The Annals of Statistics
1984, Vol. 12, No. 4, 1551–1557

TURNING PROBABILITIES INTO EXPECTATIONS

BY MICHAEL GOLDSTEIN

University of Hull

Suppose that you specify your prior probability that an unknown quantity θ lies in each member of a disjoint partition of the values of θ. What does this imply about your prior mean and variance for θ, and your posterior mean and variance, given sample information? We provide a partial answer by modifying a suggestion of Manski for incorporating the cost of specification of prior probabilities into the analysis of decision problems. This modification leads to a simple explicit solution in the problem of estimating the mean of a distribution, with quadratic loss, in the class of linear functions of the sample, and this solution is related to the problem of turning probabilities into expectations.

1. Introduction. Suppose that θ is an unknown quantity and that you specify your prior beliefs concerning θ by stating your prior probability P_i that θ lies in I_i, for $i = 1, \cdots n$, where the I_j form a disjoint partition of the set of possible values of θ. You are specifying these quantities as a starting point for updating your beliefs when you observe a quantity S for which $E(S \mid \theta) = \theta$. What do the values P_i imply about your prior and posterior expectation and variance for θ? In particular, is it worth the additional effort required to specify further probabilities over a refinement of the existing partition of θ values?

Our starting point will be a general consideration of the problem of subjective choice of probability domains. In Section 2, we consider a suggestion of Manski (1981) for incorporating the cost of specification of prior probabilities into the analysis of decision problems. We point out various problems with this approach and suggest possible modifications. In Section 3, we apply the suggested modifications to estimate a single random quantity based on a sample estimate. This yields a simple solution which provides a partial answer to the questions raised above concerning how you should turn probabilities into expectations. Finally, in Section 4, we consider the problem of whether to specify further prior probabilities.

2. Subjective choice of probability domains. Manski (1981) considers the following problem. Suppose θ is an unknown quantity. You have specified a disjoint partition $R_1, \cdots, R_k$, for θ and the probabilities $P_i = \Pr(\theta \in R_i)$. There is a specified utility function $U(d, \theta)$ corresponding to each decision d. There is a cost involved in further prior specifications. Should you stop now and choose a decision (and if so, which decision)? Or should you continue to make further prior specifications (and if so which)? Manski suggests that for each rule d, and each set R_i, you should evaluate the quantities $\bar{U}(d, R_i) = \sup_{\theta \in R_i} U(d, \theta)$ and

Received August 1983; revised February 1984.
AMS 1980 *subject classifications.* Primary 62F15; secondary 62G05.
Key words and phrases. Midrisk, linear Bayes rules, elicitation of subjective probabilities.

$\underline{U}(d, R_i) = \inf_{\theta \in R_i} U(d, \theta)$. Then defining $\underline{V}(d)$, $\bar{V}(d)$ by

$$\underline{V}(d) = \sum_i \underline{U}(d, R_i)P_i, \quad \bar{V}(d) = \sum_i \bar{U}(d, R_i)P_i,$$

find the rules d_0, d_1 maximizing $\underline{V}(d)$, $\bar{V}(d)$ respectively. The maximal value of refining the partition is less than $\bar{V}(d_1) - \underline{V}(d_0)$. Thus if the cost of further prior specification is greater than this quantity, you should immediately stop and choose rule d_0. Otherwise you should make further prior specifications. If you observe data y, with likelihood, given θ, $f(y \mid \theta)$, you make similar calculations, based instead on the quantities $\sup_{\theta \in R_i} U(d, \theta) f(y \mid \theta)$, $\inf_{\theta \in R_i} U(d, \theta) f(y \mid \theta)$.

There are several difficulties with the above procedures:

1. The procedure may be computationally difficult to apply.
2. You have no guidance as to what further specifications you should make and the gain you might reasonably expect.
3. In the case where you observe data y, you must specify your *prior* beliefs about θ *after* having observed y.
4. You are not artificially assigning a full prior distribution for θ. You may wish to apply these methods when, in similar fashion, there is no fully-specified family of likelihood functions for y given θ. This suggests a quasinonparametric framework. However, the maximization and minimization of $U(d, \theta) f(y \mid \theta)$ over any set R_i are now relatively uninformative. (For a strict nonparametric framework, the respective values would be infinity and zero, over any set R).
5. You may not wish to choose the rule d_0 when you stop, as avoiding the worst possible case is not the only property you would like your rule to possess. (As in (4), the maximization may concentrate on pathological cases).

We can partially overcome these difficulties by making two practical modifications to the above procedure.

(A) If we shall observe data s, then, before observing s, we are choosing between decision functions $\delta(s)$. Given the limited prior specification, we may not be overly concerned with hypothetical optimality. Instead, it may be more reasonable to restrict attention to a limited class of decision functions A, which is rich enough to contain reasonably good rules for each possible prior distribution consistent with the prior constraints, while only containing rules whose properties can be easily investigated given the restricted prior specification.

(B) For any specified decision function δ, the quantities which must be computed to carry out the suggested analysis are

$$S(\delta) = \sup_{P \in G} E(U(\delta(S), \theta)), \quad I(\delta) = \inf_{P \in G} E(U(\delta(S), \theta))$$

where G is the set of all prior distributions consistent with any constraints that you might impose. There are several criteria we might propose for the choice of an element of A based on these quantities. You might choose the rule δ which maximizes $I(\delta)$ over all $\delta \in A$. This is essentially the restriction of the minimax rule, suggested by Manski, to the class A. However, an obvious modification is

to seek a rule $\delta \in A$ for which $S(\delta)$ and $I(\delta)$ are both large, to take advantage of favourable cases without incurring too large a penalty under unfavourable cases. The simplest way to attempt this is to choose the rule which maximizes $M_\alpha(\delta)$ over $\delta \in A$, where for any $\alpha \in (0, 1)$,

$$M_\alpha(\delta) = \alpha S(\delta) + (1 - \alpha)I(\delta).$$

(This is similar in intention to the Hurwicz-α criterion. See, for example Fishburn, 1966.) It will be shown, in the example in Section 3, that a particular choice, $\alpha = \frac{1}{2}$, leads to an extremely simple solution. Thus, use of $M_{1/2}(\delta)$ (which we shall term the "midrisk") as well as providing an intuitively plausible decision criterion, is also an important simplifying assumption. We now consider within the above framework the problem described in Section 1.

3. Choice of linear Bayes rule with minimum midrisk. You wish to estimate the unknown quantity θ with quadratic loss. Suppose that θ is known to lie in the bounded interval $[\theta_0, \theta_\infty)$. Suppose that the only prior values you have chosen to specify are your prior probabilities over a finite partition of this interval. Thus you have specified values $\theta_0 < \theta_1 < \theta_2 < \cdots < \theta_n < \theta_{n+1}$, where $\theta_{n+1} = \theta_\infty$. You have also specified prior probabilities $P_0, P_1, \cdots, P_n$, where $P_i = P(\theta_i \leq \theta < \theta_{i+1})$. You are about to sample data S, from a distribution F of unknown form, and you intend to calculate a statistic X, based on S, which is unbiased for θ. The variance of X, for given F, is $\sigma^2(F)$. We will suppose that you can specify lower and upper bounds V_1, V_2 for your prior expectation of $\sigma^2(F)$ i.e. so that $V_1 \leq E\sigma^2(F) \leq V_2$. You decide to restrict attention to the class E of estimators of θ which are of the form $aX + b$. Thus, in terms of the criterion proposed in Section 2, you must choose values a^*, b^* to minimize, over all a, b, the midrisk

$$M(aX + b) = \tfrac{1}{2}(\sup_{P \in G} E((aX + b - \theta)^2 \mid P) + \inf_{P \in G} E((aX + b - \theta)^2 \mid P))$$

where G is the set of all prior distributions on F which satisfy the constraints $P(\theta_i \leq \theta < \theta_{i+1}) = P_i$, $i = 0, 1, \cdots, n$; $E\sigma^2(F) \in [V_1, V_2]$; we will impose one further constraint on G, a form of "weak" independence between θ and σ^2, namely that $\sup_{P \in G} E(\sigma^2(F) \mid \theta, P) = V_2$ and $\inf_{P \in G} E(\sigma^2(F) \mid \theta, P) = V_1$, for each θ.

We will now find the values a^*, b^*. We first define the prior distribution $P^* \in G$ which satisfies the following constraints.

(i) $P_0^* = P^*(\theta = \theta_0) = P_0/2$

$$P_i^* = P^*(\theta = \theta_i) = (P_{i-1} + P_i)/2, \; i = 1, \cdots, n$$

$$P_{n+1}^* = P^*(\theta = \theta_{n+1}) = P_n/2$$

(ii) $E(\sigma^2(F) \mid P^*) = V_m = (V_1 + V_2)/2.$

Denote the expected value and variance of θ with respect to P^* by $\bar{\theta}^*$ and $V^*(\theta)$ respectively, so that

$$\bar{\theta}^* = \sum_{i=0}^{n+1} \theta_i P_i^*, \quad V^*(\theta) = \sum_{i=0}^{n+1} \theta_i^2 P_i^* - (\bar{\theta}^*)^2.$$

What we shall show is that the optimal rule $a^*X + b^*$ is "almost" the actual linear Bayes rule with respect to the prior distribution P^*. The modifications we must make are as follows. Let r^* be the integer (between 0 and n) for which $\theta_{r^*} \leq \bar{\theta}^* < \theta_{r^*+1}$. Let $\bar{P} = P_{r^*}$. Let $\hat{\theta}^*$ be the value of θ_i which minimizes $|\bar{\theta}^* - \theta_i|$ over $i = 0, \cdots, n+1$.

Now define $\hat{\theta}^{**}$, $\bar{\theta}$, $\bar{\theta}^{**}$ by

$$\hat{\theta}^{**} = \bar{\theta}^* + (\bar{P}/(2 - \bar{P}))(\bar{\theta}^* - \hat{\theta}^*)$$

$$\bar{\theta} = (\theta_{r^*} + \theta_{r^*+1})/2$$

$$\bar{\theta}^{**} = \hat{\theta}^{**} \quad \text{if} \quad \min\{\bar{\theta}, \hat{\theta}^*\} \leq \hat{\theta}^{**} \leq \max\{\bar{\theta}, \hat{\theta}^*\}$$

$$= \bar{\theta} \qquad \text{otherwise.}$$

Finally define $V^{**}(\theta)$ by

$$V^{**}(\theta) = V^*(\theta) - (\bar{P}/(2 - \bar{P}))(\bar{\theta}^* - \hat{\theta}^*)^2, \qquad \text{if} \quad \bar{\theta}^{**} = \hat{\theta}^{**}$$

$$= V^*(\theta) + (\bar{\theta} - \bar{\theta}^*)^2 - \bar{P}(\bar{\theta} - \hat{\theta}^*)^2/2, \quad \text{if} \quad \bar{\theta}^{**} = \bar{\theta}.$$

We have the following results.

THEOREM 1. *The values a^*, c^* which minimize the expression*

$$M(aX + (1 - a)c)$$

are

$$a^* = \frac{V^{**}(\theta)}{(V^{**}(\theta) + V_m)}, \quad c^* = \bar{\theta}^{**}.$$

The minimum value of $M(aX + (1 - a)c)$ is

$$\frac{V^{**}(\theta)V_m}{(V^{**}(\theta) + V_m)}.$$

PROOF. Consider a particular estimator $aX + (1 - a)c$. For any prior distribution P for F, the expected loss of $aX + (1 - a)c$, given θ, is

$$\begin{aligned} L(\theta) &= E((a(X - \theta) + (1 - a)(c - \theta))^2 \mid \theta) \\ &= a^2V(\theta) + (1 - a)^2(c - \theta)^2 \end{aligned} \tag{1}$$

where $V(\theta) = E(\sigma^2 \mid P, \theta)$.

For each $r = 0, 1, 2, \cdots, n$, we define S_r, I_r to be the supremum and infinum of $L(\theta)$ over $\theta_r \leq \theta < \theta_{r+1}$ and $P \in G$. We have from (1) that

$$S_r \begin{cases} = a^2V_2 + (1 - a)^2(c - \theta_r)^2 & \text{if} \quad c \geq \theta_{r+1} \\ = a^2V_2 + (1 - a)^2\max((c - \theta_r)^2, (c - \theta_{r+1})^2) & \text{if} \quad \theta_{r+1} > c \geq \theta_r \\ = a^2V_2 + (1 - a)^2(c - \theta_{r+1})^2 & \text{if} \quad c < \theta_r \end{cases} \tag{2}$$

and

$$(3)\qquad I_r \begin{cases} = a^2V_1 + (1-a)^2(c-\theta_{r+1})^2 & \text{if } c \geq \theta_{r+1} \\ = a^2V_1 & \text{if } \theta_{r+1} > c \geq \theta_r \\ = a^2V_1 + (1-a)^2(c-\theta_r)^2 & \text{if } c < \theta_r. \end{cases}$$

The supremum and infinum of the quantity $E((aX + (1-a)c - \theta)^2 \mid P)$ over all $P \in G$ are $S(aX + (1-a)c) = \sum_{r=0}^{n} S_r P_r$, $I(aX + (1-a)c) = \sum_{r=0}^{n} I_r P_r$. Thus the midrisk of $aX + (1-a)c$, which we denote $M(a, c)$ is

$$M(a, c) = (\tfrac{1}{2})(I(aX + (1-a)c) + S(aX + (1-a)c)).$$

Suppose that $\theta_r \leq c < \theta_{r+1}$ and let $\theta_r^+ = \theta_r$ if $c - \theta_r \leq \theta_{r+1} - c$, otherwise let $\theta_r^+ = \theta_{r+1}$. From (2) and (3),

$$(4)\qquad \begin{aligned} M(a, c) &= M(aX + (1-a)c) \\ &= a^2V_m + (1-a)^2V^*(\theta) + (1-a)^2((c-\bar{\theta}^*)^2 - P_r(c-\theta_r^+)^2/2). \end{aligned}$$

As $\bar{\theta}^*$ lies in the interval $[\theta_{r^*}, \theta_{r^*+1})$, the value of c minimizing $M(a, c)$ must also lie in this interval; as for any value of c outside $[\theta_{r^*}, \theta_{r^*+1})$, the expression $(c-\bar{\theta}^*)^2 - P_r(c-\theta_r^+)^2/2$ in relation (4) will be positive. By a similar argument, the value of c minimizing $M(a, c)$ must be nearer to the value $\hat{\theta}^*$ than to any other element of the set $\{\theta_0, \theta_1, \theta_2, \cdots, \theta_{n+1}\}$. Thus, c^* is the value between $\min\{\bar{\theta}, \hat{\theta}^*\}$ and $\max\{\bar{\theta}, \hat{\theta}^*\}$ which minimizes the function.

$$h(c) = (c-\bar{\theta}^*)^2 - \bar{P}(c-\hat{\theta}^*)^2/2.$$

The value c^* minimizing $h(c)$ in this interval is $c^* = \bar{\theta}^{**}$. Further, if $\bar{\theta}^{**} = \hat{\theta}^{**}$ then

$$h(\bar{\theta}^{**}) = -(\bar{P}/(2-\bar{P}))(\bar{\theta}^* - \hat{\theta}^*)^2,$$

with a similar expression if $\bar{\theta}^{**} = \bar{\theta}$, so that we have, from (4), that

$$M(a, c^*) = a^2V_m + (1-a)^2V^{**}(\theta).$$

Thus the optimal choice a^*, and the value of $M(a^*, c^*)$ follow immediately. □

Recall that for any prior distribution P, the linear Bayes rule for θ, i.e. the rule $aX + (1-a)c$ which minimizes

$$R(a, c, P) = E((aX + (1-a)c - \theta)^2 \mid P),$$

is given (Goldstein, 1975) by

$$c = E(\theta \mid P),\ a = \frac{\operatorname{var}(\theta \mid P)}{\operatorname{var}(\theta \mid P) + E(\sigma^2(F) \mid P)}.$$

Thus, by Theorem 1, if you minimize midrisk, you will act as if you had specified the prior mean for θ as $\bar{\theta}^{**}$, and your prior variance for θ as $V^{**}(\theta)$.

This provides a possible solution to the problem posed in the introduction, namely how should you use a prior probability specification on a disjoint partition of θ values to specify a prior expectation and a prior variance, since if you choose the value $\bar{\theta}^{**}$ and $V^{**}(\theta)$ then linear Bayes methods will carry the additional justification of possessing the minimum midrisk property. Except for very coarse partitions, $\bar{\theta}^{**}$, $V^{**}(\theta)$ will be nearly equal to $\bar{\theta}^{*}$, $V^{*}(\theta)$.

Thus, in practice it will often be adequate to use

$$\hat{a} = (V^{*}(\theta)/(V^{*}(\theta) + V_m)) \quad \hat{c} = \bar{\theta}^{*}.$$

From relation (4),

$$M(a^{*}, c^{*}) = R(a^{*}, c^{*}, P^{*}) = \frac{V^{**}(\theta)V_m}{(V^{**}(\theta) + V_m)} \leq M(\hat{a}, \hat{c})$$

$$\leq \frac{V^{*}(\theta)V_m}{(V^{*}(\theta) + V_m)} = R(\hat{a}, \hat{c}, P^{*})$$

where the difference between the left- and right-hand expressions will be small unless most of the probability is concentrated in $[\theta_{r^{*}}, \theta_{r^{*}+1})$. Thus, for most purposes, it is adequate to use $\bar{\theta}^{*}$ and $V^{*}(\theta)$ as your prior mean and variance.

4. Refining the partition. We now briefly consider whether you should further refine the partition, if so in what way, and how much can you reasonably expect to gain by doing so. Suppose you refine the partition sequentially, considering at each stage whether you will stop or make a single further refinement. The effect of the prior specification upon the midrisk depends mainly on the value $V^{*}(\theta)$. Thus, we shall consider the change in $V^{*}(\theta)$. With notation as in Section 3, suppose for simplicity you decide that if you refine the partition, you shall do so by specifying the probabilities $P_{i1} = P(\theta_i \leq \theta < \theta_i^{*})$, $P_{i2} = P(\theta_i^{*} \leq \theta < \theta_{i+1})$, for some value i, where $\theta_i^{*} = (\theta_i + \theta_{i+1})/2$. Denote the value of $V^{*}(\theta)$ after refining the partition at value θ_i^{*}, by $V_i^{*}(\theta)$, and denote $d_i = \theta_{i+1} - \theta_i$. We have the following result.

THEOREM 2.

$$V^{*}(\theta) - V_i^{*}(\theta) = \tfrac{1}{16}d_i^2(P_{i2} - P_{i1})^2 + \tfrac{11}{64}d_i^2P_i + d_i(\theta_i^{*} - \bar{\theta}^{*})(P_{i1} - P_{i2})/2.$$

PROOF. Write $V^{*}(\theta)$ as

$$V^{*}(\theta) = \textstyle\sum_{j=0}^{n} (\theta_j^{*} - \bar{\theta}^{*})^2P_j + \tfrac{1}{4} \sum_{j=0}^{n} d_j^2P_j.$$

Denote by $\bar{\theta}_i^{*}$ the value of $\bar{\theta}^{*}$ after the partition is refined at θ_i^{*}. Thus

$$V_i^{*}(\theta) = \textstyle\sum_{j\neq i} (\theta_j^{*} - \bar{\theta}_i^{*})^2P_j + (\theta_i^{*} + d_i/4 - \bar{\theta}_i^{*})^2P_{i2} + (\theta_i^{*} - d_i/4 - \bar{\theta}_i^{*})^2P_{i1} + \tfrac{1}{4} \sum_{j\neq i} d_j^2P_j + \tfrac{1}{4}(d_i/4)^2P_i.$$

Expanding the above formula, and using the relations $P_i = P_{i1} + P_{i2}$ and $\bar{\theta}_i^{*} - \bar{\theta} = (d_i/4)(P_{i2} - P_{i1})$, gives the required result.

Thus, you should choose to split a wide interval, with large prior probability. Further, you should choose an interval for which you feel that it is plausible that the probability is concentrated in one half of the interval. The effect is largest if the interval is far from the current mean $\bar{\theta}^*$ (this decreases $V_i^*(\theta)$ if $P_{i1} - P_{i2}$ has the same sign as $\theta_i^* - \bar{\theta}^*$, i.e. if your beliefs are concentrated towards the centre of the distribution).

REFERENCES

FISHBURN P. C. (1966). Decision under uncertainty. An introductory exposition. *J. Industrial Engineering* **17** 341–353.

GOLDSTEIN, M. (1975). A note on some Bayesian nonparametric estimates. *Ann. Statist.* **3** 736–740.

MANSKI, C. F. (1981). Learning and decision making when subjective probabilities have subjective domains. *Ann. Statist.* **9** 59–65.

DEPARTMENT OF STATISTICS
THE UNIVERSITY OF HULL
COTTINGHAM ROAD
HULL HU6 7RX ENGLAND