HOW TO DEAL WITH THE A PRIORI PROBABILITY ON THE "TWO-ARMED" BANDIT PROBLEM

Kono, Kazumasa
General Education Department of Kyushu Univerity

https://doi.org/10.15017/1448947

出版情報：九州大学教養部数学雑誌. 4 (2), pp.27-34, 1967-02. General Education Department, Kyushu University
バージョン：

Math. Rep.
IV-2, 1966.

HOW TO DEAL WITH THE A PRIORI PROBABILITY ON THE "TWO-ARMED" BANDIT PROBLEM

by

Kazumasa Kôno

[1966. 11. 10]

1. Introduction. The type of problem known as the "Two-Armed Bandit" holds interest for it to be one of the simplest type in sequential design of experiment. Some results on this problem and its generalizations have been noted in papers by the authors [1], [2], [3] and [4]. A simple formulation of this problem is as follows: We have two binomial random variables, X and Y, having parameters under the two hypotheses, H_1 and H_2, given by

		X	Y
(ξ)	H_1	p	q
$(1-\xi)$	H_2	q	p

where ξ is the a priori probability that H_1 is true. We wish to find a sequential procedure for taking observations so as to maximize the expected sum of the observations. The authors have studied the class of Bayes procedures, when the a priori probability ξ had been known. The procedure considered by them is the stepwise maximumization design termed "S_1-procedure", which, after the j-th observation, the a posterior probability, ξ_j, is computed, at next step, the variable corresponding to the maximum of the expected values

$$\xi_j p+(1-\xi_j)q$$

and

$$\xi_j q+(1-\xi_j)p$$

is observed. This means that we select the variable X or Y at the $(j+1)$-th step according as $\xi_j \geq \frac{1}{2}$ or $\xi_j < \frac{1}{2}$, under the assumption $p>q$. They have mentioned that the S_1-procedure has the optimal property for a certain criterion.

Define $W_n(\xi, S_1)$ as follows: Assume $p>q$,

$$
(1.1) \qquad W_1(\xi, S_1) = \begin{cases} q+(p-q)\xi & \text{for } \xi \geq \frac{1}{2} \\ p-(p-q)\xi & \text{for } \xi < \frac{1}{2} \end{cases}
$$

and if $\xi \geq \frac{1}{2}$,

$$
(1.2) \qquad \begin{aligned} W_n(\xi, S_1) = W_1(\xi, S_1) + W_{n-1}(\xi_1^{X=1}, S_1)P_\xi(X=1) \\ + W_{n-1}(\xi_1^{X=0}, S_1)P_\xi(X=0), \end{aligned}
$$

while, if $\xi < \frac{1}{2}$,

$$
(1.3) \qquad \begin{aligned} W_n(\xi, S_1) = W_1(\xi, S_1) + W_{n-1}(\xi_1^{Y=1}, S_1)P_\xi(Y=1) \\ + W_{n-1}(\xi_1^{Y=0}, S_1)P_\xi(Y=0) \end{aligned}
$$

where

$$
(1.4) \qquad \xi_1^{Z=c} = \frac{\xi Pr(Z=c/H_1)}{P_\xi(Z=c)}
$$

and

$$
(1.5) \qquad P_\xi(Z=c) = \xi Pr(Z=c/H_1) + (1-\xi)Pr(Z=c/H_2).
$$

Then we have the expected value of the average of the first n observations

$$
(1.6) \qquad a_n(\xi, S_1) = \frac{1}{n} W_n(\xi, S_1).
$$

The following two theorems have been shown by them.

THEOREM [Bradt, Johnson and Karlin] *Following the S_1-procedure, the expected value of average $a_n(\xi, S_1)$ converges to max (p, q) as $n \to \infty$.*

THEOREM [Feldman] *For every n the S_1-procedure is optimal.*

D. Feldman has defined the "risk function" by $R_n(\xi, S_1)$ which is equivalent essentially to $W_n(\xi, S_1)$ and suggested to minimize the expected number of "mistakes", where, under H_1, it is correct to select the variable X with higher probability p of success and wrong to select the other Y, and, under H_2, it is correct to select Y and wrong to select X. This theorem means that, for every n, $R_n(\xi, S_1)$ is equal to the risk function $R_n^*(\xi)$ of optimum procedure satisfying

$$
(1.7) \qquad \begin{aligned} &R_n^*(\xi) = \min[R_n^X(\xi), R_n^Y(\xi)] \\ &R_n^X(\xi) = (1-\xi) + R_{n-1}(\xi_1^{X=1})P_\xi(X=1) + R_{n-1}(\xi_1^{X=0})P_\xi(X=0) \\ &R_n^Y(\xi) = \xi + R_{n-1}(\xi_1^{Y=1})P_\xi(Y=1) + R_{n-1}(\xi_1^{Y=0})P_\xi(Y=0) \end{aligned}
$$

where $R_n^Z(\xi)$ is the risk function when we observe the variable Z at the first step.

2. Discussion about a priori probability ξ.

We shall deal with the mechanic on which the problem stands up. What is that the a priori probability ξ means? We have two situations: The first is one that it exists such chance mechanic that ξ is a mixed strategy of manager or nature and the second is one that ξ represents an amount of imformation for hypothesis H_1 and so the chance mechanic does not exist or ξ is an initial plausible value for true probability which is a mixed strategy of manager. In the first situation, let the manager (or nature) set up at each step two undistinguishable slot machines corresponding to parameters p and q on two places X and Y or Y and X with the probability ξ or $1-\xi$. Then H_1 occurs with the probability ξ and the other occurs with $1-\xi$ at eact stage.

Let assume that we know the true probability ξ which H_1 occurs, that is, X is correct. Then we can make the procedure which continues to observe only X or Y through all playing game according as $\xi\geq\frac{1}{2}$ or $\xi<\frac{1}{2}$. Let denote such procedure by C, and one for selecting repeatedly only X inspite of value ξ by C_x and one for Y by C_y.

We can define the risk functions by $\bar{R}_n(\xi, C_x)$, $\bar{R}_n(\xi, C_y)$ and $R_n(\xi, C)$ for these procedures and we have, for every n and fixed ξ,

$$\text{(2.1)}\qquad \begin{aligned} \bar{R}_n(\xi, C_x) &= n(1-\xi), \\ \bar{R}_n(\xi, C_y) &= n\xi \end{aligned}$$

and

$$R_n(\xi, C)=n\cdot\min[(1-\xi), \xi],$$

while they are unequal to the "risk function" defined by D. Feldman. Further, we have the following remark.

REMARK. *When we can know the true probability ξ, the procedure C is the best one. Hence such assumption that we can know ξ leads the problem to a trivial case.*

We can define any procedure for every n trials by $\Delta_n=(\lambda_1, \lambda_2, \ldots\ldots, \lambda_n)$ whose i-th conponent λ_i means any mixed strategy at the i-th step such that the variable X is observed with a probability λ_i and λ_i may be a function of proceeding i-1 selections and observations and of ξ.

Then we have the risk function of such procedure for every n as follows:

(2.2) $$\overline{R}_n(\xi,\ \Delta_n)=\sum_{i=1}^{n}EE'[\lambda_i(1-\xi)+(1-\lambda_i)\xi]$$

where E' means the conditional expectation under hypothesis which the preceeding i-1 selections and observations occurred. Since

$$E'[\lambda_i(1-\xi)+(1-\lambda_i)\xi]\geq\min[(1-\xi),\ \xi],$$

we have the following inequality for any procedure Δ_n.

(2.3) $$R_n(\xi,\ \Delta_n)\geq n\cdot\min[(1-\xi),\ \xi]$$

Using (2.1) and (2.3), we can prove the remark. Now, let consider an another case in the first situation. Once the manager or nature set up a slot machine, such setting does not change through playing game. In this case, unless we restrict ourselves to consider under the condition which H_1 or H_2 occurred, we can identify the essential body of problem with the case mentioned above. If we restrict ourselves to consider under such condition, then we can reduce this case to the second situation. The preceeding discussion leads us to the following conclusion.

CONCLUSION. *In order to avoid a trivial case, we have to start from the situation where we can not know a true probability* $(0\leq\xi\leq1)$ *but have its initial plausible value* $\eta_0(0<\eta_0<1)$. *The case that unknown parameter* ξ *is equal to* 0 *or* 1 *is the degenerate one that the chance machanic does not exist.*

3. Behavior of S_1-procedure in special case $p+q=1$.

In general, that is, in either situation existing chance mechanic or not, it is sufficient to give an unknown parameter ξ which lies in an interval [0,1], while η_0 is an initial plausible value for ξ and lies an interval (0,1). We study the a posterior η_j which is computed after the j-th observation under S_1-procedure. For simplicity, suppose the special case such that holds $p+q=1$. Define the states of η as follows:

$$E_0=\left\{\eta\ ;\ \frac{1}{2}\leq\eta<p\right\}$$

and

$$E_1=\left\{\eta\ ;\ p\leq\eta<\frac{p^2}{p^2+q^2}\right\}$$

$$E_2=\left\{\eta\ ;\ \frac{p^2}{p^2+q^2}\leq\eta<\frac{p^3}{p^3+q^3}\right\}$$

$$E_{-1}=\left\{\eta\ ;\ q\leq\eta<\frac{1}{2}\right\}$$

$$E_{-2}=\left\{\eta\,;\ \frac{q^2}{p^2+q^2}\leq\eta<q\right\}$$

$$\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\cdots\cdots\cdots\cdots\cdots\cdots$$

$$E_n=\left\{\eta\,;\ \frac{p^n}{p^n+q^n}\leq\eta<\frac{p^{n+1}}{p^{n+1}+q^{n+1}}\right\} \tag{3.1}$$

$$E_{n+1}=\left\{\eta\,;\ \frac{p^{n+1}}{p^{n+1}+q^{n+1}}\leq\eta<\frac{p^{n+2}}{p^{n+2}+q^{n+2}}\right\}$$

$$E_{-n}=\left\{\eta\,;\ \frac{q^n}{p^n+q^n}\leq\eta<\frac{q^{n-1}}{p^{n-1}+q^{n-1}}\right\}$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

for $n=1,\ 2,\ 3,\ \ldots\ldots\ldots$ and $p>q$.

LEMMA 1. *E_0, E_1, E_{-1}, ... E_n, E_{-n}, ... are mutually disjoint.*

FROOF. For all $n\geq 0$, since $p>q$, we have

$$\frac{p^{n+1}}{p^{n+1}+q^{n+1}}-\frac{p^n}{p^n+q^n}=\frac{p^nq^n(p-q)}{(p^{n+1}+q^{n+1})(p^n+q^n)}>0$$

and for $n<0$, we have

$$\frac{q^{n-1}}{p^{n-1}+q^{n-1}}-\frac{q^n}{p^n+q^n}=\frac{q^{n-1}p^{n-1}(p-q)}{(p^{n-1}+q^{n-1})(p^n+q^n)}>0\,.$$

Hence they show clearly lemma.

LEMMA 2. *Let η be in state $E_n(n=0,\ \pm 1,\ \pm 2\ldots)$. Then, following S_1-procedure at next step, an a posterior value η' of η lies in state E_{n+1} with probability P and in state E_{n-1} with probability Q, where*

$$P=\xi p+(1-\xi)q,\ \ Q=1-P. \tag{3.2}$$

PROOF. Let η be in E_n for $n\geq 0$. Then we select at next step a variable X. If an event $\{X=1\}$ occurs, then an a posterior value η' of η is

$$\eta'=\frac{\eta p}{\eta p+(1-\eta)q}=\frac{1}{1+\dfrac{1-\eta}{\eta}\dfrac{q}{p}}, \tag{3.3}$$

and is a monotone increasing function of η, while η satisfies the inequality

$$\frac{p_n}{p_n+q_n}\leq\eta<\frac{p^{n+1}}{p^{n+1}+q^{n+1}},\ \ \text{for } n\geq 0. \tag{3.4}$$

Using (3.4), we have the inequality

$$\frac{p^{n+1}}{p^{n+1}+q^{n+1}}\leq\eta'<\frac{p^{n+2}}{p^{n+2}+q^{n+2}}, \tag{3.5}$$

that is, η' lies in state E_{n+1}.

If an event $\{X=0\}$ occurs, then, since we have

(3.6) $$\eta' = \frac{q}{q+(1-\eta)p},$$

we have the inequalities

(3.7) $$\frac{p^{n-1}}{p^{n-1}+q^{n-1}} \leq \eta' < \frac{p^n}{p^n+q^n} \quad \text{for } n \geq 1$$

and

(3.8) $$q \leq \eta' < \frac{1}{2} \quad \text{for } n=0,$$

that is, η' lies in state E_{n-1}. On other hand,

(3.9) $$P_\xi(X=1) = \xi p + (1-\xi)q$$

and $$P_\xi(X=0) = \xi q + (1-\xi)p$$

hold.

Similarly, for all $n \leq 1$, we can prove the similar proposition, using

(3.10) $$P_\xi(Y=0) = \xi p + (1-\xi)q$$

and

$$P_\xi(Y=1) = \xi q + (1-\xi)p.$$

From lemma 1, we can define an unrestricted random walk with infinite many states E_0, E_1, E_{-1}, E_2, E_{-2}, ... and if $P \neq 0$, that is, $(p-q)(2\xi-1) \neq 0$, then all states are transient. This fact means that if $P > Q$, then $Pr\left\{\eta_n \geq \frac{1}{2}\right\}$ converges to 1 as $n \to \infty$ and, if $P < Q$, then $Pr\left\{\eta_n < \frac{1}{2}\right\}$ converges to 1 as $n \to \infty$. We have the following theorem.

THEOREM. *Assume $p > q$. Following S_1-procedure with any initial plausible η_0 for unknown ξ, the expected avearage number of "mistakes" for the first n observations converges to $min[(1-\xi), \xi]$ as $n \to \infty$.*

PROOF. CASE 1. $P > Q$, i. e., $\xi > \frac{1}{2}$. For sufficiently large N and arbitrary $\delta (0 < \delta < 1)$, we have

(3.10) $$Pr\left\{\eta_n \geq \frac{1}{2}\right\} > 1-\delta$$
$$Pr\left\{\eta_n < \frac{1}{2}\right\} \leq \delta \qquad \text{for } n > N.$$

Let denote by $\frac{1}{n}\overline{R}_n\ (\xi/\eta_0,\ S_1)$ the risk function for S_1-procedure which starts an initial value η_0 and which means the expected average number of mistakes and then the risk function for n trials is

(3.11) $$\frac{1}{n} R_n(\xi/\eta_0,\ S_1) = \frac{1}{n}\sum_{j=1}^{n}\left[(1-\xi)Pr\left\{\eta_j \geq \frac{1}{2}\right\} + \xi Pr\left\{\eta_j < \frac{1}{2}\right\}\right],$$

because the a posterior value η_j is computed after the j-th observation and, if $\eta_j \geq \frac{1}{2}$, then we select the variable X at next step and, if $\eta_j < \frac{1}{2}$, then we select the variable Y. For sufficiently large N and arbitrary $\delta > 0$, we have $\frac{1}{n}\bar{R}_n(\xi/\eta_0, S_1)$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{j=1}^{N}\left[(1-\xi)Pr\left\{\eta_j \geq \frac{1}{2}\right\} + \xi Pr\left\{\eta_j < \frac{1}{2}\right\}\right] \\
&\qquad + \frac{1}{n}\sum_{j=N+1}^{n}\left[(1-\xi)Pr\left\{\eta_j \geq \frac{1}{2}\right\} + \xi\ Pr\left\{\eta_j < \frac{1}{2}\right\}\right] \\
&> \frac{1}{n}\sum_{j=1}^{N}\left[(1-\xi)Pr\left\{\eta_j \geq \frac{1}{2}\right\} + \xi Pr\left\{\eta_j < \frac{1}{2}\right\}\right] \\
&\qquad + \frac{n-N}{n}(1-\xi)(1-\delta) + \frac{1}{n}\sum_{j=N+1}^{n}\xi Pr\left\{\eta_j < \frac{1}{2}\right\} \\
&> \frac{n-N}{n}(1-\xi)(1-\delta), \quad \text{for } n > N.
\end{aligned} \tag{3.12}
$$

In the other hand, for sufficiently large M, and arbitrary $\varepsilon > 0$, we have

$$
\frac{1}{n}\sum_{j=1}^{M}\left[(1-\xi)Pr\left\{\eta_i \geq \frac{1}{2}\right\} + \xi Pr\left\{\eta_j < \frac{1}{2}\right\}\right] < \varepsilon \tag{3.13}
$$

for $n > M$ $(M > N)$.

Using (3.12) and (3.13), for sufficiently large M, N $(M > N)$ and arbitrary ε, δ, we have the inequality

$$
\begin{aligned}
\varepsilon + (1-\xi) + \frac{n-M}{n}\xi\delta &> \frac{1}{n}R_n(\xi/\eta_0, S_1) \\
&> \frac{n-N}{n}(1-\xi)(1-\delta) \quad \text{for } n > M.
\end{aligned} \tag{3.14}
$$

The inequality (3.14) give the fact that $\frac{1}{n}R_n(\xi/\eta_0, S_1)$ converges to $(1-\xi)$ as $n \to \infty$.

Case 2. $P < Q$ i.e., $\xi < \frac{1}{2}$. By the similar way, we have the fact that $\frac{1}{n}\bar{R}_n(\xi/\eta_0, S_1)$ converges to ξ as $n \to \infty$.

Case 3. $P = Q$, i. e., $\xi = \frac{1}{2}$. In this case, either selection of X and Y gives the same risk and any procedure leads the risk function to tend $\frac{1}{2}$.

For $P>Q$, $P=Q$, or $P<Q$, min $[(1-\xi), \xi]$ gives respectively $(1-\xi)$, $\frac{1}{2}$, or ξ and theorem is proved.

General Education Department of Kyushu Univerity.

References

[1] Bradt, R. N. Johnson, S. M. and Karlin, S. (1956). On sequential designs for maximizing the sum of n observations. Ann. Math Statist. 27 1060–1074.

[2] Bradt, R. N. and Karlin, S. (1956). On the design and comparison of certain dichotomous experiments. Ann. Math. Statist. 27 390–409.

[3] Feldman, D. (1962). Contributions to the "two-Armed Bandit" problem. Ann. Math. Statist. 33. 847–856.

[4] Robbins, H. (1952). Some aspects of the sequential designs. Bull. Amer. Math. Soc. 58 527–535.