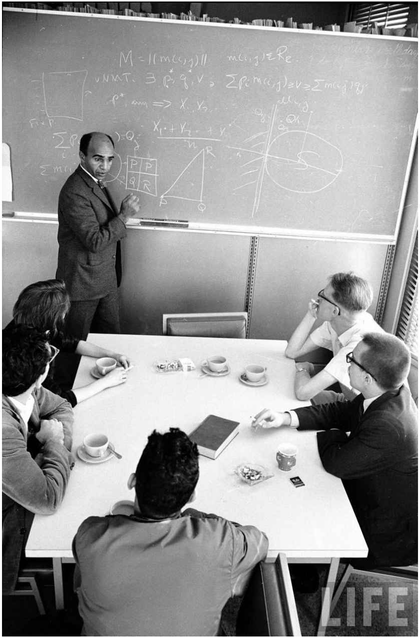
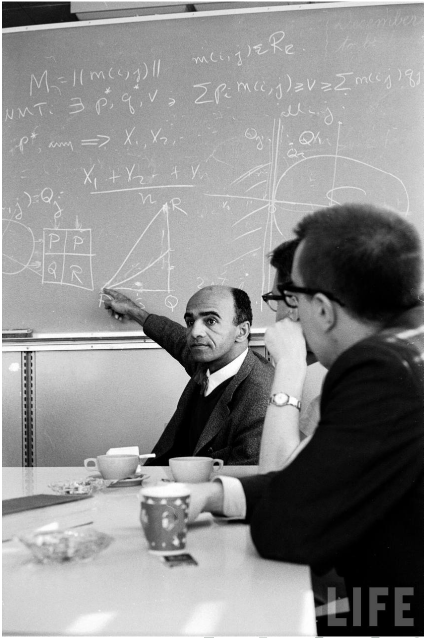
The Contributions of David Blackwell to Bayesian Decision Theory

giovanni_parmigiani@dfci.harvard.edu

Blackwell Memorial Conference
Howard University; April 2012

by Life Magazine photographer Alfred Eisenstaed





Three Cornerstones by Blackwell

- It is rational to summarize data via sufficient statistics.
  Blackwell *Annals Math Stat 1947*
- The general structure how to optimally learn from sequentially accruing information
  Arrow Blackwell & Girshick
  Econometrica 1949
- All rational (and not too opinionated) agents will agree given enough common data
  Blackwell & Dubins *Annals Math Stat 1962*

CONDITIONAL EXPECTATION AND UNBIASED SEQUENTIAL ESTIMATION[1]

By David Blackwell

Howard University

1. Summary. It is shown that $E[f(x)\ E(y \mid x)] = E(fy)$ whenever $E(fy)$ is finite, and that $\sigma^2 E(y \mid x) \leq \sigma^2 y$, where $E(y \mid x)$ denotes the conditional expectation of y with respect to x. These results imply that whenever there is a sufficient statistic u and an unbiased estimate t, not a function of u only, for a parameter θ, the function $E(t \mid u)$, which is a function of u only, is an unbiased estimate for θ with a variance smaller than that of t. A sequential unbiased estimate for a parameter is obtained, such that when the sequential test terminates after i observations, the estimate is a function of a sufficient statistic for the parameter with respect to these observations. A special case of this estimate is that obtained by Girshick, Mosteller, and Savage [4] for the parameter of a binomial distribution.

Theorem (Rao-Blackwell Inequality)

Suppose that $\mathcal{A}$ is a convex subset of $\Re^m$ and that $L(\theta, a)$ is a convex function of a for all $\theta \in \Theta$. Suppose also that t is a sufficient statistic for θ, and that δ_0 is a nonrandomized decision rule such that $E_{x|\theta}[|\delta_0(x)|] < \infty$. The decision rule defined as

$$\delta_1(t) = E_{x|t}[\delta_0(x)] \tag{1}$$

is R–equivalent to or R–better than δ_0.

Proof

Jensen's inequality states that if a function g is convex, then $g(E(x)) \leq E(g(x))$. Therefore,

$$L(\theta, \delta_1(t)) = L(\theta, E_{x|t}[\delta_0(x)]) \leq E_{x|t}[L(\theta, \delta_0(x))] \tag{2}$$

and

$$\begin{aligned} R(\theta, \delta_1) &= E_{t|\theta}[L(\theta, \delta_1(t))] \\ &\leq E_{t|\theta}[E_{x|t}\{L(\theta, \delta_0(x))\}] \\ &= E_{x|\theta}[L(\theta, \delta_0(x))] \\ &= R(\theta, \delta_0) \end{aligned}$$

$\square$

An example

Suppose that $x_1, \ldots, x_n$ are independent and identically distributed as $N(\theta, 1)$ and that we wish to estimate the tail area to the left of $c - \theta$ with square error loss

$$L(\theta, a) = (a - \Phi(c - \theta))^2.$$

Here c is some fixed real number, and $a \in \mathcal{A} \equiv [0, 1]$. A possible decision rule is the empirical tail frequency

$$\delta_0(x) = \sum_{i=1}^{n} I_{(-\infty, c]}(x_i)/n.$$

$t = \bar{x}$ is a sufficient statistic for θ. Since $\mathcal{A}$ is convex and the loss function is a convex function of a,

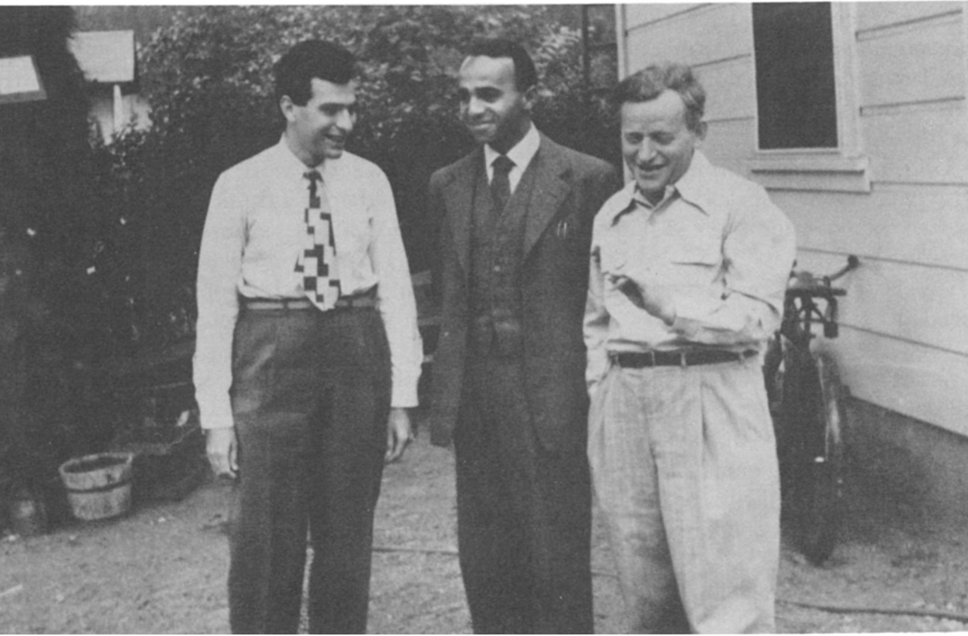
$$E_{x|t}[\delta_0(x)] \quad = \quad \frac{1}{n}\sum_{i=1}^{n} E_{x_i|t}[I_{(-\infty, c]}(x_i)] = \Phi\left(\frac{c - t}{\sqrt{\frac{n-1}{n}}}\right),$$

Because of the Rao–Blackwell theorem, the rule $\delta_1(t) = E_{x|t}[\delta_0(x)]$, is R–better than δ_0.

BAYES AND MINIMAX SOLUTIONS OF SEQUENTIAL DECISION PROBLEMS[1]

By K. J. Arrow, D. Blackwell, M. A. Girshick

The present paper deals with the general problem of sequential choice among several actions, where at each stage the options available are to stop and take a definite action or to continue sampling for more information. There are costs attached to taking inappropriate action and to sampling. A characterization of the optimum solution is obtained first under very general assumptions as to the distribution of the successive observations and the costs of sampling; then more detailed results are given for the case where the alternative actions are finite in number, the observations are drawn under conditions of random sampling, and the cost depends only on the number of observations. Explicit solutions are given for the case of two actions, random sampling, and linear cost functions.



Long term implication: Irrelevance of stopping rule

The Bayesian optimal terminal decision is not affected by the stopping rule. The reason for this result is a general factorization of the likelihood: for any stopping rule ζ for sampling from a sequence of observations $x_1, x_2, \ldots$ having fixed sample size parametric model $f(x^n|n,\theta) = f(x^n|\theta)$, the likelihood function is

$$f(n, x^n|\zeta, \theta) = \zeta_n(x_n) \prod_{i=1}^{n-1} (1 - \zeta_i(x_i)) f(x^n|\theta) \propto f(x^n|\theta), \quad \theta \in \Theta, \tag{3}$$

for all (n, x^n) such that $f(n, x^n|\zeta, \theta) \neq 0$.

MERGING OF OPINIONS WITH INCREASING INFORMATION[1]

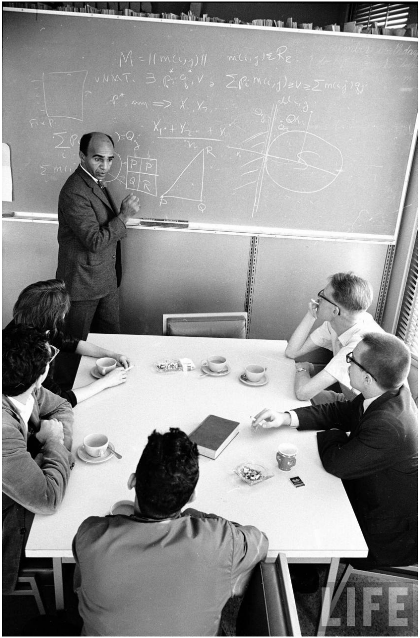
BY DAVID BLACKWELL AND LESTER DUBINS

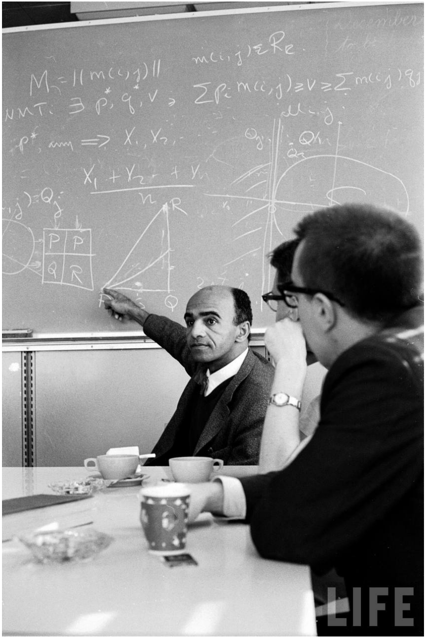
University of California, Berkeley

1. History. One of us [1] has shown that if Zn, $n = 1, 2, \cdots$ is a stochastic process with D states, $0, 1, \cdots, D - 1$ such that $X = \sum_{n=1}^{\infty} Z_n/D^n$ has an absolutely continuous distribution with respect to Lebesgue measure, then the conditional distribution of $R_k = \sum_{n=1}^{\infty} Z_{k+n}/D^n$ given $Z_1, \cdots, Z_k$ converges with probability one as $k \to \infty$ to the uniform distribution on the unit interval, in the sense that for each λ, $0 < \lambda \leqq 1$, $P(R_k < \lambda \mid Z_1, \cdots, Z_k) \to \lambda$ with probability 1 as $k \to \infty$. It follows that the unconditional distribution of R_k converges to the uniform distribution as $k \to \infty$. If $\{Z_n\}$ is stationary, the distribution of R_k is independent of k, and hence uniform, a result obtained earlier by Harris [3]. Earlier work relevant to convergence of opinion can be found in [4, Chap. 3, Sect. 6].

Here we generalize these results and also show that the conditional distribution of R_k given $Z_1, \cdots, Z_k$ converges in a much stronger sense. All probabilities in this paper are countably additive.

5. Interpretation. Usually, there is essentially only one conditional distribution Q^n of the future given the past. Therefore, our theorem may be interpreted to imply that if the opinions of two individuals, as summarized by P and Q, agree only in that $P(D) > 0 \leftrightarrow Q(D) > 0$, then they are certain that after a sufficiently large finite number of observations $x_1, \cdots, x_n$, their opinions will become and remain close to each other, where close means that for every event E the probability that one man assigns to E differs by at most ϵ from the probability that the other man assigns to it, where ϵ does not depend on E. Leonard J. Savage observed that our theorem applies to the particularly interesting case in which P and Q are symmetric (or exchangeable). That is, if the measures P and Q on the sequences x_i are those that arise when the x_i are, for a fixed parameter value, independent and identically distributed observations, with prior distributions p and q on the parameter, then the relations of absolute continuity between P and Q are precisely those between p and q.


LIFE


LIFE