5-1969

Bayesian Inference for Decision Making

Ming-yih Kao
Utah State University

1-1-1969

Bayesian Inference for Decision Making

Ming-yih Kao

BAYESIAN INFERENCE FOR DECISION MAKING

by

Ming-yih Kao

A plan B report submitted in partial fulfillment
of the requirements for the degree

of

MASTER OF SCIENCE

in

Statistics

Approved:

UTAH STATE UNIVERSITY
Logan, Utah

1969

ACKNOWLEDGMENTS

I would like to express my appreciation to Dr. David White, my major professor, for his providing this subject and suggestions in preparation of this report.

I would also like to thank Dr. Rex L. Hurst and Dr. Eugene C. Kartchner for their willingness to serve as members of my graduate committee.

Ming-yih Kao

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER I

INTRODUCTION

Objectives of study

In recent years, Bayesian inference has become very popular in applied statistics. This study will present the fundamental concept of Bayesian inference and the basic techniques of application to statistical quality control, marketing research, and other related fields.

Historical development

The name of Bayesian statistics comes from a mathematical theorem by an eighteenth-century English philosopher and Presbyterian minister, Thomas Bayes (1702-1761). The theorem states that if a certain probability is found, some events will occur; and if later additional information is secured, the revised probability can be estimated by combining these two preliminary probabilities. Sometimes this is referred to as the "inverse probability method."

For instance, we can combine the statistical probability derived from a sample with a probability estimated personally by people experienced in the discipline being considered. An actual example is the marketing research group in Du Pont fibers division. They used management's judgment about the probable sales of production combined into a probability curve of potential sales along with the cost data. The results told Du Pont the demand in the market and what size plant should be built.[1]

[1] Theodore J. Sielaff, *Statistics in Action* (San Jose, California: San Jose State College, 1963), p. 155-158.

To businessmen, the translation of their subjective forecast into mathematical probability is a new experience; but to the statistician, it departs from the classical statistical theory that probabilities are not available unless they are based on repeated observations. The relevance of subjective probability in Bayesian statistics is still a controversial topic among statisticians, who explain their favorite concepts of "significance level," "confidence coefficient," "unbiased estimates," etc., in terms of objective probability; i.e., the frequency of occurrence in repeated trials such as a game of chance. According to the classical approach, elements of personal judgment or subjective belief should be excluded from statistical calculation as much as possible. The classical school believes that the statistician can exercise his judgment, but he should be careful about it, and it had better be separated from statistical theory.

In the past 30 or 40 years mathematical statistics has been treated increasingly rigidly. Bayesian statistics was neglected. One of the reasons is that one would seldom have enough information about the states of nature (prior probability). Without this information the Bayes theorem is not applicable. Even then, classical statisticians show a great diversity. R. A. Fisher, who contributed so much to the development of classical statistics, held an unclassical viewpoint, not far removed from the Bayesian. Especially has A. Wald made much use of the formal Bayesian approach to which no probabilistic significance is attached.

In the past two decades, L. J. Savage's interpretation[2] of the works

[2] Leonard J. Savage, *The Foundations of Statistics* (New York: John Wiley and Sons, Inc., 1954), p. 27-30.

of Bruns de Finetti on subjective probability has established the foundation of Bayesian statistics. Subjective probability may differ from individual to individual. The members of the Bayesian school also are divided on how the prior subjective probability is to be determined. For example, R. V. Juises thought it should be on the basis of prior experience, while H. Jeffreys uses certain canonical distribution. Some others claim that the prior probability distribution may be based upon either subjective beliefs or upon previous objective frequency.

The relevance of this prior probability would depend upon the similarity between their additional information and that which has been previously experienced. In other words, if the additional information being undertaken is from an entirely different population, then the prior information may be of little relevance. As an illustration, we consider an experienced statistical quality controller's estimate of the defective proportion of a particular production process. According to his previous experience in production process, he would have some notion about the defective proportion to be produced. The similarity between this particular process and the previous process needs to be considered. If it is an entirely new production being undertaken using a new process, the prior information may be of little relevance. It departs from the assumption that posterior probability is consistent with his prior probability and likelihood in accordance with Bayes' theorem.

Contributions in Bayesian statistics have been made by V. Neumann,[3]

[3] Von Neumann and O. Morgenstern, *Theory of Game and Economic Behavior* (Princeton, New Jersey: Princeton University Press, 1947), p. 15-29.

A. Wald,[4] D. Blackwell and M. A. Girshick,[5] until the work of H. Raiffa and R. Schlaifer,[6] which presents a rigid mathematical theory of statistical decisions suitable for application.

Generally speaking, Bayesian inference assesses some underlying "states of nature" that are uncertain. These states of nature are a set of mutually exclusive and collectively exhaustive events that are considered to be a random variable, and it is known in advance that one, and only one, of these events will actually occur, but there is uncertainty about which one will occur. Bayesian inference starts by assigning a probability to each of these events on the basis of whatever prior probability is available under current investigation. If additional information is subsequently obtained, the initial probabilities are revised on the additional information through the Bayes theorem.

The importance of relative consequence

In testing hypotheses, a type I error is committed if H_1 is accepted when H_o is true. It means rejecting a true hypothesis; i.e., $P(a_2|H_o) = \alpha$. Conversely, a type II error is committed if H_o is accepted when H_1 is true. It means accepting a false hypothesis; i.e., $P(a_1|H_1) = \beta$. To determine the optimum selection, it is necessary to measure the risk of committing these two kinds of errors. It is already known that to eliminate errors of these two types is impossible, since for a given size sample, the type I error and type II error have

[4] Abraham Wald, *Statistical Decision Function* (New York: John Wiley and Sons, Inc., 1950), p. 103-122.

[5] David Blackwell and M. A. Girshick, *Theory of Games and Statistical Decisions* (New York: John Wiley and Sons, Inc., 1954), p. 147-169.

[6] Howard Raiffa and Robert Schlaifer, *Applied Statistical Decision Theory* (Boston, Massachusetts: Harvard University Press, 1961), p. 132-174.

an inverse relationship. That is, if one tries to eliminate a type I error by shifting a critical value outward (eliminating α), this will relatively increase the committing of a type II error (increasing β). Therefore, the only way to reduce both α and β is to increase the sample size. Hence the classical testing hypotheses use the comprised procedure in selecting the optimum decision. They set up a favorable prespecified value called the "significance level α," and select a left-hand, a right-hand, or two-hand tail test according to the different alternative hypotheses to minimize the value of a type II error.

But this still leaves the problem unsolved. Since the more null hypothesis is close to the alternative hypothesis, the type II error will be committed more often. When the value of the alternative hypothesis approaches the null hypothesis, the alternative hypothesis becomes ignored, although the probability of committing the type II error approaches probability 1. (See Table 1.) For example, let $H_0: \mu = 45$, $\sigma_{\bar{x}} = 3$, the probabilities of committing the type II error for various H_1's are shown in Table 1.

Table 1. Probabilities of committing the type II error for various H_1's

H_1	25	30	35	40	42	43	45	47	48	50	55	60	65
$P(a_1\|H_1)$	0.0000	0.0012	0.0853	0.6130	0.8300	0.8971	0.9500	0.8971	0.8300	0.6131	0.853	0.0012	0.0000

Let $H_0{:}\mu = 45$, $H_1{:}\mu = 47$, $\sigma_{\bar{x}} = 3$, the probabilities of committing type I error and type II error are shown in Figure 1.

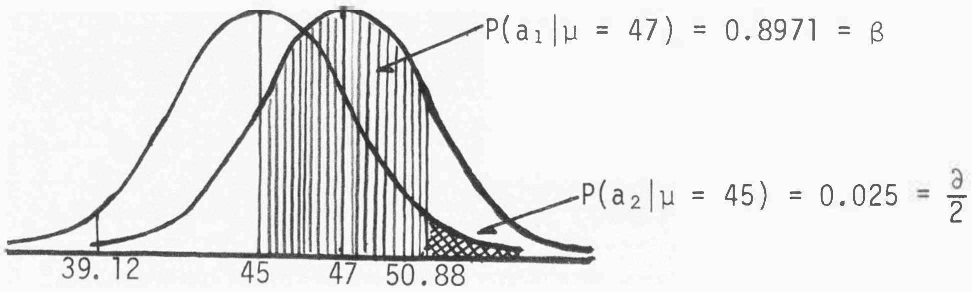


Figure 1. Probabilities of committing type I and type II errors.

This means that errors of these two kinds should be considered from not only the standpoint of the probability of occurrence but also from the standpoint of the relative consequence of loss or utility. If the statistician feels that the loss incurred from committing a type II error is larger than that from a type I error, he would like to decrease the probability of taking action a_1. In other words, the statistician should look at the probability of occurrence as well as the consequence of making a wrong decision.

Comparison of classical and Bayesian statistics

From the Bayesian point of view, classical statistics is commented upon as follows: The preassignment of null hypothesis is arbitrary. Moreover, the limiting of the analysis to only two numerical values for the states of nature (parameters) in order to get a unique α and a unique β is either arbitrary or even dangerous. Here they use only two possible actions: To accept or reject the hypothesis. There are only two possible states of nature: The null hypothesis or the alternative hypothesis. Indeed, there exist many possible states of nature. It avoids any probability distribution for the unknown parameter and attempts to arrive at the decision purely on the basis of the objective evidence. At this point, classical statistics treats the statistic of samples as the random variable, while Bayesian statistics treats the parameter itself as a random variable. It attaches to the values of parameter its probability, and revises this random variable when additional information is obtained. There are various treatments of this random variable such as uniform, binomial, normal, β distributions, etc. It depends on the different types of phenomena. If the random variable fits with a uniform (prior) probability function, then the Bayesian inference is close to the classical inference. This means that posterior probabilities can be calculated from sample evidence alone. This is why some of the Bayesian statisticians accuse the classical school of implicitly assuming the uniform prior function in its analysis, even when prior information might be available. Hence, when Bayesian analysis assumes that the prior probability is uniform, the numerical result will be the same as that of the classical approach, although the interpretation of the results is somewhat different. A Bayesian decision is to establish the "optimum" or the "best" action on the basis of all available information while some other possible decision often ignores some information (see Table 5).

The following are some important relations between the prior evidence and additional information:

1. The greater the amount of additional information obtained, the less is the uncertainty.

2. If the prior evidence is taken into account, the size of

sample necessary to achieve a given relative degree of certainty will be smaller.

3. The greater the cost of acquiring sample size, the greater the importance of this prior evidence.[7]

[7] Bruce W. Morgan, *An Introduction to Bayesian Statistical Decision Processes* (Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1968), p. 3.

CHAPTER II

BAYESIAN DECISION THEORY

The objective of the Bayesian inference, like that of classical inference, is to establish an optimal decision under uncertainty. In the introduction, we talked about the basic difference between the classical and Bayesian inferences. Bayesian inference is a revolutionary movement forward. Also, it is a movement backward, since it comes back to an approach ignored by the statisticians for centuries and makes use of Bayes' theorem.

Bayesian decision theory is a mathematical structure formulated for the statistician in choosing a course of action under uncertainty. Before we mention the decision rules, some probability theorems might be reviewed:

Conditional probability and the Bayes' theorem

Theorem 1. If $P(B) > 0$, then

(a) $P(A|B) > 0$.

(b) $P(\Omega|B) = 1$ where Ω is an arbitary fundamental probability set.

(c) $P(\sum_{k}^{\infty} A_k|B) = \sum_{k}^{\infty} P(A_k|B)$ for $A_i \cap A_j = \emptyset$ where $i \neq j$.

Theorem 2. If $P(A_o A_1 \ldots A_{n-1}) > 0$, then

$P(A_o A_1 \ldots A_n) = P(A_o)P(A_1|A_o)P(A_2|A_o A_1) \ldots P(A_n|A_o A_1 \ldots A_{n-1})$.

Theorem 3. If $P(\sum_{n}^{N} H_n) = 1$ and $P(H_n) > 0$, then

$P(A) = \sum_{n}^{N} P(A|H_n)P(H_n)$.

Theorem 4. If $P(\sum_{n}^{N} H_n) = 1$, $P(A) > 0$ and $P(H_n) > 0$ for every n, then

$$P(H_j|A) = \frac{P(A|H_j)P(H_j)}{\sum_{n}^{N} P(A|H_n)P(H_n)} = \frac{P(A|H_j)P(H_j)}{P(A)} \qquad (2.1)$$

This theorem is called Bayes' Theorem.[8]

Some assumptions, definitions, and theorems in Bayesian decision theory

If there exist some decision rules:

Assumption 1. The statistician will be able to decide whether he prefers action a_1 to action a_2, or if he prefers action a_2 to action a_1, or both of the actions are equivalent.

Assumption 2. If action a_1 is preferred to action a_2, and action a_2 is preferred to action a_3, then action a_1 is preferred to action a_3.

Assumption 3. If action a_1 is preferred to action a_2, which in turn is preferred to action a_3, then there is a mixture of action a_1 and action a_3 which is preferred to action a_2, and there is a mixture of action a_1 and action a_3, over which action a_2 is preferred.

Assumption 4. If the statistician prefers action a_1 to action a_2, and action a_3 is another action, then we assume that he will prefer a mixture of action a_1 and action a_3 to the same mixture of action a_2 and action a_3.

The statistician can also express his preference for consequence by a real-value function $U(a_i)$, called utility function, such that $U(a_1) > U(a_2)$ if, and only if, action a_1 is preferred to action a_2.

Further, if the statistician faces action a_1 with probability p and action a_2 with probability (1 - p), then

[8] Howard G. Tucker, *An Introduction to Probability and Mathematical Statistics* (New York: Academic Press, Inc., 1962), p. 15-17.

$$U(a) = pU(a_1) = (1 - p)U(a_2). \qquad (2.2)$$

Most parts of the payoff matrices in Bayesian statistics are expressed in terms of monetary value; but the monetary value is not a good measure of a gain or a loss, because the value of money to the individual varies from one person to another.

For example:

(1) Player A receives \$2 if a fair coin falls heads and player B pays \$1 if it falls tail.

(2) Player A has an entire fortune of \$100,000 cash, player A receives \$200,000 extra if the coin falls heads and player A loses his fortune otherwise.

In situations (1) and (2) the odds favored player A two to one. But our reactions to these situations would be different. In situation (1), the chance to win is one-half, the amount to be gained is twice as much as the amount to be lost. In situation (2), this is also true; but the winning of \$200,000 would increase our happiness very little while the loss of our \$100,000 would lead to considerable misery. Hence in situation (1) we would like to bet, but we would not in situation (2). This example indicates the value of money to the individual is not proportional to the amount of money.[9]

The following are some essential elements in decision-making:

1. A space of possible actions available to the statistician $A = \{a_1, a_2, ... a_n\}$. One of these alternative actions is chosen upon the state of nature which is not known. These actions are sometimes referred to as "terminal" actions.

[9] H. Chermoff and L. E. Moses, *Elementary Decision Theory* (New York: John Wiley & Sons, Inc., 1959), p. 70-89.

2. A space of possible state of nature $\theta = \{\theta_1, \theta_2, \ldots \theta_m\}$. The state of nature summarizes those aspects of the world that are relevant to the decision problem and about which the statistician is not certain. Nature exists in exactly one, and only one, of these states, $\theta_i \in \theta$.

3. The loss matrix or utility table measures the consequence of taking actions in monetary or other terms, while their corresponding states of nature are $\{\theta_1, \theta_2, \ldots \theta_m\}$, respectively.

4. A set of possible experiments, $E = \{e_1, e_2, \ldots e_\ell\}$. The statistician can use one of these experiments to obtain information about the state of nature. E includes making decisions with experimentation or with no experiments in E.

5. A space of possible outcome $X = \{x_1, x_2, \ldots x_i, \ldots\}$ for the experiments in E. Each combination $(a, \theta, e, x) \in A \times \theta \times E \times X$ determines a consequence for the statistician.

The statistician can express his judgments about the relative likelihood of the states of nature and the experimental outcome by measures of a probability function $P(\theta, x)$ on $\theta \times X$. From $P(\theta, x)$, we can obtain the marginal probability function $P(\theta)$ on θ, called the prior probability function of the state of nature. If experiment e results in an outcome x, the statistician's prior evidence is revised by Bayes' theorem to get the posterior probability function of the states of nature $P(\theta|x)$, i.e.,

$$P(\theta_j|x) = \frac{P(\theta_j)P(x|\theta_j)}{\sum_i P(\theta_i)P(x|\theta_i)} \tag{2.3}$$

in a discrete case, and

$$f(\theta|x) = \frac{f(\theta)l(x|\theta)}{\int_\theta f(\theta)l(x|\theta)d\theta} \quad (2.4)$$

in a continuous case. $l(x|\theta)$ is called the likelihood function: the conditional distribution of the outcome x, given that θ.

The sample outcome is a point in a multidimentional sample space, and we often could express the essential information of the sample in a space of fewer dimensions. Any function y(x) which maps the space of outcome x onto another space Y is called a statistic. A statistic is said to be sufficient if use of y in place of x does not affect the decision made by the statistician; that is, y(x) is a sufficient statistic if, for all $y_i \in Y$ and $x_i \in X$, $P[\theta|y(x)] = P(\theta|x)$. It is equivalent to the definition of a sufficient statistic in classical statistics, that y is a sufficient statistic if, and only if,

$$l(x|\theta) = k[y(x)|\theta]r(x) \quad (2.5)$$

where $k[y(x)|\theta]$ is a function of y and θ only, while r(x) is a function of x only.

A statistical decision problem is a special game (θ, A, L) combined with an experiment involving random observations $X = \{x_1, x_2, \ldots x_l\}$, whose distribution $P(X|\theta)$ depends on the state of nature $\theta_i \in \theta$.

On the basis of the possible outcome of a certain experiment $X = \{x_1, x_2, \ldots x_l\}$, the statistician chooses an action $d(x_1, x_2, \ldots x_l) \in A$. The function d which maps the sample space into the action space is called a decision function. The consequence in making a wrong decision is the random loss denoted by $L(\theta, d(X))$. The expectation of $L(\theta, d(X))$ when θ is the state of nature is called the risk function:

$$R(\theta, d) = E[L(\theta, d(X))] = \int L(\theta, d(X))dF(X|\theta). \quad (2.6)$$

When the true state of nature is not known, the statistician employs this expected risk function to make the decision.

Definition 1. If the risk function $R(\theta, d)$ is finite for all $\theta_i \in \theta$, any function $d(X)$ which maps the sample space into the action space A is called a nonrandomized decision function.

Suppose in action space A, the statistician leaves the choice of action to a random mechanism, such as to toss a fair coin to decide it. This decision is called a randomized decision and is denoted by δ. In game theory, δ would be called a mixed strategy, since this kind of strategy combining the original nonrandomized strategy with random mechanism, while the nonrandomized strategy d is called a pure strategy.

Definition 2. If the risk function $R(\theta, d)$ is finite for all θ_i θ, any probability distribution d on the space of nonrandomized decision functions D is called a randomized decision function (rules). The space of all randomized decision functions is denoted by D'.

The space D of non-randomized decision functions (rules) may be considered as a subset of the space D' of randomized decision functions. That is: $D \subset D'$. Hence in speaking of randomized decision functions (rules), we just say decision functions (rules), since it also contains the non-randomized functions (rules). Also, we use A as a nonrandomized action, while A' is referred to as randomized action.

The advantage in extending the definition from $L(\theta,a)$ to $L(\theta,\alpha)$ and the definition from $R(\theta, D)$ to $R(\theta, D')$ is that these functions (rules) become linear on α and D', respectively. That is, if

$\alpha_1 \in A'$, also $\alpha_2 \in A'$, and $0 \leq p \leq 1$,

then $p\alpha_1 + (1 - p)\alpha_2 \in A'$ and

$L(\theta, p\alpha_1 + (1 - p)\alpha_2) = pL(\theta,\alpha_1) + (1 - p)\, L(\theta,\alpha_2)$.

Also, if $\delta_1 \in D'$, $\delta_2 \in D'$ and $0 \leq p \leq 1$ then

$p\delta_1 + (1 - p)\delta_2 \in D'$ and

$R[\theta, p\delta_1 + (1 - p)\delta_2] = pR(\theta, \delta_1) + (1 - p)R(\theta, \delta_2)$.

Optimal decision rules

The decision theory is designed to provide a "good" decision if the statistician is given the states of nature, actions, and the payoff (loss function); i.e., (θ, A, L), and a random variable X which distributes on $\theta_i \in \theta$, then what decision rule δ should be the best one. The best decision rule undoubtedly should have the smallest risk for every state of nature in θ. Usually in only a few cases does such a best decision rule exist. In all other cases, the best decision rule to the state of nature θ_i is not the best decision rule to the state of nature θ_j, where i is not equal to j, since a uniformly best decision rule usually does not exist.

Bayesian decision rule

The statistician may set up some principles (criteria) in selecting a decision rule. The most important and useful decision principle is the Bayesian decision rule.

The Bayesian decision rule involves the concept of the prior distribution. The following conditions are needed:

1. Bayesian risk of a decision rule δ corresponding to a prior distribution t,

$$r(t, \delta) = E[R(T, \delta)]. \tag{2.7}$$

T denotes a random variable over the parameter space θ having the distribution t.

2. The posterior distribution of the parameter, given the sample observations.

It is clear that with the definition of expectation, any finite distribution t on the parameter space θ satisfies these two conditions.

For specific purposes, we use θ' as distribution t on θ that satisfies the above-mentioned two conditions. In addition, θ' is a set of finite distribution on θ; i.e., the states of nature are finite and θ' is linear. As we have mentioned before, in Bayesian inference the statistician looks at the parameter as a random variable whose distribution be previously known. Given a certain distribution, the statistician prefers a decision rule δ_i to another decision rule δ_j if the former has a smaller risk. Hence we might say that the Bayesian decision rule is that which minimizes the expected losses.

Definition 1. A decision rule δ_o is said to be Bayesian with respect to the prior distribution $t \in \theta'$ if

$$r(t, \delta_o) = \inf_{\delta \in D'} r(t, \delta).^{10} \qquad (2.8)$$

The value on the right hand side of (2.8) is known as the minimum Bayesian risk.

Bayesian decision rules may not exist even if the minimum Bayesian risk is defined and finite for the same reason that a smallest positive number does not exist. In such a case the statistician uses the approximate which is close to minimum Bayesian risk.

Definition 2. Let $\varepsilon > 0$, a decision rule δ_o is said to be ε-Bayes

[10] Let S be a set of numbers. A lower bound for a Set S is a number W such that $W \leq X$ whenever $X \in S$. The greatest lower bound of S is a lower bound that is greater than all other lower bounds of S. Common abbreviation for "greatest lower bound of S" is inf(S). The abbreviation "inf" is derived from "infimum."

with respect to the prior distribution $t \in \theta'$ if

$$r(t, \delta_o) \leq \inf_{\delta \quad D'} r(t, \delta) + \varepsilon. \tag{2.9}$$

A set of risk functions constitutes a risk set. In other words, the risk set is

$S = \{R(\theta_1, \delta), R(\theta_2, \delta), \ldots\ldots\ldots, R(\theta_k, \delta)\}$,

where δ ranges through D'.

Theorem 1. The risk set is a convex set.

Suppose that θ is a finite state of nature that consists of k points, $\theta = \{\theta_1, \theta_2, \ldots \theta_k\}$, let the set S be a risk set in k-dimensional Euclidean space E_k

$$S = \{R(\theta_1, \delta), R(\theta_2, \delta), \ldots R(\theta_k, \delta)\} \tag{2.10}$$

where $\delta \in D'$

then a risk set must be convex.

Proof. A subset A of Euclidean k-dimensional space is said to be convex if whenever $Y = (y_1, y_2, \ldots, y_k)$ and $Y' = (y_1', y_2', \ldots, y_k')$ are elements of A, the points

$pY + (1 - p)Y' = [pY_1 + (1 - p)Y_1', \ldots, pY_k + (1 - p)Y_k'], 0 \leq p \leq 1$

are also elements of A.

Let Y and Y' be arbitrary points of the risk set S. Since $y_j = R(\theta_j, \delta_1)$ and $y_j' = R(\theta_j, \delta_2)$ where $j = 1, 2, \ldots k$, $pR(\theta_j, \delta_1) + (1 - p)R(\theta_j, \delta_2) = R[\theta_j, p\delta_1 + (1 - p)\delta_2] = R(\theta_j, \delta_c)$, $\delta_c \in D'$ if $R(\theta_j, \theta_c)$ is denoted by z, then $z = [pY_j + (1 - p)Y_j'] \in S$. Further S is the convex hull, the smallest convex set containing S_o which is the non-randomized risk set, where

$$S_o = \{R(\theta_1, d), R(\theta_2, d), \ldots R(\theta_k, d)\} \tag{2.11}$$

where $d \in D$.

Since the risk function contains all the information about a decision rule, the risk set S contains all the information about the decision problem. For a given decision problem (θ, D', R), the risk set S is convex. Conversely, for any convex set S in E_k, there is a decision problem.

A prior distribution for k finite states of nature is merely a k-tuple of non-negative numbers $(p_1,p_2, \ldots p_k)$, such that $\Sigma p_i = 1$. P_i is the prior probability with respect to the specific state of nature θ_i.

The expectation of the risk is $p_i R(\theta_i, \delta) = b$, where b is any real number. One advantage that Bayesian approach has over the minimax approach to decision theory is that in the Bayesian case, "good" decision rules are restricted to the class of nonrandomized decision rules.

Suppose $\delta_o \in D'$ is Bayesian with respect to a distribution t over θ, let X denote the random variable with value in D whose distribution is given by δ_o, then $r(t, \delta_o) = E[r(t, X)]$, but δ_o is Bayesian with respect to t, $r(t, \delta_o) \leq r(t, d)$ for all $d \in D$. This entails $r(t, X) = r(t, \delta_o)$ with probability 1. So that any $d \in D$ that X chooses with $p = 1$, satisfies the equality $r(t, d) = r(t, \delta_o)$, implying that d is Bayesian with respect to t.

Given the prior distribution t, we want to choose a nonrandomized decision rule $d \in D$ that minimizes the Bayesian risk

$r(t, d) = \int R(\theta, d)P(\theta)d\theta$

where $R(\theta, d)$ is the risk function

$R(\theta, d) = \int L(\theta, d(x))f(x|\theta)dx.$

The joint distribution of θ and x is

$h(\theta, x) = P(\theta)f(x|\theta)$

$k(x) = \int_\theta h(\theta, x)d\theta.$

Choosing θ according to the conditional distribution of θ, given $X = x$

$$\begin{aligned} r(t, d) &= \int R(\theta, d)P(\theta)d\theta \\ &= \int\int L(\theta, d(x))f(x|\theta)dxp(\theta)d\theta \\ &= \int\int L(\theta, d(x))p(\theta)f(x|\theta)dxd\theta \\ &= \int\int L(\theta, d(x))k(x)g(\theta|x)d\theta dx \\ &= \int[\int L(\theta, d(x))g(\theta|x)d\theta]k(x)dx. \qquad (2.12) \end{aligned}$$

The decision rule is to find a function $d(x)$ that minimizes the double integral (2.12). First, we may find the value $d(x)$ inside the parenthesis for each x that minimizes the value of the bracket; i.e.,

$$\int L(\theta, d(x))g(\theta|x)d\theta \qquad (2.13)$$

That is to say that Bayesian decision rule minimizes the posterior conditional expected loss, given the observation. When the infimum of (2.13) does not exist, we may find a decision rule $d(x)$ within ε. Then we call it ε-Bayes.

Decision rule and loss function

For simplification, we use a linear loss junction to estimate the real parameter. But using a quadratic loss junction $L(\theta,a) = C(\theta)$ $(\theta - a)^2$ where $C(\theta) > 0$, to estimate the real parameter is more frequent. This function implies that as the loss increases, the further $\hat{\theta}$ is from the true state of nature θ. It may be quite difficult to find a $C(\theta)$. But experience has indicated that $c(\theta)$ plays a minor role in determining the decision rules. If we let $c(\theta) = 1$, then the loss function is called a squared-error loss function. The posterior expected loss, given $X = x$, for a squared-error loss function is as follows:

$$E[L(\theta, a) \mid X = x] = \int c(\theta)(\theta - a)^2 g(\theta|x)d\theta. \qquad (2.14)$$

This quantity is minimized by taking $a = E(\theta)$. Hence the Bayesian decision rule is simply

$$d(x) = E[\theta | X = x]. \tag{2.15}$$

This generates the following general rules:

1. Given a certain prior distribution for θ, with quadratic loss function (squared error loss), the Bayesian estimation of the true state of nature (parameter) θ is the expectation of the posterior distribution of θ, given the observation X. A greater generalization is the weighted squared loss

$L(\theta, a) = w(\theta_i)(\theta_i - a)^2$

where $w(\theta_i) > 0$, for all $\theta_i \in \theta$.

Then the Bayesian decision rule (function) is

$$d(x) = \frac{E[\theta w(\theta) \mid X = x]}{E[w(\theta) \mid X = x]} = \frac{\int \theta w(\theta) g(\theta \mid x) d\theta}{\int w(\theta) g(\theta \mid x) d\theta} \tag{2.16}$$

Another loss function is the absolute error loss

$L(\theta, a) = c(\theta) \mid \theta - a \mid$. For a given observed value X = x, the Bayesian decision rule d(x) is the action a that minimizes

$$E[L(\theta, a) \mid X = x] = \int c(\theta) \mid \theta - a \mid g(\theta|x) d\theta. \tag{2.17}$$

This quantity is minimized by taking $a = Me(\theta)$, given X = x. This generates the second rule:

2. Given a certain prior distribution for θ with absolute error, the Bayesian estimation of a true state of nature (parameter) is the median of the posterior distribution of θ, given the observation X.

It is difficult to specify a loss function. But in most statistical problems with a reasonable amount of sample size, small variations in loss function on the decisions selected are negligible. However, gross variations in loss function should be avoided.

Extensions to the Bayesian decision rule[11]

There exist some extensions to the concept of the Bayesian decision rule.

Definition 1. A decision rule δ is said to be a limit of the Bayesian decision rule δ_n, if for almost all x, $\delta_n(x) \to \delta(x)$, for non-randomized decision rules $d_n \to d$ if $d_n(x) \to d(x)$ for almost all x.

For example: Let the distribution of x, given θ, be normal with mean θ and unity variance; i.e., $X \sim N(\theta, 1)$, and the prior distribution t is normal with mean zero and variance σ^2. The joint distribution of x and θ has density

$$h(\theta, x) = \frac{1}{2\pi\sigma} e^{-\frac{1}{2}[(x-\theta)^2 + \frac{\theta^2}{\sigma^2}]}.$$

The marginal density of x is therefore

$$f(x) = \frac{1}{\sqrt{2\pi(1+\sigma^2)}} e^{-\frac{x^2}{2(1+\sigma^2)}}$$

and the posterior density of θ given X = x is

$$g(\theta|x) = \left(\frac{1+\sigma^2}{2\pi\sigma^2}\right)^{\frac{1}{2}} e^{-\frac{(1+\sigma^2)}{2\sigma^2}(\theta - \frac{x\sigma^2}{1+\sigma^2})^2}$$

normal with mean $\frac{\sigma^2 x}{(1+\sigma^2)}$ and variance $\frac{\sigma^2}{(1+\sigma^2)}$.

According to (2.15), $d(x) = E[\theta|X = x]$, we know that the Bayesian decision rule with respect to t is

$$d_\sigma(x) = \frac{x\sigma^2}{1+\sigma^2}$$

$$\lim_{\sigma\to\infty} d_\sigma(x) = \lim_{\sigma\to\infty} \frac{x\sigma^2}{1+\sigma^2} = x = d(x)$$

so that d is a limit of Bayesian decision rules.

Definition 2. The decision rule δ_o is said to be a generalized Bayesian rule if there exists a measure t on θ such that

[11] Thomas S. Ferguson, *Mathematical Statistics: A Decision Theoretic Approach* (New York: Academic Press, Inc., 1967), p. 47-49.

$L(\theta, \delta)f(x \mid \theta)dt(\theta)$

takes on a finite minimum value when $\delta = \delta_0$.

For example: The posterior distribution of θ

$$f(\theta \mid x)d\theta = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\theta - x)^2}$$

with mean x and varience unity; i.e., $\theta \sim N(x, 1)$. The generalized Bayesian decision rule is therefore $d(x) = x$.

<u>Definition 3.</u> A decision rule δ_0 is said to be an extended Bayesian rule if δ_0 is ε-Bayes for every $\varepsilon > 0$.

For example:

$$r(t_\sigma, d) = E(\theta - X)^2 = E[E\,(\theta - X)^2 \mid \theta] = 1$$

$$\text{but } \inf_\sigma r(t_\sigma, \delta) = r(t_\sigma, d_\sigma) = \frac{\sigma^2}{1 + \sigma^2}$$

$$r(t_\sigma, d) = \inf_\sigma r(t_\sigma, \delta) + \varepsilon \text{ for } \varepsilon = \frac{1}{1 + \sigma^2}.$$

<u>Minimax decision rule</u>

One different approach in decision problems is the minimax decision principle. The rule is to select the action for which the maximum amount which can be lost is minimized. It involves a decision rule δ_1 preferred to a rule δ_2 if

$\sup_{\theta_i \in \Theta} R(\theta_i, \theta_1) < \sup_{\theta_i \in \Theta} R(\theta_i, \theta_2)$.[12]

<u>Definition 1.</u> A decision rule $_o$ is said to be minimax if

$$\sup_{\theta_i \in \Theta} R(\theta_i, \delta_o) = \inf_{\theta \in D'} \sup_{\theta_i \in \Theta} R(\theta_i, \delta). \tag{2.18}$$

The value on the right side of (2.18) is known as the minimax value.

[12] Let S be a set of numbers. An upper bound for S is a number W such that $W \geq X$ whenever $X \in S$. The least upper bound of S is an upper bound that is less than all other upper bounds of S. Common abbreviation for "least upper bound of S" is Sup(S) which is derived from "Supremum."

Minimax rules and the Bayesian decision rule

When are minimax rules also Bayesian rules with respect to some prior distribution? The answer is if

(a) $\sup_{t\in\theta'} \inf_{\delta\in D'} r(t, \delta) = \inf_{\delta\in D'} \sup_{t\in\theta'} r(t, \delta)$ and if

(b) $\inf_{\delta\in D'} r(t_0, \delta) = \sup_{t\in\theta'} \inf_{\delta\in D'} r(t, \delta)$.

A least favorable distribution t_0 exists then any minimax rule δ_0 is Bayes with respect to t_0.

Proof: Since $\sup_{t\in\theta} r(t, \delta) = \sup_{\theta_i\in\theta} R(\theta_i, \delta)$

and δ_0 is said to be minimax if

$$\sup_{\theta_i\in\theta} R(\theta_i, \delta_0) = \inf_{\delta\in D} \sup_{\theta_i\in\theta} R(\theta_i, \delta)$$

hence $\sup_{t\in\theta'} r(t, \delta_0) = \inf_{\delta\in D'} \sup_{t\in\theta'} r(t, \delta)$

$$\sup_{t\in\theta'} r(t, \delta_0) = \sup_{t\in\theta'} \inf_{\delta\in D'} r(t, \delta)$$

$$r(t, \delta_0) = \inf_{\delta\in D'} r(t, \delta).$$

Admissible decision rule

Decision rules which are not dominated are called admissible.

Definition 1. A decision rule δ_1 is said to be as good as δ_2 if $R(\theta_i, \theta_1) \leq R(\theta_i, \delta_2)$ for all $\theta_i \in \theta$. A decision rule δ_1 is said to be better than a rule δ_2 if

$R(\theta_i, \theta_1) \leq R(\theta_i, \delta_2)$ for all $\theta_i \in \theta$, and

$R(\theta_i, \delta_1) < R(\theta_i, \delta_2)$ for at least one $\theta_i \in \theta$. A rule δ_1 is said to be equivalent to a rule δ_2 if $R(\theta_i, \delta_1) = R(\theta_i, \delta_2)$ for all $\theta_i \in \theta$.

Definition 2. A decision rule δ is said to be admissible if there exists no rule better than δ. A rule is said to be inadmissible if it is not admissible.

The word "admissible" is a synonym for the word "optimal." In a given decision problem every rule may be inadmissible.

For example: when the risk set S does not contain its boundary points, there exists no admissible rule.

Complete class of decision rule

Decision rules which are in set C and not dominated by the decision rules which are not in set C, are called complete.

Definition 1. A set C of decision rules, $C \subset D'$ is said to be complete, if given any decision rule $\delta \in D'$ not in C, there exists a decision rule $\delta_o \in C$ that is better than δ.

Definition 2. A set of decision rules C is said to be essentially complete, if given any decision rule δ not in C, there exists a decision rule $\delta_o \in C$ that is as good as δ.

Definition 3. A set of decision rules C is said to be minimal complete, if no proper subset of C is complete.

Definition 4. A set of decision rules is said to be minimal essentially complete, if no proper subset of C is essentially complete.

It is not necessary that a minimal complete or a minimal essentially complete set exists. The concept of complete set is to simplify the decision rule by finding a small complete set in decision-making. A smallest set may not exist, but if it exists that would largely simplify the decision problem.

Likelihood ratio test and Bayesian[13] decision rule

A particular case in Bayesian decision rule can be included in

[13] Alexander M. Mood and Franklin A. Graybill, *Introduction to the Theory of Statistics* (New York: McGraw-Hill Book Co., Inc., 1963), p. 276-290.

classical likelihood ratio test. This particular case means the space of possible states of nature θ is decomposed into two parts: $\theta = \{\theta_1, \theta_2\}$. Also the space of possible actions is decomposed into two parts: $A = \{a_1, a_2\}$. The appropriate action to take depends on the value of the unknown state of nature (parameter). The loss associated with the states of nature θ and action a_1 is denoted by $L(\theta, a_1)$, where $L(\theta, a_1) \geq 0$.

Let $X = \{x_1, x_2, \ldots x_n\}$ be a random space from $f(x|\theta)$, and let S be the n-dimensional sample space which can be partitioned into two disjoint sets S_1 and S_2. A decision rule is a function d which assigns an action of A to each possible sample; i.e.,

$a = d(x_1, x_2, \ldots x_n)$.

The risk corresponding to decision rule d is given by:

$$\begin{aligned}
R(\theta, d) &= \int\int\ldots\int_S L[\theta, d(x_1,x_2,\ldots x_n)]f(x_1|\theta)\ldots f(x_n|\theta)dx_1dx_2\ldots dx_n \\
&= \int\int\ldots\int_{S_1} L(\theta,a_1)f(x_1|\theta)\ldots f(x_n|\theta)dx_1dx_2\ldots.dx_n \\
&+ \int\int\ldots\int_{S_2} L(\theta,a_2)f(x_1|\theta)\ldots f(x_n|\theta)dx_1dx_2\ldots dx_n \\
&= L(\theta,a_1)\int\int\ldots\int_{S_1} f(x_1|\theta)\ldots f(x_n|\theta)dx_1dx_2\ldots.dx_n \\
&+ L(\theta,a_2)\int\int\ldots\int_{S_2} f(x_1|\theta)\ldots f(x_n|\theta)dx_1dx_2\ldots dx_n \\
&= L(\theta,a_1)P(S_1|\theta) + L(\theta,a_2)P(S_2|\theta). \qquad (2.19)
\end{aligned}$$

If we assume $\theta = \theta_1$, then the above equation (2.19) is as follows:

$$\begin{aligned}
R(\theta_1,d) &= L(\theta_1,a_1)P(S_1|\theta_1) + L(\theta_1,a_2)P(S_2|\theta_1) \\
&= L(\theta_1,a_2)P(S_2|\theta_1) \\
&= L(\theta_1,a_2)P(I), \text{ where } L(\theta_1,a_1) = 0.
\end{aligned}$$

Similarly: If $\theta = \theta_2$;

$$\begin{aligned}
R(\theta_2,d) &= L(\theta_2,a_1)P(S_1|\theta_2) + L(\theta_2,a_2)P(S_2|\theta_2) \\
&= L(\theta_2,a_1)P(S_1|\theta_2).
\end{aligned}$$

The expected risk is:

$$\begin{aligned} r(\theta,d) &= P(\theta_1)R(\theta_1,d) + P(\theta_2)R(\theta_2,d) \\ &= P(\theta_1)L(\theta_1,a_2)P(I) + P(\theta_2)L(\theta_2,a_1)P(II) \\ &= P(\theta_1)L(\theta_1,a_2)[1 - \int\int\ldots\int_{S_1} f(x_1|\theta_1)\ldots f(x_n|\theta_1)]dx_1dx_2..dx_n \\ &+ P(\theta_2)L(\theta_2,a_1)[\int\int\ldots\int_{S_1} f(x_1|\theta_2)\ldots f(x_n|\theta_2)]dx_1dx_2..dx_n \\ &= P(\theta_1)L(\theta_1,a_2) + \int\int\ldots\int_{S_1}[- P(\theta_1)L(\theta_1,a_2)\pi f(x_i|\theta_1) \\ &+ P(\theta_2)L(\theta_2,a_1)\pi f(x_i|\theta_2)]dx_1dx_2\ldots dx_n. \qquad (2.20) \end{aligned}$$

Since Bayesian decision rule is a decision rule which minimizes:

$r(\theta,d) = E[R(\theta,d)] = P(\theta_1)R(\theta_1,d) + P(\theta_2)R(\theta_2,d)$, as defined previously.

This can be done by letting the value of bracket in (2.20) be negative:

$$-P(\theta_1)L(\theta_1,a_2)\prod_{i=1}^{n} f(x_i|\theta_1) + P(\theta_2)L(\theta_2,a_1)\prod_{i=1}^{n} f(x_i|\theta_2) < 0.$$

That is:

$$P(\theta_2)L(\theta_2,a_1)\prod_{i=1}^{n} f(x_i|\theta_2) < P(\theta_1)L(\theta_1,a_2)\prod_{i=1}^{n} f(x_i|\theta_1).$$

Taking action 1, if

$$\frac{\pi f(x_i|\theta_1)}{\pi f(x_i|\theta_2} > \frac{P(\theta_2)L(\theta_2,a_1)}{P(\theta_1)L(\theta_1,a_2)} = k.$$

Taking action 2, if

$$\frac{\pi f(x_i|\theta_1)}{\pi f(x_i|\theta_2)} < k.$$

Taking either action, if

$$\frac{\pi f(x_i|\theta_1)}{\pi f(x_i|\theta_2)} = k.$$

Bayesian inference for decision-making is a generalization of classical inference. But this doesn't mean that there is no role for classical statistical inference and that all statistical inference can be solved by Bayesian decision theory.

The main difficulties in applying the method of decision theory are:

1. The statistician has difficulty in obtaining sufficient information for knowing the prior probability, or difficulty in calculating appropriate payoff.

2. Most two-tail tests are not action oriented and it is difficult to give them a Bayesian interpretation.

The convexity and decision-making

We know that the risk set is a convex set. Now we will further discuss how we apply the convexity to decision-making.

For example: A coin is tossed once to test the state of nature of falling heads as either $\theta_1 = 0.5$ or $\theta_2 = 0.3$. Two actions are action a_1 accepting $\theta_1 = 0.5$ and action a_2 accepting $\theta_2 = 0.3$. The loss matrix is as follows (see Table 2):

Table 2. Loss table for coin tossing

	a_1	a_2
θ_1	0	1
θ_2	2	0

A coin is allowed to be tossed only once. The sample space contains only two points--heads and tails. There are four possible decision rules. These are:

d_1: $d_1(H) = a_1$ $d_1(T) = a_1$

d_2: $d_2(H) = a_1$ $d_2(T) = a_2$

d_3: $d_3(H) = a_2$ $d_3(T) = a_1$

d_4: $d_4(H) = a_2$ $d_4(T) = a_2$

Where H indicates the toss is heads, T indicates tails. Decision rule d_1 means that action a_1 (accepting $\theta_1 = 0.5$) is taken regardless of whether the toss of the coin is heads or tails. Decision rule d_2 means that action a_1 (accepting $\theta_1 = 0.5$) is taken if the toss of the coin is heads, action a_2 is taken (accepting $\theta_2 = 0.3$) if the toss of the coin is tails. The error probabilities for d_2 and d_3 are calculated as follows:

$$d_2: \quad P(I) = P(a_2|\theta_1 = 0.5) = \binom{1}{0}\cdot(0.5)^0\cdot(0.5)^1 = 0.5$$

$$P(II) = P(a_1|\theta_2 = 0.3) = \binom{1}{1}\cdot(0.3)^1\cdot(0.7)^0 = 0.3$$

$$d_3: \quad P(I) = P(a_1|\theta_1 = 0.5) = \binom{1}{1}\cdot(0.5)^1\cdot(0.5)^0 = 0.5$$

$$P(II) = P(a_2|\theta_2 = 0.3) = \binom{1}{0}\cdot(0.3)^0\cdot(0.7)^1 = 0.7$$

The corresponding risk function for d_2 and d_3 are calculated as follows:

$$R(\theta_1,d_2) = L(\theta_1,a_1)P(a_1|\theta_1 = 0.5) + L(\theta_1,a_2)P(a_2|\theta_1 = 0.5)$$
$$= 0 + (1)\cdot(0.5) = 0.5$$

$$R(\theta_2,d_2) = L(\theta_2,a_1)P(a_1|\theta_2 = 0.3) + L(\theta_2,a_2)P(a_2|\theta_2 = 0.3)$$
$$= (2)\cdot(0.3) + 0 = 0.6$$

$$R(\theta_1,d_3) = L(\theta_1,a_1)P(a_1|\theta_1 = 0.5) + L(\theta_1,a_2)P(a_2|\theta_1 = 0.5)$$
$$= 0 + (1)\cdot(0.5) = 0.5$$

$$R(\theta_2,d_3) = L(\theta_2,a_1)P(a_1|\theta_2 = 0.3) + L(\theta_2,a_2)P(a_2|\theta_2 = 0.3)$$
$$= (2)\cdot(0.7) + 0 = 1.4$$

The risk functions are given in Table 3 and Figure 2.

Obviously, the risk set is a convex set. From this convex set, we find that d_2 is preferred over d_3. Since $R(\theta_i,d\ ) \le R(\theta_i,d_3)$ for all $\theta_i \in \theta$ and $R(\theta_i,d_2) < R(\theta_i,d_3)$ for $\theta_i = \theta_2$.

Table 3. Risk function for coin tossing

	Risk functions	
	$R(\theta_1,d)$	$R(\theta_2,d)$
d_1	0.	2.
d_2	0.5	0.6
d_3	0.5	1.4
d_4	1.0	0.

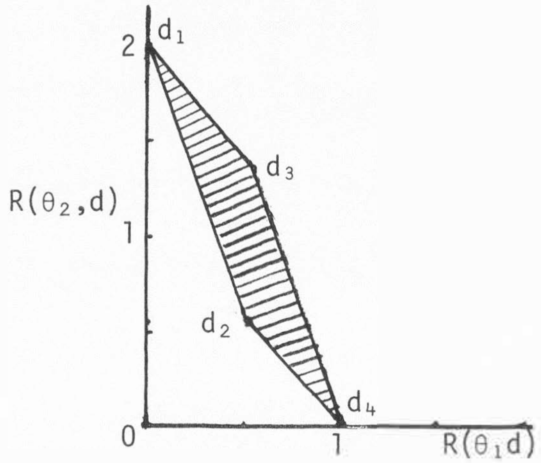


Figure 2. Risk set for coin tossing.

Hence we would discard d_3 as a possible decision rule. We also see that d_1 is better than d_2 if θ_1 is the true state of nature; d_4 is better than d_2 if θ_2 is the true state of nature. It is clear from Figure 3 that, of all the decision rules, the only ones entitled to serious consideration are d_1, d_2, and d_4. Thus the lower boundary of convex set constitutes the admissible decision rules. The Bayesian decision rule is to use the prior probabilities to find the optimal solution from these admissible decision rules. If we assume that $P(\theta_1) = \frac{1}{4}$ and $P(\theta_2) = \frac{3}{4}$, the Bayesian decision rule corresponding to $P(\theta_1)$ and $P(\theta_2)$ can be represented geometrically by drawing the line $P(\theta_1)R(\theta_1,d) + P(\theta_2)R(\theta_2,d) = C$ and moving it parallel to itself by changing C until it touches the convex set. The point or points where it just touches the convex set is then the Bayesian solution. Let $C = \frac{1}{8}$, we get the line $\frac{1}{4}R(\theta_1,d) + \frac{3}{4}R(\theta_2,d) = \frac{1}{8}$. If we let $C = \frac{1}{4}$, we get another line $\frac{1}{4}R(\theta_1,d) + \frac{3}{4}R(\theta_2,d) = \frac{1}{4}$ which parallels the first line and touches the convex set at d_4. Thus d_4 is called the Bayesian solution. These are shown in Figure 3.[14]

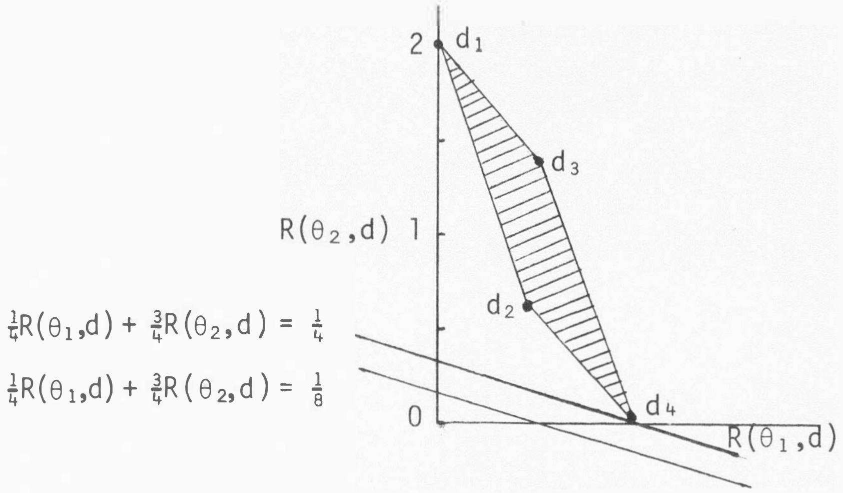

2
d_1
d_3
$R(\theta_2,d)$ 1
d_2
$\frac{1}{4}R(\theta_1,d) + \frac{3}{4}R(\theta_2,d) = \frac{1}{4}$
$\frac{1}{4}R(\theta_1,d) + \frac{3}{4}R(\theta_2,d) = \frac{1}{8}$
0
d_4
$R(\theta_1,d)$

Figure 3. Risk set and support lines for the Bayesian solution.

We make some important points as follows:

1. The Bayes' solution corresponding to prior probabilities $(P\theta_1)$ and $P(\theta_2)$ is to minimize the expected risk function.

2. Admissible solutions are easy to get. If we can identify the Bayes' solutions with admissible solutions, we can then restrict our search to the latter; now, it is a fact that any admissible solution is a Bayes' solution, which fact depends on convexity.

[14] *Ibid.*

3. Almost all Bayes solutions are admissible. Hence within the class of admissible solutions, we can hunt with confidence for the appropriate Bayes solution. This is a much easier task.

CHAPTER III

BAYESIAN DECISION PROCESSES

Classification of decision-making

Following the introduction and the Bayesian decision theory, we now refer to decision processes in applied statistics. These fall into two categories:

1. Bayesian decision processes without sampling.
2. Bayesian decision processes with sampling.

The essential components in decision problems are (θ,A,L): θ is the possible states of nature, $\theta = \{\theta_1,\theta_2,\ldots\theta_m\}$. A is the possible actions, $A = \{a_1,a_2,\ldots a_n\}$. L is a loss function (or loss table) which measures the consequence of taking actions $a_1,a_2,\ldots a_n$, respectively, when the states of nature are $\theta_1,\theta_2,\ldots\theta_m$, respectively.

In decision processes the statistician has some prior evidences, but he does not know which one of the possible states of nature is the true one. If the states of nature were known, it would be easy to select the optimal action.

Bayesian decision without sampling

A decision is made by the statistician without any additional information. In other words, no additional information on the states of nature is collected by sampling or performing an experiment.

There are three kinds of decision-making:

1. If a particular state of nature is sure to occur, this decision

process is called decision-making under certainty. Linear programming is decision-making under certainty.[15]

2. When a particular state of nature to occur is not sure, but there exists a distribution for the states of nature, this decision process is called decision-making under risk.

3. When no information about the states of nature is available, this decision process is called decision-making under uncertainty.

Now we explain these three kinds as follows:

1. Decision-making under certainty: Since the particular state of nature that will occur is certain, if the class of action is finite, there is no difficulty in finding an optimal action (decision).

For example:

$\theta = \{\theta_1, \theta_2\}$ and $A = \{a_1, a_2, a_3\}$.

Suppose the loss function (negative of utility) is given by Table 4.

Table 4. Loss table for decision-making under certainty

	a_1	a_2	a_3
θ_1	5	1	4
θ_2	2	5	4

When the state of nature is known for certain to be θ_1; i.e., $P(\theta_1) = 1$, we will take action a_2 because this action will result in the minimum loss. Similarly, if the state of nature is known to be $P(\theta_2) = 1$, we will take action a_1.

[15] Kyohei Sasaki, *Statistics for Modern Business Decision Making* (Belmont, California: Wadsworth Publishing Company, Inc., 1968), p. 220-223.

If there exists an infinite number of strategies (actions) which constitute a convex set, we use the linear programming method to maximize profit (or minimize the loss).[16]

2. Decision-making under risk: A particular state of nature is not sure to occur, but the objective probability distribution of the states of nature is known. In this case we might calculate the expected loss for each strategy (action) and determine the optimal action.

For example: The probability distribution of the possible states of nature θ_1 and θ_2 are 0.3 and 0.7, respectively. The expectation for a_1, a_2, and a_3 are calculated as follows:

Suppose the payoff table was given in Table 4:

$$R(\theta,a_1) = P(\theta_1)L(\theta_1,a_1) + P(\theta_2)L(\theta_2,a_1)$$
$$= (0.3)\cdot 5 + (0.7)\cdot 2 = 2.9$$
$$R(\theta,a_2) = P(\theta_1)L(\theta_1,a_2) + P(\theta_2)L(\theta_2,a_\theta)$$
$$= (0.3)\cdot 1 + (0.7)\cdot 5 = 3.8$$
$$R(\theta,a_3) = P(\theta_1)L(\theta_1,a_3) + P(\theta_2)L(\theta_2,a_3)$$
$$= (0.3)\cdot 4 + (0.7)\cdot 4 = 4.0$$
$$R(\theta,a_1) < R(\theta,a_2) < R(\theta,a_3).$$

The risk for action a_1 is smaller than for any others. Hence action a_1 will be selected as the optimal action. One thing we should note is that in game theory the statistician would take action a_2 rather than action a_1 or action a_3.

3. Decision-making under uncertainty: Neither the true state of nature nor an objective probability about the states of nature are known.

[16] *Ibid.*

Three criteria are used to decide the optimal action:

a. Maximin criterion: This is one of the most conservative approaches. The payoff matrix is expressed in terms of profit (utility). The statistician selects the strategy (action) for which the minimum profit is as great as possible. In other words, maximizing the minimum profit.

b. Minimax criterion: This approach is the same as maximin except the payoff matrix is expressed in terms of loss. The statistician selects the strategy for which the maximum loss is as small as possible. In other words, minimizing the maximum loss.

c. Bayesian criterion: This method is identical to the decision-making under risk, except that it uses subjective probability with respect to the states of nature. Given the subjective prior probability for the states of nature, the statistician might calculate the expected loss and choose the strategy which minimizes the expected loss.

Bayesian decision with sampling

Given the states of nature θ, we assume a prior probability. This state of nature θ acts as though it were a random variable. If an experiment E was conducted, the outcome of this experiment X is then a sample. Using this sample, we are led to revise the probabilities of the states of nature. This revised probability is the conditional probability of the state of nature, given the result of the experiment, and is called a posterior probability. In the case of a Bayesian decision with sample, we might use this posterior probability together with the payoff matrix to derive the Bayesian strategies (actions).

There are a great many different types of theoretical distributions, each used to represent a specified state of nature, such as the uniform distribution, binominal distribution, Beta distribution, Poisson distribution, and normal distribution, etc.

The uniform distribution is much the easiest of these distributions but is also less useful from the Bayesian point of view. Roughly speaking, the uniform distribution is a probability function which specifies that every possible value of the random variable is equally possible within its interval. The uniform distribution is also a special case of the Beta distribution.

To use the uniform distribution as a prior distribution has been considered by the statistician to represent ignorance about the true value of a random variable. But in the case of the absence of any prior probability about the true value of a random variable, the assumption of the uniform prior distribution for all possible values will minimize the maximum error.

The greater the discrepancy between the prior distribution and the sample distribution, the greater the posterior variance of the random variable. The assumption of a uniform distribution minimizes the possibility of such discrepancy. One way of viewing this is in terms of the variance of a probability distribution; i.e.,
$\sigma^2 = \Sigma[\theta_i - E(\theta_i|X)]^2P(\theta_i|X)$.
The assignment of equally probable probabilities as prior beliefs minimizes the possibility of such a discrepancy.

The uniform prior probability brings the Bayesian inference close to the classical inference, although the interpretation of the results is different.

For example: The states of nature are assumed to be $\theta_1 = 0.05$, $\theta_2 = 0.10$, $\theta_3 = 0.20$, and $\theta_4 = 0.35$. Suppose the prior probabilities for these states of nature are $P(\theta_1 = 0.05) = \frac{1}{4}$, $P(\theta_2 = 0.10) = \frac{1}{4}$, $P(\theta_3 = 0.20) = \frac{1}{4}$ and $P(\theta_4 = 0.35) = \frac{1}{4}$. In other words, we are assuming the prior probability function is uniform. Also we assume that the states of nature are binominally distributed. Since we draw a sample of size 10 with replacement, from a state of nature of $\theta_1 = 0.05$, $\theta_2 = 0.10$, $\theta_3 = 0.20$, and $\theta_4 = 0.35$, respectively, the outcome x = 4 in this sample is shown in Table 5.

Table 5. Effect of uniform prior probabilities on the posterior probabilities

State of nature	Prior probability	Likelihood $P(x = 4 \mid \theta_i)$	Joint probability	Posterior probability $P(\theta = \theta_1 \mid x = 4)$	Relative likelihood
$\theta_1 = 0.05$	$\frac{1}{4}$	0.00101	0.00025	0.0026	0.0026
$\theta_2. = 0.10$	$\frac{1}{4}$	0.01116	0.00279	0.0288	0.0288
$\theta_3 = 0.20$	$\frac{1}{4}$	0.13763	0.03441	0.3552	0.3552
$\theta_4 = 0.35$	$\frac{1}{4}$	0.23767	0.05942	0.6134	0.6134

The last two columns of this table show the posterior probability and the relative likelihood. Even though the prior probabilities are not taken into account in calculating the relative likelihood, the numerical results between the last two columns of Table 5 are the same.

The prior probability is equally probable for all θ_i; this would be looked upon as calculating the posterior probability simply on the basis of sample information. This is why some of the Bayesian statisticians may accuse the classical statisticians of implicitly assuming the prior uniform distribution in all cases, even when some other prior distribution is available.

If the sample information is sufficiently large, we shall use the prior uniform distribution to approximate the posterior distribution. When the sample size is sufficiently large compared to the prior probability, the quantity of information obtained from sample (I_s) would overwhelm the quantity of information obtained from prior evidence (I_0). Hence we can obtain a good approximation of the exact posterior probability by assuming the uniform prior distribution. (See Figure 4.)

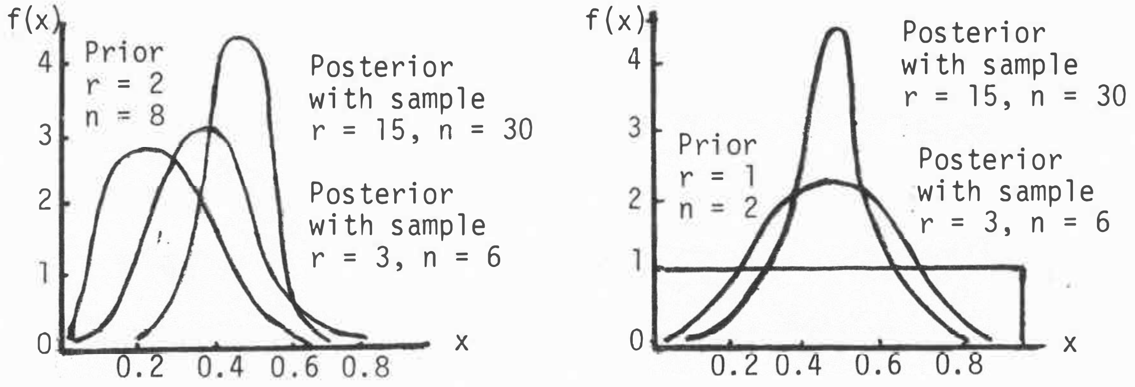


Figure 4. Comparison of uniform and β prior distributions with different sample sizes.[17]

Bayesian decision with binomial sampling

Suppose we have finite numbers of the states of nature, $\theta = \{\theta_1, \theta_2, \ldots \theta_m\}$, which are represented by the proportion of successes, subject to a certain prior probability density function. Assuming we draw a sample of size n from a binomial population, we can then combine this

[17] *Ibid.*, p. 336.

prior probability with a binomial sample to obtain the posterior probability function. This kind of Bayesian decision with binomial sample is widely applied in statistical quality control, marketing research, and production. The application will be illustrated in detail later (see Chapter IV).

Bayesian decision with normal sampling

In any probability function, the posterior probabilities are derived directly from the prior probabilities and the likelihoods. It could be shown by means of calculus that, if the prior probabilities and the likelihoods are both normally distributed, then the posterior distribution is normal. In normal distribution, we always use the mean and the variance to specify its probability function.

Let $\theta \sim N(\theta_0, \sigma_0^2)$ represent the prior normal distribution of the states of nature, and $X \sim N(\theta, \sigma^2)$ be the likelihoods, $\overline{X} \sim N(\theta, \frac{\sigma^2}{n})$, $X = \{x_1, x_2, \ldots x_n\}$, then the posterior function of the states of nature is also normally distributed with

$$\text{mean} = \frac{\sigma_0^2 \bar{x} + \sigma_{\bar{x}}^2 \theta_0}{\sigma_0^2 + \sigma_{\bar{x}}^2}, \text{ and variance} = \frac{\sigma_0^2 \sigma_{\bar{x}}^2}{\sigma_{\bar{x}}^2 + \sigma_0^2}$$

$$f(\bar{x}|\theta) = \frac{1}{\sqrt{2\pi\frac{\sigma^2}{n}}} e^{-\frac{1}{\frac{\sigma^2}{n}}(\bar{x} - \theta)^2}$$

$$P(\theta) = \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{1}{2\sigma_0^2}(\theta - \theta_0)^2}$$

$$h(\bar{x},\theta) = p(\theta)f(\bar{x}|\theta) = \frac{1}{2\pi\sigma_0\frac{\sigma}{\sqrt{n}}} e^{-\frac{(\theta_0 - \bar{x})^2}{2(\sigma_0^2 + \sigma^2/n)}} e^{-\frac{1}{2\left[\frac{\sigma_0 \cdot \sigma^2/n}{\sigma_0^2 + \sigma^2/n}\right]}\{\theta - \frac{\theta_0^2 \bar{x} + \sigma^2/n \cdot \theta_0}{\sigma_0^2 + \sigma^2/n}\}^2}$$

By the marginal distribution:

$$k(\bar{x}) = \int_{-\infty}^{\infty} h(\bar{x},\theta)d\theta = \int_{-\infty}^{\infty} \frac{1}{(2\pi)\sqrt{\sigma_o^2\sigma_o^2/n}} e^{-\frac{1}{2}\frac{(\theta - \bar{x})^2}{(\sigma_o^2 + \sigma^2/n)}} e^{-\frac{1}{2\left[\frac{\sigma^2/n\cdot\sigma_o^2}{\sigma_o^2 + \sigma^2/n}\right]}\{\theta - \frac{\sigma_o^2\bar{x} + \sigma^2/n\cdot\theta_o}{\sigma_o^2 + \sigma^2/n}\}^2} d\theta$$

$$= \frac{1}{\sqrt{2\pi(\sigma_o^2 + \sigma^2/n)}} e^{-\frac{1}{2}\frac{(\theta_o - \bar{x})^2}{(\sigma_o^2 + \sigma^2/n)}}$$

$$g(\theta|\bar{x}) = \frac{h(\bar{x},\theta)}{k(\bar{x})} = \frac{\frac{1}{2\pi\sqrt{\sigma_o^2\cdot\sigma^2/n}} e^{-\frac{1}{2}\frac{(\theta_o - \bar{x})^2}{(\sigma_o^2 + \sigma^2/n)}} e^{-\frac{1}{2\left[\frac{\sigma_o\cdot\sigma^2/n}{\sigma_o^2 + \sigma^2/n}\right]}\{\theta - \frac{\sigma_o^2\bar{x} + \sigma^2/n\cdot\theta_o}{\sigma_o^2 + \sigma^2/n}\}^2}}{\frac{1}{\sqrt{2\pi(\sigma_o^2 + \sigma^2/n)}} e^{-\frac{1}{2}\frac{(\theta_o - \bar{x})^2}{(\sigma_o^2 + \sigma^2/n)}}}$$

$$= \frac{1}{\sqrt{2\pi\frac{\sigma_o^2\sigma^2/n}{\sigma_o^2 + \sigma^2/n}}} e^{-\frac{1}{2\left[\frac{\sigma^2\cdot\sigma^2/n}{\sigma_o^2 + \sigma^2/n}\right]}\{\theta - \frac{\sigma_o^2\bar{x} + \sigma^2/n\cdot\theta_o}{\sigma_o^2 + \sigma^2/n}\}^2} \tag{3.1}$$

$$g \sim N\left(\frac{\sigma_o^2\bar{x} + \sigma_{\bar{x}}^2\theta_o}{\sigma_o^2 + \sigma_{\bar{x}}^2}, \frac{\sigma_o^2\cdot\frac{\sigma^2}{n}}{\sigma_o^2 + \frac{\sigma^2}{n}}\right)$$

$$E(g) = \frac{\sigma_o^2\bar{x} + \sigma_{\bar{x}}^2\theta_o}{\sigma_o^2 + \sigma_{\bar{x}}^2} = \frac{\bar{x}\cdot\frac{1}{\sigma_{\bar{x}}^2} + \theta_o\cdot\frac{1}{\sigma_o^2}}{\frac{1}{\sigma_o^2} + \frac{1}{\sigma_{\bar{x}}^2}} \tag{3.2}$$

$$V(g) = \frac{\sigma_o^2\cdot\frac{\sigma^2}{N}}{\sigma_o^2 + \frac{\sigma^2}{n}} = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\sigma^2}} \tag{3.3}$$

Let $I_1 = \frac{1}{\sigma_1^2}$, $I_o = \frac{1}{\sigma_o^2}$, $I_s = \frac{n}{\sigma^2}$

$$E(g) = \frac{\bar{x}\cdot I_s + \theta_o I_o}{I_o + I_s} \tag{3.4}$$

$$I_1 = I_o + I_s \tag{3.5}$$

This comes out to be a rather interesting concept called the quantity of information. $\sigma_o^2, \frac{\sigma^2}{n}$ and σ_1^2 represents the variance of prior probability, sample, and posterior probability, respectively. Its reciprocal $\frac{1}{\sigma_o^2}$, $\frac{n}{\sigma^2}$, and $\frac{1}{\sigma_1^2}$ might be looked at as the quantity of information. There exists this relation: the less the variance, the greater the quantity of information. The prior information available is then the sum of prior information and additional information (sample). Hence the posterior expected value might be interpreted as the weighted mean of prior population mean and sample mean where the weights are the information.

Further interpretation of the relation among the quantity of information for various estimates would be helpful for the concept of the quantity of information.

1. The greater the variance of the prior probability, the less the prior information; hence the total information comes mostly from the sample information.

2. Given the quantity of sample information, the total information would be increased if the prior information is increased.

3. As the prior information becomes increasingly small and tends to be zero, the more the prior probability tends to be the prior uniform probability function. Hence the prior uniform probability function is regarded as having the limit of obtaining no information from the prior probabilities.

Expected opportunity loss for the optimum action

Opportunity loss of a decision is the difference between the loss (or profit) realized by the decision and the loss (or profit) which would have been realized if the decision had been the best one possible for the true state of nature. The expected loss (expected opportunity loss) is calculated by multiplying the conditional expected loss for each possible outcome of θ by its corresponding probability. If the expected loss of the optimum action is great, we must try to decrease the loss by securing additional information before we make a final decision on the terminal action. If a sample can be obtained without cost, the expected loss of the optimum action can always be decreased by sampling additional information. Hence it is desirable to obtain an additional sample with no cost involved. But, in fact, cost is always involved in sampling procedure. The cost factor should therefore be taken into consideration.

Since the sample outcome is a random variable, there are many possible outcomes for any given sample size. Hence the expected loss (payoff) of the optimal action also becomes a random variable. This leads us to the analysis of the preposterior distribution to find out the optimal terminal action or the net gain from sampling. The preposterior analysis is illustrated in Chapter IV.

Comparison of normal prior distribution without sampling and with normal sampling for optimum action[18]

1. Normal prior distribution without sampling is discussed as follows:

If the profit (or loss) function is linear, a simple rule can be derived for selecting the optimum action. Suppose we have the following profit functions for taking action a_1 and action a_2:

For action a_1: $K_1 = A_1 + B_1\theta$ (3.6)

For action a_2: $K_2 = A_2 + B_2\theta$ (3.7)

where B_1 and B_2 are assumed to be positive values and θ is the population

[18] *Ibid.*, p. 366-370.

mean. The breakeven point for action a_1 and action a_2 can be obtained.

Let $K_1 = K_2$

$$A_1 + B_1\theta = A_2 + B_2\theta,$$

therefore,

$$\theta = \frac{A_2 - A_1}{B_1 - B_2} = \theta_b \qquad (3.8)$$

where θ_b is the breakeven point.

The expected profit for taking action a_1 is

$$E(K_1) = E(A_1) + E(B_1\theta)$$
$$= A_1 + B_1E(\theta). \qquad (3.9)$$

Similarly, the expected profit for taking action a_2 is

$$E(K_2) = E(A_2) + E(B_2\theta)$$
$$= A_2 + B_2E(\theta). \qquad (3.10)$$

a. Action a_1 is preferred over action a_2, if $E(K_1) > E(K_2)$. That is,

$$A_1 + B_1E(\theta) > A_2 + B_2E(\theta)$$

$$E(\theta)(B_1 - B_2) > A_2 - A_1.$$

Hence

$E(\theta) > \frac{A_2 - A_1}{B_1 - B_2} = \theta_b$ where

$B_1 - B_2 > 0$ (see Figure 5a) or

$E(\theta) < \frac{A_2 - A_1}{B_1 - B_2} - \theta_b$ where

$B_1 - B_2 < 0$ (see Figure 5b).

b. Action a_2 is preferred over action a_1 if $E(K_2) > E(K_1)$

$$A_2 + B_2E(\theta) > A_1 + B_1E(\theta)$$

$$E(\theta)(-B_1 + B_2) > -A_2 + A_1$$

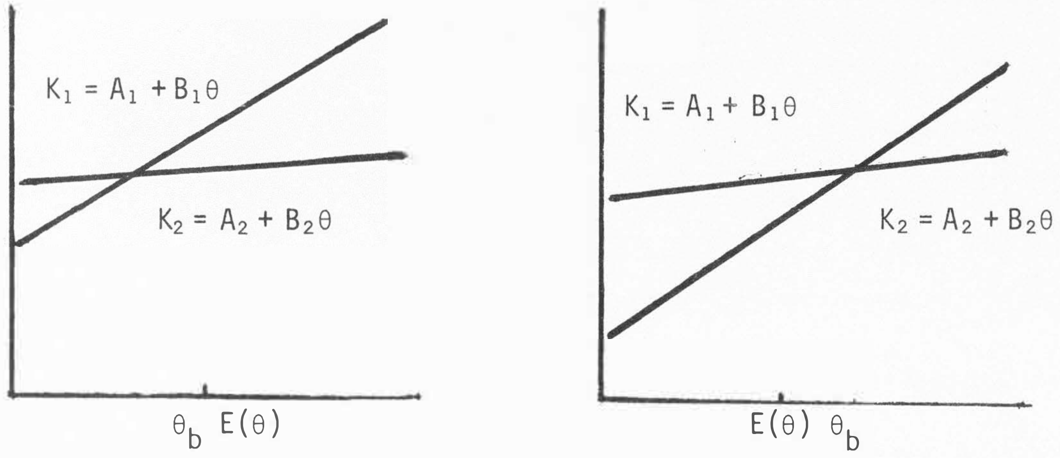
$$E(\theta)(B_1 - B_2) < A_2 - A_1$$

$$E(\theta) < \frac{A_2 - A_1}{B_1 - B_2} = \theta_b$$

where $B_1 - B_2 > 0$ (see Figure 6a)

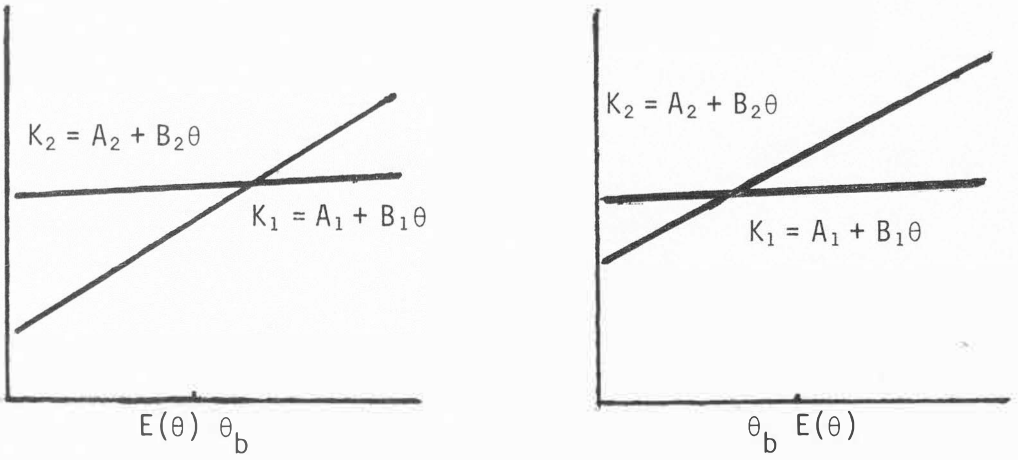
or $E(\theta) > \frac{A_2 - A_1}{B_1 - B_2} = \theta_b$

where $B_1 - B_2 < 0$, (see Figure 6b).



(a) (b)

Figure 5. Action a_1 preferred over action a_2.



(a) (b)

Figure 6. Action a_2 preferred over action a_1.

c. Either action a_1 or action a_2 makes no difference if

$E(K_1) = E(K_2)$

$$A_1 + B_1E(\theta) = A_2 + B_2E(\theta)$$

$$E(\theta)(B_1 - B_2) = A_2 - A_1$$

$$E(\theta) = \frac{A_2 - A_1}{B_1 - B_2} = \theta_b.$$

Expected loss for the optimum action is sometimes referred to as "expected value with perfect information" (EVPI).

$$\text{EVPI} = \int_{\theta_b}^{\infty} (K_2 - K_1)f(\theta)d\theta$$

$$= |B_2 - B_1|\int_{\theta_b}^{\infty}(\theta - \theta_b)f(\theta)d\theta$$

$$= |B_2 - B_1|\sigma_o[\frac{1}{\sqrt{2\pi}\,\sigma_o} e^{-\frac{1}{2}(\frac{\theta_b - \theta_o}{\sigma_o})^2}$$

$$- \frac{(\theta_b - \theta_o)}{\sigma_o} \int_{\theta_b}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma_o} e^{-\frac{1}{2}(\frac{\theta - \theta_o}{\theta_o})^2} d\theta]$$

$$= |B_2 - B_1|\sigma_o[\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}Z_b^2} - Z_bP(Z > Z_b)]$$

$$= |B_2 - B_1|\sigma_o L_n(D_o) \qquad (3.11)$$

where $Z_b = \dfrac{\theta_b - \theta_o}{\sigma_o}$,

$$L_n(D_o) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}Z_b^2} - Z_bP(Z > Z_b),$$

and $D_o = \dfrac{\theta_b - E_o(\theta_i)}{\sigma_o(\theta_i)}$

The value in brackets denoted by $L_n(D_o)$ represents the normal loss function and is presented in tables in some decision textbooks such as those by Raiffa and Schlaifer[19] and Schlaifer.[20]

[19] Howard Raiffa and Robert Schlaifer, *Applied Statistical Decision Theory* (Boston, Massachusetts: Harvard University Press, 1961), p. 356.

[20] Robert Schlaifer, *Introduction to Statistics for Business Decisions* (New York: McGraw-Hill Book Company, Inc., 1961), p. 370-371.

The loss function for the optimum action a_1 is shown in Figure 7.

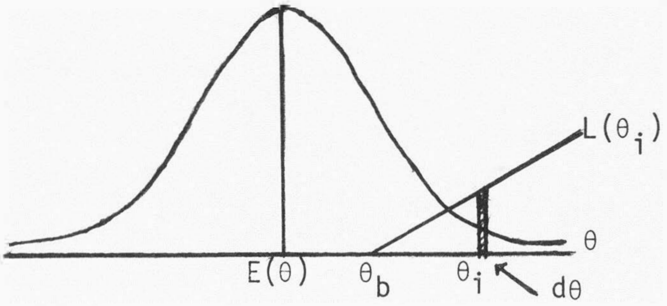


Figure 7. Loss function for the optimum action a_1 in normal distribution.

2. Normal prior distribution with normal sampling is discussed as follows:

When the sampling is available, we should make use of this additional information. Hence the expected loss (payoff) for the optimum action is as follows:

$$EVPI = |B_2 - B_1|\sigma_1(\theta_i)L_n(D_1) \qquad (3.12)$$

This formula is obtained directly from (3.11) by substituting the expected value and variance of the prior normal distribution for the expected value and variance of the posterior normal distribution, respectively, where

$$D_1 = \frac{\theta_b - E_1(\theta_i)}{\sigma_1(\theta_i)}.$$

The above-mentioned decision process was limited to a two-action process, but we can extend it to many-action problems following exactly the same process.

CHAPTER IV

APPLICATION

Some problems in application

One of the difficulties in using decision theory in applied problems is that of specifying a realistic loss (payoff) function. It is impossible to specify accurately the consequence in making a wrong decision in asserting the true state of nature. In a two-person, zero-sum game, the loss function is the real numerical loss, but it is still questionable whether the mathematical expectation of loss is an appropriate measure of the random losses when the statistical experiment is performed only once.

These difficulties may be partially solved as follows:

1. Experience with statistical problems shows that "good" processes are not sensitive to small changes in loss function, especially when sample sizes are quite large. Hence the precise values of the loss matrix are not so serious in application.

2. The statistician might measure the random loss by taking an expected value if the loss matrix is measured in terms of utility function rather than in terms of monetary value, since monetary value is not a good evaluation of loss or profit.

3. Usually the states of nature are uncertain. There exist two types of uncertainty:

   a. The randomness of probability, and

   b. The absence of knowledge of the probability distribution.

If the probability distribution of the states of nature is known, the randomness is the only type of uncertainty left. The problem is then what principles (criteria) does the statistician use to make decisions? If the states of nature are not known, then what should the statistician do?

This difficulty in many applications is not serious, since in many industrial applications, the frequency with which the states of nature distribute is known approximately by previous experimentation.

Bayesian statistical estimate

In classical statistics we use "unbiasedness," "efficiency," "consistent," and "sufficiency" as criteria of "good" estimators. For example, maximum likelihood method, the method of moment, and the least squares method are different methods to find a point estimator. Their properties are measured by these criteria.

The Bayesian method is another way of finding an estimator. Indeed, the estimate problem is the same as the decision problem. We may call the decision rule the estimator and the action the estimate.

Generally speaking, the decision rule with minimum expected risk is the Bayesian estimator. Tables 6 and 7 show the loss matrix and outcomes of experiment in a Bayesian estimate.

For example: Let the states of nature; i.e., the parameters, be $\{\theta_1, \theta_2, \theta_3\}$.

$R(\theta_1, d_i) = f_{11}L_{11} + f_{12}L_{12} + f_{13}L_{13} + f_{14}L_{14}$

$R(\theta_2, d_i) = f_{21}L_{21} + f_{22}L_{22} + f_{23}L_{23} + f_{24}L_{24}$

$R(\theta_3, d_i) = f_{31}L_{31} + f_{32}L_{32} + f_{33}L_{33} + f_{34}L_{34}$.

If the prior probabilities $P(\theta_1)$, $P(\theta_2)$, and $P(\theta_3)$ are known, we can use these prior probabilities to calculate the expected risk for each decision rule and to select the smallest expected risk. That is the optimum decision rule; i.e., Bayesian solution.

Table 6. Loss table for Bayesian estimate

State of nature	Action			
	$d(X_1)$	$d(X_2)$	$d(X_3)$	$d(X_4)$
θ_1	L_{11}	L_{12}	L_{13}	L_{14}
θ_2	L_{21}	L_{22}	L_{23}	L_{24}
θ_3	L_{31}	L_{32}	L_{33}	L_{34}

Table 7. Probabilities of sample outcomes for various states of nature

State of nature	Outcomes			
	X_1	X_2	X_3	X_4
θ_1	f_{11}	f_{12}	f_{13}	f_{14}
θ_2	f_{21}	f_{22}	f_{23}	f_{24}
θ_3	f_{31}	f_{32}	f_{33}	f_{34}

$E[R(\theta,d_i)] = P(\theta_1)R(\theta_1,d_i) + P(\theta_2)R(\theta_2,d_i) + P(\theta_3)R(\theta_3,d_i).$

This is the expected risk when the estimator d_i is used. If there exist expected risks for all possible d_i, the smallest expected risk is then called the Bayesian estimator. One numerical example in statistical quality control is illustrated as follows:

The Statistical Quality Control Division of a company is considering whether or not to accept a lot of a certain production from the production division. Before it can make a decision, the Statistical

Quality Control Division gets some defective fraction (states of nature) 0.05, 0.10, and 0.20 with the prior probabilities 0.50, 0.35, and 0.15, respectively. Assume that the quality controller draws a sample of size four and finds that the sample contains three defectives. Should he accept the lot? The quality controller chooses the action to accept the lot (action a_1) or reject the lot (action a_2) on the outcomes of the experiment. In this illustration, there are five possible outcomes and the two possible actions are associated with each. Hence, there exist $2^5 = 32$ ways of associating outcomes and actions as shown in Table 8.

In this illustration, the quality controller assumes the percent defectives are binomially distributed, since the quality controller draws a sample of size four with replacement from a lot of 0.05, 0.10, and 0.20 percent defectives, respectively.

The likelihood probability function (the conditional probability of obtaining a particular sample outcome given the state of nature) is shown in Table 9.

Given the likelihood probability function shown in Table 9, and the decision rules shown in Table 8, we can calculate the action probabilities for taking action a_1 and action a_2 for the given states of nature, $\theta_1 = 0.05$, $\theta_2 = 0.10$, and $\theta_3 = 0.20$. These can be expressed as $P(a_1|\theta_1)$, $P(a_2|\theta_1)$, $P(a_1|\theta_2)$, $P(a_2|\theta_2)$, $P(a_1|\theta_3)$, and $P(a_2|\theta_3)$. $P(a_2|\theta_1)$, $P(a_1|\theta_2)$, and $P(a_1|\theta_3)$ are called error probabilities. The action probabilities are given in Table 10. Sample size is four, and binomial distribution is assumed.

To illustrate how these action probabilities have been calculated, let us consider d_{17}. If the quality controller chooses d_{17} as decision

Table 8. All possible decision rules associating sample outcomes and actions

Decision rules d_i	Outcomes (number of defectives) 0	1	2	3	4
d_1	a_1	a_1	a_1	a_1	a_1
d_2	a_1	a_1	a_1	a_1	a_2
d_3	a_1	a_1	a_1	a_2	a_1
d_4	a_1	a_1	a_2	a_1	a_1
d_5	a_1	a_2	a_1	a_1	a_1
d_6	a_2	a_1	a_1	a_1	a_1
d_7	a_1	a_1	a_1	a_2	a_2
d_8	a_1	a_1	a_2	a_2	a_1
d_9	a_1	a_2	a_2	a_1	a_1
d_{10}	a_2	a_2	a_1	a_1	a_1
d_{11}	a_2	a_1	a_1	a_1	a_2
d_{12}	a_2	a_1	a_1	a_2	a_1
d_{13}	a_2	a_1	a_2	a_1	a_1
d_{14}	a_1	a_2	a_1	a_1	a_2
d_{15}	a_1	a_2	a_1	a_2	a_1
d_{16}	a_1	a_1	a_2	a_1	a_2
d_{17}	a_1	a_1	a_2	a_2	a_2
d_{18}	a_1	a_2	a_2	a_2	a_1
d_{19}	a_2	a_2	a_2	a_1	a_1
d_{20}	a_2	a_2	a_1	a_1	a_2
d_{21}	a_2	a_1	a_1	a_2	a_2
d_{22}	a_2	a_1	a_2	a_2	a_1
d_{23}	a_2	a_1	a_2	a_1	a_2
d_{24}	a_2	a_2	a_1	a_2	a_1
d_{25}	a_1	a_2	a_1	a_2	a_2
d_{26}	a_1	a_2	a_2	a_1	a_2
d_{27}	a_1	a_2	a_2	a_2	a_2
d_{28}	a_2	a_2	a_2	a_2	a_1
d_{29}	a_2	a_2	a_2	a_1	a_2
d_{30}	a_2	a_2	a_1	a_2	a_2
d_{31}	a_2	a_1	a_2	a_2	a_2
d_{32}	a_2	a_2	a_2	a_2	a_2

Table 9. Probabilities of sample outcomes for various defective fractions

State of nature (defective fractions)	Outcome (number of possible defectives) 0	1	2	3	4
0.05	0.8145	0.1715	0.0135	0.0005	0.0000
0.10	0.6561	0.2916	0.0486	0.0036	0.0001
0.20	0.4096	0.4096	0.1536	0.0256	0.0016

Table 10. Action probabilities of all possible decision rules

Decision	Action probabilities					
rules d_i	$P(a_1\|\theta_1)$	$P(a_2\|\theta_1)$	$P(a_1\|\theta_2)$	$P(a_2\|\theta_2)$	$P(a_1\|\theta_3)$	$P(a_2\|\theta_3)$
d_1	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000
d_2	1.0000	0.0000	0.9999	0.0001	0.9984	0.0016
d_3	0.9995	0.0005	0.9964	0.0036	0.9744	0.0256
d_4	0.9865	0.0135	0.9314	0.0486	0.8464	0.1536
d_5	0.8285	0.1715	0.7084	0.2916	0.5904	0.4096
d_6	0.1855	0.8145	0.3439	0.6561	0.5004	0.4096
d_7	0.9995	0.0005	0.9963	0.0037	0.9728	0.0272
d_8	0.9860	0.0140	0.9478	0.0522	0.8208	0.1792
d_9	0.8150	0.1850	0.6598	0.3402	0.4368	0.5632
d_{10}	0.0140	0.9860	0.0523	0.9477	0.1808	0.8172
d_{11}	0.1855	0.8145	0.3438	0.6562	0.5888	0.4112
d_{12}	0.1850	0.8150	0.3403	0.6597	0.5675	0.4325
d_{13}	0.1720	0.8380	0.2953	0.7047	0.4368	0.5632
d_{14}	0.8285	0.1715	0.5368	0.4632	0.5888	0.4112
d_{15}	0.8280	0.1720	0.7048	0.2952	0.5648	0.4352
d_{16}	0.9865	0.0135	0.9513	0.0487	0.8448	0.1552
d_{17}	0.9860	0.0140	0.9477	0.0523	0.8192	0.1808
d_{18}	0.9145	0.1855	0.6562	0.3438	0.4112	0.5888
d_{19}	0.0005	0.9995	0.0037	0.9963	0.0272	0.9728
d_{20}	0.0140	0.9860	0.0522	0.9478	0.1792	0.8208
d_{21}	0.1850	0.8150	0.3402	0.6598	0.5632	0.4368
d_{22}	0.1715	0.8285	0.2917	0.7083	0.4112	0.5888
d_{23}	0.1720	0.8280	0.2952	0.7048	0.4325	0.5675
d_{24}	0.0135	0.9865	0.0487	0.9513	0.1687	0.8313
d_{25}	0.8280	0.1720	0.7047	0.2953	0.5632	0.4368
d_{26}	0.8150	0.1850	0.6597	0.3403	0.4352	0.5648
d_{27}	0.8145	0.1855	0.6561	0.3439	0.4096	0.5904
d_{28}	0.0000	1.0000	0.0001	0.9999	0.0016	0.9984
d_{29}	0.0005	0.9995	0.0036	0.9964	0.0256	0.9744
d_{30}	0.0135	0.9865	0.0486	0.9514	0.1536	0.8464
d_{31}	0.1715	0.8285	0.2916	0.7084	0.4096	0.5904
d_{32}	0.0000	1.0000	0.0000	1.0000	0.0000	1.0000

rule, he will use this decision rule as the criterion to decide whether or not to accept the lot. This decision rule (d_{17}) states that if the number of defectives in the sample of size four is 0, or 1, the quality controller must accept the lot (a_1), and that if the number of defectives in this sample is 2, 3, or 4, he must reject the lot (a_2). Action a_1 is selected when the number of defectives is 0 or 1; the probability of taking a_1 will be $P(a_1) = P(x = 0 \cup x = 1)$, where x is a random variable denoting the number of defectives.

There exist three different states of nature; i.e., the percent defective of the lot is $\theta_1 = 0.05$, $\theta_2 = 0.10$, and $\theta_3 = 0.20$, respectively. Thus the probabilities of taking action a_1, and action a_2 become:

$$P(a_1|\theta_1) = P(x = o \cup x = 1|\theta_1) = P(x = 0|\theta_1) + P(x = 1|\theta_1)$$

$$= 0.8145 + 0.1715$$

$$= 0.9860$$

$$P(a_2|\theta_1) = P(x = 2 \cup x = 3 \cup x = 4|\theta_1)$$

$$= P(x = 2|\theta_1) + P(x = 3|\theta_1) + P(x = 4|\theta_1)$$

$$= 0.0135 + 0.0005 + 0.0000 = 0.0140$$

$$P(a_1|\theta_2) = P(x = 0 \cup x = 1|\theta_2) = P(x = 0|\theta_2) + P(x = 1|\theta_2)$$

$$= 0.6561 + 0.2916$$

$$= 0.9477$$

$$P(a_2|\theta_2) = 1 - P(a_1|\theta_2) = 0.0523$$

$$P(a_1|\theta_3) = P(x = 0 \cup x = 1|\theta_3) = P(x = 0|\theta_3) + P(x = 1|\theta_3)$$

$$= 0.4096 + 0.4096$$

$$= 0.8192$$

$$P(a_2|\theta_3) = 1 - P(a_1|\theta_3) = 1 - 0.8192 = 0.1808$$

The loss matrix for action a_1 and action a_2 when the states of nature are $\theta_1 = 0.05$, $\theta_2 = 0.10$, and $\theta_3 = 0.20$ is shown in Table 11.

Table 11. Loss table for statistical quality control

State of nature (percent defective of the lot)	a_1 (accept the lot)	a_2 (reject the lot)
$\theta_1 = 0.05$	0	90
$\theta_2 = 0.10$	20	0
$\theta_3 = 0.20$	80	0

Given the action probabilities (Table 10) and the corresponding loss matrix (Table 11), the expected losses (weighted average of the losses) will be calculated for each of the decision rules. These expected losses are called the risk. The risk for any d_i, when the state of nature is θ_1 is designated by:

$R(\theta_1,d_i) = \Sigma L(\theta_1,d(X))P(d(X)|\theta_1)$.

Hence

$$
\begin{aligned}
R(\theta_1,d_{17}) &= \Sigma L(\theta_1,d(X))P(d(X)|\theta_1)\\
&= L(\theta_1,d_1)P(a_1|\theta_1) + L(\theta_1,a_2)P(a_2|\theta_1)\\
&= 0 + (90)(0.0140)\\
&= 1.260\\
R(\theta_2,d_{17}) &= \Sigma L(\theta_2,d(X))P(d(X)|\theta_2)\\
&= L(\theta_2,a_1)P(a_1|\theta_2) + L(\theta_2,a_2)P(a_2|\theta_2)\\
&= (20)(0.9477) + 0\\
&= 18.954\\
R(\theta_3,d_{17}) &= \Sigma L(\theta_3,d(X))P(d(X)|\theta_3)\\
&= L(\theta_3,a_1)P(a_1|\theta_3) + L(\theta_3,a_2)P(a_2|\theta_3)\\
&= (80)(0.8192)\\
&= 65.536
\end{aligned}
$$

Suppose the prior probabilities of the states of nature; i.e., $\theta_1 = 0.05$, $\theta_2 = 0.10$, and $\theta_3 = 0.20$ are 0.50, 0.35, and 0.15, respectively. That is,

$P(\theta_1 = 0.05) = 0.50$

$P(\theta_2 = 0.10) = 0.35$

$P(\theta_3 = 0.20) = 0.15$.

Then the expected risk for decision rule d_{17} can be calculated by taking the weighted average of the risks with the corresponding prior probability as weight. That is,

$$\begin{aligned} r(\theta,d_{17}) &= P(\theta_i)R(\theta_i,d_{17}) \\ &= P(\theta_1)R(\theta_1,d_{17}) + P(\theta_2)R(\theta_2,d_{17}) \\ &+ P(\theta_3)R(\theta_3,d_{17}) \\ &= (0.50)(1.260) + (0.35)(18.954) + (0.15)(65.536) \\ &= 0.630 + 6.634 + 9.830 \\ &= 17.094. \end{aligned}$$

The risk and the expected risk for each of the remaining decision rules can be similarly calculated as shown in Table 12.

The expected risks are given in the last column of Table 12. The decision rule d_{17} has the smallest expected risk. Hence it is called the optimum decision rule. This optimum decision rule is the Bayesian solution.

We have assumed that the quality controller drew a sample of size four and found that the sample contained three defectives. According to this solution, the lot of this certain production should be rejected.

The above-mentioned example is for a discrete case. In the continuous case, it follows exactly the same theory.

Table 12. Computation of expected risk

Decision rule	State of nature			Expected
d_i	$\theta_1 = 0.05$	$\theta_2 = 0.10$	$\theta_3 = 0.20$	risk
d_1	0.000	20.000	80.000	19.000
d_2	0.000	19.998	79.872	18.980
d_3	0.045	19.928	77.952	18.690
d_4	1.215	18.628	67.712	17.284
d_5	15.435	14.168	47.232	19.761
d_6	73.305	6.878	40.032	45.065
d_7	0.045	19.926	77.824	18.670
d_8	1.260	18.956	65.664	17.177
d_9	16.380	13.196	34.944	18.050
d_{10}	88.740	1.046	14.464	46.906
d_{11}	73.305	6.876	47.104	46.125
d_{12}	73.350	6.806	45.400	45.867
d_{13}	74.520	5.906	34.944	44.569
d_{14}	15.435	10.736	47.104	18.541
d_{15}	15.480	14.096	45.184	19.451
d_{16}	1.215	19.026	67.584	19.404
d_{17}	1.260	18.954	65.536	17.094
d_{18}	16.695	13.124	32.896	17.875
d_{19}	89.955	0.074	2.176	45.330
d_{20}	88.740	1.044	14.336	46.886
d_{21}	73.350	6.804	45.056	45.815
d_{22}	74.565	5.834	32.896	44.259
d_{23}	74.520	5.904	34.600	44.516
d_{24}	88.785	0.974	1.496	44.958
d_{25}	15.480	14.094	45.056	19.431
d_{26}	16.650	13.194	34.816	18.165
d_{27}	16.695	13.122	32.768	26.203
d_{28}	90.000	0.002	0.128	45.020
d_{29}	89.955	0.072	2.048	45.310
d_{30}	88.785	0.972	12.288	46.576
d_{31}	74.565	5.832	32.768	44.239
d_{32}	90.000	0.000	0.000	45.000

$$E[R(\theta,d_i)] = \int_{-\infty}^{\infty} R(\theta,d_i)P(\theta)d\theta$$

$$= \int_{-\infty}^{\infty}\{\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} L[\theta,d(x_1,x_2,\ldots x_n)]g(x_1,x_2,\ldots x_n|\theta)dx_1dx_2$$

$$\ldots dx_n\}P(\theta)d\theta$$

$$= \int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty}\{\int_{-\infty}^{\infty} L[\theta,d(x_1,x_2,\ldots x_n)]g(x_1,x_2,\ldots x_n|\theta)P(\theta)d\theta\}$$

$$dx_1dx_2\ldots dx_n \qquad (4.1)$$

A "good" estimator will be an estimator which minimizes the expected risk. To satisfy this condition we can minimize the quantity in brackets (4.1); i.e., minimizing

$\int_{-\infty}^{\infty} L[\theta,d(x_1,x_2,\ldots x_n)]g(x_1,x_2,\ldots x_n|\theta)P(\theta)d\theta.$

Since

$$\int_{-\infty}^{\infty} L[\theta,d(x_1,x_2,\ldots x_n)]g(x_1,x_2,\ldots x_n|\theta)P(\theta)d\theta$$

$$= \int_{-\infty}^{\infty} L[\theta,d(x_1,x_2,\ldots x_n)]g(x_1,x_2,\ldots x_n,\theta)d\theta$$

$$= \int_{-\infty}^{\infty} L[\theta,d(x_1,x_2,\ldots x_n)]k(x_1,x_2,\ldots x_n)\cdot h(\theta|x_1,x_2,\ldots x_n)D\theta$$

where

$$g(x_1,x_2,\ldots x_n,\theta) = k(x_1,x_2,\ldots x_n)\cdot h(\theta|x_1,x_2,\ldots x_n)$$

$$= k(x_1,x_2,\ldots x_n)\int_{-\infty}^{\infty} L[\theta,d(x_1,x_2,\ldots x_n)]h(\theta|x_1,x_2,\ldots x_n)d\theta. \qquad (4.2)$$

Hence a Bayesian estimator is the state of nature (parameter) $\hat{\theta}$ which minimizes the above equation for all possible samples, $X = \{x_1,x_2,\ldots x_n\}$. In other words, if $\hat{\theta} = d(x_1,x_2,\ldots x_n)$, then $\int_{-\infty}^{\infty} L(\theta,\hat{\theta})h(\theta|x_1,x_2,\ldots x_n)d\theta$ will be the smallest value. $\hat{\theta}$ is then called the Bayesian estimator.

For example:[21]

1. If $f(x|\theta) = \frac{2x}{\theta^2}$ $\qquad 0 < x < \theta$

and $P(\theta) = 1$ $\qquad 0 < \theta < 1$

[21] Alexander M. Mood and Franklin A. Graybill, *Introduction to the Theory of Statistics* (New York: McGraw-Hill Book Company, Inc., 1963), p. 196.

using the loss function $L(\theta,\hat{\theta}) = \theta^2(\theta - \hat{\theta})^2$, the Bayesian estimate is

$$\frac{\partial}{\partial\hat{\theta}}\int_0^1 \theta^2(\theta - \hat{\theta})h(\theta|x)d\theta = 0$$

by conditional and marginal distribution theorems:

$$g(\theta,x) = P(\theta)f(x|\theta) = \frac{2x}{\theta^2},\ k(x) = \int_x^1 g(\theta,x)d\theta = \int_x^1 \frac{2x}{\theta^2}d\theta = 2$$

$$h(\theta|x) = \frac{g(\theta,x)}{k(x)} = \frac{x}{\theta^2}$$

$$\frac{\partial}{\partial\hat{\theta}}\int_0^1 \theta^2(\theta - \hat{\theta})\ \frac{2x}{\theta^2}d\theta = 0.$$

Solving this equation, we get $\hat{\theta} = \frac{1}{2}$, the Bayesian solution.

2. Let $X = \{x_1,x_2,\ldots x_n\}$ be a random sample of size n from the Poisson density functions:

$$f(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \qquad x = 0,\ 1,\ 2,\ \ldots$$

λ has the probability density

$$P(\lambda) = e^{-\lambda} \qquad 0 < \lambda < \infty.$$

Using the loss function $L(\lambda,\hat{\lambda}) = (\lambda - \hat{\lambda})^2$

the Bayesian estimate is the solution of:

$$\frac{\partial}{\partial\hat{\lambda}}\int_0^\infty L(\lambda,\hat{\lambda})h(\lambda|x_1,x_2,\ldots x_n)d\lambda = 0$$

Similarly, by conditional and marginal distribution theorems,

$$g(\lambda,x_1,x_2,\ldots x_n) = P(\lambda)f(x_1,x_2,\ldots x_n|\lambda)$$

$$= \frac{\lambda^{\Sigma x_i}e^{-\lambda(n+1)}}{x_1!\ x_2!\ \ldots x_n!}$$

$$k(x_1,x_2,\ldots x_n) = \int_0^\infty g(\lambda,x_1,x_2,\ldots x_n)d\lambda$$

$$= \int_0^\infty \frac{\lambda^{\Sigma x_i}e^{-\lambda(n+1)}}{x_1!\ x_2!\ \ldots x_n!}$$

$$= \frac{1}{x_1!\, x_2! \ldots x_n!} \int_0^\infty \lambda^{\Sigma x_i} e^{-\lambda(n+1)} d\lambda$$

$$= \frac{\Gamma(\Sigma x_i + 1)}{(n+1)^{\Sigma x_i + 1}(x_1!\, x_2! \ldots x_n!)}$$

$h(\lambda | x_1, x_2, \ldots x_n)$

$$= \frac{g(\lambda, x_1, x_2, \ldots x_n)}{k(x_1, x_2, \ldots x_n)} = \frac{\lambda^{\Sigma x} e^{-\lambda(n+1)}(n+1)^{\Sigma x + 1}}{\Gamma(\Sigma x_i + 1)}$$

$$\frac{\partial}{\partial \hat{\lambda}} \int_0^\infty L(\lambda, \hat{\lambda}) h(\lambda | x_1, x_2, \ldots x_n) d\lambda = 0$$

$$\frac{\partial}{\partial \hat{\lambda}} \int_0^\infty (\lambda - \hat{\lambda})^2 \frac{\lambda^{\Sigma x} e^{-\lambda(n+1)}(n+1)^{\Sigma x + 1}}{\Gamma(\Sigma x_i + 1)} d\lambda = 0$$

$$\int_0^\infty 2(\lambda - \hat{\lambda}) \frac{\lambda^{\Sigma x} e^{-\lambda(n+1)}(n+1)^{\Sigma x + 1}}{\Gamma(\Sigma x_i + 1)} d\lambda = 0$$

$$\frac{2(n+1)^{\Sigma x + 1}}{\Gamma(\Sigma x_i + 1)} \left[\int_0^\infty (\lambda - \hat{\lambda}) \lambda^{\Sigma x_i} e^{-\lambda(n+1)} d\lambda\right] = 0$$

$$\int_0^\infty \lambda^{\Sigma x + 1} e^{-\lambda(n+1)} d\lambda - \hat{\lambda} \int_0^\infty \lambda^{\Sigma x_i} e^{-\lambda(n+1)} d\lambda = 0$$

$$\frac{\Gamma(\Sigma x_i + 2)}{(n+1)^{\Sigma x + 2}} - \hat{\lambda} \frac{\Gamma(\Sigma x_i + 1)}{(n+1)^{\Sigma x_i + 1}} = 0$$

$$\hat{\lambda} = \frac{(n+1)^{\Sigma x_i + 1}\Gamma(\Sigma x_i + 2)}{(n+1)^{\Sigma x_i + 2}\Gamma(\Sigma x_i + 1)} = \frac{\Sigma x_i + 1}{n + 1}, \text{ the Bayesian solution.}$$

The state of nature (parameter) is determined by the decision rule (estimator) $\hat{\lambda} = \frac{\Sigma x + 1}{n + 1}$.

In conclusion, in classical statistical estimates, the "confidence interval" is attached in "interval estimate"; also "the significance level" is attached in "point estimate." In Bayesian inference we neither use "confidence interval" nor "the significance level," since from the Bayesian viewpoint the implications of "the confidence interval" and "the significance level" is still left entirely to the judgment of the statistician. R. Schlaifer has shown that only with a prior uniform distribution can the value of the confidence interval estimate be interpreted as the central area in the posterior distribution. Confidence intervals are a "good" indication only if the prior probabilities are roughly the same for all possible values. Now, $I_1 = I_o + I_s$; i.e., $\frac{1}{\sigma_1^2} = \frac{1}{\sigma_o^2} + \frac{n}{\sigma^2}$. If σ_o^2 approaches infinity, $I_1 = I_s$; i.e., $\frac{1}{\sigma_1^2} = \frac{n}{\sigma^2}$. This means that if $\sigma_o^2 \to \infty$ or the prior probability is uniform distribution, we can calculate the posterior probability that $\theta < \theta_b$ by finding the one tail level of significance with the population mean θ_b. Figure 8 will be helpful in illustrating the relation between the prior and the posterior distributions.

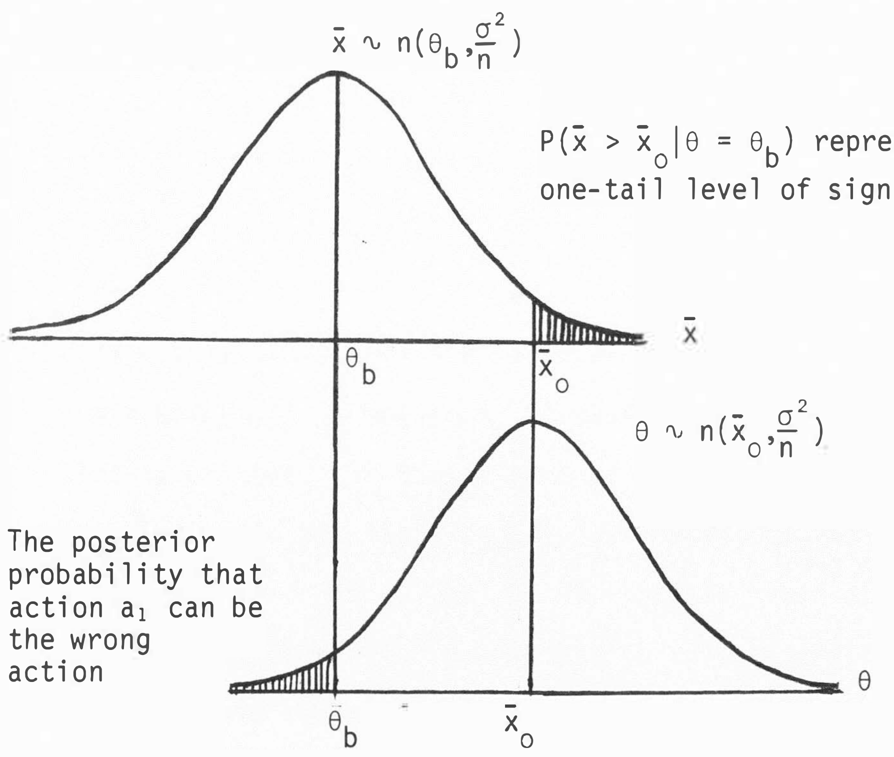

$\bar{x} \sim n(\theta_b, \frac{\sigma^2}{n})$
$P(\bar{x} > \bar{x}_o | \theta = \theta_b)$ represents one-tail level of significance
$\bar{x}$
θ_b
$\bar{x}_o$
$\theta \sim n(\bar{x}_o, \frac{\sigma^2}{n})$
The posterior probability that action a_1 can be the wrong action
θ
θ_b
$\bar{x}_o$

Figure 8. Relationship between a prior uniform distribution and a posterior distribution.

The upper portion of the figure is the conditional distribution of statistic $\bar{x}$, given $\mu = \theta_b$. The lower portion is the posterior distribution of the basic random variable θ, given the observed statistic $\bar{x}_o$, and a prior distribution with $I_o = 0$.[22]

Decisions on acquiring the sample size

It is supposed to be worthwhile to acquire additional samples if any are available. But even if some additional information might be available, there is a further question concerning the quantity to be obtained. If the information about the states of nature can be obtained without increasing the sampling cost, any sample survey would imply surveying the entire population. And the statistician would sample as widely as possible, since this could reduce the risk in the decision problem without extra cost. But such is not the case, because the cost of total information increases with the sample size. There exist two types of models for the cost function, which is expressed by $c(n) = c \cdot n$, where c is cost and n is sample size. Another model is $c(n) = a + c \cdot n$, where the cost is divided into two parts, fixed cost and variable cost.

Since the expected payoff increases as the sample size increases, the optimal sample size can be calculated by the expected payoff. But there is a difficulty that the value of additional information is uncertain before it is obtained, because the outcome of this information (sample) is unknown. If the outcome of the sample were known, the statistician would not take the sample. Hence the decision on acquiring additional information must be made on the basis of all the possible outcomes of a given size of sample and on calculating the expected value of these possible outcomes. Since this process is undertaken before

[22] Robert Schlaifer, *Introduction to Statistics for Business Decision* (New York: McGraw-Hill Book Company, Inc., 1961), p. 296-315.

the sample is taken and also before the corresponding posterior probabilities can be computed, it is called the "preposterior analysis." The optimal sample size to be obtained may then be determined by this preposterior analysis for varying quantities.

The computation of the expected terminal payoffs (or expected terminal losses) for varying sample size is a rather burdensome job in even a relatively simple problem. It can only be obtained by "trial and error" method. An electronic computer can make this job much easier.

Now we refer to the problem of the optimal sample itself. What constitutes the optimum sample?

For example: How many times should a coin be tossed before deciding whether it is a fair coin? It is very hard to say, since the answer would depend on both the cost of tossing the coin and the consequence of making the wrong decision. This problem can be treated like the microeconomic theory of production. The principle is that additional information should be acquired as long as the marginal value of this information exceeds the marginal cost of acquiring it. In other words, if the expected value based on a sample size, less the cost of sampling, is greater than the expected value without sampling, it would be worthwhile to secure an additional sample. And the optimal sample size will be the sample maximizing these expected values.

The computer program was written to derive the expected payoff of optimal terminal action and optimal sample size for preposterior analysis. (See Appendix.)

One numerical example to illustrate a market research application follows:

The planning division of a bus service is studying whether or not a new commuter bus service is to be made. Before it can make a decision, the division gets some frequency of proportions of commuters using bus service daily from prior experience, as shown in Table 13.

Table 13. Frequency of proportions of commuters using bus service from prior experience

Proportion of commuters using bus service	Relative frequency
0.20	0.60
0.25	0.25
0.30	0.15

The payoff matrix in terms of the daily profit is shown in Table 14.

Table 14. Profit table for setting up new bus service

Proportion of commuters using bus service	a_1 service	a_2 no service
0.20	-8	0
0.25	5	0
0.30	16	0

The payoff matrix is also a utility function $U(\theta_i,a_j)$. For example, the profit associating θ_2 and a_1 in payoff matrix is 5. It also can be expressed in terms of utility function $U(\theta_2, a_1) = 5$.

The sample is selected at random from the suburban community for which the new bus service is planned. The sampling unit is individual persons in the community. For the sampling cost, it is assumed that there is a fixed cost of $50 and a variable cost of $5 per sampling unit. Terminal action is any action that puts a final end to the decision process. Optimal terminal action is the action which optimizes the expected payoff.

For example: Suppose the payoff matrix was given in Table 14. The optimal terminal action of x = 2, where x is the observation representing the number of persons who prefer a new bus service, in a given sample of size 10 is calculated as follows:

$$P(x = 2|\theta_1) = \frac{10!}{2!8!}(0.20)^2(0.80)^8 = 0.301995$$

$$P(x = 2|\theta_3) = \frac{10!}{2!8!}(0.25)^2(0.75)^8 = 0.281565$$

$$P(x = 2|\theta_3) = \frac{10!}{2!8!}(0.30)^2(0.70)^8 = 0.233474$$

By Bayes' theorem:

$$P(\theta_j|x = 2) = \frac{P(\theta_j)\cdot P(x = 2|\theta_j)}{\Sigma P(\theta_i)\cdot P(x = 2|\theta_i)}$$

$$P(\theta_1|x = 2) = \frac{0.181197}{0.286609} = 0.6322$$

$$P(\theta_2|x = 2) = \frac{0.070391}{0.286609} = 0.2456$$

$P(\theta_3|x = 2) = \frac{0.035021}{0.286609} = 0.1222.$ (See Table 15.)

The optimal terminal action is then no bus service. This posterior expected payoff given the sample outcome x = 2 is also the conditional payoff in the sense of being conditional upon this particular sample outcome. The expected conditional payoff of optimal action is simply the conditional payoff of optimal terminal action multiplied by probability of the particular sample outcome. The expected terminal payoff of particular sample size is the sum of the expected conditional payoffs:

$$\text{The expected terminal payoff of the particular sample size} = \sum_{x=0}^{n} P(x) \sum_{i=1}^{m} f(\theta_i|x) \cdot U(\theta_i, a_o)$$

where a_o is optimal terminal action. Since the payoff matrix is shown in daily basis, if the planning division decides that the sampling cost must be covered in at least one year's operation of the new service, it must take the working days in the year into consideration. Here, we assume there being 255 working days in the year on which the bus will be served.

Table 15. Expected payoffs of actions--posterior probabilities

State of nature	$P(\theta_i \| x = 2)$	Payoff (a_1) bus service	$P(\theta_i \| x)$ $U(\theta_i, a_1)$	Payoff (a_2) no bus service	$P(\theta_i \| x)$ $U(\theta_i, a_2)$
$\theta_1 = 0.20$	0.6322	-8	-5.0576	0	0
$\theta_2 = 0.25$	0.2456	5	1.2280	0	0
$\theta_3 = 0.30$	0.1222	16	1.9552	0	0
			-1.8744		0

The expected payoff of optimal terminal action multiplied by the working days reduced by sampling costs for the particular sample size is then the expected net gain for the year.

The data in Table 16 were put into the electronic computer and the output from these data are shown in Tables 17 and 18 and in Figure 9.

Table 16. State of nature, prior probability, and loss matrix for setting up new bus service

State of nature	Prior probability	Payoff matrix	
		a_1	a_2
$\theta_1 = 0.20$	0.60	-8	0
$\theta_2 = 0.25$	0.25	5	0
$\theta_3 = 0.30$	0.15	16	0

In conclusion; we had calculated the expected payoff of optimal terminal action for various sample sizes. We also assumed a fixed cost of $50 and a variable cost per sampling unit of $5. We reduced the sampling cost from expected payoff for a given sample size. The expected payoff increased as the sampling cost decreased. Hence the expected net gain increased; but until the sample size was 37, it decreased. This point is then called the optimum point since it represents the maximum expected net gain. The best action is then to take a sample of size 37 and to take action a_1 (new bus service) if sample outcomes are greater than or equal to 10 $(x > 10)$; otherwise, to take action a_2. It should also be pointed out that the preposterior analysis assumes that any information is obtained by sampling. Furthermore, any information sampling is at random. In practice, this randomness may not easily be achieved. Hence we must be careful in interpreting the results.[23]

[23] Bruce W. Morgan, *An Introduction to Bayesian Statistical Decision Processes* (Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1968), p. 80-86.

Table 17. Expected net gain for various sample sizes

Sample size	Expected net gain
1	- 36.51
2	- 33.96
3	- 10.30
4	5.28
5	9.20
6	9.46
7	27.45
8	38.21
9	40.58
10	37.93
11	52.58
12	61.08
13	62.83
14	57.93
15	70.43
16	77.59
17	79.01
18	74.65
19	83.31
20	89.56
21	90.78
22	86.96
23	92.49
24	98.07
25	99.18
26	95.78
27	98.76
28	103.83
29	104.86
30	101.82
31	102.67

Table 17. Continued

Sample size	Expected net gain
32	107.33
33	108.31
34	105.57
35	104.61
36	108.93
37	109.87
38	107.40
39	104.86
40	108.89
41	109.81
42	107.56
43	103.64
44	107.44
45	108.32
46	106.29
47	101.33
48	104.73
49	105.60
50	103.74
51	99.16
52	100.90
53	101.76
54	100.06
55	95.81
56	96.08
57	....
.	....
.	....

Table 18. Preposterior expected payoff of optimal terminal action for sample of size 37[a]

Sample outcome x	P(x)	Optimal terminal action[b]	Optimal terminal action	
			Conditional	Expected
0	0.000162	A02	0.0	0.0
1	0.001519	A02	0.0	0.0
2	0.006959	A02	0.0	0.0
3	0.020797	A02	0.0	0.0
4	0.045666	A02	0.0	0.0
5	0.078752	A02	0.0	0.0
6	0.111426	A02	0.0	0.0
7	0.133550	A02	0.0	0.0
8	0.139067	A02	0.0	0.0
9	0.128483	A02	0.0	0.0
10	0.107174	A01	0.8043	0.086196
11	0.081843	A01	2.8228	0.231027
12	0.057776	A01	4.9005	0.283132
13	0.037907	A01	6.9005	0.261578
14	0.023145	A01	8.7078	0.201559
15	0.013128	A01	10.2577	0.134660
16	0.006893	A01	11.5296	0.079475
17	0.003337	A01	12.5429	0.041851
18	0.001483	A01	13.3345	0.019769
19	0.000602	A01	13.9461	0.008400
20	0.000223	A01	14.4162	0.003214
21	0.000075	A01	14.7771	0.001107
22	0.000023	A01	15.0543	0.000343
23	0.000006	A01	15.2676	0.000095
24	0.000002	A01	15.4321	0.000024
25	0.000000	A01	15.5593	0.000005
26	0.000000	A01	15.6577	0.000001
.	....	...	....	....
.	....	...	....	....
Total	1.000000			1.352432

[a] Expected net gain = 109.870117.

[b] The expected conditional payoff for sample outcomes of x > 26 are so small as to be insignificant. A01 means action a_1 (bus service). A02 means action a_2 (no bus service).

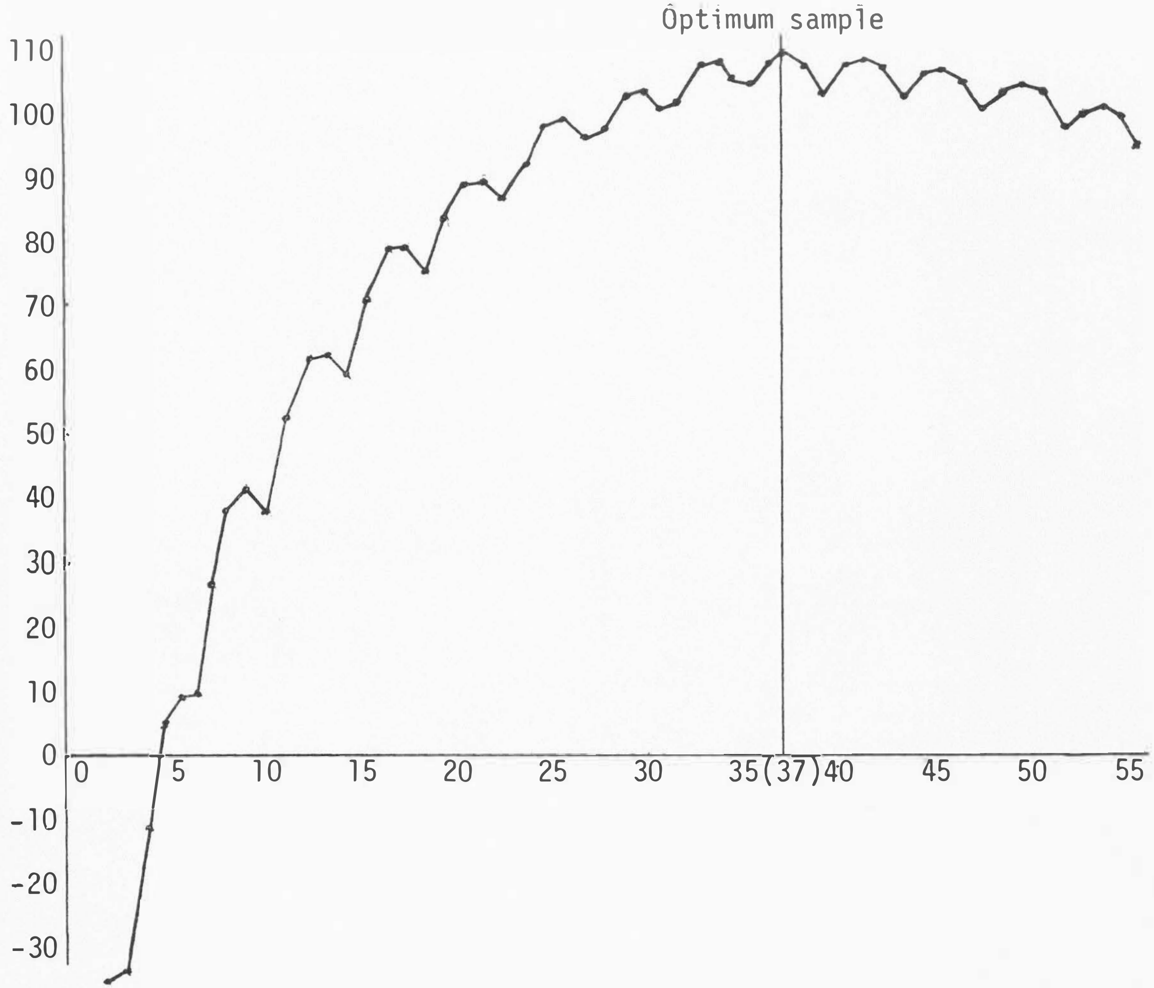


Figure 9. Expected net gain for various sample sizes.

BIBLIOGRAPHY

Blackwell, David, and M. A. Girshick. *Theory of Game and Statistical Decisions.* New York: John Wiley & Sons, Inc., 1954.

Chernoff, H., and L. E. Moses. *Elementary Decision Theory.* New York: John Wiley & Sons, Inc., 1959.

Ferguson, Thomas S. *Mathematical Statistics: A Decision Theoretic Approach.* New York: Academic Press, Inc., 1967.

Machol, R. E., and P. Gray. *Recent Developments in Information and Decision Processes.* New York: Macmillan and Company, 1962.

Mood, Alexander M., and Franklin A. Graybill. *Introduction to the Theory of Statistics.* New York: McGraw-Hill Book Company, Inc., 1963.

Morgan, Bruce W. *An Introduction to Bayesian Statistical Decision Processes.* Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1968.

Neumann, Von, and O. Morgenstern. *Theory of Game and Economic Behavior.* Princeton, New Jersey: Princeton University Press, 1947.

Raiffa, Howard, and Robert Schlaifer. *Applied Statistical Decision Theory.* Boston, Massachusetts: Harvard University Press, 1961.

Sasaki, Kyohei. *Statistics for Modern Business Decision Making.* Belmont, California: Wadsworth Publishing Company, Inc., 1968.

Savage, Leonard J. *The Foundation of Statistics.* New York: John Wiley & Sons, Inc., 1954.

Schlaifer, Robert. *Probability and Statistics for Business Decision.* New York: McGraw-Hill Book Company, Inc., 1959.

Schlaifer, Robert. *Introduction to Statistics for Business Decision.* New York: McGraw-Hill Book Company, Inc., 1961.

Sielaff, Theodore J. *Statistics in Action.* San Jose, California: San Jose State College, 1963.

Tucker, Howard G. *An Introduction to Probability and Mathematical Statistics.* New York: Academic Press, Inc., 1962.

Wald, Abraham. *Statistical Decision Functions.* New York: John Wiley & Sons., Inc., 1950.

APPENDIX

Computer Program to Derive Optimum Sample Size and Optimum Action for Preposterior Analysis

```
C     BAYESIAN STATISTICAL DECISION PROCESS
C     PREPOSTERIOR ANALYSIS
C     OPTIMUM SAMPLE SIZE AND OPTIMUM ACTION
C     PRIOR(K) IS THE PRIOR PROBABILITY OF THE STATES OF NATURE
C     CDP(K) IS CONDITIONAL PROBABILITY
C     AJNT(K) IS JOINT PROBABILITY
C     NAME(J) IS ALPHAMETIC VARIABLE FROM ACTION A01 TO A10
C     P(K) IS THE VARIABLE OF STATES OF NATURE
C     P(K,J) IS PAYOFF MATRIX
C     POST(K) IS POSTERIOR PROBABILITY
C     ACTION(J) IS CONDITIONAL ACTION
C     ACT(II) IS OPTIMUM CONDITIONAL ACTION
C     NA(II) IS ALPHAMETIC VARIABLE FOR OPTIMUM CONDITIONAL ACTION
C     EPTA(II) IS EXPECTED OPTIMUM TERMINAL ACTION
      DIMENSION PRIOR(10),CDP(10),ALKH(10),NAME(10), P(10),PAY(10,10),
     1POST(10),ACTION(10),SUML(100),EPTA(100),OPTM(100),NA(100),ACT(100)
      READ(5,90)M1,M2,DAY,FC,VC
   90 FORMAT(I2,I2,F5.0,F4.0,F3.0)
      DO 5 J=1,M2,1
      READ(5,200)NAME(J)
      DO 5 K=1,M1,1
    5 READ(5,300)PAY(K,J)
  200 FORMAT(A3)
  300 FORMAT(F11.6)
      DO 15 K=1,M1,1
   15 READ(5,100) P(K),PRIOR(K)
  100 FORMAT(2F11.6)
      WRITE(6,500)
  500 FORMAT(1H ,28X,55HPREPOSTERIOR EXPECTED PAYOFF OF OPTIMAL TERMINAL
     1 ACTION/)
      WRITE(6,600)
  600 FORMAT(26X,7H SAMPLE,15X,7HOPTIMAL,5X,23HOPTIMAL TERMINAL ACTION)
      WRITE(6,700)
  700 FORMAT(26X,8H OUTCOME,14X,8HTERMINAL)
      WRITE (6,800)
  800 FORMAT(29X,1HX,5X,8H P(X)  ,5X,6HACTION,6X,11HCONDITIONAL,5X,
     18HEXPECTED)
      DO 99 N=1,60,1
      WRITE(6,130) N
  130 FORMAT(44X,13H SAMPLE SIZE=,I3)
      M=N+1
      SUMI=0
      SUM2=0
      DO 10 I=1,M,1
      II=I-1
      SUM0=0
      DO 20 K=1,M1,1
      CDP(K)=TOR(N)/(TOR(II)*TOR(N-II))*(P(K)**II)*(1.-P(K))**(N-II)
      ALKH(K)=PRIOR(K)*CDP(K)
   20 SUMO=SUMO+ALKH(K)
      SUML(II)=SUM)
      DO 30 J=1, M2, 1
      ACTION(J)=0
      DO 30 K=1,M1,1
      POST(K)=ALKH(K)/SUMO
   30 ACTION( J)= ACTION(J)+PAY(K,J)*POST(K)
      CALL BEST (OPTMV,IK,ACTION,M2)
      ACT(II)=OPTIMV
      EPTA(II)=SUML(II)*ACT(II)
      SUM1=SUM1+SUML(II)
      NA(II)=NAME(IK)
      SUM2=SUM2+EPTA(II)
      WRITE(6,900) II,SUML(II),NA(II),ACT(II),EPTA(II)
   10 CONTINUE
  900 FORMAT(26X,I4,4X,F11.6,6X,A3,6X,F11.4,5X,F11.6)
      CN=N
      OPTM(N)=DAY*SUM2-(FC+VC*CN)
      WRITE(6,120) SUM1,SUM2
      WRITE(6,110) OPTM(M)
   99 CONTINUE
  120 FORMAT(26X,6H TOTAL,2X,F11.6,31X,F11.6)
  110 FORMAT(38X,19H EXPECTED NET GAIN=,F11.6///)
      STOP
      END

      FUNCTION TOR(IL)
      IF(IL) 65,65,75
   65 TOR=1.
      GO TO 85
   75 S=1.
      DO 95 I=1,IL,1
      X=I
   95 S=S*X
      TOR=S
   85 RETURN
      END

      SUBROUTINE BEST(OPTMV,IK,DECIDE,MN)
      DIMENSION DECIDE(10)
      IK=1
      OPTMV=DECIDE(IK)
      M =MN-1
      DO 30 K=1,M3,1
      L=K+1
      IF(OPTMV-DECIDE(L)) 20,20,30
   20 OPTMV=DECIDE(L)
      IK=L
   30 CONTINUE
      RETURN
      END
```