Some remarks on the foundations of statistics

Frank Hampel

ETH Zürich

Preface

This short research report contains the 2-page contributed paper for the 48th I.S.I. Session in Cairo, 9-17 September, 1991 (to appear in Vol.4 of the Proceedings), transcriptions of the posters displayed at the Session, and as an appendix (for ease of reference) the 2-page paper for the 47th I.S.I. Session in Paris, 1989, and the 1-page abstract for the 2nd Bernoulli Society World Meeting in Uppsala, 1990.

F. Hampel

Some remarks on the foundations of statistics

Frank Hampel

Seminar für Statistik, ETH Zürich, CH-8092 Zürich, Switzerland

Key words: philosophical foundations of statistics, parametric models, prediction, intersubjective inference, partial knowledge, upper and lower probabilities, Bayesian paradigm, state of total ignorance, successful bets, surprise, optimal successful bets, "possibility," "fiducial probability," fair bets.

One of the simplest situations to be considered in the philosophical foundations of statistics is the following: Let a parametric model $(\Omega, \mathcal{A}, \{P_\theta, \theta \in \Theta\})$ for possible observations be given, such as the binomial model. There may or may not be some prior knowledge on θ. What can we infer on the event $\{Y \in A\}$, namely that a future observation Y lies in an arbitrary predetermined set A, given that a presently observed random variable X equals some value x?

Two variants of this prediction problem are: inference on the parameter θ instead of some future Y, and inference on a set A which may depend on the random variable X, as in usual prediction and confidence intervals.

The aim considered here is to arrive at an intersubjective description of the state of our knowledge about the occurrence of A, given $X = x$. "Intersubjective" means that it is the same for all scientists with the same factual background knowledge. It is to be distinguished both from the true, unknown state of Nature and from subjective (personal) opinions about this state. The description of the state of partial knowledge (where several facts may be equally and highly plausible) is also to be distinguished from decision making (where only one course of action is allowed).

It has frequently been assumed that the description of the state of partial knowledge has to be done in terms of a (conditional) probability distribution on the sample space of Y. But this is not only unnecessary and unjustifiable (unless we can start with an apriori probability distribution), it seems also implausible because it would mean perfect (instead of imperfect) knowledge of the stochastic mechanism generating Y.

Instead, we can work with some kind of "upper probability" $m(Y \in A | X = x,\ M_0)$, where M_0 describes the prior knowledge about θ. If M_0 is a Bayesian prior probability distribution, we should recover the Bayesian solution with our m. But what properties can we reasonably demand of m, if M_0 is the state of total ignorance about θ (denoted by $M_0 \equiv 1$)?

If we want to get out of the state of total ignorance, we cannot be always "sufficiently cautious" in some sense, but only in some stochastic average. Since A is considered fixed (or perhaps a fixed function of X), one has to average over the distribution of X.

One reasonable looking requirement is that of "successful betting:"
$E_\theta\ m(Y \in A|X,\ M_0 \equiv 1) \geq P_\theta(Y \in A),\quad \forall\theta$. The interpretation is that a bet $(1-m) : m$ on A^c (with $m = m(Y \in A|X = x,\ M_0 \equiv 1)$) has a nonnegative expected gain and thus is successful "in the long run." (Nothing is said about bets on A, which depend on $m(A^c)$! We only get, by a wellknown argument, $m(A) + m(A^c) \geq 1$.)

A stronger requirement would be that of "smaller average surprise:"
$E_\theta\ (\log\ m(Y \in A|X,\ M_0 \equiv 1)) \geq \log\ P_\theta(Y \in A),\quad \forall\theta$, where the intersubjective and the appropriate surprise are $-\log\ m$ and $-\log\ P_\theta$, respectively. The idea is, not to be unnecessarily surprised (because surprise implies assumed knowledge). But let us return to successful bets.

Among all functions m (depending on x, but not on θ) which lead to successful bets, we may select the one which minimizes $\sup_\theta(E_\theta m(\ldots) - P_\theta(A))$. (In cases of nonuniqueness, additional criteria might be used.)

In some simple binomial examples, the smallest sup is actually even zero.

It is conjectured that the "possibility" M defined in Hampel (1989) is a good first approximation to the optimal m if X contains enough information about θ, and that a correction applied to the "possibility" using a Taylor expansion gives an even better approximation.

If we want to make inference about θ, if X (and θ) are continuous and one-dimensional, and if $A = A(x)$ is a suitably chosen function of x, we arrive at the "fiducial probability" of $A(x)$, given x. This concept can in our framework obviously be interpreted by means of a successful bet with the special property that the "opposite" bet, namely $m : (1 - m)$ on A, is also (barely) successful and hence the pair of bets derived from m happens to be fair in the Bayesian sense.

BIBLIOGRAPHY

Hampel, F. (1989): Zu den Grundlagen der Statistik. Research Report No.58, Seminar für Statistik, ETH, Zürich. English translation (On the foundations of statistics): contributed paper, Bulletin of the ISI, 47th Session, Paris 1989, Book 1, 423-424.

RESUME

Quelques arguments et aspects des fondements philosophiques de la statistique sont expliqués, conduisant à la notion de "pari favorable". Si l'on commence à "l'état d'ignorance totale," ces paris n'entrent pas dans le paradigme bayesien, sauf dans le cas des "probabilités fiducielles."

Some remarks on the

foundations of statistics

Frank Hampel

ETH Zürich

Work in progress.

Attempt to combine:

Frequentist approach
Bayesian approach
Likelihood approach
Fiducial approach
Information theoretic approach

Puzzle. Not yet complete.

Main aim:

Description of *intersubjective* knowledge
(not of subjective beliefs, nor of optimal decisions) based on observed random variables
(in known stochastic models with unknown parameters).

Basic situation considered:

Given a parametric model $(\Omega, \mathcal{A}, \{P_\theta, \theta \in \Theta\})$

with random variables $X, Y, \ldots$;

given the *"state of total ignorance"* about θ

(here denoted as "$M_0 \equiv 1$");

given the observation $X = x$;

what can we *infer* about the possibility that $Y \in A$ for some predetermined set A ?

Remarks:

- We seek a numerical representation

$$m(Y \in A \mid X = x,\ M_0 \equiv 1), \quad 0 \leq m \leq 1.$$

Definition, interpretation, and properties of m ?

- We can always be "unlucky" with x. A is fixed, hence m can have good properties only averaged over the full distribution of X.

- $m(Y \in A \mid \ldots)$ and $m(Y \in A^c \mid \ldots)$ are basically unrelated. A standard Bayesian argument implies often

$$m(Y \in A \mid \ldots) + m\ (Y \in A^c \mid \ldots) \geq 1,\ \text{but } \textbf{not} = 1.$$

Variants of basic situation:

- Some prior knowledge M_0 about θ. If $M_0 \equiv \pi$ for some Bayesian prior π, m should coincide with the Bayesian solution.

- $A = A(X)$ $\quad\leadsto$ prediction intervals

- $A = A(X)\ c\ \Theta$ (no Y) $\quad\leadsto$ confidence intervals, "fiducial probabilities."

Some specific likelihood-based proposals

Let π always denote a Bayesian prior on Θ (thought discretized), let M_0 ($1 \geq M_0 \geq \pi$ for some π) denote an arbitrary prior knowledge on Θ, let $\hat{\pi}$ and π^* be implicitly defined as arg $\sup_{\pi \leq M_0} \cdots$

Define

$$m_1(Y \in A | X = x, M_0) \;=\; \frac{\sum_\theta \hat{\pi}(\theta) P_\theta(X = x) P_\theta(Y \in A | X = x)}{\sum_\theta \hat{\pi}(\theta) P_\theta(X = x) \;=\; \sup_{\pi \leq M_0} \sum_\theta \pi(\theta) P_\theta(X = x)}$$

$$m_2(Y \in A | X = x, M_0) \;=\; \frac{\sup_{\pi \leq M_0} \sum_\theta \pi(\theta) P_\theta(X = x) P_\theta(Y \in A | X = x)}{\sup_{\pi \leq M_0} \sum_\theta \pi(\theta) P_\theta(X = x)}$$

$$m_3(Y \in A | X = x, M_0)$$

$$= \frac{\sum_\theta \pi^*(\theta) P_\theta(X = x) P_\theta(Y \in A | X = x) = \sup_{\pi \leq M_0} \sum_\theta \pi(\theta) P_\theta(X = x) P_\theta(Y \in A | X = x)}{\sum_\theta \pi^*(\theta) P_\theta(X = x)}$$

We have $0 \leq m_1 \leq m_2 \leq m_3 \leq 1$,

for $M_\theta \equiv 1$ is $m_1 = P_{\hat{\theta}_{\mathrm{ML}}}(Y \in A | X = x)$ with $\hat{\theta}_{\mathrm{ML}}$ = maximum likelihood estimator based on X,

m_2 seems more suitable for parameters and fiducial probabilities,

m_3 seems more suitable for prediction and equals the “Möglichkeit” M (“possibility”) introduced in Hampel (1989).

The m_i lead out of the state of total ignorance towards perfect knowledge of the stochastic mechanism,

and if a Bayesian prior is known, they coincide with the Bayesian solution.

Interpretation of m by means of successful bets

Bayesian theory: probabilities $\longleftrightarrow$ "fair" bets

$$P(A) = p \iff \text{ willing to bet } \begin{cases} p : (1-p) & \text{on} \quad A \\ (1-p) : p & \text{on} \quad A^c \end{cases}$$

Definition: A bet is called *"successful"* (in the long run; or "favorable") if its expected gain is ≥ 0.

Let $m(A) \leq m_0 \iff$ willing to bet (at least) $(1-m_0) : m_0$ on A^c.

Aim:
Define m such that it leads to successful bets for all $\pi \leq M_0$, and for all θ if $M_0 \equiv 1$.

Try to minimize among "successful" m's; obtain only "admissible class."

Canonical choice??

Tentative ("minimax") choice: write $m = m_{A,M_0}(X) = m(X)$
and select $m = \text{ arg min }_{m:E_\pi m \geq P_\pi(Y \in A), \forall \pi \leq M_0} \sup_{\pi \leq M_0}(E_\pi m(X) - P_\pi(Y \in A))$

Notes:

- Bayesian bets are often not successful (unless there exists a true prior and the Bayesian prior coincides with it).
- There exist (probably rather generally) nontrivial successful m's converging to the true $P(Y \in A)$ as X contains more and more information.
- In several examples with X and Y independent binomials, m_3 is successful. In others, it is not quite so, but may be used as a starting point in the search for successful m's.

Interpretation of m by means of cautious mean surprise

Information theory suggests: probabilities $\longleftrightarrow$ "surprises."

$P(A) = p \iff -\log p =$ surprise when A occurs.

Surprise means (assumed) knowledge about what is how possible.

Let $m(A) = m \iff$ my surprise about A is $-\log m$.

Definition: m is called *"cautious"* if its mean or average surprise is at most the true (average) surprise.

Theorem: m is cautious $\Longrightarrow$ m is successful.

Proof:

$$E_\pi(-\log m_{A,M_0}(X)) \leq E_\pi(-\log P_\theta(Y \in A)) \Longrightarrow$$

$$E_\pi m_{A,M_0}(X) \geq E_\pi P_\theta(Y \in A),\ \forall \pi \leq M_0 \text{ (Jensen's inequality).}$$

Again: Seek small cautious m; get "admissible class"; canonical choice??

Notes:

- The "minimax" choice does not always work properly since the sup may be $\equiv \infty$.
- Minimal cautious m's tend to be strictly larger than minimal successful m's, but again there usually seem to be cautious m's which are asymptotically fully informative.
- Since both m_3 and cautious m's tend to shun away from zero, m_3 may again be used as a starting point in the search for cautious m's.

Interpretation of fiducial probabilities

If in the basic situation A is allowed to depend on X (as with prediction intervals), there is sometimes a successful pair of m's with $m(Y \in A(X)|X = x, M_0 \equiv 1) + m(Y \in (A(X))^c|X = x, M_0 \equiv 1) \equiv 1$; that is, the pair of successful bets is even fair in the Bayesian sense.

The same may hold if $A\ c\ \Theta$ and Y is replaced by θ (confidence intervals which keep their level exactly).

But this interpretation by means of successful **and** fair bets appears to be the essence of the *"fiducial argument."*

Notes:

- The fiducial argument refers not to (unknown) Nature, but to our knowledge about Nature – as desired!
- Its bets – one at a time – can be imbedded in known sequences of bets with known long-range properties (which is not true for Neyman-Pearson tests).
- The fiducial argument basically refers only to a 2-point probability space. Any extension – or even unique extension – would need additional arguments.
- A clearer way of writing might be, e.g., $P(\text{“}\theta > x\text{” is true}) = \frac{1}{2}$. In the realized random statement, x is known and θ unknown; the probability of its truth derives from the full distribution of X.
- It is possible to approximate fiducial "probabilities" arbitrarily by successful but not quite fair pairs of bets in discretized models.

Some numerical examples

Let $X \sim B(n,\theta)$ and $Y \sim B(k,\theta)$ be independently binomially distributed.

Successful bets:

For $k=1$, the "Möglichkeit" $M(=m_3)$ leads to successful bets, but not necessarily for $k>1$.

Counterexamples: $n=1,\ k=2,\ A=\{1\}$ or $A=\{0,1\}$.

Consider $n=2, k=2, A=\{1\}$ in detail.

(Task: Find function $m(x),\ 0 \le m \le 1, \quad \ni$

$$P_\theta(A) = 2\theta(1-\theta) \overset{!}{\le} (1-\theta)^2\ m(0) + 2\theta(1-\theta)m(1) + \theta^2 m(2),\ \forall\theta.)$$

Some choices for m :

$x=0$	1	2	Remarks
(3/8	1/2	3/8)	$M=m_3$; not successful
0	1	0	successful; $E_\theta m(X) \equiv P_\theta(A)$
1	0	1	successful
7/16	9/16	7/16	successful; $m_3 + \frac{1}{16}$
3/8	5/8	3/8	successful; only $m(1)$ raised
3/7	4/7	3/7	successful; $m_3 \cdot 8/7$

Cautious surprises:

For $k=1,\ m_3$ (and hence also Laplace's "rule of succession") is (only) asymptotically cautious.

Consider $A=\{1\}$. Cautious are: For $n=k=1:\ m(0)=e^{-1},\ m(1)=1$

For $\quad n=2,\ k=1: m(0)=e^{-3/2},\ m(1)=e^{-1/2},\ m(2)=1$

For $\quad n=3,\ k=1: m(0)=e^{-11/6},\ m(1)=e^{-5/6},\ m(2)=e^{-2/6},\ m(3)=1$

For $\quad n=k=2:$

$m(0)=m(2)=$	1/4	1/2	1	7/16	3/8	3/7	$3/8+1/15$
$m(1)=$	1	1/2	1/4	4/7	2/3	7/12	$4/8+1/15$

Summary and outlook

The definitions and methods described lead to objectively correct statements about a stochastic system converging to the true probability distribution – even starting from the state of total ignorance about the parameter. They also provide a natural framework for so-called fiducial probabilities.

Major open problems within this framework are the choice between successful bets and cautious surprises, the selection among "admissible" solutions, the finding of simple, nearly admissible solutions based on, e.g., the m_i ($i = 1, 2, 3$), and an investigation and comparison of the asymptotic properties of the m_i (and of some other procedures).

So far, we have assumed the parametric model to be known and to be exact. What about the questions of model choice and of robustness?

Appendix

(I.S.I. Session, Paris 1989)

On the foundations of statistics

English translation of: **Zu den Grundlagen der Statistik**

by **Frank Hampel**

Key words: Foundations of statistics, inductive logic, state of total ignorance, principle of minimum total surprise, "possibility"; Fisher's likelihood, Bayesian statistics, fair bets, maximum entropy, rule of succession.

Let a parametric model $(\Omega, \mathcal{A}, \{P_\theta, \theta \varepsilon \Theta\})$ be given (discrete, for simplicity, including Θ). There may also be given an apriori distribution $\pi(\cdot)$ on Θ and/or the observed value x of a (vector-valued) random variable X on the model. What can we infer about the values of another, still unobserved random variable Y on the model?

Many authors refer in their proposals, in rather different ways, to the famous paper by Bayes (1763), and hence it seems worthwhile to have another look at it. At first, it contains clearly a frequentist foundation with a physical uniform distribution (Postulate 1), but later (Scholium) we find only a vague conclusion by analogy about the state of total ignorance. In retrospect, the arbitrariness of this argument is however somewhat mitigated by the fact that the resulting discrete distribution in the sample space is the distribution with maximum entropy.

A critical hidden assumption in the paper by Bayes appears to be the belief that empirical partial knowledge about probabilities can again be described by probabilities. Similarly, in the foundation arguments using fair bets by Neobayesians we find the hidden assumption of perfect (at least subjective) knowledge of the overall stochastic mechanism (including the random selection of θ), which precludes any discussion or investigation of the state of total ignorance about θ and of a learning process leading out of it. Yet it is easily possible, with less complete knowledge, to offer more cautious bets (e.g., $1:1$ for A and $1:2$ for A^c) without losing in the long run.

In the following I'd like to submit another approach to discussion.

We introduce the concept of "possibility" ("Möglichkeit") $M_0(\cdot)$ as a function on Θ, $0 \leq M_0 \leq 1$, which describes the state of our apriori knowledge about θ. $\ddot{U}(\theta) := -\log M_0(\theta)$ is our "surprise" ("Überraschung") when we learn that θ is the true parameter (in generalization of $\ddot{U} := -\log P$ for a probability P). Now $M_0 \equiv 1$ is the state of total ignorance; $M_0 \equiv \pi$ for an apriori distribution π is the model of the Bayesians. A consistency condition is $M_0 \geq \pi$ for some π, otherwise M_0 is arbitrary.

Using a "principle of the least total surprise" ("Prinzip der minimalen Gesamtüberraschung"), we define

$$\pi^* := \pi^*(A|x, M_0) := \text{ arg } \sup_{\pi \leq M_0} \sum_{\theta} \pi(\theta) P_\theta(X = x) \; P_\theta(Y \varepsilon A | X = x).$$

From there we obtain the "possibility"

$$\begin{aligned} & M(Y \varepsilon A | X = x, M_0) : \\ & = \textstyle\sum_\theta \pi^*(\theta) P_\theta(X = x) \; P_\theta(Y \varepsilon A | X = x) \left(\sum_\theta \pi^*(\theta) \; P_\theta(X = x)\right)^{-1} \end{aligned}$$

(or $M = \sup_{\{\pi^*\}} \sum \ldots$, if π^* is not unique). If $M_0 \equiv \pi$ for some π, we obtain a probability distribution and the solution of the Bayesians. If $M_0 \equiv 1$, we can always choose π^* concentrated on some θ^*. If in addition Y is independent of X, we obtain

$$M(Y \varepsilon A | X = x, M_0) = P_{\theta^*}(Y \varepsilon A), \text{ with } \theta^* = \theta^*(A|x, M_0).$$

An example is given by two independent binomial distributions $X \sim B(n, p)$ and $Y \sim B(m, p)$. For $M_0 \equiv 1$, $X = x$, and $Y = y$ the resulting $p^* = (x+y)/(n+m)$, and Fisher's (1959, p.130 ff.) likelihood of the corresponding 2×2 table is proportional to exp(minus the least total surprise). If $n >> m$, then $M(Y \varepsilon A | X = x, M_0 \equiv 1)$ is in first approximation the probability distribution of Y under the maximum likelihood estimate x/n as parameter. If $m = 1$ and $X = n$, then $M(Y = 1 | X = n, M_0 \equiv 1) = 1$ and $M(Y = 0 | X = n, M_0 \equiv 1) = 1/(n + 1)$, but there is now no reason to renorm these values to $(n+1)/(n+2)$ and $1/(n+2)$ as in the well-known rule of succession.

References:

Bayes, T. (1763): An essay towards solving a problem in the doctrine of chances. Phil. Trans. Roy. Soc., Vol. 53, p. 370-418. Reprinted in Biometrika 45 (1958), p.293-315.

Fisher, R.A. (1959): Statistical Methods and Scientific Inference. 2nd edition. Oliver and Boyd, Edinburgh and London.

Résumé: Quelques aspects des fondements de la statistique sont discutés. Une structure formelle est proposée qui permet de sortir de l'état de l'ignorance totale et qui contient aussi la solution bayesienne.

Abstract: Several aspects of the foundations of statistics are discussed. A formalism is proposed which allows to get out of the state of total ignorance and which contains also the Bayesian solution.

(Bernoulli meeting, Uppsala 1990)

FAIR BETS, SUCCESSFUL BETS, AND THE FOUNDATIONS OF STATISTICS

Frank Hampel

ETH Zürich, Switzerland

One of several equivalent paradigms of Bayesian statistics is centered around the concept of a "fair bet," which entails that if one offers a bet $p : q$ on some event A, one must also be willing to offer a bet $q : p$ on the complement $\bar{A}$.

The "fairness" is rather limited. For example, there are situations with two Bayesians such that whoever offers a "fair bet" is expected to lose.

We now consider the following definition. Given a parametric model and an observation $X = x$ from an (unknown) model distribution, we define a bet (depending on x) on some event A (without restriction on the bet on $\bar{A}$) to be "successful," if the expected gain, averaged over the distribution of X, is nonnegative for every model distribution. The class of "successful bets" on A is nonempty and usually very rich, so that we could define "admissibility" etc. as in Wald's statistical decision theory. Yet it is possible to pick out a fairly natural choice.

The "possibility" defined in Hampel (1989) in general leads not quite to "successful bets," but they can be made "successful" by a slight modification, such as by multiplying the "possibilities" with a correction factor and, when necessary, by relaxing the maximum likelihood condition in the definition of the "possibility."

If the event A is allowed to depend on x, as with prediction and confidence intervals, we obtain a natural framework for Fisher's "fiducial probabilities," in which their meaning, interpretation and basic restriction become much clearer and in which also the pertaining gap between discrete and continuous probability models can be bridged.

Reference:

F. Hampel (1989): Zu den Grundlagen der Statistik. Res.Rep.58, Seminar für Statistik, ETH Zürich. English translation: On the foundations of statistics. Proc. 47th Session ISI, Paris 1989, Contrib. Papers, Book 1, 423-424.