Universität Bielefeld/IMW

Working Papers
Institute of Mathematical Economics

Arbeiten aus dem
Institut für Mathematische Wirtschaftsforschung

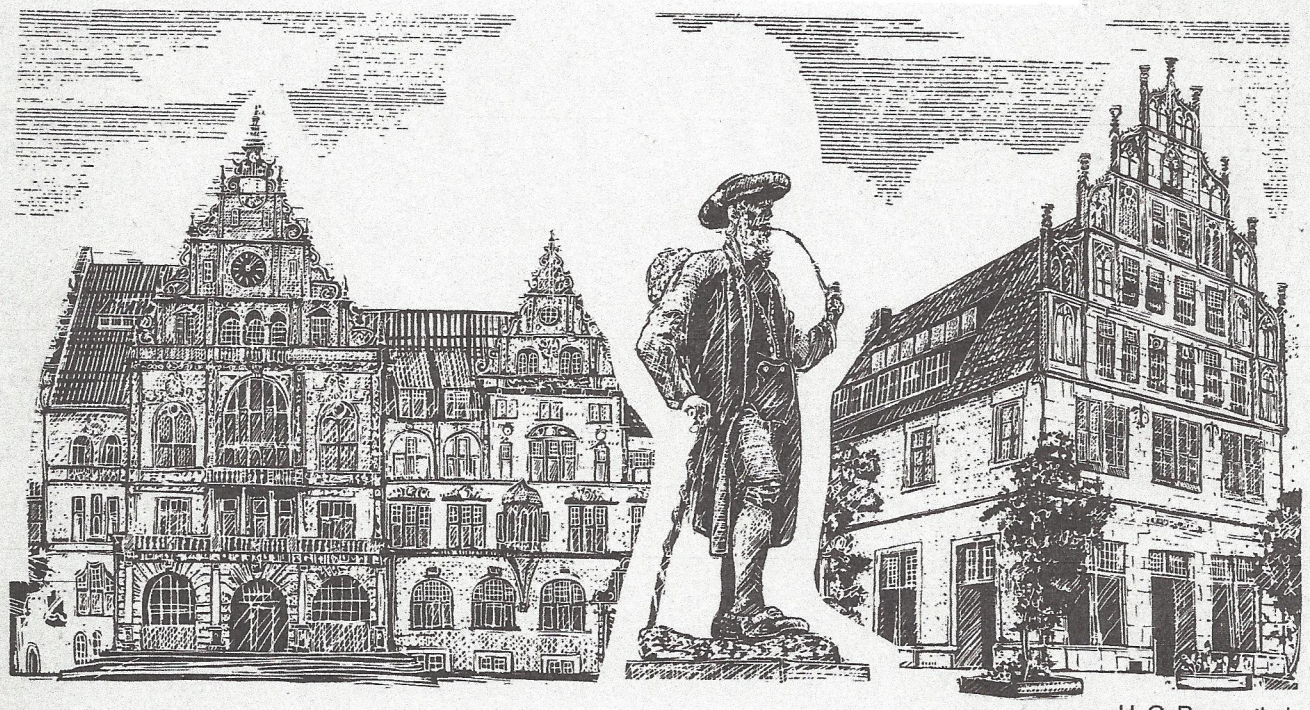
Nr. 53

Hans W. Gottinger

Simple Decision Procedures

An Expository Paper

February 1977



H. G. Bergenthal

Institut für Mathematische Wirtschaftsforschung
an der
Universität Bielefeld
Adresse/Address:
Universitätsstraße
4800 Bielefeld 1
Bundesrepublik Deutschland
Federal Republic of Germany

Simple Decision Procedures

An Expository Paper

Hans W. Gottinger

The modern approach to statistics can be characterized by the words inference and decision.
In many statistical contexts the two concepts are used in a mutually exclusive sense, but at least one important development in modern statistics has convincingly shown that the key ideas of statistical inference fit into the broader framework of Bayesian decision theory. Thus we will argue here that it will deepen our understanding of inference if we explore certain facets of decision theory by examining several simple decision procedures.
This viewpoint has been consistently emphasized and further developed in R. Schlaifer (1959), H. Raiffa and R. Schlaifer (1961) and recently been summarized in conjunction with statistical methodology, Gottinger (1975).

1. An Example: A Quality Control Problem.

Consider an automatic machine that has just been adjusted by an operator, and we are uncertain as to how good an adjustment has been made.
In principle it is possible to make an exhaustive and mutually exclusive list of events or states of the world that are relevant to the problem: one of these events surely obtains but we are uncertain as to which one. The events of the example can be described by the probability p that the machine will turn out a defective part. Suppose the adjustments of the machine can be described by four values of p: p = 0.1, .05, .15, .25. One can think of p in terms of betting odds. If p = .25, for example, you would be indifferent as to which side you took of a 3 to 1 bet against a defective item, assuming that the wager involves stakes comparable to those involved in a friendly poker game in which the maximum potential loss or gain is not so large as to drastically impair or improve your total assets, yet still large enough for you to take it seriously.
Whichever p is of the four possibilities - .01, .05, .15, .25 - we assume that it will remain constant during the production run which consists of 500 parts.
If we knew that p = .01, which represents the best possible adjustment, we would not tinker with the operator's adjustment. If, on the other hand, we knew that p = .25, we might be tempted to change the adjustment in the hopes of improvement. Suppose there is a master mechanic who can, without fail, put the machine in the best possible adjustment. The time needed by the mechanic to make the necessary adjustment should be valued at $ 10. The problem is to decide whether or not to incur this $ 10 cost.

We shall in this case assume that just two acts or decisions might be taken:
(1) acceptance of the adjustment, that is, do not check it,
(2) rejection of the adjustment, that is, have it checked by the master mechanic.
All the information can be summarized in a payoff table the entries of which show the expected incremental profits or costs for each event-act combination.

Table 1
Payoff Table

Event	Act	
p	Acceptance	Rejection
.01	$ 2*	$ 12
.05	10*	12
.15	30	12*
.25	50	12*

* Best act for given event.

Acceptance is clearly the better act if p = .01 or .05, but rejection is better otherwise, as is indicated by the asterisks in Table 1.
If the event is known, the best decision is obvious, but the problem is a problem because of uncertainty as to which event obtains. Your decision depends on your assessment of the probabilities to be attached to the four possible events. To make that task easy, suppose that there is extensive evidence on the history of the fraction of defective parts in 1000 previous long production runs under similar conditions in the past, and that this history is summarized in Table 2. Other information being judged negligible by comparison, the needed probabilities are assessed by the relative frequencies.

Table 2.

History of Fraction Defectives

Fraction Defective	Relative Frequency
.01	.70
.05	.10
.15	.10
.25	.10
	1.00
(Number)	(1000)

The basic criterion for decision can now be applied: choose that act for which expected cost is lowest (or, for which expected net revenue is highest). For each act, we take costs from Table 1 and probabilities from Table 2, and weight the costs by the probabilities. The expected cost for acceptance is

(.70) \$2 + (.10) \$10 + (.10) \$30 + (.10) \$50 = \$10.40.

Similarly, the expected cost for rejection is

(.70) \$12 + (.10) \$12 + (.10)\$12 + (.10) \$12 = \$12.00.

According to the decision criterion, the better act is to accept. (We assume that a decision must be made without getting more evidence.)

2. Definitions and Concepts

Let us first recall a brief but formal description of decision theory. Start with the payoff table, which gives acts that might be taken, events that might obtain, and utilities for each act-event combination. For simplicity, consider events that can be described by the possible values θ of a discrete-valued parameter $\tilde{\theta}$; the tilde distinguishes the random variable or function from a particular value of the function. Denote any possible act by a. The utility of taking act a if event θ obtains denoted by $U(a,\theta)$.

Strictly speaking, utility is not directly determined by a parameter but rather by things that happen - future observations.

Thus, in the example above, utility is determined by the number r of defective items in a production run of 500 items, not by p.

Hence we really should write U(a,r) instead of U(a,p). But p defines a binomial distribution of $\tilde{r}$, $f_b(r|500,p)$, and expected utility is $\sum_r f_b(r|500,p)U(a,r)$, which we denote by U(a,p).

For simplicity, however, we call U(a,p) - and the general expression U(a, θ) - 'utility' rather than 'expected utility'.

Besides the payoff table we require a (prior) probability distribution of $\tilde{\theta}$, $P(\theta)$.

Assume first that an immediate terminal decision is to be made. For any act a compute its expected utility $\sum_\theta P(\theta)U(a,\theta)$.

Finally, choose that act for which expected utility is maximized. The maximum expected utility is written

(1) $$\max_a \sum_\theta P(\theta)U(a,\theta).$$

(For simplicity, assume in this section that a unique maximum exists in all cases).

Suppose now that sample evidence, represented by the symbol x, is obtained before a terminal decision is made. By application of Bayes' theorem, the prior distribution $P(\theta)$ of $\tilde{\theta}$ becomes the posterior distribution $P(\theta|x)$. Then any act a is evaluated by its expected utility $\sum_\theta P(\theta|x)U(a,\theta,x)$. Choose the act for which this is maximized, and call the maximum expected utility.

(2) $$\max_a \sum_\theta P(\theta|x)U(a,\theta,x).$$

We write $U(a,\theta,x)$ instead of $U(a,\theta)$ to emphasize that it may cost something, directly or indirectly, to obtain x. This cost of sampling is a sunk cost when the final decision is made; that is, it is the same for all a and θ, and so can be either included or ignored without affecting the decision. For the next problem, however, the cost is not yet sunk.

Next in order of complexity, consider a specific sampling plan that promises an observation of the random variable $\tilde{x}$. What is the expected utility of carrying out this sample and then making a terminal decision in the light of $P(\theta|x)$? Work backwards from the solution to the previous problem. The prior distribution $P(\theta)$ in conjunction with the proposed sampling plan implies a predictive distribution P(x) for $\tilde{x}$ in the usual way; that is $P(x) = \sum_\theta P(\theta)P(x|\theta)$, where $P(x|\theta)$ is the conditional distribution of $\tilde{x}$ given θ for the sampling plan. For any x the maximum utility is given by (2), that is $\max_a \sum_\theta P(\theta|x)U(a,\theta,x)$.

Now take the expectation of (2) with respect to P(x):

(3) $$\sum_x P(x)\max_a \sum_\theta P(\theta|x)U(a,\theta,x).$$

This is the expected utility, as seen in advance, of executing the sampling plan in question and then taking the best act after the sample evidence x is available.

Now recognize explicity that $U(a,\theta,x)$ has two components: $U(a,\theta)$, as originally defined (ignoring sampling costs), and an expected cost of sampling (not necessarily measured in monetary units), denoted $C(x)$, where $C(0) = 0$ and $x = 0$ represents the dummy outcome of a sample of size 0, that is, no sample at all. If, as is often reasonable, $U(a,\theta,x) = U(a,\theta) - C(x)$, we can decompose (3) as

$$(3a) \qquad \sum_x P(x) \max_a \sum_\theta P(\theta|x)U(a,\theta) - \sum_x P(x)\sum_\theta P(\theta|x)C(x).$$

The first term of (3a) is the expected utility, ignoring sampling costs, of carrying out the sample plan. The second term which simplifies to $\sum_x P(x)C(x)$ is the expected cost of sampling. The expected value of sample information, EVSI, is defined as (1) subtracted from the first term of (3a):

$$(4) \qquad \sum_x P(x) \max_a \sum_\theta P(\theta|x)U(a,\theta) - \max_a \sum_\theta P(\theta)U(a,\theta).$$

The summation of the second term can be written $\sum_\theta P(\theta)U(a,\theta) = \sum_x P(x)\sum_\theta P(\theta|x)U(a,\theta)$; by substitution in (4) the EVSI can be expressed as

$$(4a) \qquad \sum_x P(x) \max_a \sum_\theta P(\theta|x)U(a,\theta) - \max_a \sum_x P(x)\sum_\theta P(\theta|x)U(a,\theta).$$

From (4a) it is apparent that the EVSI can never be negative.That is, by tailoring the act a to the sample outcome x, we cannot lose expected utility: a posteriori we would always be free to take the act that was best a priori, in which case (4a) would be 0. If in the light of an observed x we choose a different act from the one preferred a priori, we do so only because the expected utility is larger. Moreover, we can gain in expected utility only if some sample outcome x will change the choice of acts. If no sample outcome could change the best a priori decision, the EVSI is 0. In words, the EVSI is the weighted average posterior expected utility, the weights being given by $P(x)$, minus the prior expected utility. The fact that the EVSI can never be negative can be expressed by saying that the weighted average posterior expected utility cannot be less than the prior expected utility. But this does not rule out the possibility that the actual posterior expected utility can be less than the prior. Suppose, for example, that no sample outcome could change the best act, that is, the EVSI is zero. Then unless sampling is completely uniformative--$P(\theta|x) = P(\theta)$ for all x--there will typically be some outcomes for thich posterior expected utility $\sum_\theta P(\theta|x)U(a,\theta)$ is lower than prior expected utility $\max_a \sum_\theta P(\theta)U(a,\theta)$, and others for which it is higher.

Finally, examine the problem of choice of a sampling plan or, as it is often called, the problem of sample design. For all proposed sampling plans-- all methods of drawing the sample, all sample sizes--compute (3a) and choose that plan for which the result is largest, that is, the expected value of sample information minus the expected cost of sampling is the greatest.

Among the sample plans contemplated, there is a dummy plan in which sample size is zero, that is, no sample at all is taken. For this dummy plan, (3) reduces to (1). (Alternatively, we could avoid reference to a dummy plan by saying, "choose the sampling plan for which the expected value of sample information exceeds expected cost of sampling by the largest amount, assuming that this difference is nonnegative. Otherwise the optimal act is to take no sample at all, and make an immediate terminal decision." That is, there is no point in taking a sample if even the best sampling plan has an expected cost in excess of its expected value).

The problem solved by (2), which is called terminal analysis, is simpler than the problem of sample design solved by (3) or (3a). For analysis, we start with the given $\tilde{x} = x$. The fact that $\tilde{x}$ might have exhibited other values is of no interest, either in calculating $P(\theta|x)$ or in carrying out (2). Given the payoff table and prior distribution we must make the inferential step that carries from $P(\theta)$ to $P(\theta|x)$, and carry out the expected value computation of (2) for each act. For evaluating even one proposed design by (3), we have to do this for every possible x, and also calculate $P(x)$ and take an expectation over all possible values that might be exhibited by $\tilde{x}$. There are special devices for specific problems that can make both analysis and design less cumbersome than this abstract description makes them sound. Moreover, the analytical framework may be of value even when it cannot be carried out completely, that is, when informal analysis is used to extend a partial formal analysis to a final decision.

In the remainder of the paper we examine special cases.

3. Two-Action Problems with Linear Utilities

In the machine-setup problem of Sec. .1, the conditional utilities (negative of costs) for each act were linear functions of p. If a_1 denotes rejection, we write

$$U(a_1,p) = -12 + 0.p = -12. \tag{1}$$

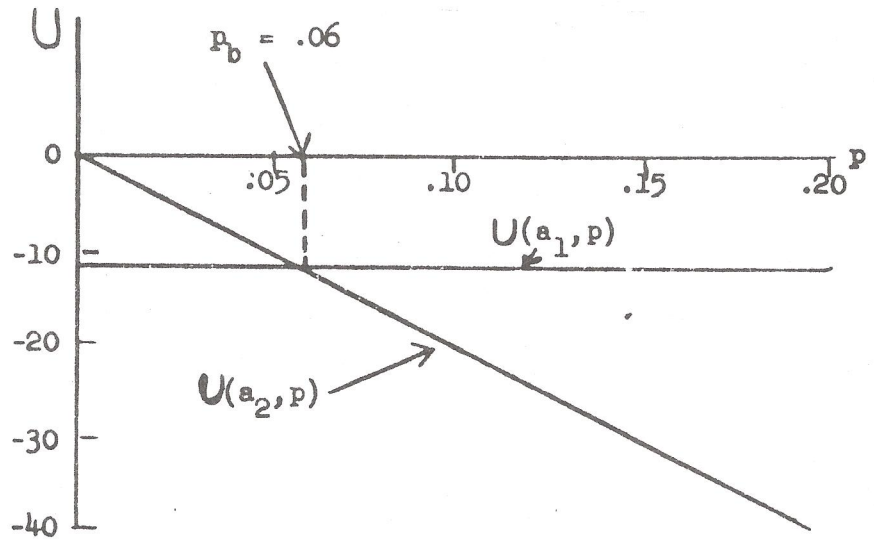
If a_2 denotes acceptance, we write

$$U(a_2,p) = 0 - 200\,p = -200\,p. \tag{2}$$

Each utility function is a linear function of p, as shown in Fig. 1. The horizontal coordinate of the point of intersection of these two curves, $p = p_b = .06$, has a significance that we shall see in a moment.

Fig. 1

Utility Functions for Machine-Setup Problem



To choose a terminal action in the light of a prior distribution P(p), apply Sec. 1 (1), to compute expected utility for a_1 and a_2 in the light of this distribution. Recalling the properties of expectation,

(3) $$E\,U(a_1,\tilde{p}) = -12 + 0 \cdot E(\tilde{p}) = -12$$

and

(4) $$E\,U(a_2,\tilde{p}) = 0 - 200\,E(\tilde{p}).$$

For the prior distribution of Table 2, $E(\tilde{p}) = .052$, so $E\,U(a_2,\tilde{p}) = -10.40$. Since -10.40 exceeds -12, a_2 is indicated.

In comparing (3) with (1) and (4) with (2), we see that in each case p has been replaced by $E(\tilde{p})$. The expected utilities of (3) and (4) are evaluated by treating $E(\tilde{p})$ as if it were p. This is possible because the utility functions $U(a_1,p)$ and $U(a_2,p)$ are linear in p.

We can view the analysis even more simply. From (1) and (2), the breakeven or indifference point p_b between values of p indicating a_1 and those indicating a_2 occurs when $-12 = -200\,p_b$, or $p_b = .06$. We prefer a_1 if $p>.06$, a_2 if $p<.06$. Since we can treat $E(\tilde{p}) = .052$ as if it were p, we see that a_2 is preferred since $.052<.06$.

The analysis would proceed in the same way for a posterior distribution $P(p|r,n)$.

In general, suppose that there are two acts 1 and 2 with conditional utilities given by $U(a_1,\theta) = c_1 + b_1\theta$, and $U(a_2,\theta) = c_2 + b_2\theta$, where the c's and b's are real numbers, $b_1>b_2$, and θ is a symbol for a value taken by an uncertain parameter $\tilde{\theta}$. We want to choose a_1 if $E\,U(a_1,\tilde{\theta}) > E\,U(a_2,\tilde{\theta})$; otherwise, unless the two expectations are equal and either act is optimal, we want to choose a_2. Since both utility functions are linear functions of θ, making use of the expectation to obtain $E\,U(a_1,\tilde{\theta}) = c_1 + b_1\,E(\tilde{\theta})$ and $E\,U(a_2,\tilde{\theta}) = c_2 + b_2\,E(\tilde{\theta})$.

Therefore choose a_1 if $c_1 + b_1 E(\tilde{\theta}) > c_2 + b_2 E(\tilde{\theta})$, that is, if

$$(5) \qquad E(\tilde{\theta}) > \frac{c_2 - c_1}{b_1 - b_2} = \theta_b .$$

We use θ_b to denote the breakeven point; we prefer a_1 to a_2, are indifferent to, or prefer, a_2 to a_1, according as $E(\tilde{\theta}) > \theta_b$, $E(\tilde{\theta}) = \theta_b$, or $E(\tilde{\theta}) < \theta_b$. (it must be assumed that θ_b lies within the interval of values of θ that $\tilde{\theta}$ can exhibit; otherwise one act is dominated by the other and can be discardedfrom consideration without regard for the probability distribution. If $b_1 < b_2$, the direction of the inequality (5) is reversed).

The direct calculation of expected utilities for each act is more laborious. It does not exploit the fact that the mean is the only fact about a distribution--prior or posterior--needed for a terminal decision. This follows from the assumed linearity of the conditional utility functions. Since for linear utilities the mean is just as good as the entire distribution for the choice of a terminal act, we speak of the mean as a certainty equivalent. Whenever for a specific decision we can replace a probability distribution by a number computed from that distribution, that number serves as a certainty equivalent. There is a parallel between the concept of certainty equivalent and that of sufficient statistic. A sufficient statistic tells all we need to know about a sample in order to reach the same posterior distribution that would have been obtained from analysis of all the sample evidence. A certainty equivalent tells all we need to know about a distribution in order to reach the same decision that would have been indicated by a direct analysis of the entire distribution.

Two key quantities must be computed for the choice of a terminal action in a two-action problem with linear utilities: $E(\tilde{\theta})$ and θ_b. The first comes from the probability distribution and has nothing to do with utility; the second comes from the utilities and has nothing to do with probability.

It is meaningless for the purpose of terminal decision to inquire about the standard error of a certainty equivalent. Suppose, for illustration, that the utilities are linear in μ, the mean of a normal process, and that the breakeven utility is μ_b. Suppose also that the prior distribution of $\tilde{\mu}$ is diffuse, so that the mean of the approximate posterior distribution is $\bar{x}$. Then the decision is made simply by comparing $\bar{x}$ with μ_b. The standard error of $\tilde{\mu}$ --$\sigma/\sqrt{n}$ or $s/\sqrt{n}$ -- is irrelevant.

We now turn to the more complicated problem of evaluating the expected value of sample information, the EVSI. Repeating eqs. (4) and (4a) of Sec. 2, for convenience, for the discrete case the EVSI is

$$(6) \qquad \sum_x P(x) \max_a \sum_\theta P(\theta|x)\, U(a,\theta) - \max_a \sum_\theta P(\theta)\, U(a,\theta),$$

or

$$(7) \qquad \sum_x P(x) \max_a \sum_\theta P(\theta|x)\, U(a,\theta) - \max_a \sum_x P(x) \sum_\theta P(\theta|x)\, U(a,\theta).$$

In the application of this section, two special features facilitate evaluation of (6) of (7). First, there are just two acts, a_1 and a_2. Supposing for concreteness that a_1 is optimal a priori, then for all x such that a_1 is still optimal a posteriori, the corresponding terms of (7) are zero.

We may therefore rewrite (7) as

$$(8) \quad \sum_{\{x:a_2 opt\}} P(x) \sum_\theta P(\theta|x)\, U(a_2,\theta) - \sum_{\{x:a_2 opt\}} P(x) \sum_\theta P(\theta|x)\, U(a_1,\theta)$$
$$= \sum_{\{x:a_2 opt\}} P(x) \sum P(\theta|x) \left[U(a_2,\theta) - U(a_1,\theta)\right] .$$

The second special feature is the linearity of the utility functions $U(a_1,\theta)$ and $U(a_2,\theta)$, which implies

$$(9) \qquad U(a_2,\theta) - U(a_1,\theta) = (c_2 - c_1) + (b_2 - b_1)\theta$$
$$= \frac{c_2 - c_1}{b_1 - b_2} (b_1 - b_2) + (b_2 - b_1)\theta$$
$$= (b_1 - b_2)(\theta_b - \theta).$$

Substituting (9) in (8) we obtain the EVSI

$$(10) \quad \sum_{\{x:a_2 opt\}} P(x) \sum_\theta P(\theta|x)(b_1 - b_2)(\theta_b - \theta) = (b_1 - b_2) \sum_{\{x:a_2 opt\}} P(x)(\theta_b - E(\tilde{\theta}|x)),$$

where $E(\tilde{\theta}|x) = \sum_\theta \theta P(\theta|x)$.

We can express (10) in a convenient computational form. Each possible x defines a posterior distribution of $\tilde{\theta}$, $P(\theta|x)$, by Bayes' theorem. Denote the mean of this distribution by θ''. Before $\tilde{x}$ is observed θ'' is a random variable, with distribution $P(\theta'')$ induced by $P(x)$. The prior distribution of $\tilde{\theta}$ and conditional distribution of $\tilde{x}$ given θ serve to determine $P(x)$, also $P(\theta|x)$ for each x, and therefore $P(\theta'')$. From the work on posterior terminal analysis, we know that for $b_1 > b_2$, a_2 is optimal if and only if $\theta'' < \theta_b$.

Hence we can rewrite (10) as

$$(11)\quad (b_1 - b_2) \sum_{\theta'' \le \theta_b} P(\theta'')\,(\theta_b - \theta'') = (b_1 - b_2)\left[\theta_b\, P(\tilde{\theta}'' < \theta_b) - \sum_{\theta'' < \theta_b} \theta'' P(\theta'')\right].$$

We have derived (11) on the assumption that a_1 was best a priori. Had we assumed a_2 best, the same reasoning would have led to

$$(11a)\quad (b_1 - b_2)\left[\sum_{\theta'' > \theta_b} \theta'' P(\theta'') - \theta_b\, P(\tilde{\theta}'' > \theta_b)\right].$$

To evaluate (11) numerically, we need to deduce the distribution of $\tilde{\theta}''$, and evaluate $P(\tilde{\theta}'' < \theta_b)$ and $\sum_{\theta'' < \theta_b} \theta'' P(\theta'')$. To illustrate how this is done we consider a two-action problem on the mean μ of a normal process of known variance σ^2, with a normal prior distribution for $\tilde{\mu}$, $f_N(\mu|\bar{x}', \sigma/\sqrt{n'})$. Using $\bar{x}''$ to denote the posterior mean, the counterpart of the left-hand side of (11) is

$$(12)\quad (b_1 - b_2) \int_{-\infty}^{\mu_b} (\mu_b - \bar{x}'')\, D(\bar{x}'')\, d\bar{x}''.$$

We wish to deduce $D(\bar{x}'')$. A sample of size n is considered. The predictive density for $\tilde{\bar{x}}$ is therefore $f_N(\bar{x}|\bar{x}', \sigma\sqrt{\frac{1}{n'} + \frac{1}{n}})$. For any value $\bar{x}$ that $\tilde{\bar{x}}$ can exhibit, we would be led by the usual formula to a normal posterior density $f_N(\mu|\bar{x}'', \sigma/\sqrt{n''})$. In advance of sampling the only uncertainty about $f_N(\mu|\bar{x}'', \sigma/\sqrt{n''})$ is the uncertainty about $\tilde{\bar{x}}''$. The uncertainty about $\tilde{\bar{x}}''$, in turn, stems from the prior uncertainty about x; that is, $\tilde{\bar{x}}'' = (n'\bar{x}' + n\,\tilde{\bar{x}})/n''$. We see that $\tilde{\bar{x}}''$ is a linear function of the normally-distributed random variable $\tilde{\bar{x}}$, which has predictive density $f_N(\bar{x}|\bar{x}', \sigma\sqrt{\frac{1}{n'} + \frac{1}{n}})$. Therefore the prior distribution of $\tilde{\bar{x}}''$ is normal with mean

$$(13)\quad E(\tilde{\bar{x}}'') = \frac{n'\bar{x}' + n\,\bar{x}'}{n''} = \bar{x}'$$

and variance

$$(14)\quad \sigma^2(\tilde{\bar{x}}'') = \left(\frac{n}{n''}\right)^2 \sigma^2\left(\frac{1}{n'} + \frac{1}{n}\right) = \frac{n}{n''} \cdot \frac{\sigma^2}{n'}.$$

Summarizing, the prior distribution of $\tilde{\bar{x}}''$ is given by the density $f_N(\bar{x}''|\bar{x}', \sigma(\tilde{\bar{x}}''))$, where $\sigma(\tilde{\bar{x}}'') = \sqrt{\frac{n}{n''}}\,\frac{\sigma}{\sqrt{n'}}$. Substituting in (12), we have

$$(15)\quad (b_1 - b_2) \int_{-\infty}^{\mu_b} (\mu_b - \bar{x}'') f_N(\bar{x}''|\bar{x}', \sigma(\tilde{\bar{x}}''))\, d\bar{x}''.$$

Using the substitution $u = (\bar{x}'' - \bar{x}') / \sigma(\tilde{\bar{x}}'')$, $d\bar{x}'' = \sigma(\tilde{\bar{x}}'')du$, and writing

$$\mu_b - \bar{x}'' = \sigma(\tilde{\bar{x}}'')\left[\frac{\mu_b - \bar{x}'}{\sigma(\tilde{\bar{x}}'')} - \frac{\bar{x}'' - \bar{x}'}{\sigma(\tilde{\bar{x}}'')}\right] = \sigma(\tilde{\bar{x}}'')(u_b - u),$$

we can express (15) as

$$(16) \qquad (b_1 - b_2)\,\sigma(\tilde{\bar{x}}'') \int_{-\infty}^{u_b} (u_b - u) f_N(u|0,1)du,$$

where $u_b = (\mu_b - \bar{x}') / \sigma(\tilde{\bar{x}}'')$. The integral of (16) is easily evaluated as $u_b F_N(u_b|0,1) + f_N(u_b|0,1)$. In conclusion, the EVSI for the normal case, assuming a_1 best a priori, is

$$(17) \qquad (b_1 - b_2)\,\sigma(\tilde{\bar{x}}'')\left[u_b F_n(u_b|0,1) + f_N(u_b|0,1)\right].$$

Had a_2 been best a priori, the same argument would have led to

$$(17a) \qquad |b_1 - b_2|\,\sigma(\tilde{\bar{x}}'')\left[f_N(u_b|0,1) - u_b G_N(u_b|0,1)\right].$$

Both cases, and the corresponding result when it is assumed that $b_2 > b_1$, can be subsumed under

$$(17b) \qquad |b_1 - b_2|\,\sigma(\tilde{\bar{x}}'')\left[f_N(u_b|0,1) - |u_b| G_N(|u_b||0,1)\right].$$

The expression in brackets in (17b) can be evaluated numerically by Tables II and III of Schlaifer (1959). Alternatively, Table IV, Schlaifer (1959), gives it as a single function called G(D), where $D = |u_b|$.

In this notation the EVSI is

$$(17c) \qquad |b_1 - b_2|\,\sigma(\tilde{\bar{x}}'')\,G(|u_b|).$$

By examining (17c) we observe:

(1) The EVSI is directly proportional to the absolute difference of the slopes of the utility functions, $|b_1 - b_2|$.

(2) One effect of increased uncertainty, as measured by the standard deviation $\sigma(\tilde{\bar{x}}'')$, is a proportional increase in the EVSI. A second effect of increased uncertainty is a decrease of $|u_b| = |\mu_b - \bar{x}'|/\sigma(\tilde{\bar{x}}'')$ in inverse proportion. $G(|u_b|)$ is a decreasing function of $|u_b|$, so this effect also increases the EVSI. In summary, increased dispersion of the distribution of $\tilde{\bar{x}}''$ entails a larger EVSI, other things the same.

(3) $|u_b|$ is directly proportional to the distance between the breakeven point μ_b and the prior mean $\bar{x}'$. Since $G(|u_b|)$ is a decreasing function of $|u_b|$, the EVSI decreases as this distance increases, other things the same.

It is interesting to examine what happens to the EVSI as $n \to \infty$.

Since $\sigma^2(\tilde{\bar{x}}'') = \frac{n}{n''}\frac{\sigma^2}{n'}$, the limiting standard deviation is $\frac{\sigma}{n'}$ since $\frac{n}{n''} \longrightarrow 1$ as $n \to \infty$. The limiting distribution of $\tilde{\bar{x}}''$ is thus normal with mean $\bar{x}'$ and standard deviation $\sigma/\sqrt{n'}$. It is, in fact, the prior distribution of $\tilde{\mu}$. Equating $n \to \infty$ with perfect information, we can call this limiting EVSI the expected value of perfect information, or EVPI.

In general, the EVPI gives the maximum EVSI, and so sets an upper bound on the amount we would be willing to pay for sample information. If the EVPI is smaller than the cost of even a small sample, this upper bound gives the useful information that sampling is unlikely to be worthwhile. If the EVPI is large compared to the cost of a small sample, detailed investigation of the EVSI may be warranted.

In the two-action problem on a normal mean, the EVPI is obtained from (17c) by replacing $(\bar{x}'')$ by the prior standard deviation $\sigma(\tilde{\mu}) = \sigma/\sqrt{n'}$:

$$(18) \qquad |b_1 - b_2|(\sigma/\sqrt{n'})\, G(|u_b|), \text{ where } u_b = \frac{|\mu_b - \bar{x}'|}{\sigma/\sqrt{n'}} .$$

An interpretation of EVPI in terms of loss functions can be based on (15); for $n \to \infty$, we can replace $\bar{x}''$ by μ to obtain:

$$(19) \qquad (b_1 - b_2) \int_{-\infty}^{\mu_b} (\mu_b - \mu) f_N(\mu|\bar{x}', \sigma/\sqrt{n'})\, d\mu.$$

Assuming that $b_1 - b_2 > 0$ and that a_1 is best a priori, the loss function for a_1 and $\mu < \mu_b$ is

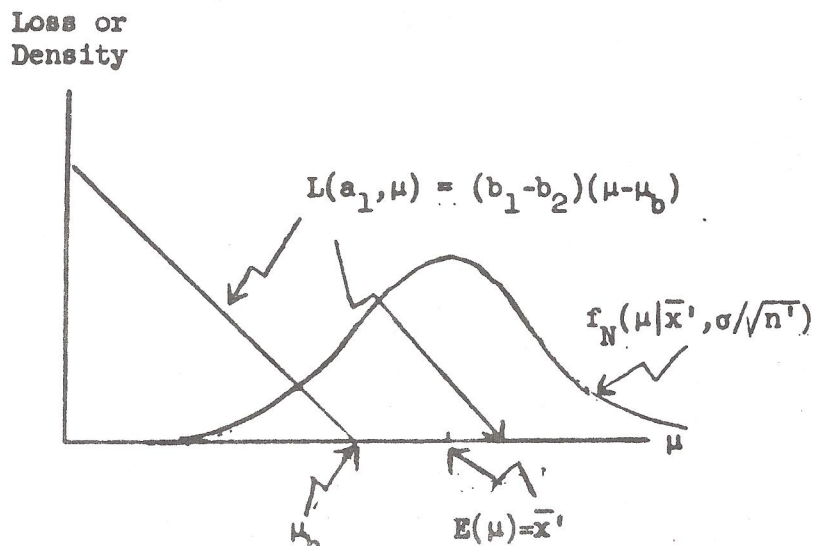
$$(20) \qquad \begin{aligned} L(a_1,\mu) &= U(a_1,\mu) - U(a_2,\mu) \\ &= (c_1 - c_2) + (b_1 - b_2)\mu \\ &= (b_1 - b_2)(\mu - \mu_b) \end{aligned}$$

since $\mu_b = (c_2 - c_1)/(b_1 - b_2)$. For $\mu > \mu_b$, a_1 is the desired act, so $L(a_1,\mu) = 0$. Combining (19) and (20), we see that the EVPI is

$$(21) \int_{-\infty}^{\mu_b} L(a_1\mu) f_N(\mu|\bar{x}', \sigma/\sqrt{n'})\, du = (b_1 - b_2) \int_{-\infty}^{\mu_b} (\mu - \mu_b) f_N(\mu|\bar{x}', \sigma/\sqrt{n'})\, d\mu,$$

that is, the expected loss, in the light of the prior distribution, of the act that seems best a priori. Roughly, for each $\mu < \mu_b$, we multiply the loss of taking a_1 by the probability of incurring such a loss, and add up these products for all $\mu < \mu_b$. A geometrical interpretation is given by Fig. 2.

Figure 2

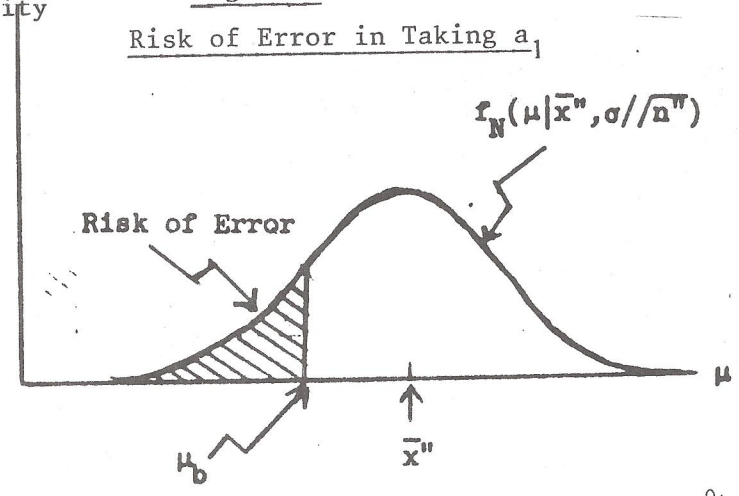


Since a posterior distribution serves as a prior distribution for a new sample, we can also interpret the EVPI by replacing the prior density $f_N(\mu|\bar{x}',\sigma/\sqrt{n'})$ by the posterior density $f_N(\mu|\bar{x}'',\sigma/\sqrt{n''})$. Computed from a posterior distribution, the EVPI serves in exactly the same way as before to set an upper bound on the value of <u>further</u> sample information. Note that while $\sigma/\sqrt{n''}$ is irrelevant for determining the best decision in the light of the posterior distribution, it <u>is</u> needed to compute the EVPI or EVSI. An even cruder indication of the value of further information is given by the <u>risk of error</u> in taking the act that looks best in the light of the posterior density. If $\bar{x}'' > \mu_b$, the risk of error is the <u>tail area</u> of $f_N(\mu|\bar{x}'',\sigma/\sqrt{n''})$ to the left of μ_b:

$$(22) \qquad \int_{-\infty}^{\mu_b} f_N(\mu|\bar{x}'',\sigma/\sqrt{n''})du = F_N(\mu_b|\bar{x}'',\sigma/\sqrt{n''}).$$

This is illustrated in Fig. 3. In terms of EVPI, the risk of error <u>would be</u> the EVPI <u>if</u>, contrary to assumption, $L(a_1,\mu) = 1$ for $\mu < \mu_b$, $L(a_1,\mu) = 0$ for $\mu \geq \mu_b$. In other words, if μ is on the wrong side of the breakeven value μ_b, the actual loss is the same regardless of the distance $|\mu-\mu_b|$. We now give two examples of computation of the risk of error.

<u>Figure 3.</u>



(1) Under the original prior distibution of the machine set-up example, $E(\tilde{p}) = .052$, $p_b = .06$, so acceptance was the best terminal decision. The probability of error is $P(\tilde{p} > p_b) = P(.15) + P(.25) = .10 + .10 = .20$.

(2) Suppose that the approximate posterior distribution of $\tilde{\mu}$ is normal with mean $\bar{x}$ and standard error $s/\sqrt{n}$, and that $\bar{x} > \mu_b$. Hence a_1 is indicated if an immediate terminal act is to be taken. What is the probability that a_1 is really not best? We seek $F_N(\mu_b|\bar{x}, s/\sqrt{n})$, that is, a left-tail area of the posterior distribution of $\tilde{\mu}$. We have

$$F_N(\mu_b|\bar{x}, s/\sqrt{n}) = F_N\left(\frac{\mu_b - \bar{x}}{s/\sqrt{n}}\,|0,1\right).$$

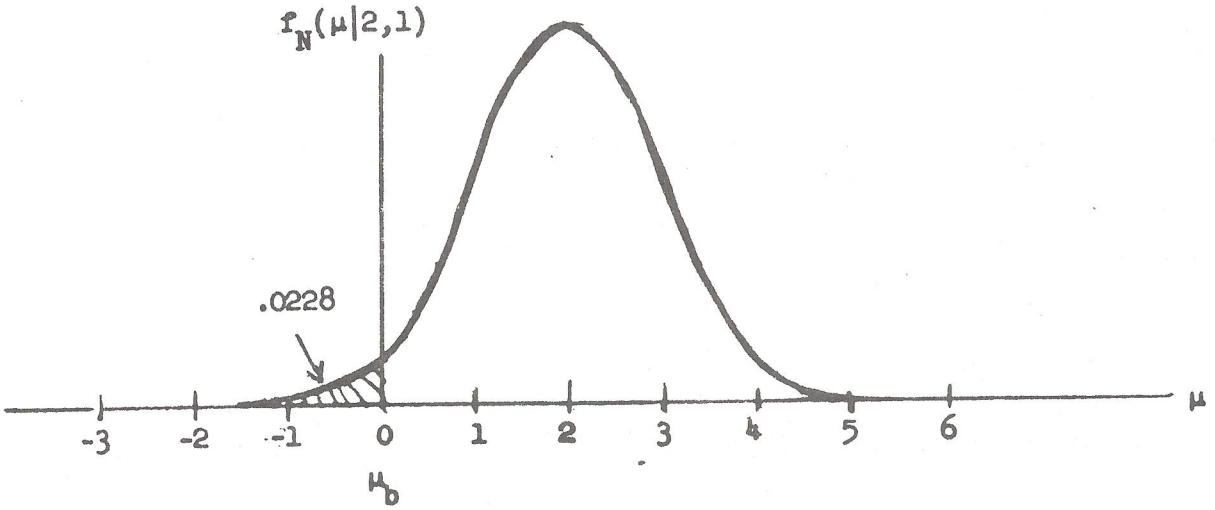
If $\mu_b = 0$, $\bar{x} = 2$, $s = 10$, $n = 100$,, we have

$$F_N\left(\frac{\mu_b - \bar{x}}{s/\sqrt{n}} \,|\, 0,1\right) = F_N\left(\frac{0-2}{10/\sqrt{100}} \,|\, 0,1\right)$$

$$= F_N(-2|0,1) = .0228,$$

Using the Table III, Schlaifer (1959). Geometrically, we have found the left-tail area shown in Fig. 4.

Figure 4



4. Bayesian Point Estimation

Suppose that each value x of a random variable $\tilde{x}$ corresponds with an action that might be taken. For example, $\tilde{x}$ might be the number of perishable items demanded. Each possible x is not only a possible realization of $\tilde{x}$, but it is also a possible stock level of a retailer. How do we determine the best stock level in the light of a probability distribution of $\tilde{x}$? Denote the probability distribution of $\tilde{x}$ by P(x) and a proposed stock level by $\hat{x}$.

The retailer would be happy--incur no opportunity loss--if it happens that the value x exhibited by $\tilde{x}$ is exactly the same as $\hat{x}$. In this event, the number of units demanded would exactly equal the number stocked. If the number demanded exceeds the number stocked, it is assumed that the opportunity loss is the lost profit per unit times the excess demand. If the number demanded falls short of the stock level, it is assumed that the opportunity loss is the cost per unit (less salvage value) times the deficiency of demand. This simple inventory problem has been called the "newsboy problem" because of the following kind of illustration. A newsboy buys each paper for 4 cents, sells for 5 cents. Each unsold copy has salvage value of 1 cent. Hence the loss per unit shortage of stock is 5-4 = 1 cent.

The loss per unit overage of stock is 4-1 = 3 cents. Given a distribution of demand P(x), we shall show that the best stock level is any .25-fractile of the distribution of $\tilde{x}$, the .25 being computed as 1/(3+1) = .25. That is, the best stock level $\hat{x}^*$ ("x-hat star") is any $x_{.25}$ on the distribution of $\tilde{x}$. (If $\tilde{x}$ were continuous, we would say <u>the</u> .25-fractile; discontinuity raises slight complications).

We now give an abstract discussion of the newsboy problem in terms of "point estimation". The discussion applies not only to the newsboy problem but to many others that are formally identical. We couch the exposition in terms of an observable random variable $\tilde{x}$, but the development applies equally well to an unobservable parameter $\tilde{\theta}$ and point estimation thereof. The distribution of $\tilde{x}$ is a predictive distribution: it is not a true, unknown distribution, but a distribution assessed by the person making the point estimate.

Let x denote the best act and $\hat{x}$ an act we can choose--a <u>point estimate</u>. Unless we choose $\hat{x} = x$, we incur a positive opportunity loss. If $\hat{x} > x$,, we describe this as an opportunity loss of overestimation; if $\hat{x} < x$, we describe it as a loss of underestimation. In many problems, of which the newsboy problem is one, the conditional loss function for any proposed $\hat{x}$ can be described by

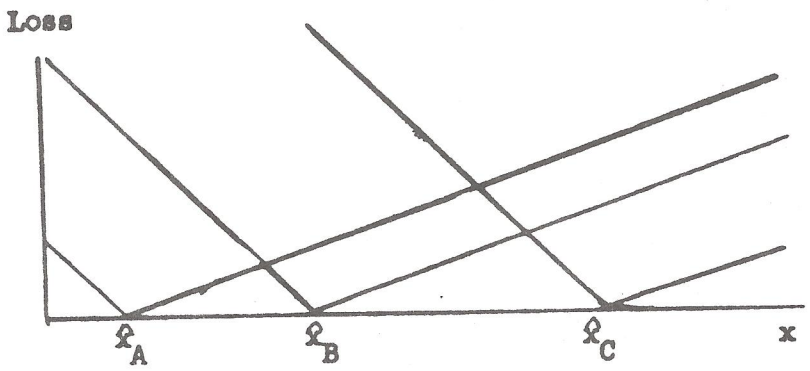
$$(1) \qquad L(\hat{x},x) = k_o(\hat{x} - x) \quad \text{if } \hat{x} \geq x,$$

$$L(\hat{x},x) = k_u(x - \hat{x}) \quad \text{if } \hat{x} \leq x, \qquad k_u,\ k_o \geq 0.$$

The losses are proportional to the overestimate or underestimate, and the constants of proportionality are called k_o and k_u, the subscripts "o" and "u" suggesting "overestimate" and "underestimate".

Geometrically, there is a "V-shaped" loss function for <u>each possible point estimate</u> $\hat{x}$, and any one such function can be obtained from any other by horizontal displacement or translation. Fig. 5 shows schematically three such functions, corresponding with point estimates $\hat{x}_A$, $\hat{x}_B$, and $\hat{x}_C$.

Figure 5



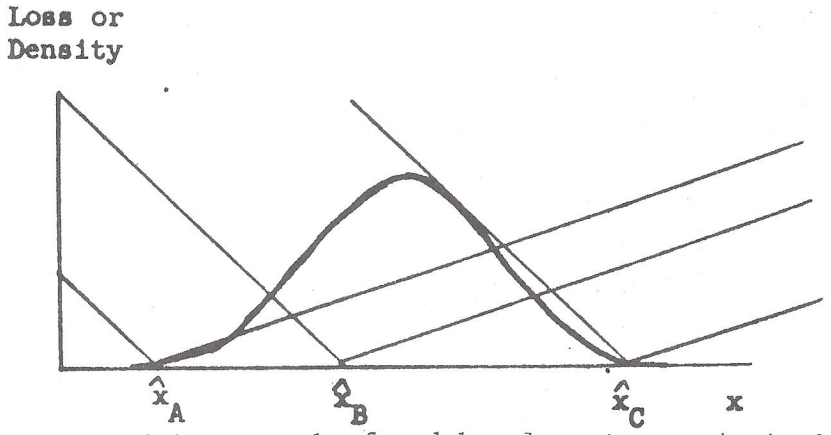
The absolute slope of the left-hand branch is k_o, the loss per unit overestimate; the absolute slope of the right-hand branch is k_u, the loss per unit underestimate. The loss functions of Fig. 5 are so drawn that $k_u = 1/3\ k_o$, as in the numerical example above with k_u = 1 cent and k_o = 3 cents.

The choice of an estimate $\hat{x}$ depends on the assessment of the distribution of $\tilde{x}$, the uncertain best act. In Fig. 6 such a probability distribution--a normal distribution, for illustration--is superimposed on the loss functions of Fig. 5.

Consider first the continuous case. Denoting the density of $\tilde{x}$ by D(x), we define expected loss for a possible point estimate x:

$$(2) \qquad EL(\hat{x},\tilde{x}) = \int_{-\infty}^{\hat{x}} k_o(\hat{x}-x)\ D(x)dx + \int_{\hat{x}}^{\infty} k_u(x-\hat{x})\ D(x)dx.$$

Figure 6



The minimum expected loss can be found by elementary calculus:

$$(3) \qquad \frac{d}{d\hat{x}} E\ L(\hat{x},\tilde{x}) = k_o\left[\hat{x}D(\hat{x}) + P(\tilde{x}< \hat{x}) - \hat{x}D(\hat{x})\right]$$

$$+ k_u\left[-\hat{x}D(\hat{x}) + \hat{x}D(\hat{x}) - (1-P(\tilde{x}< \hat{x}))\right]$$

$$= k_o\ P(\tilde{x}< \hat{x}) - k_u\ (1-P(\tilde{x}< \hat{x}))]$$

$$= (k_u + k_o)\ P(\tilde{x}< \hat{x}) - k_u.$$

We find the minimizing $\hat{x}$, call it $\hat{x}^*$, by setting (3) equal to zero; $\hat{x}^*$ is thus computed from

$$(4) \qquad P(\tilde{x}< \hat{x}^*) = \frac{k_u}{k_u + k_o}\ .$$

(That $\hat{x}^*$ really gives a relative minimum can be seen by noting that the second derivative of (2), $(k_u - k_o)\, D(\hat{x})$, is positive since k_u, k_o, and $D(\hat{x})$ are positive). The best point estimate is the $[k_u/(k_u + k_o)]$ - fractile of the distribution of $\tilde{x}$. For example, in Fig. 6 $\hat{x}_B$ is the best point estimate because $3k_u = k_o$ (remember that k_u and k_o define <u>absolute</u> slopes), whence $k_u/(k_o + k_u) = k_u/(3k_u + k_u) = 1/4$, and $\hat{x}_B$ is the .25 fractile of the distribution of Fig. 6.
The discrete case is more tedious, but it frequently occurs and the argument gives added insight. Begin with the expected loss

$$(5) \qquad E\,L(\hat{x},\tilde{x}) = \sum_{x \leq \hat{x}} k_o(\hat{x}-x)\,P(x) + \sum_{x \geq \hat{x}} k_u(x-\hat{x})\,P(x).$$

We want to find the $\hat{x}$, call it $\hat{x}^*$, that minimizes (5). Consider first a tentative $\hat{x}_1$ that is certainly not too large. Suppose we now contemplate $\hat{x}_2 = \hat{x}_1 + 1$.
The change in expected loss in going from $\hat{x}_1$ to $\hat{x}_2$ will reflect two effects:

(1) For each $x \leq \hat{x}_1$, the new conditional loss of overestimation is k_o units greater than the old, so the first summation will be <u>increased</u> by $\sum_{x \leq \hat{x}_1} k_o\,P(x) = k_o\,P(\tilde{x} \leq \hat{x}_1)$.

(2) For each $x > \hat{x}_1$, the new conditional loss of underestimation is k_u units smaller than the old, so the second summation will be <u>decreased</u> by $\sum_{x > \hat{x}_1} k_u\,P(x) = k_u\,P(\tilde{x} > \hat{x}_1)$. It therefore pays to stop at $\hat{x}_1$ if $k_u\,P(\tilde{x} > \hat{x}_1) < k_o\,P(\tilde{x} \leq \hat{x}_1)$; and, in general, it pays to stop at $\hat{x}_i = \hat{x}_1 + i - 1$ if for the first time $k_u\,P(\tilde{x} > \hat{x}_i) < k_o\,P(\tilde{x} \leq \hat{x}_i)$. Since $P(\tilde{x} > \hat{x}_i) = 1 - P(\tilde{x} \leq \hat{x}_i)$, we may substitute in this inequality and solve for $P(\tilde{x} \leq \hat{x}_i)$:

$$(6) \qquad P(\tilde{x} \leq \hat{x}_i) > \frac{k_u}{k_u + k_o}\ .$$

Thus $\hat{x}_i$ is the desired $\hat{x}^*$ if it corresponds with a jump in the cdf that for the first time takes the height of the cdf above the height $k_u/(k_u + k_o)$.
Thus $\hat{x}^*$ is a $k_u/(k_u + k_o)$ - fractile of the distribution of $\tilde{x}$.
The previous paragraph does not mention the possibilities that before $\hat{x}^*$ is reached there can be one or more preceding values for which expected loss is the same as it is at $\hat{x}^*$. At such a value, say $\hat{x}_{i-1}$, we have $k_u P(\tilde{x} > \hat{x}_{i-1}) = k_o\,P(\tilde{x} \leq \hat{x}_{i-1})$, which implies $P(\tilde{x} \leq \hat{x}_{i-1}) = k_u/(k_u + k_o)$. The sample cdf has height exactly $k_u/(k_u + k_o)$ at $\hat{x}_{i-1}$, so $\hat{x}_{i-1}$ is also a $k_u/(k_u + k_o)$ - fractile: the sample cdf has height $k_u/(k_u + k_o)$ at $\hat{x}_{i-1}$, and jumps higher when $\hat{x}^* = \hat{x}_i$ is reached.

In this kind of problem the $k_u/(k_o + k_u)$ - fractile is a certainty-equivalent: the fractile is all we need to know about the distribution. The point estimate is the best point estimate; "best" not in a universal sense but best for the problem at hand. Moreover, had the x's been transformed into some other scale, such as log x then the logarithm of the point estimate would have been the best estimate in the transformed scale. In general, if $\hat{x}^*$ is the optimal Bayesian estimate of x, then $f(\hat{x}^*)$ is the optimal estimate of f(x); hence Bayesian estimates are said to be invariant. This is true because a Bayesian point estimate is essentially a description of an act, and the best act is in no way altered by changing its description from one language to another.

Thus if 100 units is the best stock level, then the common logarithm of 100, or 2, is the best stock level in units of the common logarithm.

Returning to the continuous case, we consider, as an example of an EVPI calculation, a normal predictive density $f_N(x|\bar{x}'', \sigma\sqrt{\frac{1}{n''} + 1})$. To avoid cumbersome notation we temporarily denote this density as $f_N(x|\mu_p, \sigma_p)$, the subscript "p" suggesting "predictive"; when there is no danger of misunderstanding we shorten this to $f_N(x)$. The EVPI is the expected loss of the best point estimate $\hat{x}^*$. Since $f_N(x|\mu_p, \sigma_p)$ is the density of $\tilde{x}$, we substitute in (2):

$$(7)\quad E\,L(\hat{x},\tilde{x}) = \int_{-\infty}^{\hat{x}^*} k_o(\hat{x}^* - x) f_N(x|\mu_p,\sigma_p)dx + \int_{\hat{x}^*}^{\infty} k_u(x-\hat{x}^*) f_N(x|\mu_p,\sigma_p)dx$$

$$= k_o\left[\hat{x}^* F_N(\hat{x}^*) - \int_{-\infty}^{\hat{x}^*} x\, f_N(x)dx\right] + k_u\left[\int_{\hat{x}^*}^{\infty} x\, f_N(x)dx - \hat{x}^*(1-F_N(\hat{x}^*))\right].$$

We now show how to evaluate the integrals on the right hand side. First,

$$\int_{-\infty}^{\hat{x}^*} x\, f_N(x)dx = \sigma_p \int_{-\infty}^{\hat{x}^*} \left(\frac{x-\mu_p}{\sigma_p}\right) \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu_p}{\sigma_p}\right)} \frac{dx}{\sigma_p} + \mu_p F_N(\hat{u}^*|0,1).$$

With the substitution $u = (x - \mu_p)/\sigma_p$, and the further substitution $t = \frac{1}{2}u^2$, and using the notation $\hat{u}^* = (\hat{x}^* - \mu_p)/\sigma_p$, we conclude

$$(8)\qquad \int_{-\infty}^{\hat{x}^*} x\, f_N(x)dx = -\sigma_p f_N(\hat{u}^*|0,1) + \mu_p F_N(\hat{u}^*|0,1).$$

The second integral of (7) is μ_p - (8), so we have

$$(9)\qquad \int_{\hat{x}^*}^{\infty} x\, f_N(x)dx = \sigma_p f_N(\hat{u}^*|0,1) + \mu_p(1 - F_N(\hat{u}^*|0,1)).$$

Substituting (8) and (9) into (7) and rearranging, we write

(10) $E\ L(\hat{x},\tilde{x}) = (k_u + k_o)\sigma_p f_N(\hat{u}^*|0,1) + \hat{u}^*\sigma_p[k_o F_N(\hat{u}^*|0,1) - k_u(1-F_N(\hat{u}^*|0,1)]$.

but $\hat{u}^*$ is determined so that the expression in brackets is zero, so we conclude

(11) $E\ L(\hat{x},\tilde{x}) = (k_u + k_o)\sigma_p f_N(\hat{u}^*|0,1)$, $\qquad \hat{u}^* = \dfrac{\hat{x}^* - \mu_p}{\sigma_p}$.

The EVPI is proportional to the sum of the unit losses of overestimation and underestimation. The larger the standard deviation of the predictive distribution, the larger the EVPI: (1) σ_p appears as a factor in the formula; (2) for a given $\hat{x}^* - \mu_p$, the standardized normal density increases as σ_p increases unless $\hat{x}^* = \mu_p$, that is, unless the median is the best point estimate. Finally, for fixed $k_u + k_o$, the EVPI is smaller as $\dfrac{k_u}{k_u + k_o}$ departs from $\frac{1}{2}$, since this increases $|\hat{x}^* - \mu_p|$ and hence the normal density decreases.

A word about the meaning of EVPI with respect to a predictive distribution: the "perfect information" in question refers to x, the realization of $\tilde{x}$. If we knew this, we would set $\hat{x}^* = x$ and incur no opportunity loss. An infinite sample from the normal process would remove uncertainty about the parameters $\tilde{\mu}$ and $\tilde{\sigma}$ of the process, but not about $\tilde{x}$: the predictive density $f_N(x|\mu_p,\sigma_p) = f_N(x|\bar{x}'',\sigma\sqrt{\frac{1}{n''} + 1})$ would become $f_N(x|\mu,\sigma)$, but we would be uncertain about $\tilde{x}$.

So far the loss of underestimation of overestimation has been taken as proportional to the amount of underestimate or overestimate. Consider next the loss function

(12) $$L(\hat{x},x) = k(\hat{x} - x)^2 ,$$

where k is a positive constant. In words, the loss of $\hat{x}$ at x is proportional to the square of the discrepancy between $\hat{x}$ and x. Roughly, small discrepancies are not serious, but big ones extremely serious. It can be shown that $\hat{x}^*$--the $\hat{x}$ for which expected loss is least--is the mean or expected value of the probability distribution. We show the development for the continuous case, letting D(x) be an arbitrary predictive density for which the mean exists. Since $E\ L(\hat{x},\tilde{x}) = \int_{-\infty}^{\infty} k(\hat{x}-x)^2\ D(x)dx$, we have

(13) $$\frac{d}{d\hat{x}}\ E\ L(\hat{x},\tilde{x}) = 2k \int_{-\infty}^{\infty} (\hat{x}-x)\ D(x)dx.$$

Setting (13) equal to zero, we see that $\hat{x}^*$, the minimizing value of $\hat{x}$, is

(14) $$\hat{x}^* = \int_{-\infty}^{\infty} xD(x)dx = E(\tilde{x}).$$

The mean serves as a certainty equivalent in two different decision problems: The two-action problem with linear utilities (Sec. 3) and the point estimation problem with loss proportional to the square of the discrepancy of the point estimate.

The EVPI for $L(\hat{x},x) = k(\hat{x}-x)^2$ is easily calculated. For any predictive distribution for which the mean $E(\tilde{x})$ exists we have, remembering that $\hat{x}^* = E(\tilde{x})$,

$$(15) \qquad E\,L(\hat{x}^*,\tilde{x}) = kE[\hat{x}^* - \tilde{x}]^2 = kE[\tilde{x} - E(\tilde{x})]^2 = k\sigma^2(\tilde{x}),$$

where $\sigma^2(\tilde{x})$ is the variance of the predictive distribution of $\tilde{x}$. The EVPI is simply k times the variance.

So far we have considered point estimation for an immediate decision problem. Now examine another kind of point estimation problem in which the decision is at least one step removed from the point estimation. To illustrate, suppose that we have a random sample of n from a normal process with known variance σ^2. The unknown mean μ is of interest because we need to know μ, and others parameters as well, in a subsequent decision problem. Since μ is unknown, we can assess a joint distribution for $\tilde{\mu}$ and the other parameters.

But the use of the entire joint distribution in the subsequent decision problem may unduly complicate that analysis. In order to simplify the analysis we might replace the marginal distribution of $\tilde{\mu}$ by a single number $\hat{\mu}$ and act as if μ were $\hat{\mu}$ Ideally $\hat{\mu}$ would be a strict certainty equivalent. Under some conditions it may be possible to find such a certainty equivalent. For example, if the utility of each possible act in the final problem is linear in μ, and if $\tilde{\mu}$ is independent of all the other random variables in the final problem, then $E(\tilde{\mu})$, the expectation of the marginal distribution of $\tilde{\mu}$, is a strict certainty equivalent. A special case of this is given by the two-action problem of Sec.10-3 . Often, however, the substitution of a point estimate $\hat{\theta}$ for a distribution of $\tilde{\theta}$ does not lead to the same result as a full analysis: a strict certainty equivalent cannot be found. Even so, the practical difficulty of a complete analysis may be so great that we take the short-cut anyway. We still seek the "best" point estimate $\hat{\theta}^*$ of θ. The criterion of "best" is as follows,

Assess subjectively an <u>estimation loss function</u> $L(\hat{\theta}, \theta)$ that depends for any proposed $\hat{\theta}$ on the discrepancy between $\hat{\theta}$ and θ. This loss function together with the distribution of $\tilde{\theta}$ permits calculation of expected loss for the proposed $\hat{\theta}$. Then pick that $\hat{\theta}$, say $\hat{\theta}^*$, for which expected loss is least.

<u>Formally</u> the problem is identical with those we have been discussing. For example, if we have an estimation loss function of the form

$$(16) \qquad L(\hat{\theta},\theta) = k_o(\hat{\theta} - \theta) \qquad \text{if} \quad \hat{\theta} \geq \theta$$

$$= k_u(\theta - \hat{\theta}) \qquad \text{if} \quad \hat{\theta} \leq \theta,$$

$\hat{\theta}^*$ is the $k_u/(k_u + k_o)$ fractile of the distribution of $\tilde{\theta}$. For an estimation loss function of the form $L(\hat{\theta},\theta) = k(\hat{\theta} - \theta)^2$, $\hat{\theta}^*$ is the mean of the distribution of $\tilde{\theta}$.

The difference between the present problem and the earlier one lies in the difficulty of assessing loss functions. If the point estimation and the ultimate decision are identical, as in the newsboy problem, the loss function follows immediately from the decision problem. But if the ultimate decision is one or more steps removed from the point estimation, it may be very hard to compute the estimation loss function implied by the ultimate decision problem, even when the loss structure of the ultimate problem is clearly defined.

And often the ultimate decision problem may be ill-defined, as in Schlaifer's illustration of the soap manufacturer (Sec. 37.7 of Schlaifer (1959)), who needed to form some judgement as to the total number of automatic dishwashers as one component of a decision about development and marketing of a special detergent for dishwashers. We can always assess the estimation loss function informally without attempting to make a full formal analysis of the ultimate decision problem, as is also illustrated by Schlaifer's soap manufacturer. The manafacturer was willing to make the judgement that a loss associated with an error in estimating the number of dishwashers was proportional to the size of the error, regardless of the direction of the error. For a normal posterior distribution for the number of dishwashers, this implies that the .50 fractile, which is also the mean, is the best point estimate.

The manufacturer may be willing to go one step further and assess the constant of proportionality in his loss function. He may, as Schlaifer has him do, assess the loss of a unit error in either direction as \$0.02. That is, he would be willing to spend \$0.02 te reduce the error in an estimate by one unit, regardless of whether the estimate was high or low. This, then, defines the EVPI of an actual point estimate in the normal case by substitution in (11), where x now represents the total number of dishwashers.

The soap manuafcturer's problem is symbolic of many problems in statistics. People want point estimates as an aid to some ultimate decision problem, perhaps ill-defined.

They are willing to judge that the losses of overestimation or underestimation are symmetrical (and not so badly behaved as to make the expected loss integrals or summations diverge). The prior distributions are diffuse, the posterior distributions are approximately normal. Then the $\bar{x}$'s or r/n's are the desired pointestimates: they are numbers people can carry around in their heads and treat as if they were the"true" numbers. The uncertainties attached to these numbers, measured say by their standard errors, are irrelevant except for two circumstances:

(1) The ill-defined decision problem becomes well-defined and the entire posterior distribution is needed.

(2) Whether or not the decision problem becomes well-defined, the question is raised as to whether it is worthwhile to spend more money to reduce the uncertainties of the estimates.

In complicated problems point estimates will often be used as a practical matter even though such a satisfactory rationale for their choice is lacking. In such problems procedures of sampling-theory point estimation may be useful in suggesting rule-of-thumb approximations.

We conclude this section with a point of terminology that can be made by an example. Suppose that we are sampling from a normal process with known variance. Under many circumstances, $\bar{x}$ will be used as an estimate of the unknown μ. Frequently, however, we may wish to speak of the random variable $\tilde{\bar{x}}$ instead of its particular realization $\bar{x}$. For example, conditional on μ and σ, $\tilde{\bar{x}}$ has a sampling distribution.

We then refer to $\tilde{\bar{x}}$ as a point estimator, while $\bar{x}$ is a point estimate. The usefulness of this distiction can be seen when we study sampling-theory point estimation.

5. Classification, Diagnosis, or Discrimination: Simple Dichotomy

Another type of decision problem is a special case of the two-action problem of Sec. 3. We illustrate this by a cancer diagnosis example.

Suppose that for some reason it is possible to make only one test; perhaps the test always gives the same answer, right or wrong, when repeated on the same individual. The present problem is this: given the result of the test, + or −, should we diagnose that the person has cancer? By "diagnose that the person has cancer", a_C, we mean "pursue a further course of diagnosis or treatment". By "diagnose not cancer," $a_{\bar{C}}$, we mean simply to take no further action.

In this problem there are just two, incompatible events: "has cancer," (C) and "does not have cancer" ($\bar{C}$), hence the expression simple dichotomy. We define the losses as follows:

	Act	
Event	a_C	$a_{\bar{C}}$
C	0	$L(a_{\bar{C}},C)$
$\bar{C}$	$L(a_C,\bar{C})$	0

$L(a_C,\bar{C})$ is the loss of a false positive diagnosis; $L(a_{\bar{C}},C)$ is the loss of a false negative. Suppose now that the test says +. By application of Bayes' theorem, the expected loss of a_C is

$$(1) \qquad \frac{P(C)P(+|C)}{P(+)} \cdot 0 + \frac{P(\bar{C})P(+|\bar{C})}{P(+)} L(a_C, \bar{C}) = \frac{P(\bar{C})P(+|\bar{C})}{P(+)} L(a_C,\bar{C}).$$

The expected loss of $a_{\bar{C}}$ is

$$(2) \qquad \frac{P(C)P(+|C)}{P(+)} \cdot L(a_{\bar{C}},C) + \frac{P(\bar{C})P(+|\bar{C})}{P(+)} \cdot 0 = \frac{P(C)P(+|C)}{P(+)} L(a_{\bar{C}},C).$$

We diagnose cancer, take a_C, if (1) < (2), that is, if

$$(3) \qquad \frac{L(a_{\bar{C}},C)}{L(a_C,\bar{C})} > \frac{P(\bar{C})P(+|\bar{C})}{P(C)P(+|C)} = \frac{P(\bar{C}|+)}{P(C|+)} .$$

In the numerical example let $P(\bar{C}) = .995$, $P(C) = .005$ and $P(+|\bar{C}) = .05$, $P(+|C) = .95$. Therefore the right hand side of (3) is

$$\frac{(.995)\ (.05)}{(.005)\ (.95)} = \left(\frac{199}{1}\right)\ \left(\frac{1}{19}\right) = 10.47.$$

This means that $L(a_{\bar{C}},C)$ must be at least 10.47 times $L(a_C,\bar{C})$ to warrant the diagnosis a_C; that is, the consequences of ignoring cancer when it is present must be at least 10.47 times as serious as the consequences of further diagnostic testing or treatment if cancer is not present.

The inequality (3) can also be written, diagnose a_C if

$$(4) \qquad \frac{P(+\,|C)}{P(+|\bar{C})} > \frac{P(\bar{C})}{P(C)} \cdot \frac{L(a_C,\bar{C})}{L(a_{\bar{C}},C)} \quad .$$

The factor $P(\bar{C})/P(C)$ is the prior odds ratio against cancer. The factor $L(a_C,\bar{C})/L(a_{\bar{C}},C)$ is the ratio of the loss of false positive diagnosis to that of false negative. The ratio $P(+|C)/P(+|\bar{C})$ is the <u>likelihood</u> ratio, the ratio of the data given C to the probability of the data given $\bar{C}$. In order to diagnose a_C, the likelihood ratio in <u>favor of</u> C must exceed the product of the prior odds ratio against C and the loss ratio of false positive diagnosis to that of false negative: (4) provides the criterion for terminal analysis.

In general, there are two incompatible events E_1 and E_2, and acts a_1 and a_2 such that $L(a_1,E_2)$, $L(a_2,E_1) > 0$, $L(a_1,E_1) = L(a_2,E_2) = 0$.

Given any amount of data x bearing on the problem, choose a_1 if

$$(5) \qquad \frac{P(x|E_1)}{P(x|E_2)} > \frac{P(E_2)}{P(E_1)} \cdot \frac{L(a_1,E_2)}{L(a_2,E_1)} \quad .$$

or

$$\frac{P(E_1|x)}{P(E_2|x)} > \frac{L(a_1,E_2)}{L(a_2,E_1)} \quad .$$

The EVPI is simple; for a_1, it is

$$(6) \qquad P(E_2|x)\ L(a_1,E_2).$$

for a_2, the EVPI is

$$(7) \qquad P(E_1|x)\ L(a_2,E_1).$$

In (6) and (7), $P(E_1|x)$ and $P(E_2|x)$ are the posterior probabilities computed by Bayes' theorem. For example,

$$(8) \qquad P(E_1|x) = \frac{P(E_1)\ P(x|E_1)}{P(x)} \quad .$$

Consider an application involving the normal distribution. On the basis of a large scale survey of men in a certain age group, it is found that blood cholesterol counts are normally distributed both for those who do and those who do not subsequently have heart attacks in a five year period.

For the first group (C, for coronary) the mean is 268 and the standard deviation is 50. For the second group ($\bar{C}$, for non-coronary) the mean is 268 and the standard deviation is 50. The prior probability of a heart attack within the stated period is well established by frequency evidence to be .05. A particular man in this age group has a cholesterol reading of 275. First compute the odds that he will have a heart attack:

$$\frac{.05f_N(275|268,\ 50)}{.95f_N(275|248,\ 50)} = \frac{1}{19} \cdot \frac{f_N(\frac{275 - 268}{50}\ |0,1)/50}{f_N(\frac{275 - 248}{50}\ |0,1)/50} = \frac{1}{19} \cdot \frac{f_N(.14|0,1)}{f_N(.54|0,1)}$$

$$= \frac{1}{19} \cdot \frac{.3951}{.3448} = .063$$

$$= \frac{2}{33} .$$

Suppose that it is possible, with some expense and trouble, to carry around pills that may be useful in mitigating the effects of a heart attack if one comes. In order to justify carrying the pills, the expected loss of carrying the pills if not needed would have to be no more than 2/33 the expected loss of not having the pills if they are needed.

In this example, μ_C, $\mu_{\bar{C}}$, σ_C, and $\sigma_{\bar{C}}$ are assumed to be well determined by $\bar{x}_C$, $\bar{x}_{\bar{C}}$, s_C, and $s_{\bar{C}}$ of large samples. If there is substantial uncertainty about $\tilde{\mu}_C$, $\tilde{\mu}_{\bar{C}}$, $\tilde{\sigma}_C$, and $\tilde{\sigma}_{\bar{C}}$, more elaborate methods are needed. In such cases the cruder method given here tends to give an overly optimistic impression of the degree of discrimination--that is, the departure of the posterior odds ratio from unity.

6. An example of a Sequential Decision Procedure

Suppose that we are going to buy a new automobile. Having decided on the make, model, and accessories, we want to find a dealer who will sell it to us at a good price. We regard the process of securing a price quotation from a dealer as tantamount to observing a random variable $\tilde{x}$ from a normal process with unknown mean μ and known standard deviation σ. The prior density for $\tilde{\mu}$ is $f_N(\mu|\bar{x}', \sigma/\sqrt{n'})$.

Clearly at least one observation x_1 is needed if we are to buy the car. The problem is how long to continue the search for a good price. We assume that all utilities are measured in money terms, and that the marginal cost of a price quotation is a constant c.
A natural way to go about it is to proceed sequentially, examining each price quotation and deciding whether to take the minimum price so far quoted, or to continue, and secure at least one more quotation.
Denote successive price quotations $\tilde{x}_1$, $\tilde{x}_2$, $\tilde{x}_3$, ... Denote by $x_{\min,k}$ the minimum obtained in the first k quotations. As we accumulate quotations, we modify the distribution for $\tilde{\mu}$ by Bayes' theorem. The posterior density for $\tilde{\mu}$ after k quatations is

$$(1) \qquad D(\mu|x_1, \ldots, x_k, \bar{x}', n') = f_N(\mu|\bar{x}''_k, \sigma/\sqrt{n''_k}),$$

where $\bar{x}''_k = (n'\bar{x}' + n_k\bar{x}_k)/n''_k$, $n''_k = n' + n_k$, $\bar{x}_k + \sum_{i=1}^{k} x_i/k$. The predictive density implied by (1) for the k+1-st quotation $\tilde{x}_{k+1}$ is

$$(2) \qquad D(x_{k+1}|x_1, \ldots, x_k, \bar{x}', n') = f_N(x_{k+1}|\bar{x}''_k, \sigma\sqrt{\frac{1}{n''_k} + 1}).$$

Suppose that we have observed $\tilde{x}_1$, ..., $\tilde{x}_k$ and want to decide whether or not to observe $\tilde{x}_{k+1}$ and then make a purchase. This is a two-action problem: either a_1, we stop getting quotations and buy the car for $x_{\min,k}$; or, a_2, we observe $\tilde{x}_{k+1}$ and buy the car at $x_{\min,k}$ if $\tilde{x}_{k+1} \geq x_{\min,k}$ or at x_{k+1} if $\tilde{x}_{k+1} < x_{\min,k}$.
(The present application differs from Sec. 3, however, in that now there is just one terminal action--to buy a car.) The payoff table for this decision, ignoring the cost c of the observation, is

Event	a_1	a_2
$\tilde{x}_{k+1} \geq x_{\min,k}$	$-x_{\min,k}$	$-x_{\min,k}$
$\tilde{x}_{k+1} < x_{\min,k}$	$-x_{\min,k}$	$-x_{k+1}$

The corresponding loss table is

Event	a_1	a_2
$\tilde{x}_{k+1} \geq x_{\min,k}$	0	0
$\tilde{x}_{k+1} < x_{\min,k}$	$x_{\min,k} - x_{k+1}$	0

The expected loss of a_1, ignoring cost of observation, is

$$(5)\quad E\,L(a_1, \tilde{x}_{k+1}) = \int_{-\infty}^{x_{min,k}} (x_{min,k} - x_{k+1})\, f_N(x_{k+1} | \bar{x}''_k, \sigma\sqrt{\frac{1}{n''_k} + 1})\, dx_{k+1}.$$

If we examine eq. (15), Sec. 3, we see that (5) is the same thing in different notation: in eq. (15) replace $(b_1 - b_2)$ by 1, μ_b by $x_{min,k}$, $\bar{x}''$ by x_{k+1}, and the normal density by the density in (5). Therefore we can use Sec. 3 (17c), to evaluate (5) as

$$(6)\quad E\,L(a_1, \tilde{x}_{k+1}) = \sigma\sqrt{\frac{1}{n''_k} + 1}\,.\, G\left(\left| \frac{x_{min,k} - \bar{x}''_k}{\sigma\sqrt{\frac{1}{n''_k} + 1}} \right| \right) .$$

We can interpret (6) as the EVSI for a sample of one more quotation. We take the sample if and only if

$$(7)\quad \sigma\sqrt{\frac{1}{n''_k} + 1}\;\,.\, G\left(\left| \frac{x_{min,k} - \bar{x}''_k}{\sigma\sqrt{\frac{1}{n''_k} + 1}} \right| \right) > c.$$

(If it should happen that $\bar{x}''_k < x_{min,k}$, we would want the negative of the argument of G(.). The analogous phenomenon is impossible in Sec. 3 (17c).)

This decision rule, applied repeatedly for k = 1,2, ... until it terminates sampling, defines a sequential decision rule. There are many other possible rules. We could always stop when k=1, that is, accept the first price quotation; we could decide in advance to take a fixed number of observations n >1, then accept the lowest price; or we could follow other sequential decision rules. How good is our rule compared with the others? It is not obviously best.

The EVSI of (6) is for one observation only. It does not take into account the possibility that, having observed an $\tilde{x}_{k+1} < x_{min,k}$, we might nonetheless find still additional sampling attractive.

It can be proved, however, that our sequential decision rule is in fact the best of <u>all</u> decision rules, sequential or not, for the problem formulated here. This conclusion does not turn on normality, only on the utility structure of the problem.

This simple application is given as a concrete example of one of the few cases in which an unequivocally <u>best</u> sampling plan can be found, given present knowledge.

In general, sequential decision rules can offer both advantages and disadvantages as compared with fixed-sample decision rules, and much research on them is needed.

References:

R. Schlaifer, (1959) Probability and Statistics for Business Decisions, McGraw Hill, New York.

H. Raiffa and R. Schlaifer, (1961) Applied Statistical Decision Theory, M.I.T. Press, Cambridge (Mass.).

H.W. Gottinger, (1975) Bayesian Analysis: Probability and Decision, Angewandte Statistik (Applied Statistics), G.Tintner et. al. (eds.), Vandenhoeck & Ruprecht, Göttingen.