldsymbol{U} \not\gg \Psi$ then the probability $P(\{\sigma; f(\boldsymbol{M}_\sigma) \in V_0\})$ is small. Thus f and V_0 are chosen on the base of their probabilistic properties.

What do we really observe and decide? We assume that our data are in $\mathscr{M}^{V \cap \mathbf{Q}}$. If $V \subseteq \mathbf{Q}$ no problems arise but if $V \cap (\mathbf{R} - \mathbf{Q}) \neq \emptyset$ we face the following problem: For an arbitrary V_0, we must answer the following questions: Is the probability

$P(\{\sigma; f(\boldsymbol{M}_\sigma) \in V_0\})$ well defined? How is our reasoning affected by the fact that our observation is approximate (we restrict ourselves to rational structures)? Can we really compute $f(\boldsymbol{M}_\sigma)$?

1. CONTINUOUS AND COMPUTABLE STATISTICS

1.1. First we make some preliminary requirements on sets of values. A set $X \subseteq \mathbf{R}$ is a *regular set* of values if (a) all boundary points of X are raiional and (b) $X \cap \mathbf{Q}$ is a recursive set of rationals. Examples of regular sets: **N**, **R**, intervals of arbitrary kind with rational end points, finite unions of such intervals, etc. Intervals with irrational end points, **Q** and Cantor's discontinuum can serve as examples of non-regular sets. It is clear that in practice we need to decide whether $f(\boldsymbol{M}_\sigma)$ is an element of a regular set or not.

1.2. Lemma. (a) Regular sets form a field of sets. (b) If X is regular then $X \cap \mathbf{Q}$ is dense in X. (c) If X_1, X_2 are regular sets, then $X_1 \neq X_2$ implies $X_1 \cap \mathbf{Q} \neq X_2 \cap \mathbf{Q}$. (d) Each regular set is Borel.

Proof is easy: (a) Denote the system of all regular subsets of **R** by $\mathscr{A}$. Then $\mathbf{R} \in \mathscr{A}$. If $X \in \mathscr{A}$ then its complement X^c has all boundary point rational and $\mathbf{Q} \cap X^c = \mathbf{Q} - (\mathbf{Q} \cap X)$ is recursive. Similarly for finite union of sets. (b) Note that each irrational point of a regular set X is an interior point of X. (c) follows easily from (b). (d) Each regular set X can be decomposed as follows: $X = (X - \mathbf{Q}) \cup (X \cap \mathbf{Q})$; $X - \mathbf{Q}$ is open, $X \cap \mathbf{Q}$ is clearly Borel.

1.3. In the following we restrict ourselves to regular sets.

A mapping $f : \mathscr{M}^V$ is a *cc-statistic* (*continuous and computable statistic*) on $\mathscr{M}^V$ if the following conditions hold:

(a) f is invariant under isomorphism of models.
(b) For each M finite, the function $f \restriction \mathscr{M}^V_M$ is continuous.
(c) The function $f \restriction \mathscr{M}^{V \cap \mathbf{Q}}$ is recursive (with the range included in **Q**).

1.4. Consider now regular random V-structures. If $\boldsymbol{U}$ is such a structure, and $M \subseteq U$ a finite sample, then the function $f^{\boldsymbol{U}}_M$ defined by the equality $f^{\boldsymbol{U}}_M(\sigma) = f(\boldsymbol{M}_\sigma)$ is a random variate. By $D^{\boldsymbol{U}}_{f_M}$ we shall denote its distribution function. In the following we shall write f_M instead of $f^{\boldsymbol{U}}_M$.

Note that if $V_0 \subseteq V$ is regular then, for each sample $M \subseteq U$, the probability $P(\{\sigma; f(\boldsymbol{M}_\sigma) \in V_0\})$ is well defined.

The assumption (a) in 1.3 guarantees that the value depends only on the structure but not on the particular samples. Assumption (b) guarantees, besides other properties, that small changes of values in a model $\boldsymbol{M}$ cause only small shift of $f(\boldsymbol{M})$. If we accept usual equivalence between recursivity and computability (cf. Rogers [18]) we can say that the assumption (c) in 1.3 gives us an answer to the computability question: whenever we have a rational-valued structure $\boldsymbol{M}$ we can calculate $f(\boldsymbol{M})$ and since V_0 is regular, we can decide whether $f(\boldsymbol{M}) \in V_0$. (Note that all following considerations remain valid if we substitute the notion of recursivity by another stronger notion of computability.)

The main paradoxon is avoided by the following theorem.

1.5. Theorem. Let V be a regular set and f, g two cc-statistics on $\mathcal{M}^V$. Then $f \upharpoonright \mathcal{M}^{V \cap \mathbf{Q}} = g \upharpoonright \mathcal{M}^{V \cap \mathbf{Q}}$ implies $f = g$.

Proof. $V \cap \mathbf{Q}$ is dense in V, hence $\mathcal{M}_M^{V \cap \mathbf{Q}}$ is dense in $\mathcal{M}_M^V$ for each finite M. □

1.6. Note that cc-statistics are statistics in the usual sense. Hence, for example, if we find an optimal (e.g. uniformly most powerfull) test in the class of all usual non-randomized tests and prove that this test is based on a cc-statistic and a regular set (i.e. it is an observational test, see [10]), then it is optimal in the class of all observational tests.

On the base of previous considerations we can identify some statistical tests (or statistics in general) having definite probabilistic meta-properties (as optimality) with particular computational procedures. (For further information, particularly in connection with observational functor calculi in which computational procedures are represented for purposes of Hypothesis formation, see [10]).

1.7. Convention. We require in all cases the invariancy of statistics under isomorphism of sample structures. Hence we shall consider further, for each cardinality m of samples, only one representant, namely $\{1, \ldots, m\}$. The sense of $\mathcal{M}_m^V, f_m$ etc. is clear.

1.8. How strong is the condition of recursivity? Consider now regular random $\{0, 1\}$-structures. $\boldsymbol{U} = \langle U, q_1, q_2 \rangle$. Each sample M and state σ gives and 2×2 table $\begin{array}{c|c} a & b \\ \hline c & d \end{array}$ $(\boldsymbol{M}_\sigma)$. We use, without any scruples, a statistic of the form

$$f(\boldsymbol{M}_\sigma) = \frac{\ln(ad/bc)}{\sqrt{\left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\right)}} \text{ (see [1])} .$$

It is clearly continuous on its domain but not computable (since the values need not be rational). What is the mathematical property that enables us to use such statistics without any problems?

1.9. Definition. Consider a regular set V, and a statistic f on $\mathcal{M}^V$. A three argument recursive function $b : \mathbf{N}^+ \times \mathcal{M}^{V \cap \mathbf{Q}} \times \mathbf{N}^+ \to V \cap \mathbf{Q}$ is called a *computable numerical* *approximant* of f if, for each $m \in \mathbf{N}^+$, the following conditions hold:

$$b(m, -, -) : \mathscr{M}_m^{V \cap \mathbf{Q}} \times \mathbf{N} \to V \cap \mathbf{Q}$$

and

$$(\forall n \in \mathbf{N}^+)\,(\forall M \in \mathscr{M}_m^{V \cap \mathbf{Q}})\,|b(m, M, n) - f(M)| < 1/2n\,.$$

(Cf. the notion of numerical approximant in [13]; note that our notion of approximation differs substantially from the notions used in the usual recursion theory, see [2].)

1.10. It is clear that, in many cases ($V \neq V \cap \mathbf{Q}$), it has no sense to speak about probabilistic properties of b. On the other hand, if f is, for example, an estimator (continuous statistic) then if we use b instead of f we cannot make great errors.

1.11. Notation. Let V be a regular set. Denote by $B(V)$ its boundary, and denote R_V the relation on $\mathbf{Q}^2$ defined as follows: $R_V(r_1, r_2)$ iff $[r_1, r_2] \subseteq V$ ($[r_1, r_2]$ means the closed interval with end points r_1, r_2).

1.12. Remark. (1) Note that $[r_1, r_2] \subseteq V$ iff $[r_1, r_2] \cap \mathbf{Q} \subseteq V$. (2) The relation R_V is recursive iff it is recursively enumerable. (Note that $\neg R_V$ is recursively enumerable in all cases: $\neg R_V(r_1, r_2)$ iff $(\exists r \in \mathbf{Q})\,(r_1 \leqq r \leqq r_2,\ r \in V^c)$.)

In the following we use freely notations of mathematical logic, particularly $\neg$ for negation and & for conjunction.

1.13. Definition. A regular set V is called *intervals recognisable* set (*ir-set*) if R_V is recursive.

1.14. Theorem. Consider regular sets V and $V_0 \subseteq V$ and a statistic f on $\mathscr{M}^V$. Let (1) f have a numerical approximant b, (2) $f(\mathscr{M}^{V \cap \mathbf{Q}}) \cap B(V_0) = \emptyset$ and (3) the sets V_0 and $V - V_0$ be ir-sets. Then $f \in V_0$ is decidable (i.e. the set $\{M \in \mathscr{M}^{V \cap \mathbf{Q}};\ f(M) \in V_0\}$ is recursive).

Proof. Consider an $M \in \mathscr{M}^{V \cap \mathbf{Q}}$. Two cases can occur: (i) $f(M) \in V_0 - B(V_0)$ or (ii) $f(M) \in V_0^c - B(V_0)$. In the case (i) there is an $n \in \mathbf{N}^+$ such that

$$R_{V_0}(b(m, M, n) - 1/2n\,,\quad b(m, M, n) + 1/2n)\,;$$

in the case (ii) there is an $n \in \mathbf{N}^+$ such that

$$R_{V_0^c}(b(m, M, n) - 1/2n\,,\quad b(m, M, n) + 1/2n)\,.$$

Hence both the sets $\{M \in \mathscr{M}^{V \cap \mathbf{Q}};\ f(M) \in V_0\}$ and $\{M \in \mathscr{M}^{V \cap \mathbf{Q}};\ f(M) \in V_0^c\}$ are recursively enumerable. □

1.15. The above theorem can be applied in many cases. For example, use the statistics $\ln(ad/bc)$ and put $V_0 = [r, +\infty)$ where r is a rational number. Note that for s rational $\ln(s)$ is transcendental.

In many cases, in which the condition $f(\mathscr{M}^{V \cap \mathbf{Q}}) \cap B(V_0) = \emptyset$ is not satisfied, we have a numerical approximant such that, if $f(\boldsymbol{M}) \in \mathbf{Q}$, it finds n_0 that $b(m, \boldsymbol{M}, n_0) = f(\boldsymbol{M})$ and indicates this fact. (Think of functions of the form $x^{1/n}$, $n \in \mathrm{N}^+$.) Then we obtain the decidability of $f \in V_0$ again.

In practice, we stop the work of each computable numerical approximant at a given $n_{\max}$. Hence it can happen that $f(\boldsymbol{M}) \in V_0$ but $b(m, \boldsymbol{M}, n_{\max}) \notin V_0$, but only for point close to the boundary.

1.16. What are the ir-sets? The following lemma shows some particular cases.

Lemma. The following cases of regular sets are ir-sets:
(1) Finite union of intervals of arbitrary kind.
(2) Complement of countable union of intervals with given minimal length.
(3) Such V that the sets $B(V)$, $\{p \in B(V); (\exists p' \in B(V)) (p' > p)\}$ and $\{p_1, p_2 \in B(V); (p_1, < p_2) \& \neg(\exists p \in B(V)) (p_1 < p < p_2)$ are recursive.

1.17. Lemma. (1) There is a ir-set V_1 such that $B(V_1)$ is not recursivelly enumerable.
(2) There is a regular set V_2 with $B(V_2)$ recursive which is not ir-set.

Proof. Let the variables x, y, z, e vary over natural numbers. We use the usual enconding of Turing machines and their input words by natural numbers. Let now $T(e, x, y)$ be the relation saying: the Turing machine e halts the computation on input word x exactly after y steps. This relation is recursive. Hence $\neg T(e, x, y)$ is also recursive.

(1) (Hájek) Define $T(e, z) = (\exists x, y)\, T(e, x, y) \,\&\, z = \langle x, y\rangle)$. Then $\{e; (\exists^\infty z) . T(e, z)\}$ is a universal Π_2^0 set and hence not recursively enumerable (cf. [18]; $\exists^\infty$ means – there is infinitely many ...). Define now $I_{ez} = (e + 1/(2z + 1), e + 1/(2z)]$ and $V_1 = \bigcup_{\{e,z;T(e,z)\}} I_{ez}$. Then V_1 is an ir-set. But $B(V) = \{e + 1/(2z + 1), e + 1/2z; T(e, z)\} \cup \{e; (\exists^\infty z)\, T(e, z)\}$.

(2) The relation $\neg T(e, e, z)$ is recursive, but $(\forall z) (\neg T(e, e, z))$ is not recursively enumerable ([12], § 46). Define $I_{ez} = (e + 1/(z + 1), e + 1/z]$ and $I_e^* = \{e\} \cup \bigcup_{\{z;\neg T(e,e,z)\}} I_{ez}$. Finally, put $V_2 = \bigcup_e I_e^*$. Note that

$$R_{V_2}(e, e + 1) \quad \text{iff} \quad (\forall z) (\neg T(e, e, z)) .$$

Hence if $R_{V_2}(r_1, r_2)$ is recursive then $R_{V_2}(e, e + 1/2)$ is and we obtain a contradiction. On the other hand

$$B(V) = \bigcup_e \{e\} \cup \bigcup_e \{e + 1/z;\ T(e, e, z) \vee T(e, e, z - 1)\}$$

and hence $B(V)$ is recursive. □

1.18. Discussion. Now we touch very briefly the question What functions have a compuiable numerical approximant? We shall now write $f \in$ CNA for "f has$_g$a numerical approximant" (in the natural sense: if f maps $\mathbf{R}^k$ into $\mathbf{R}$ we look for a recursive b: $\mathbf{Q}^k \times \mathbf{N}^+ \rightarrow \mathbf{Q}$ such that $|b(r, n) - f(r)| < 1/2n$ for r rational).

(1) Trivially, if f is recursive (on $\mathbf{Q}$), then $f \in$ CNA. Particularly all arithmetical operations on $\mathbf{R}$ $+, -, . ,$) are functions from CNA. Rational constants are functions from CNA.

(2) The following elementary functions are from CNA: ln, exp, sin, cos, tg, arcsin, arctg (see [16], [4]).

(3) If $t(x_1, \ldots, x_{k'}) \in$ CNA, t is continuous and $g(x_1, \ldots, x_k), \ldots, g_{k'}(x_1, \ldots, x_k) \in$ $\in$ CNA, then the function h defined by the equality

$$h(x_1, \ldots, x_k) = t(g_1(x_1, \ldots, x_k), \ldots, g_{k'}(x_1, \ldots, x_k))$$

is in CNA.

If f is now a statistic on $\mathscr{M}^V$ and if each f_m can be obtained by (1)–(3) then $f \in$ CNA in the sense of 1.9.

1.19. In the present section (namely 1.1 – 1.9) we required statistics to be continuous. This had a good reason – the representation of a statistic by its restriction to rational valued models and then by a computational procedure. As we shall see later, there are reasonable and substantially noncontinuous statistics. Hence we have one way of generalization – to cover these cases and satisfy desired properties of statistics. This is done in the following section.

Another direction that has to be investigated is the problem of concrete coputation complexity and its influence on the choice of appropriate statistical procedures. The importance of this question is obvious in connection with multiple usage of statistical procedures on computers, particularly in connection of AI-methods (cf. [10]).

Its solution requires a good hierarchy of numerical procedures (e.g. based on the number of operations in floating point arithmetic, [17]). Then if we have two statistical procedures $f_0, f_1, f_0 \lll \ldots < f_1$ in computational hierarchy and f_1 slightly better than f_0 in some statistical sense, we choose surely f_0. Similarly if f_0, f_1 are results of two statistical approaches. Nevertheless, such good hierarchy has not yet been constructed.

2. ALMOST CONTINUOUS COMPUTABLE STATISTICS

2.1. As we shall see later, the condition of continuity of statistics is too restrictive. Before discussing this topic we must be more specific as to the form of theoretical sentences in question.

We shall suppose, for the sake of simplicity, that our random structures are *d-homogeneous* (*distributionally homogeneous*). This means the following: Consider a regular random V-structure $\boldsymbol{U} = \langle U, q_1, \ldots, q_n \rangle$. Then the distribution function $D_{\boldsymbol{U},o}$ of the n-dimensional random variate $\langle q_1(o, -), \ldots, q_n(o, -) \rangle$ is independent on o. Then the probabilistic properties of $\boldsymbol{U}$ can be described by the distribution function $D_{\boldsymbol{U}}$ ($D_{\boldsymbol{U}} = D_{\boldsymbol{U},o}$ for any $o \in U$). Our theoretical sentences have now in many cases the following property: $\boldsymbol{U} \gg \Phi$ and $D_{\boldsymbol{U}} = D_{\boldsymbol{U}'}$ implies $\boldsymbol{U}' \gg \Phi$ (for any $\boldsymbol{U}, \boldsymbol{U}'$). Such sentences are called *distributional*.

2.2. Consider a random $\mathbf{R}$-structure. Let Φ be true in $\boldsymbol{U} = \langle U, q_1, q_2 \rangle$ iff (a) q_1 is independent on σ with probability 1, (b) $\boldsymbol{U}$ is d-homogeneous and (c) the distribution function $D_{\boldsymbol{U}_2}$ of $\boldsymbol{U}_2 = \langle U, q_2 \rangle$ is absolutely continuous. Define a mapping f as follows:

$f(\boldsymbol{M}_\sigma) = 1$ iff ($q_1(o_1, \sigma) \geqq 1$ and $q_1(o_2, \sigma) < 1$ implies $q_2(o_1, \sigma) \geqq q_2(o_2, \sigma)$ for each $o_1, o_2 \in M$). f is a statistic; it is the simplest example of a rank statistic. We see immediately that f is computable but not continuous.

2.3. We shall now speak about *computable statistics*, i.e. functions satisfying (a) and (c) from 1.3 and

(b1): for each M finite, $f \upharpoonright \mathscr{M}^V_M$ is Borel measurable.

What is the probabilistic property desired from the point of wiev of hypotheses testing that our computable statistics have to satisfy? It can be formulated as follows:

Let V be a regular set and Φ a distributional sentence. Consider a class $\mathscr{F}$ of computable statistics such that, $f \upharpoonright \mathscr{M}^{V \cap \mathbf{Q}} = g \upharpoonright \mathscr{M}^{V \cap \mathbf{Q}}$ for each $f, g \in \mathscr{M}$. We call $\mathscr{F}$ *d-invariant* (distributional invariant) if the following holds: If $\boldsymbol{U} \gg \Phi$, then, for each $f, g \in \mathscr{F}$ and M finite, $M \subseteq U$

$$(*) \qquad P(\{\sigma; f(\boldsymbol{M}_\sigma) \in V_0\}) = P(\{\sigma; g(\boldsymbol{M}_\sigma) \in V_0\})$$

$$\text{for each } V_0 \subseteq V,\ V_0 \text{ regular}\,.$$

2.4. It is clear that, under the frame assumption Φ, the class of d-invariant (computable) statistics is appropriate, e.g., as the class of optimal tests statistics for some problem.

If $\boldsymbol{U}$ is a regular random structure and f a statistic on $\mathscr{M}^V$, then $D^{\boldsymbol{U}}_{f_M}$ denotes the distribution function of f_M (for each finite sample $M \subseteq U$).

2.5. Lemma. The condition $(*)$ is equivalent to:

$$(\forall M)\,(\forall x \in Q \cap V)\,(D^{\boldsymbol{U}}_{f_M}(x) = D^{\boldsymbol{U}}_{g_M}(x))\,.$$

Proof. ($\Leftarrow$): Use the regularity of V and left-side continuity of the distribution function to obtain the equality of distribution functions on V. Then use the fact that regular sets are Borel.

($\Rightarrow$): Let f be a statistic on $\mathscr{M}^V$. Use sets $V_k = [-k, x)$, where $x \in \mathbf{Q}$, $k \in \mathbf{N}$. Then, under Φ, $P(f_M \in [-k, x)) = P(f_M \in [-k, x) \cap V) = P(g_M \in [-k, x) \cap V) = P(g_M \in [-k, x))$. Note that $D^{\boldsymbol{U}}_{f_M}(x) = \lim_{k \to \infty} P(f_M \in [-k, x))$. □

2.6. Remark. (1) Naturally the condition ($*$) is equivalent to the following: for each $V_0 \subseteq V$, V_0 Borel, $P(f_M \in V_0) = P(g_M \in V_0)$.

(2) Let Φ be a distributional statement such that $\boldsymbol{U} \gg \Phi$ implies: for each finite sample M, $P(\{\sigma; \boldsymbol{M}_\sigma \notin \mathscr{M}^{V \cap \mathbf{Q}}_M\}) = 0$. Then each system $\mathscr{F}$ is d-invariant w.r.t. Φ.

(3) If statistics from $\mathscr{F}$ are continuous then $\mathscr{F}$ is d-invariant w.r.t. any distributional sentence.

Now we can define a new kind of statistics that covers also rank statistics.

2.7. Definition. Let Φ be a distributional sentence and let V be a regular set of values. Let f be a computable statistic. f is called an *almost continuous computable statistic* (*acc-statistic*) w.r.t. Φ if f satisfies the following condition:

(acc): For each regular random V-structure $\boldsymbol{U}$ such that $\boldsymbol{U} \gg \Phi$ and for each finite sample $M \subseteq U$, $f \upharpoonright \mathscr{M}^V_M$ is continuous on an open set $\mathscr{M}_{cont} \subseteq \mathscr{M}^V_M$ such that $P(\{\sigma; \boldsymbol{M}_\sigma \in \mathscr{M}_{cont}\}) = 1$.

2.8. Remember that for each open set $A \subseteq \mathbf{R}$ the set $A \cap \mathbf{Q}$ is dense in A. Thus if f is an acc-statistic, then for each $\boldsymbol{M} \in \mathscr{M}_{cont}$, the value $f(\boldsymbol{M})$ can be approximated by values of f on rational models from $\mathscr{M}_{cont}$.

Note that the statistic f from 2.2 is an acc-statistic.

2.9. Theorem. Let Φ be a distributional sentence and V a regular set. Consider a class $\mathscr{F}$ of statistics on $\mathscr{M}^V$ such that (i) all statistics from $\mathscr{F}$ are acc-statistics w.r.t. Φ and (ii) for each $f, g \in \mathscr{F}$ we have $f \upharpoonright \mathscr{M}^{V \cap \mathbf{Q}} = g \upharpoonright \mathscr{M}^{V \cap \mathbf{Q}}$. Then $\mathscr{F}$ is d-invariant w.r.t. each distributional sentence implying Φ.

Proof. Consider statistics f, g from $\mathscr{F}$. Let a sample M be fixed. $f \upharpoonright \mathscr{M}^V_M$ is continuous on $\mathscr{M}^f_{cont}$, similarly for $g \upharpoonright \mathscr{M}^V_M$ and $\mathscr{M}^g_{cont}$. $P(\{\sigma; \boldsymbol{M}_\sigma \in \mathscr{M}^f_{cont}\}) = P(\{\sigma; \boldsymbol{M}_\sigma \in \mathscr{M}^g_{cont}\}) = 1$, hence for $\mathscr{M}_{cont} = \mathscr{M}^f_{cont} \cap \mathscr{M}^g_{cont}$ we have $P(\{\sigma; \boldsymbol{M}_\sigma \in \mathscr{M}_{cont}) = 1$. $\mathscr{M}_{cont} \cap \mathscr{M}^{V \cap \mathbf{Q}}$ is dense in $\mathscr{M}_{cont}$. □

2.10. Remark. (1) We have just proved that under Φ we have $P(\{\sigma; f(\boldsymbol{M}_\sigma) = g(\boldsymbol{M}_\sigma)\}) = 1$ for each $f, g \in \mathscr{F}$. Hence statistics from $\mathscr{F}$ are determined with probability 1 by their values on rational models. (2) If $f \in \mathscr{F}$ and $\boldsymbol{M} \in \mathscr{M}_{cont}$ then $f(\boldsymbol{M})$ can be approximated by values on rational models from $\mathscr{M}_{cont}$.

2.11. One could ask whether the assumption of Borel measurability (cf. 2.3) in the definition of acc-statistics is necessary.

Let U be a $\langle \Sigma, \mathscr{R}, P\rangle$-random V-structure. Then, for each $M \subseteq U$, we have the induced probability measure μ_P on Borel sets of $\langle \mathscr{M}^V_M, \varrho\rangle$. Remember the notion of μ_P-measurable sets. The following theorem shows that the condition of the Borel measurability is, in fact, superflous.

2.12. Theorem. If f is an arbitrary function satisfying conditions (a) and (c) from 1.3 and the acc-condition w.r.t. a distributional sentence, then, for each sample $M \subseteq U$, we have: for each X Borel, $f_M^{-1}(X)$ is μ_P-measurable.

Proof. If $\mathscr{M}_0 \subseteq \mathscr{M}^V_M$ is an open set such that $\mu_P(\mathscr{M}_0) = 1$, then we have the following:

For each $B \subseteq \mathscr{M}^V_M$, B is μ_P-measurable iff $B \cap \mathscr{M}_0$ is μ_P-measurable. (Note that, if μ_P^* is the outer measure generated by μ_P, then $\mu_P^*(B) = \mu_P^*(\mathscr{M}_0 \cap B)$.) Further, f is continuous on $\mathscr{M}_{cont} \subseteq \mathscr{M}^V_M$, $\mu_P(\mathscr{M}_{cont}) = 1$. Hence, for each X Borel, $f^{-1}(X) \cap \mathscr{M}_{cont} = f^{-1}(X \cap f(\mathscr{M}_{cont}))$ is Borel, and, consequently, μ_P-measurable. □

2.13. Consider the converse question: how much can a statistic be discontinuous? We obtain the following particular answer:

2.14. Theorem. Let Φ be a distributional sentence such that $U \gg \Phi$ implies that μ_P is absolutely continuous w.r.t. the Lebesgue measure λ. Then each computable statistic is an acc-statistic w.r.t. Φ.

Proof. Let k be a natural number and $\mathbf{R}^k$ the metric space of k-tuples of reals with the metric ϱ. Then the Luzin theorem (cf. [14]) holds:

A real function f on $\mathbf{R}^k$ is Lebesgue measurable iff for each $\varepsilon > 0$ there is an $A \subseteq \mathbf{R}^k$ such that $\lambda(A) < \varepsilon$ and $f \in \mathbf{R}^k - A$ is continuous.

Hence, if μ_P is absolutely continuous w.r.t. λ then the Borel measurability implies that for each $n \geqq 1$ there is an $A_n \subseteq \mathscr{M}^V_M$ such that $\mu_P(A_n) \geqq 1 - 1/n$ and $f \upharpoonright A_n$ is continuous. Define $\mathscr{M}_{cont} = \bigcup_{n=1}^{\infty} A_n$. Note that $\mu_P(A_1 \cup \ldots \cup A_n) \geqq 1 - 1/n$, hence $\mu_P(\mathscr{M}_{cont}) = 1$. □

3. APPROXIMATING STATISTICS

3.1. One can easily construct an acc-statistic f such that for each discontinuity point M there is no sequence $\{M_n\}$ of rational points such that $f(M_n) \to f(M)$. If we look more carefully to practically used rank statistics (which are the most important example of acc-statistics) we see that they have a property similar to the left-side continuity and/or the right-side continuity for univariate functions.

Hence we define: a computable statistic f is *approximating* if for each discontinuity point M there is an open subset $O(M) \subseteq \mathscr{M}^V_M$ such that M is in the closure of $O(M)$ and $f \upharpoonright (O(M) \cup \{M\})$ is continuous. If moreover, under $U \gg \Phi$, (p): $P(\{\sigma; M_\sigma \in O(M)\}) > 0$, then we say that f is well approximating w.r.t. Φ.

3.2. Remark. Let Φ be a distributional sentence. (1) If $\mathscr{F}$ is a class of d-invariant (w.r.t. Φ) and approximating statistics, then if one of them is well approximating, then all of them are well approximating. (2) Neither each approximating acc-statistic (w.r.t. Φ) is well approximating nor each well approximating statistic is an acc-statistic.

An example to the second case in (2): Suppose M be fixed. Let $\{M_n\}$ be a countable set in $\mathscr{M}_M^{\mathbf{R}}$ such that $\varrho(M_0, M_n) = n$. Put $R_n = \{M; \varrho(M_0, M) < n\}$. Let $\{p_i\}_{i \in \mathbf{N}}$ be a countable set of real numbers such that $\sum_{i=1}^{\infty} p_i = 1$. Let Φ be such a sentence that, under Φ,

$$P(M_0) = p_0, P(M_1) = p_2, \ldots, P(M_n) = 2n, \ldots$$

and on each of

$$R_1 - \{M_0\}, R_2 - R_1 - \{M_1\}, \ldots, R_n - R_{n-1} - \{M_{n-1}\}, \ldots$$

we have a uniform distribution such that

$$P(\{\sigma; M_\sigma \in R_n - R_{n-1} - \{M_{n-1}\}\}) = p_{2n-1} .$$

Consider a statistic f such that it is discontinuous on hyperspheres $\bar{R}_n = \{M; \varrho(M_0, M)\} = n$. Then f can be well approximating but not an acc-statistic.

3.3. Theorem. Let Φ be a theoretical sentence. If f is an approximating statistic and if $U = \Phi$ implies that for each $o \in U$, $D_{U,o}$ has positive density w.r.t. λ on V^n, then f is a well approximating and acc-statistic w.r.t. Φ.

Proof. First, we prove that the set of discontinuity points of an approximating statistic has the Lebesgue measure zero. Let M be fixed and let $\mathscr{M}_d \subseteq \mathscr{M}_M^V$ be the set of discontinuity points. Suppose $\lambda(\mathscr{M}_d) > 0$. Then $\mathscr{M}_d$ contains a closed cartesian product of intervals J. Each interior point of J is a discontinuity point not satisfying condition of the definition of approximating statistic. The rest of the proof is a matter of routine. □

3.4. Remark. If $f(\mathscr{M}_M^V)$ is finite then the condition (p) is equivalent to the following:

for each discontinuity point M, we have:

(i) $(\exists \delta > 0)$ $P(\{\sigma; M_\sigma \in \mathscr{M}^{V \cap \mathbf{Q}}, \varrho(M, M_\sigma) < \delta, f(M) \mp f(M_\sigma)\}) = 0$ and (ii) $P(\{\sigma; M_\sigma \in O(M)\})$, where $M_\sigma \in O(M)$ implies $f(M) = f(M_\sigma)$. These conditions show the proper power of an approximation to discontinuity points based on (i) only. If $D_{U,o}$ is continuous under Φ, then (i) holds for each statistic.

(Received December 18, 1975)

REFERENCES

[1] J. Anděl: On interactions in contingency tables. Aplikace matematiky *18* (1973), 99—109.

[2] G. Aussiello, M. Protasi: On the comparison of notions of approximation. Mathematical Foundations of Computer Science 75 (ed. J. Bečvář), Lecture Notes in Computer Science 32, Springer Heidelberg 1975, 172—178.

[3] COMPSTAT 74 — Proceedings in Computational Statistics. Ed. G. Bruckmann, F. Ferschl, L. Schmetterer, Physica Verlag, Wien 1974.

[4] B. P. Demidovich, I. A. Maron: Computational mathematics. Mir, Moskva 1973.

[5] W. Freiberger, U. Grenader: A short course in computational probability and statistics. Applied mathematical science 6, Springer, New York 1971.

[6] P. Hájek: Problém obecného pojetí metody GUHA. Kybernetika *4* (1968), 505—515.

[7] P. Hájek: Automatic listing of important observational statements I—III. Kybernetika *9* (1973) and *10* (1974), 187—205, 251—271, 95—124.

[8] P. Hájek: On logic of discovery. Mathematical Foundations of Computer Science 75 (ed. J. Bečvář), Lecture Notes in Computer Science 32, Springer, Heidelberg 1975, 30—45.

[9] P. Hájek, I. Havel, M. Chytil: GUHA — a method of automatic determination of hypotheses I, II. Kybernetika *2* (1966) and *3* (1967), 31—47, 430—437.

[10] P. Hájek, T. Havránek: Mechanized Hypothesis Formation. Book in preparation.

[11] T. Havránek: The approximation problem in computational statistics. Mathematical Foundations of Computer Science (ed. J. Bečvář), Lecture Notes in Computer Science 32, Springer, Heidelberg 1975, 258—265.

[12] S. C. Kleene: Mathematical logic. J. Wiley, New York 1967.

[13] W. Miller: Toward abstract numerical analysis. Journal of ACM *20* (1973), 399—408.

[14] J. G. Oxtoby: Measure and category. Springer, Heidelberg 1971.

[15] A. J. van Reeken: Report of the Dutch working party on statistical computing. Applied Statistics (JRSS-C) *20* (1971), 73—79.

[16] A. Ralston, H. S. Wilf, ed.: Mathematical methods for digital computers. J. Wiley, New York, vol. I 1960, vol. II 1967.

[17] A. Ralston: First course in numerical analysis. McGraw-Hill, New York 1965.

[18] M. Rogers: Theory of recursive functions and effective computability. McGraw-Hill, New York 1967.

[19] V. Strassen: Evaluation of rational functions. In Complexity of Computations (R. E. Miller, J. W. Thatcher, eds.), Plenum Press, New York. 1972, 1—10.

RNDr. Tomáš Havránek, Matematické středisko biologických ústavů ČSAV (Centre of Biomathematics — Czechoslovak Academy of Sciences), Budějovická 1083, 142 20 Praha 4. Czechoslovakia.