Statistical Decision Theory:
Concepts, Methods and Applications

(Special topics in Probabilistic Graphical Models)

FIRST COMPLETE DRAFT
November 30, 2003

Supervisor: Professor J. Rosenthal
STA4000Y

Anjali Mazumder
950116380

Part I: DECISION THEORY - Concepts and Methods

Decision theory as the name would imply is concerned with the process of making decisions. The extension to statistical decision theory includes decision making in the presence of statistical knowledge which provides some information where there is uncertainty. The elements of decision theory are quite logical and even perhaps intuitive. The classical approach to decision theory facilitates the use of sample information in making inferences about the unknown quantities. Other relevant information includes that of the possible consequences which is quantified by loss and the prior information which arises from statistical investigation. The use of Bayesian analysis in statistical decision theory is natural. Their unification provides a foundational framework for building and solving decision problems. The basic ideas of decision theory and of decision theoretic methods lend themselves to a variety of applications and computational and analytic advances.

This initial part of the report introduces the basic elements in (statistical) decision theory and reviews some of the basic concepts of both frequentist statistics and Bayesian analysis. This provides a foundational framework for developing the structure of decision problems. The second section presents the main concepts and key methods involved in decision theory. The last section of Part I extends this to statistical decision theory – that is, decision problems with some statistical knowledge about the unknown quantities. This provides a comprehensive overview of the decision theoretic framework.

Section 1: An Overview of the Decision Framework: Concepts & Preliminaries

Decision theory is concerned with the problem of making decisions. The term statistical decision theory pertains to decision making in the presence of statistical knowledge, by shedding light on some of the uncertainties involved in the problem. For most of this report, unless otherwise stated, it may be assumed that these uncertainties can be considered to be unknown numerical quantities, denoted by θ. Decision making under uncertainty draws on probability theory and graphical models. This report and more particularly this Part focuses on the methodology and mathematical and statistical concepts pertinent to statistical decision theory. This initial section presents the decisional framework and introduces the notation used to model decision problems.

Section 1.1: Rationale

A decision problem in itself is not complicated to comprehend or describe and can be simply summarized with a few basic elements. However, before proceeding any further, it is important to note that this report focuses on the rational decision or choice models based upon individual rationality. Models of strategic rationality (small-group behavior) or competitive rationality (market behavior) branch into areas of game theory and asset pricing theory, respectively. Thus for the purposes of this report, these latter models have been neglected as the interest of study is statistical decision theory based on individual rationality.

"In a conventional rational choice model, individuals strive to satisfy their preferences for the consequences of their actions given their beliefs about events, which are represented by utility functions and probability distributions, and interactions among individuals are governed by equilibrium conditions" (Nau, 2002[1]). Decision models lend themselves to a decision making process which involves the consideration of the set of possible actions from which one must choose, the circumstances that prevail and the consequences that result from taking any given action. The optimal decision is to make a choice in such a way as to make the consequences as favorable as possible.

As mentioned above, the uncertainty in decision making which is defined as an unknown quantity, θ, describing the combination of "prevailing circumstances and governing laws", is referred to as the ***state of nature*** (Lindgren, 1971). If this state is unknown, it is simple to select the action according to the favorable degree of the consequences resulting from the various actions and the known state. However, in many real problems and those most pertinent to decision theory, the state of nature is not completely known. Since these situations create ambiguity and uncertainty, the consequences and subsequent results become complicated.

Decision problems under uncertainty involve "many diverse ingredients" - loss or gain of money, security, satisfaction, etc., (Lindgren, 1971). Some of these "ingredients" can be assessed while some may be unknown. Nevertheless, in order to construct a mathematical framework in which to model decision problems, while providing a rational

basis for making decisions, a numerical scale is assumed to measure consequences. Because monetary gain is often neither an adequate nor accurate measure of consequences, the notion of ***utility*** is introduced to quantify preferences among various prospects which a decision maker may be faced with.

Usually something is known about the state of nature, allowing a consideration of a set of states as being ***admissible*** (or at least theoretically so), and thereby ruling out many that are not. It is sometimes possible to take measurements or conduct ***experiments*** in order to gain more information about the state. A decision process is referred to as "statistical" when experiments of chance related to the state of nature are performed. The results of such experiments are called ***data*** or ***observations***. These provide a basis for the selection of an ***action*** defined as a ***statistical decision rule***.

To summarize, the "ingredients" of a decision problem include (a) a set of available actions, (b) a set of admissible states of nature, and (c) a loss associated with each combination of a state if nature and action. When only these make up the elements of a decision problem, the decision problem is referred to as the "no-data" or "without experimentation" decision problem. However, if (d) observations from an experiment defined by the state of nature are included with (a) to (c), then the decision problem is known as a statistical decision problem. This initial overview of the decision framework allows for a clear presentation of the mathematical and statistical concepts, notation and structure involved in decision modeling.

Section 1.2 The Basic Elements

The previous section summarized the basic elements of decision problems. For brevity purposes, this section will not repeat the description of the two types of decision models and simply state the mathematical structure associated with each element. It is assumed that a decision maker can specify the following basic elements of a decision problem.

1. *Action Space:* $A = \{a\}$.

The single action is denoted by an a, while the set of all possible actions is denoted as A. It should be noted that the term actions is used in decision literature instead of decisions. However, they can be used somewhat interchangeably. Thus, a decision maker is to select a single action $a \in A$ from a space of all possible actions.

2. *State Space:* $\Theta = \{\theta\}$. (or *Parameter Space*)

The decision process is affected by the unknown quantity $\theta \in \Theta$ which signifies the state of nature. The set of all possible states of nature is denoted by Θ. Thus, a decision maker perceives that a particular action a results in a corresponding state θ.

3. *Consequence:* $C = \{c\}$.

The consequence of choosing a possible action and its state of nature may be multi-dimensional and can be mathematically stated as $c(a,\theta) \in C$.

4. *Loss Function:* $l(a,\theta) \in A \times \Theta$.

The objectives of a decision maker are described as a real-valued loss function $l(a,\theta)$, which measures the loss (or negative utility) of the consequence $c(a,\theta)$.

5. *Family of Experiments: E = {e}.*

Typically experiments are performed to obtain further information about each $\theta \in \Theta$. A single experiment is denoted by an *e*, while the set of all possible experiments is denoted as *E*. Thus, a decision maker may select a single experiment *e* from a family of potential experiments which can assist in determining the importance of possible actions or decisions.

6. *Sample Space: X = {x}.*

An outcome of a potential experiment $e \in E$ is denoted as $x \in X$. The importance of this outcome was explained in (3) and hence is not repeated here. However, it should be noted that when a statistical investigation (such as an experiment) is performed to obtain information about θ, the subsequent observed outcome *x* is a random variable. The set of all possible outcomes is the sample space while a particular realization of *X* is denoted as *x*. Notably, *X* is a subset of $\Re^n$.

7. *Decision Rule:* $\delta(x) \in A$.

If a decision maker is to observe an outcome $X = x$ and then choose a suitable action $\delta(x) \in A$, then the result is to use the data to minimize the loss $l(\delta(x),\theta)$. Sections 2 and 3 focus on discussing the appropriate measures of minimization in decision processes.

8. *Utility Evaluation:* $u(\cdot,\cdot,\cdot,\cdot)$ on $E \times X \times A \times \Theta$.

The quantification of a decision maker's preferences is described by a utility function $u(e,x,a,\theta)$ which is assigned to a particular conduct of *e*, a resulting observed *x*, choosing a particular action *a*, with a corresponding θ. The evaluation of the utility function *u* takes into account costs of an experiment as well as consequences of the specific action which may be monetary and/or of other forms.

Section 1.3 Probability Measures

Statistical decision theory is based on probability theory and utility theory. Focusing on the former, this sub-section presents the elementary probability theory used in decision processes. The probability distribution of a random variable, such as *X*, which is

dependent on θ, as stated above, is denoted as $P_\theta(E)$ or $P_\theta(X \in E)$ where E is an event. It should also be noted that the random variable X can be assumed to be either continuous or discrete. Although, both cases are described here, the majority of this report focuses on the discrete case. Thus, if X is continuous, then

$$P_\theta(E) = \int_E f(x \mid \theta)dx = \int_E dF^X(x \mid \theta).$$

Similarly, is X is discrete, then

$$P_\theta(E) = \sum_{x \in E} f(x \mid \theta).$$

Although, E has been used to describe a family of experiments, an event and will be used in the next sub-section to denote expectations, the meaning of E will be clear from the context.

For $\forall e \in E$ (where e represents an experiment or even an event), a joint probability measure $P_{\theta,x}(\cdot,\cdot \mid e)$ or simply denoted $P_{\theta,x|e}$ is assigned and more commonly referred to as the *possibility space*. This is used to determine other probability measures (Raiffa & Schlaifer, 2000):

(i) The marginal measure $P'_\theta(\cdot)$ on the state space Θ. Of course, here the assumption is that $P'_\theta(\cdot)$ does no depend on e.

(ii) The conditional measure $P_x(\cdot \mid e, \theta)$ on the sample space X for a given e and θ.

(iii) The marginal measure $P_x(\cdot \mid e)$ on the sample space X for a given e.

(iv) The conditional measure $P''_\theta(\cdot \mid x)$ on the state space Θ for a given e and x. The condition e is not stated as x is a result of the experiment and hence the relevant information is contained in x.

Before concluding this sub-section, it is important to make two remarks. The first is to summarize the three basic methods for assigning the above set of probability measures. That is (a) if a joint probability measure is assigned to $\Theta \times X$, then the marginal and conditional measures on Θ and X can be computed, (b) if a marginal measure (probability distribution) is assigned to Θ and the conditional for X, then the joint can be found and similarly (c) if a marginal measure (probability distribution) is assigned to X and the conditional for Θ, then the joint can be determined. These elementary "methods" or concepts of probability have a more practical importance.

The second remark is simply to clarify that the prime on the probability measure indicates a prior probability where as the double prime indicates a posterior probability. For the most part, these notations are redundant but at certain times will help to keep things clear. When it is obvious these superscripts will not required. A further discussion of priors is provided in **Section 1.5**.

Section 1.4 Random Variables and Expectations

In many instances, real numbers or *n*-tuples (of numbers) describe the states of nature {θ) and sample outcomes $\{x\}$. In the sections of this report a tilde sign may be used to distinguish a random variable or function from a particular value of the function. For example, the random variables $\widetilde{x}$ and $\widetilde{\theta}$ may be used to define $\widetilde{\theta}(\theta, x) = \theta$ and $\widetilde{x}(\theta, x) = x$, respectively.

Expectations of random variables are almost always considered necessary when dealing with decision processes such as loss functions. The expectation for a given value of θ, is defined to be

$$E_\theta[h(X) = \begin{cases} \int_{\mathrm{X}} h(x) f(x \mid \theta), \text{ (continuous)} \\ \sum_{x \in \mathrm{X}} h(x) f(x \mid \theta), \text{ (discrete)} \end{cases}$$

As before, the superscripts and subscripts on the expectation operator will perform in much the same way as for the probability measure. When necessary, such scripts will be minimized when the context is clear. Thus, with respect to points (i)-(iv) in the previous section, the notation for the following expectations is provided below:

(i) E'_θ or $E'(\widetilde{\theta})$ is taken with respect to P'_θ.
(ii) $E''_{\theta|x}$ is taken with respect to $P''_{\theta|x}$.
(iii) $E_{z|e,\theta}$ is taken with respect to $P_{z|e,\theta}$.
(iv) $E_{x|e}$ is taken with respect to $P_{x|e}$.

Section 1.5 Statistical Inference (Classical versus Bayesian)

Statistical inference is considered here within the decision framework. Both classical and Bayesian perspectives are briefly presented to show the varying approaches. Classical statistics uses the sample information to make inferences about the unknown quantity, θ. These inferences (within decision theory) are combined with other relevant information in order to choose the optimal decision. These other relevant information/sources include the knowledge of the possible consequences of the decisions and prior information which was previously mentioned. The former non-sample information, consequences, can be quantified by determining the possible loss incurred for each possible decision. The latter, prior information, is the information about θ arising from other relevant sources such as past experiences.

Bayesian analysis is the approach which "seeks to utilize prior information" (Berger, 1985). This third type of information is best described in terms of a probability distribution. The symbol $\pi(\theta)$ or simply $p(\theta)$ will be used to represent a prior density of

θ. Similar, to the other definitions, under both the continuous and discrete cases, the prior probability distributions can be written as

$$P(\theta \in E) = \int_E dF^{\pi}(\theta) = \begin{cases} \int_E \pi(\theta) d(\theta), \text{(continuous)} \\ \sum_{\theta \in E} \pi(\theta), \text{(discrete)} \end{cases}$$

The uses of prior probabilities are discussed in the proceeding sections. Both the non-Bayes and Bayesian decision theory are discussed in this report.

Section 1.6: Convex Sets

Various concepts presented throughout this report make use of the concepts of convexity and concavity; hence, the required definitions and properties are summarized below.

Definition:

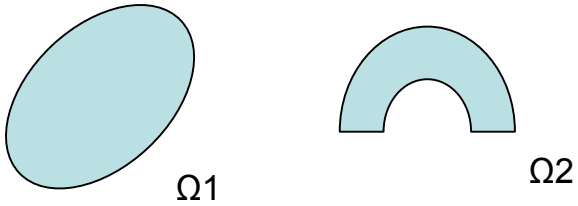
A linear combination of the form $\alpha x_1 + (1-\alpha)x_2$ with $0 \le \alpha \le 1$ is called a convex combination of x_1 and x_2. (Lindgren, 1971) If x_1 and x_2 in Ω, then it can be said that Ω is convex if the line segment between any two points in Ω is a subset of Ω. (Berger, 1985).

This definition simply suggests that the set of all combinations of two given points is precisely the set of points which make up the line segments jointing these two points. It may be conceived that the values α and $(1-\alpha)$ which fall between 0 and 1 may be interpreted as probabilities. A convex combination of two points then becomes a combination of probabilities.

Definition:

If $\{x_1, x_2, ...\}$ is a sequence of points in R^m, and $0 \le \alpha \le 1$ are numbers such that $\sum_{i=1}^{\infty} \alpha_i = 1$, then $\sum_{i=1}^{\infty} \alpha_i x_i$ (and finite) is called a convex combination. The convex hull of a set Ω is the set of all points which are convex combinations of points in Ω. (Berger, 1985).

Figure 1-1:



The set Ω_1 is convex while the Ω_2 is not, for the figure above. So a ***convex hull*** is a set with ***no holes*** in its interior and ***no indentations*** on its exterior—e.g. an egg is convex, while a doughnut or a banana or a golf ball is not. Formally, *X* is a convex set if every line segment connecting two distinct points in *X* is wholly contained in *X*, and it is ***strictly*** convex if the interior points of such a line segment are in the strict interior of *X*. Thus, a strictly convex set has no flat sides or edges: its exterior consists only of curved surfaces that are bowed outwards. A sphere is strictly convex, while a cube is convex but not strictly convex. (Nau, 2000).

Convex sets play a central role in the geometric representation of preferences (utility theory) and choices (statistical decision theory), as will be shown.

Section 2: "Non-statistical" Decision Processes

Statistical decision problems include data or observations on the unknown quantity, θ, (also referred to as the state of nature) which may be used to choose a more optimal decision. One approach to handling such problems, is to consider the non-statistical approach. Simply, this excludes the data available on the state of nature. This is reduces the problem to a simpler one. The basic elements or "ingredients" of this no-data problem were presented in the previous section.

In simple decision problems, the most important elements are the state space $\{\theta\}$, the action space $\{a\}$ and the loss function $l(\theta, a)$. These aspects form the theoretical analysis of decision making. The state space Θ and the action space A are both finite – i.e. "there are only finitely many actions [states]". The loss incurred is assumed to be measured in negative utility units where utility is defined by a function defining the prospects of a decision maker. In this instance, however unrealistic, we assume that the loss function is known. These concepts will be defined in the proceeding section within the no-data decision problem constructs.

Section 2.1 The Set of Randomized Actions

A simple way to explain the general theory of decision making is to consider a typical coin toss. This introduces an extraneous random device which is useful in providing decision rules that under some criteria are better than those that use only the given, nonrandom actions (Lindgren, 1971). In general, a random device (such as a coin (ordinary or biased)) is an experiment of chance having as many possible outcomes as there are actions from which to choose; each outcome is associated with an action, and when a particular outcome is observed, the corresponding action is taken.

The use of a random device to select an action from the set of possible actions is called a randomized or mixed action. Choosing a randomized action from among all possible actions amounts to selecting a random device from among all random devices that could be used to determine the action actually taken; and further selecting a set of probabilities for the various actions. This leads to the following definition of a random action.

Definition:

A randomized action, for a problem with action space consisting of actions $a_1, a_2, \ldots, a_k$, is a probability vector $(p_1, p_2, \ldots, p_k)$, that is a sequence of nonnegative numbers p_i whose sum is 1.

To summarize, consider constructing an experiment of chance producing outcomes $z_1, z_2, \ldots, z_k$ with probabilities $p_1, p_2, \ldots, p_k$ assigned, respectively. If the outcome of the experiment performed is z_1 then action a_1 is taken. To distinguish between pure (or original) actions $a_1, a_2, \ldots, a_k$ and randomized actions, probability vectors are used to define the latter type. A singular probability vector can be used to determine a pure action. For instance, the action a_2 can be equivalently written as the probability vector

(0,1,0,…0) that assigns all of the probability mass to action a_2. Thus, pure actions constitute a subset of the randomized actions.

Section 2.2 The Loss Function

The use of a randomized action in a decision problem with a given loss function inevitably sets the loss to be a random variable, for each state of nature. The most natural expected loss to consider when making a decision involves the uncertainty in θ. Thus, taking the expected value of the random variable (θ or $l(\theta,a)$) measures the consequence of employing a given randomized action (when nature is in a given state). In particular, if the loss of a function $l(\theta,a)$ and the randomized action (p_1, p_2,…, p_k) is used to choose among the actions a_1, a_2,…, a_k, the expected loss is the following weighted sum:

$$l(\theta,a_1)p_1 + l(\theta,a_2) + ... + l(\theta,a_k)p_k\,.$$

Notably, this can be written as taking the integral over all $l(\theta,a_i)$ and p_i, for the continuous case. The above definition of expected loss is explained in detail in Chapter 2 of Berger's *Statistical Decision Theory and Bayesian Analysis* (1985).

In a decision problem with m states θ_1, θ_2,…, θ_m, there are m L's corresponding to each randomized action. For the case of k actions, these L's are the following expected losses:

$$\begin{aligned}
L_1 &= E[l(\theta_1,a)] = l(\theta_1,a_1)p_1 + ... + l(\theta_1,a_k)p_k \\
L_2 &= E[l(\theta_2,a)] = l(\theta_2,a_1)p_1 + ... + l(\theta_2,a_k)p_k \\
&\vdots \\
L_1 &= E[l(\theta_m,a)] = l(\theta_m,a_1)p_1 + ... + l(\theta_m,a_k)p_k
\end{aligned}$$

defined by the randomized action (p_1, p_2,…, p_k). These relations can be written in the matrix form

$$\begin{pmatrix} L_1 \\ \vdots \\ L_m \end{pmatrix} = p_1 \begin{pmatrix} l(\theta_1,a_1) \\ \vdots \\ l(\theta_m,a_1) \end{pmatrix} + ... + p_k \begin{pmatrix} l(\theta_1,a_k) \\ \vdots \\ l(\theta_m,a_k) \end{pmatrix},$$

which suggests the interpretation of the vector of losses (L_1, L_2,…, L_m) is a point in m-dimensional space computed as a convex combination of the points $(l(\theta_1,a_i),...,l(\theta_m,a_i))$, for i=1,…,k. The notion of convex sets was introduced in sub-section **1.6**. The latter points are the loss vectors defined by the pure actions while those defined by randomized actions are convex combinations of those defined by the pure actions (Berger, 1985) and (Lindgren, 1971).

The closing example to this sub-section shows that the set of loss points (L_1, L_2,…, L_m) defined by all possible randomized actions (p_1, p_2,…, p_k) is a convex set in m-

dimensional space as described above. It forms a convex polyhedron where the extreme points are pure actions, although some pure actions may fall inside the set. The following example demonstrates a two dimensional case – i.e. there are two states of nature.

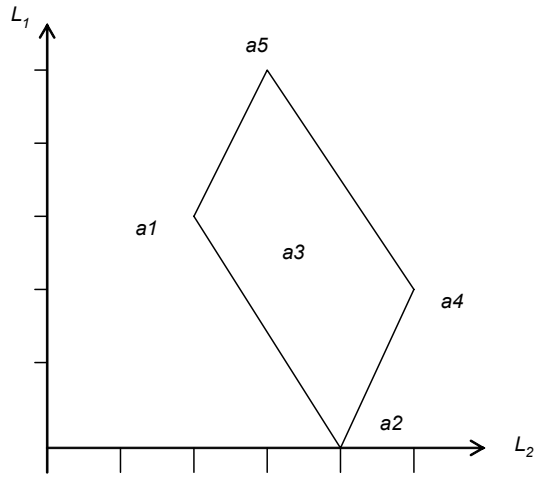
Example 2-1:
(Lindgren, 1971)

The following table describes a problem with five possible actions, two states of nature and the respective loss function.

	a_1	a_2	a_3	a_4	a_5
θ_1	2	4	3	5	3
θ_2	3	0	3	2	3

The pure actions define the loss vectors (the columns in the table) and the randomized actions define the convex set generated by these five actions which is founding the figure below. Note that one action, a_3 falls within the convex polyhedron.

Figure 2-1:



Recall from **Section 1.6** that a convex combination of a set of points corresponds to the "center of gravity of a system of point masses at those points" where these masses are proportional to the weights in the convex combination. Thus, the randomized action (p_1, p_2,…, p_k) yields a point (L_1, L_2,…, L_m), which is the center of gravity of a system of masses with p_1 units at a_1, p_2 units at a_2,…, and p_k units at a_k.

Interpreting randomized actions as centers of gravity helps geometrically interpret the state. For instance, if only p_1 and p_2 are positive, with no mass at a_3,…, a_k, then the center of gravity (and therefore the point (L_1, L_2)) must lie on the line segment joining a_1 and a_2. If these actions are represented by extreme points, then such a mixture of just these two actions lies on the edge of the convex set of points representing all randomizations (Lindgren, 1971). Thus, for this example, the action a_3 results in the same losses as a

certain mixture of a_1, a_2 and a_5, or as a mixture involving only a_1, a_4 and a_5, as well as many other mixtures of a_1, a_2, a_4 and a_5. However, it could not be obtained by mixing actions a_2, a_4 and a_5.

Section 2.3 Regret

If the state of nature is known then the action which results in minimal lost would be taken. Thus, if it were known that θ_i were the state of nature, the action a for which $l(\theta_i, a)$ is smallest should be taken, and the minimum loss

$$m_i = \min_A l(\theta_i, a) \ ,$$

is a loss that could not be avoided with even the best decision. Suppose, one takes the action a_j, which does not produce this minimum, and then discovers that nature is indeed in state θ_i, the decision maker would regret not having chosen the action that produces the minimum; the amount of loss that could have saved by knowing that state of nature is called the regret. Regret is defined for each state θ_i and action a_i as follows:

$$r(\theta_i, a_j) = l(\theta_i, a_j) - \min_A l(\theta_i, a) .$$

So for each state of nature, subtract the minimum loss m_i from the losses involving that state to obtain the regret.

Regret is often referred to as opportunity loss and represents a "thinking" in terms of gain rather than loss (Lindgren, 1971). The gain resulting from taking action a when nature is in state θ is the negative of the loss:

$$-g(\theta, a) = l(\theta, a) .$$

Hence, the minimum loss is the negative of the maximum gain:

$$\min_A l(\theta, a) = -\max_A g(\theta, a)$$

and the regret can be re-expressed in terms of gain as follows:

$$-r(\theta_i, a_j) = l(\theta_i, a_j) - \min_A l(\theta_i, a) = \max_A g(\theta_i, a) - g(\theta_i, a_j) .$$

This represents the maximum that could have been gained if the state of nature had been known, minus the amount that actually was gained by taking action a_j.

Example 2-1: (cont'd.)
(Lindgren, 1971)

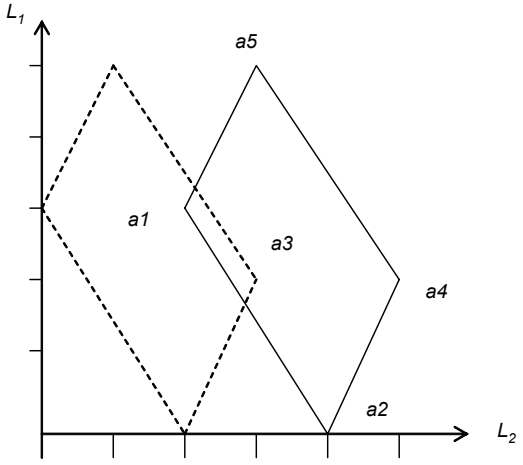
Recall the example presented in the previous sub-section. The table has been replicated below with the addition of another column representing the minimum loss over all five actions. The regret table is provided below and computed by subtracting the minimum loss from each of the losses in their respective rows.

	a_1	a_2	a_3	a_4	a_5	$\min_A l(\theta, a)$
θ_1	2	4	3	5	3	2
θ_2	3	0	3	2	3	0

	a_1	a_2	a_3	a_4	a_5
θ_1	0	2	1	3	1
θ_2	3	0	3	2	3

There is at least one zero in each state and the remaining entries are positive in the regret table. Geometrically, the objective or effect is to translate the set of points so that at least one action is on each axis. This is demonstrated in the figure below.

Figure 2-2:



Notice that the whole convex set of randomized actions shifts along with the five pure actions. This is generally the case because "the amount subtracted from each loss is independent of the action *a*, upon which a probability distribution is imposed in a mixed action" (Lindgren, 1971):

$$R_i = E[r(\theta_i, a)] = E[l(\theta_i, a)] - \min_A l(\theta_i, a) \equiv L_i - m_i .$$

Question: Would it make any difference, in studying a decision problem, if one used regret instead of loss? In some instances, the regret may be more painful then the losses, depending on the role of the decision maker and his or her stakes. The classical treatment of statistical problems in statistical decision theory, usually results in assuming a loss

function that is already a regret function; however this is not always the case and leads to a discussion in cases where it does make a difference.

Section 2.4 The Minimax Principle

Decision problems present a difficulty in determining the best decision because an action that is best under one state of nature is not necessarily the best under the other states of nature. Although, various schemes have been proposed - decision principles that lead to the selection of one or more actions as "best" according to the principle used – none is universally accepted. (See Berger, 1985 §1.5.)

By linearly ordering the available actions, assigning "values" to each action according to its desirability is a frequentist principle. The minimax principle places a value on each action according to the worst that can happen with that action. For each action *a*, the maximum loss over the various possible states of nature:

$$M(a) = \max_{\Theta} l(\theta, a),$$

is determined and provides an ordering among the possible actions (Lindgren, 1971 and French & Insua, 2000). Taking the action *a* for which the maximum loss *M*(*a*) is a minimum lends itself to the name minimax.

Berger states the same principle within the context of a decision rule. If $\delta \in \Delta$ is a randomized rule then the quantity $\sup_{\theta\in\Theta} R(\theta, \delta)$ represents the worst that could happen in the decision δ is used. Furthermore, the decision rule δ_1 is preferred to a rule δ_2 if

$$\sup_{\theta\in\Theta} R(\theta, \delta_1) < \sup_{\theta\in\Theta} R(\theta, \delta_2).$$

Similarly, a minimax decision rule is a minimax decision rule if it minimizes $\sup_{\theta\in\Theta} R(\theta, \delta)$ among all randomized rules in Δ.

Example 2-1: (cont'd.)
(Lindgren, 1971)

Again the loss table is repeated below with the addition a row stating the maximum loss for the various actions. The smallest maximum loss is determined by action a_1 to be 1.

	a_1	a_2	a_3	a_4	a_5
θ_1	2	4	3	5	3
θ_2	3	0	3	2	3
$\max_{\Theta} l(\theta, a)$	3	4	3	5	5

The table of regrets is produced below and shows different results then that of the maximum loss table. The table shows that the minimum maximum regret is determined by action a_2.

	a_1	a_2	a_3	a_4	a_5
θ_1	2	4	3	5	3
θ_2	3	0	3	2	3
$\max_{\Theta} l(\theta, a)$	3	2	3	3	5

Recall the question posed at the end of the last sub-section:

> *Question: Would it make any difference, in studying a decision problem, if one used regret instead of loss?*

A graphical approach of determining the minimax point-which is feasible when there are two states of nature is a useful in approaching the previously stated question. For a given action *a* with losses (L_1, L_2) under (θ_1, θ_2), respectively, the maximum of these losses is the first co-ordinate if the point lies below the bisector of the first quadrant while if the point lies above that bisector, or 45˚ line, the maximum loss is the second co-ordinate L_2. If two points lie above the bisector, the lower one has the smaller maximum; and if two points both lie below it, the left-most point has the smaller maximum. This approach can be generalized to many states, but becomes more messy in deciphering all the combinations.

The graphical approach to the minimax action is simply that the minimax process is related to the location of the origin of the co-ordinate system, and that moving the action points relative to the co-ordinate system can alter the process of finding the minimax point. Of course, in some instances, the minimax loss action and the minimax regret action will not differ.

To determine the minimax action among the set of all randomized actions is generally more complicated, because instead of choosing an action from a finite set of actions one must choose a probability vector from a set of possible probability vectors that is infinite in number (even if the set of actions is finite). There are two cases that can be considered for brevity sake at this point - when there are just two states of nature, and when there are just two actions. When there are just two states of nature, a graphical solution to the problem of determining a minimax mixed action can be carried out by representing the randomized actions in terms of (L_1, L_2) and (R_1, R_2) . This can be shown by returning to the example.

Example 2-1: (cont'd.)
(Lindgren, 1971)

Continuing with the same example and looking at the figure below, it is clear that this point (where the bisector meets the convex polyhedron) lies on the segment joining the

points representing a_1 and a_2, and so represents a mixture involving only those two actions. The point (x,y) in question is a convex combination of a_1 and a_2:

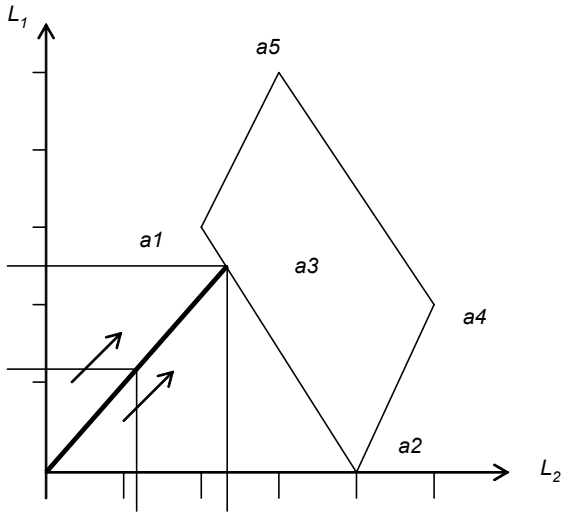
$$\begin{pmatrix} x \\ y \end{pmatrix} = p\begin{pmatrix} 2 \\ 3 \end{pmatrix} + (1-p)\begin{pmatrix} 4 \\ 0 \end{pmatrix}.$$

Moreover, it is the point on that line segment with equal co-ordinates:

$$2p + 4(1-p) = x = y = 3p + 0(1-p),$$

and equating these two functions of p yields $p = \frac{4}{5}$.

Figure 2-3:



The minimax action therefore puts $\frac{4}{5}$ of the total probability (1) at a_1, $\frac{1}{5}$ at a_2, and none at any other action, producing the probability vector $(\frac{4}{5},\frac{1}{5},0,0,0)$. The minimax loss, that is, the smallest maximum expected loss, is the value obtained by setting $p = \frac{4}{5}$, namely, $3p = \frac{12}{5}$. The point on the graph representing the minimax actions is thus $(\frac{12}{5},\frac{12}{5})$. Notice that this minimax expected loss is actually less that the minimum maximum loss achieved when only pure actions are admitted, 3, as determined in the "earlier example" presented in the previous section.

The case of two action is discussed only briefly as it can be determined in a similar fashion as above. The randomized actions are vectors of the form (p, 1 - p), defined by a single variable p. The expected losses under the various states of nature can be computed

as functions of p, and from this one can determine (at least graphically) the function $\max_{\Theta} E[l(\theta, a)]$. The minimum point of the latter function then defines the minimax action. See both Berger (1985) and Lindgren (1971) for examples and further discussion.

Section 2.5 Bayes Solutions

The frequentist approach has considered the states of nature to be fixed. In contrast, considering concept of randomness for the state of nature is the Bayes decision principle. This stance considers nature as random or not, nevertheless, it is rational to incorporate into the decision making process one's prior "hunches, convictions, or information about the state of nature-and how "likely" (whatever that means) the various states of nature are to be governing the situation" (Lindgren, 1971). This is accomplished by weighting the states and ordering the actions to permit the selection of a "best" action. Thus, if a large loss can occur for a given action when nature is in a state that the decision maker feels is highly unlikely, the extreme loss is minimized slightly by the state of nature that would have produced it.

The role of prior probabilities was briefly introduced in **Section 1** and will be expanded upon here. In a general decision problem, a probability weight $g(\theta)$ [1] is assigned to each state of nature θ, where these assigned weights are nonnegative and add up to 1. Such a set of probabilities or weights is called a *prior distribution* for θ. Given the distribution $g(\theta)$, the loss incurred for a given action a is a random variable, with expected value

$$B(a) = \sum_i g(\theta_i) l(\theta_i, a) .$$

This is referred to as the Bayes loss corresponding to action a (Lindgren, 1971). The Bayes action is then defined to be the action a that minimizes the Bayes loss $B(a)$. That is, the computation of the expected loss according to a given prior distribution provides a means of arranging or ordering the available actions on a scale (namely, $B(a)$) such that the action farthest to the left on that scale is the most desirable, and is to be taken.

When randomized actions are considered, a *Bayes* loss can be defined as the expectation with respect to a given prior distribution of the expected loss (with respect to the randomized action): For a randomized action $\mathbf{p} = (p_1, p_2, ..., p_k)$, which assigns probability p_i to action a_i, the expected loss for a given state θ is

$$E[l(\theta, a)] = \sum_j l(\theta, a_j) p_j ,$$

and the Bayes loss is obtained by averaging these (for the various θ's) with respect to $g(\theta)$:

[1] This notation is used so as to diffuse any confusion from the probability vectors assigned to each action.

$$B(\mathbf{p}) = \sum_i g(\theta_i) E[l(\theta_i, a)] = \sum_i g(\theta_i) \left\{ \sum_i l(\theta_i, a_j) p_j \right\}.$$

Since this Bayes loss is a function of $(p_1, p_2, ..., p_k)$, or $k - 1$ variables (p_k is determined as soon as $p_1, p_2, ..., p_{k-1}$ are specified), the problem of determining the minimum Bayes loss is that of minimizing a function of $(k - 1)$ variables.

Example 2-2:
(Lindgren, 1971)

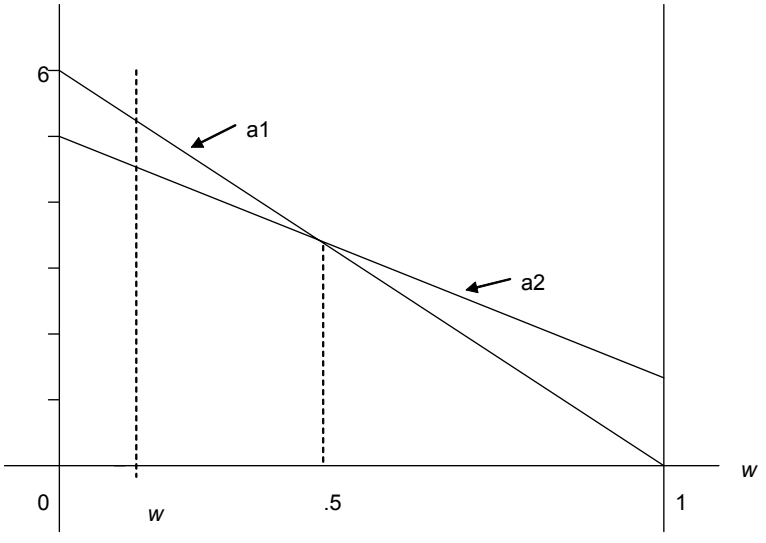
Consider the following loss table where there are just two states of nature and two actions. The prior probabilities are given and $g(\theta_1) = w$, $g(\theta_2) = 1\text{-}w$ and the expected losses are simply calculated as

$$B(a_1) = 0 \bullet w + 6(1 - w) = 6 - 6w$$
$$B(a_2) = 1 \bullet w + 5(1 - w) = 5 - 4w.$$

	θ_1	θ_2	$B(a)$
a_1	0	6	$6 - 6w$
a_2	1	5	$5 - 4w$
$g(\theta)$	w	$1 - w$	

The figure below illustrates that for any w to the left of 0.5, the value of w for which the Bayes losses are equal, the smaller Bayes loss is incurred by taking action a_2. For any w to the right of 0.5, the Bayes action is a_1 as it yields the smaller Bayes loss. When w = 0.5 then it is irrelevant whether action a_1 or a_2 is taken. Examination of that figure shows that for a prior distribution defined by a w in the range $0 \le w < w_1$ the Bayes action is a_1; for $w_1 < w < w_2$ it is a_2; for $w_2 < w \le 1$ it is a_3.

Figure 2-4:



The interpretation of small or large *w* is of course dependent on the problem; however, it can be generalized that small *w* refers to a least favorable outcome where as *w* refers to a more favorable outcome. In general, the Bayes losses for the various actions will be linear functions of *w*, in a problem with two states of nature and $g(\theta_1) = w$, $g(\theta_2) = 1-w$. These functions of *w* will be represented by straight lines, and for a given value of *w* the action corresponding to the line whose ordinate at that *w* is smallest is the Bayes action.

Now when considering randomized actions (p_1, p_2,…, p_k) for actions a_1,… a_k, respectively, produces the expected losses

$$B(\mathbf{p}) = \sum_i g(\theta_i)\left\{\sum_j l(\theta_i, a_j)p_j\right\} = \sum_j p_j\left\{\sum_i g(\theta_i)l(\theta_i, a_j)\right\}.$$

Thus, the value of $B(\mathbf{p})$ is a convex combination of the values of the Bayes losses for the various pure actions, and is at least as great as the smallest of those values (Lindgren, 1971). This means that there is no gain possible in the use of randomized actions; a pure action can always be found which yields the minimum Bayes loss.

The result that the Bayes actions are the same using the regret as using loss-is evident for the problem involving only two states of nature; however, it is also true in general. Substitution of

$$r(\theta_i, a_j) = l(\theta_i, a_j) - \min\ l(\theta_i, a)$$

into the expression for expected regret:

$$E[r(\theta, a_j)] = \sum_i r(\theta_i, a_j)g(\theta_i)$$

yields

$$E[r(\theta, a_j)] = \sum_i l(\theta_i, a_j)g(\theta_i) - \sum_i g(\theta_i)\min_a l(\theta_i, a)$$
$$= E[l(\theta, a_j)] - \sum_i g(\theta_i)\min_a l(\theta_i, a).$$

Thus, $E[r(\theta, a_j)]$ differs from $E[l(\theta, a_j)]$ by a term that does not involve a_j. The action that minimizes one must therefore minimize the other. This is illustrated the following example.

Example 2-3:
(Lindgren, 1971)

Consider the following decision problem with its corresponding loss table.

	a_1	a_2	a_3
θ_1	2	5	3
θ_2	3	1	5

The regrets are

$$r(\theta_1, a) = l(\theta_1, a) - 2$$
$$r(\theta_2, a) = l(\theta_2, a) - 1$$

The expected regrets, given prior weights $g(\theta_1)$ and $g(\theta_2)$, are

$$E[r(\theta, a_1)] = 2g(\theta_1) + 3g(\theta_2) - [2g(\theta_1) + g(\theta_2)]$$
$$E[r(\theta, a_2)] = 5g(\theta_1) + g(\theta_2) - [2g(\theta_1) + g(\theta_2)]$$
$$E[r(\theta, a_3)] = 3g(\theta_1) + 5g(\theta_2) - [2g(\theta_1) + g(\theta_2)]$$

so that $E[r(\theta, a_i)] = E[l(\theta, a)] - [2g(\theta_1) + g(\theta_2)]$.

Thus the action that minimizes $E[r(\theta, a_j)]$ also minimizes $E[l(\theta, a_j)]$ by a difference of, irrespective of the action.

Section 2.6 Dominance and Admissibility

The concepts of dominance and admissibility are counterparts to decision rules. Some examples have been encountered in which certain of the available actions would never be used because there are others for which losses are always less. It is important to state some definitions before discussing their importance any further.

Definition:

An action a^* (pure or randomized) is said to dominate an action a if the loss incurred by using action a is always at least as great as that incurred by using a:

$$l(\theta, a) \geq l(\theta, a^*), \text{ for all } \theta.$$

Definition:

An action a^* (pure or randomized) is said to dominate strictly an action a if it dominates action a and if, in addition, there is some state of nature for which the loss inequality is strict:

$$l(\theta, a) > l(\theta, a^*), \text{ for some } \theta.$$

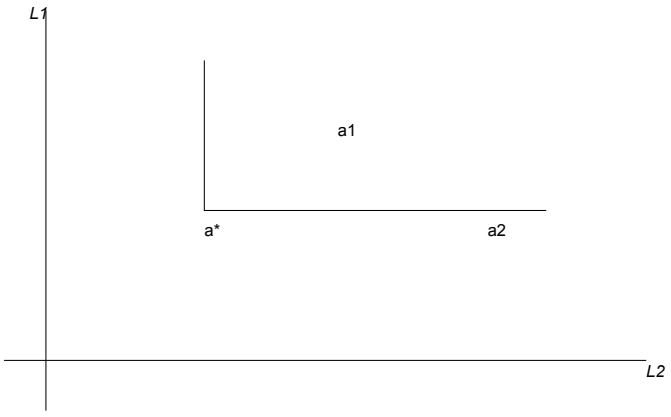
Definition:

An action (pure or randomized) is said to be admissible if no other action dominates it strictly.

The proceeding paragraph was extracted from Lindgren (1971).

Returning to the case of two states of nature, an action is represented as a point in the (L_1, L_2) plane, where L_i is the loss (or expected loss, in the case of a randomized action) incurred when that action is taken and θ_i is the state of nature. The figure below shows points corresponding to actions a^*, a_1 and a_2, for which losses are such that a^* dominates both a_1 and a_2. In the case of a_1 the losses are greater than for a^* for both states of nature; in the case of a_2 only the loss for θ_1 is strictly greater than with a^*. Incidentally the action a^* would strictly dominate every action represented by points in the shaded quadrant (including the boundaries), with the exception of the point a^* itself. (In this kind of representation, an action that dominates action a but does not strictly dominate it would be represented by the same point as action a, since the losses would be the same.)

Figure 2-5:



If an action a^* dominates an action a there is no need to leave a in the competition for a best action. If the dominance is not strict, then the losses are the same for a as for a^*, and a can be dispensed with; if it is strict, then one may actually do worse by using a than by using a^*.

Before ending this section, we can re-iterate this concept in terms of risk functions and decision rules. In Chapter 5 of French and Insua (2000), "it is stated that risk functions induce a natural ordering among decision rule: a decision rule which performs uniformly better in terms of risk than another for each value of θ seems better overall". Let $\delta_1(\cdot)$ and $\delta_2(\cdot)$ be two decision rules. Then $\delta_1(\cdot)$ dominates $\delta_2(\cdot)$ if $R(\delta_1,\theta) \le R(\delta_2,\theta), \forall\theta,$ with strict inequality. Similarly, it can be stated that the $\delta_1(\cdot)$ and $\delta_2(\cdot)$ are equivalent if $R(\delta_1,\theta) = R(\delta_2,\theta), \forall\theta.$

Definition:

A decision rule (or action) is admissible if there exists no *R*-better decision rule where *R* is referring to the risk function. A decision rule is inadmissible is inadmissible if there exists an *R*-better decision rule.

The above definition was provided by Berger (1985). An action that is not admissible is said to be inadmissible, and can be dispensed with because there is an action that does at least as well under the circumstances. On the other hand, an action that is close to the lower left boundary of a set of mixed actions (in the L_1, L_2 representation) may not be so bad as the name inadmissible would imply (Lindgren, 1971). That is, there are degrees of inadmissibility which the terminology ignores. In statistical problems, for example, there may be solutions that are slightly inadmissible but are preferred for some reason to those that dominate-because of computability, for instance.

Section 2.7 Bayesian and Classical Approaches – Bayes versus Minimax Principle

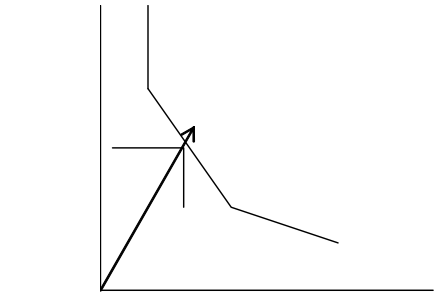
This section explores the connections between Bayesian and classical approaches through their respective forms of decision principles, namely the Bayes and Minimax Principle. The latter is essentially based on the risk function $R(\delta,\theta)$[2] which induces a partial ordering among the decision rules[3], leading to the concept of admissibility. Since this is only a partial ordering, the Bayesian approach introduces a prior distribution which may be used to weight the risk function and orders according to Bayes risk (French and Insua, 2000). Much of this discussion centers around the concept of admissibility. From the preceding three sub-sections, some possible generalities can be made – (1) Bayes solutions are usually admissible; (2) A minimax action is a Bayes action; and (3) Admissible actions are Bayes, for some prior distribution (Lindgren, 1971). These are non-formal statements; however they are discussed in detail with proofs in 4.8 (Bayes admissibility) and 5.5 (Minimax admissibility and Comparison with Bayes) of Berger (1985). (See also Chapter 6 of French and Insua (2000).

Again a geometric representation of just two states of nature is instructive as the set of all possible (randomized) actions is represented in the L_1L_2 plane as a convex set. That a minimax solution is Bayes for some prior distribution is evident, for the case of two states of nature, from the geometrical representation in the L_1L_2 plane (Lindgren, 1971). The figures below illustrate four cases that may arise.

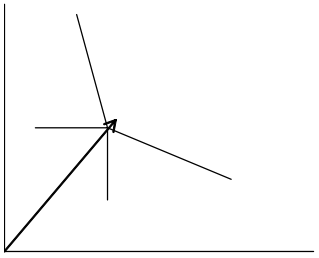
[2] It is important to recall here that risk function is the expected loss (as defined by the Frequentist or classical approach). The risk function is further explored in Section 3.

[3] It is important to note that decisions and actions are sometimes interchangeable, especially in the no-data decision problems. Thus, although the notion of decision rules has been brought in this section it is discussed in more detail in the next section.

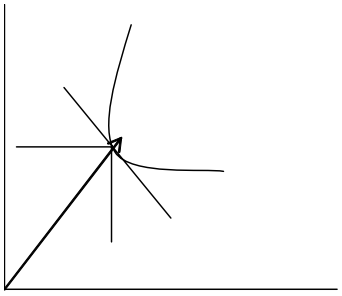
Figure 2.6:



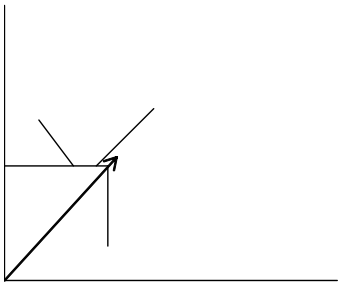
(a) The minimax action occurs on the boundary of the set of actions, and is a mixture of two pure actions; this action would be Bayes for the prior distribution $(w, 1\text{-}w)$ such that $-w/(1-w)$ is equal to the slope of the line through the two pure actions involved.



(b) The minimax action, which is a pure action, would be Bayes for any prior distribution such that $-w/(1-w)$ is a slope between the slopes of the line segments which meet in that pure action.



(c) The prior distribution which produces the minimax action as a Bayes action corresponds to the tangent line at $L_1 = L_2$; this kind of set of randomized actions would only occur if there are infinitely many pure actions at the outset.



(d) The minimax action would be any that yields an (L_1, L_2) on the bottom edge of the action set; these are Bayes for the prior distribution, which assigns probability 1 to θ_2.

Example 2-3:
(Lindgren, 1971)

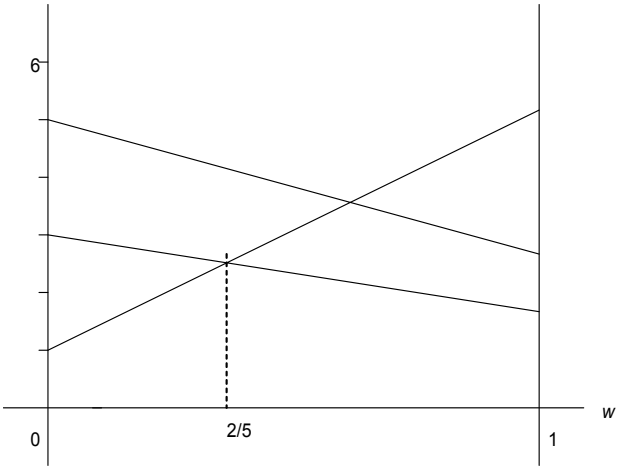
Revisiting the problem introduced in **Section 2.5** produces the following Bayes losses with given prior probabilities $(w, 1-w)$. The loss table has been repeated for ease.

	a_1	a_2	a_3
θ_1	2	5	3
θ_2	3	1	5

$$B(a_1) = 2w + 3(1-w) = 3 - w$$
$$B(a_2) = 5w + (1-w) = 1 + 4w$$
$$B(a_3) = 3w + 5(1-w) = 5 - 2w.$$

The graphs of these are shown in ***figure 2.7*** below. The lowest point of intersection and the highest minimum is at $w = \frac{2}{5}$. That is, of all the prior distributions that Nature might choose, the highest minimum Bayes loss (loss incurred by using the Bayes action) is achieved for the prior distribution $(\frac{2}{5}, \frac{3}{5})$[4].

Figure 2.7:



The minimax randomized action is easily found (by the graphical procedure described in **Section 2.4**) to be $(\frac{4}{5}, \frac{1}{5}, 0)$ with losses $L_1 = L_2 = 13/5$. The point on the line through (2, 3) and (5, 1) with equal co-ordinates:

[4] This is said to be a least favorable prior distribution (Lindgren, 1971).

$$\frac{L_1 - 3}{L_1 - 2} = -\frac{2}{3}, \text{ or } L_1 = \frac{13}{5}.$$

A prior distribution that would yield this randomized action as Bayes action is defined by a w such that

$$\frac{-w}{1-w} = -\frac{2}{3},$$

where -2/3 is the slope of the line through (5, 1) and (2, 3). Notice that this w, which is $w = 2/5$, is precisely the w of the least favorable distribution[5] determined above. A graphical representation of this has not been shown here but can be achieved in much the same manner described in **Section 2.4** and **2.5**.

The Bayes Perspective…

Mathematically the Bayes approach is just postulating a weighting function that provides an ordering among the actions; practically, the decision maker is incorporating into that weighting function personal preferences about what that unknown state of nature is likely to be. The concept of subjective probability is discussed in **Part C** on utility theory. In the role of rational decision making, the inclusion of such personal preferences deems reasonable. However, non-Bayesians do criticize this approach due to its view of subjectivity. In defence of this approach, it is not terribly sensitive; therefore, in small inaccuracies in specification is not treacherous. Furthermore, the Bayes solutions are usually admissible.

The Minimax Approach…

The minimax approach, on the other hand, is not nearly so easy to defend. Berger (1985) goes on to say that "when considered from a Bayesain viewpoint, it is clear that the minimax approach can be unreasonable. It is pessimistic view - making the assumption that the worst will happen. Although it is frequently admissible (and Bayes for some prior distribution), the distribution is least favorable. A more significant objection is that the minimax principle, by considering $\sup_{\theta} R(\theta, \delta)$, may violate the rationality principles (Berger, 1985).

Sometimes a minimax solution can be computed by determining among the Bayes solutions one for which the losses under the various states of nature are equal. That is, if $\mathbf{p^*} = (p_1^*, ..., p_k^*)$ is a randomized action that is Bayes with respect to some prior $g(\theta)$ and is such that the (expected) loss function

[5] The notion of a least favorable distribution giving rise to the minimax action as a Bayes action is general, but not trivial to establish. This idea is discussed more indepthly in sources on game theory which is not included in this report.

$$L(\theta, \mathbf{p}^*) = E[l(\theta, a)] = \sum_i l(\theta, a_i) p_i^*$$

is constant in θ, then $\mathbf{p}^*$ is minimax. The development follows (Lindgren, 1971): The assumption that $\mathbf{p}^*$ is Bayes for $g(\theta)$ means that for any randomized action $\mathbf{p}$,

$$B(\mathbf{p}^*) = \min_P B(\mathbf{p}) \le B(\mathbf{p}).$$

But since $L(\theta, \mathbf{p}^*)$ is constant, its mean value with respect to (wrt) the weighting $g(\theta)$ is just that constant value:

$$B(\mathbf{p}^*) = E[L(\theta, \mathbf{p}^*)] = L(\theta, \mathbf{p}^*).$$

On the other hand, since $B(\boldsymbol{p})$ is the expected value of $L(\theta, \mathbf{p})$, which cannot exceed the maximum value of $L(\theta, \mathbf{p}) \rightarrow B(\mathbf{p}) \le \max_\theta L(\theta, \mathbf{p})$ and it follows for any $\mathbf{p}$ that

$$L(\theta, \mathbf{p}^*) \le \max_\theta L(\theta, \mathbf{p}).$$

Similarly, the constant $L(\theta, \mathbf{p}^*)$ cannot exceed the smallest of these maxima

$$L(\theta, \mathbf{p}^*) \le \min_P \max_\theta L(\theta, \mathbf{p}),$$

where the right side of this inequality is the minimax loss.

Since $L(\theta, \mathbf{p}^*)$ is defined to be constant in θ, it is equal to its maximum value, which in turn is greater than or equal to the smallest such maximum:

$$L(\theta, \mathbf{p}^*) = \max_\theta L(\theta, \mathbf{p}^*) \ge \min_P \max_\theta L(\theta, \mathbf{p}).$$

And then because $L(\theta, \mathbf{p}^*)$ is neither less than nor greater than the minimax loss, it must be equal to it-and this means that $\mathbf{p}^*$ is a minimax solution, as was asserted. This can also be visually easing with just two states of nature. Lindgren (1985) has described this phenomenon more in depth both algebraically and graphically. A similar proof can be found in Berger (1985) and is discussed in brief in French and Insua (2000).

Before closing this section, I would like to make one note. Two of three key decision rules have been discussed here. The third, the Invariance Principle, has not been included in a theoretical discussion but for completeness is mentioned. The Invariance Principle, as the name would have it imply, basically states that if two problems have identical formal structures (sample size, parameter space, densities and loss function), then the same decision rule should be used in each problem. An entire chapter in Berger's 1985 book *Statistical Decision Theory and Bayesian Analysis* is devoted to this principle.

Section 3: "Statistical" Decision Processes

Statistical decision problems include data - meaning "the results of making one or more observations on some random quantity that is thought to be intimately related to the state of nature in some decision problem" (Lindgren, 1971). The availability of such data provides some illumination in the selection of an action such that the state of nature is *not completely* unknown. It will be shown that the use of data will define procedures which result in an expected loss that is lower than what would be incurred if the data were not available. However, even the availability of data will not avoid completely the kind of situation encountered in the no-data case, in which there is no clear-cut criterion for rating the various candidate procedures as a basis for choosing one of them as best.

Section 3.1 Data and the State of Nature

To obtain data for use in making decisions an appropriate experiment of chance should be performed – one in which the state of nature determines the generation of the data, and so the probability distribution for the data depends on that state of nature.

The data of a given problem may consist of a single number (value of a random variable), or a sequence of numbers - usually resulting from performing the same experiment repeatedly, or sometimes a result or results that are not numerical. In general, a random variable X will be employed to refer to the data, and in the problems considered here X will denote either a single random variable, or a sequence of random variables: $(X_1,\ldots,X_n)$. In any case X will have certain possible "values" and a probability for each, according to the state of nature. Thus, for each value of x of the random quantity X there is a probability

$$f(x;\theta) = P_\theta(X = x)$$

assigned to that value x by the state of nature θ. (The notation $P_\theta(E)$ will mean the probability of the event E when the state of nature is θ.)[6]

Section 3.2 Decision Functions

Section 2 introduced the concept of a decision rule within the no-data context. It's relevance develops in this section as a procedure for using data as an aid to decision making involving a rule, or set of instructions, that assigns one of the available actions to each possible value of the data X. Thus, when the pertinent experiment is performed and a value of X obtained, say $X = x$, an action has been assigned by the rule to that value, and

[6] It is important to realize that in order for the data to be of value in making a decision, the dependence of the probability distribution for X on the state of nature must be known. That is, $f(x;\theta)$ is assumed to be given or known.

that action is taken. A decision rule is a function $a = \delta(x)$, and is called a decision function, or statistical decision function. Berger (1985) gives the following definition.

Definition:

A (nonrandomized) decision rule $\delta(x)$ is a function from X into A. (It is always assumed that the functions are measureable.) If $X = x$ is the observed value of the sample information, then $\delta(x)$ is the action that will be taken. (Recall for a no-data problem, a decision rule is simply an action.) Two decision rules, $\delta_1(x)$ and $\delta_2(x)$, are considered equivalent if $P_\theta(\delta_1(X) = \delta_2(X)) = 1$ for all θ.

Question: How many distinct rules are there? If there are just k available actions (a_1, a_2,…, a_k), and if the data X can have one of just m possible values (x_1, x_2,…, x_m), then there are precisely k^m distinct decision functions that can be specified. "Of the k^m possible decision functions, some are sensible, some are foolish; some ignore the data, and some will use it wrongly" (Lindgren, 1971).

Example 3-1:
(Lindgren, 1971)

Consider the following table with 8 decision rules to between two actions and where the observed value X={0,1,2}. The first decision rule ignores the data as it takes action a_2 regardless; the last decision rule performs in much the same way. The second decision rule suggests taking action a_2 is X = 0 or 1.

x	δ_1	δ_1	δ_1	δ_1	δ_1	δ_1	δ_1	δ_1
0	a_2	a_2	a_2	a_1	a_2	a_1	a_1	a_1
1	a_2	a_2	a_1	a_2	a_1	a_2	a_1	a_1
2	a_2	a_1	a_2	a_2	a_1	a_1	a_2	a_1

Decision rules can be randomized in much the way actions are randomized – an extraneous random device to choose among the available rules. A randomized decision function is then a probability distribution over the set of pure decision functions, assigning probability p_i to decision function $\delta_i(x)$. A randomized decision rule can be defined alternatively by attaching an outcome x_i of the data X (from an experiment), which selects one of the actions (a_1, a_2,…, a_k) according to some probability distribution (g_1, g_2,…, g_k) (Lindgren, 1971). This is rule is equivalent to a randomization of pure decision functions, when the numbers of actions and possible values of X are finite. The following definition provided by Berger (1985) closes this part.

Definition:

A randomized decision rule $\delta(x,\cdot)$ is, for each x, a probability distribution on A, with the interpretation that if x is observed, $\delta(x,a)$ is the probability that ac action in A will be chosen. (Again a randomized decision rule in no-data problems is simply referred to as a

randomized action.) Nonrandomized rules can be considered a special case of randomized rules, in that they correspond to the randomized rules which, for each x, so that a specific action chosen has probability one. Let $\langle\delta\rangle$ denote the equivalent randomized rule (at this time) for the nonrandomized rule $\delta(x)$ given by

$$\langle\delta\rangle(x,a) = I_A(\delta(x)) = \begin{cases} 1 \text{ if} \delta(x) \in A, \\ 0 \text{ if} \delta(x) \notin A. \end{cases}$$

Section 3.3 The Risk Function

Section 2 first introduced the concept of risk function as that understood by the classical approach. This section defines and explores this concept more explicitly. Following from the previous sub-section, when a given decision function $\delta(x)$ is used, the loss incurred depends not only on the state of nature that governs it, but also on the value of X that is observed. Since X is random, the loss incurred can be restated as,

$$l(\theta, \delta(x)),$$

and is a random variable. The frequentist decision-theoretic evaluates, for each θ, the expected value of loss if such a decision rule $\delta(x)$ was used repeatedly with varying X in the decision problem.

Definition:

The risk function of a decision rule $\delta(x)$ is defined by

$$R(\theta, \delta) = E[l(\theta, \delta(x))]$$

which depends on the state of nature θ and on the decision rule δ. (Notably, for a no-data problem, $R(\theta, \delta) = L(\theta, a)$).

Assuming a distribution for X defined by $P_\theta(X = x) = f(x;\theta)$, the risk function would be calculated as

$$R(\theta, \delta) = \sum_i l(\theta, \delta(x_i)) f(x_i;\theta).$$

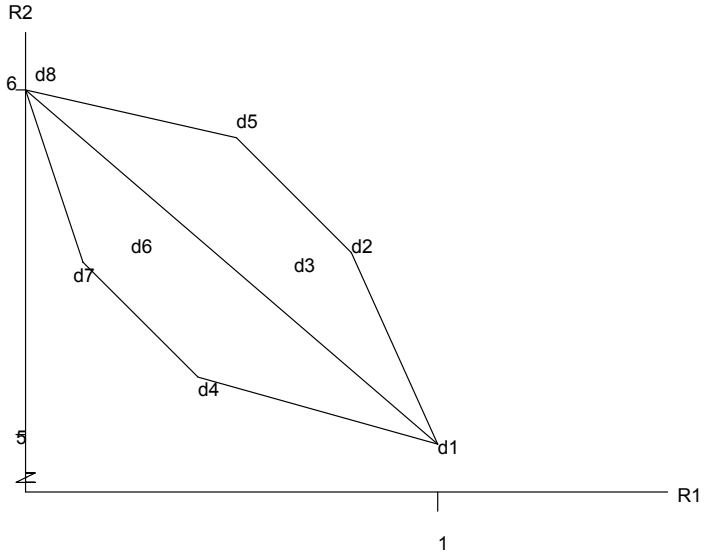
Or simply stated, the risk function is the weighted average of the various values of the random loss. Notice that the dependence of risk on the state θ arises because of two facts: the loss for a given action depends on θ, and the probability weights used in computing the expected loss depends on θ.

Thus, when the decision is based on observed results of an experiment, a decision rule (pure or randomized) can be selected from those available, knowing the risk function but ignorant of the true state of nature. The problem of selecting a decision rule knowing the risk function is mathematically exactly the same as that of selecting an action in the

absence of data, knowing the loss function. The methods (minimax, Bayes) used in attacking the no-data problem can be used for the statistical problem (with data).

In a similar fashion to a loss table, a risk table can be created in much the same way with the number of rows representing the possible states of nature and the number of columns producing the number of distinct decision rules. The decision rules are also representable graphically as before, where (R_1, R_2) is plotted, with $R_i = R(\theta_i, \delta)$, for each rule δ. The plot in the figure below shows in addition to the pure decision rules the set of randomized of the pure rules, as the convex set generated by them (French and Insua, 2000).

Figure 3-1:



Several observations can be made (Lindgren, 1971) when using or neglecting the data:

- The pure decision functions δ_1 and δ_8, which ignore the data, give points (R_1, R_2) which are exactly the same as the corresponding (L_1, L_2) for the no-data problem.
- The straight line joining δ_1 and δ_8 would consist of randomizations that are exactly equivalent (as regards risk) to randomizations of the pure actions $a_1$and a_2, respectively, in the absence of data.
- The data provides risk points which clearly dominate these no-data points; in particular, δ_4, δ_6 and δ_7 are all better, and indeed δ_4 and δ_7 are admissible.
- There are several rules (δ_2, δ_3 and δ_5) which are worse than rules that make no use of the data, and consequently make poor use of the data.

The availability of data permits a reduction of losses if the data is used properly and if the set of (R_1, R_2) points representing the various procedures is pulled in toward the zero regret point. In this instance, the data cannot be independent of the state of nature. If the distribution of X is independent of θ:

$$P_\theta(X = x) = f(x;\theta) = k(x),$$

then the risks for a given rule $\delta(x)$ are

$$R_i = E[l(\theta_i, \delta(X))] = \sum_j l(\theta_i, \delta(x_j))k(x_j) \qquad \text{for } i = 1,...,s.$$

In vector form this becomes

$$\begin{pmatrix} R_1 \\ \vdots \\ R_s \end{pmatrix} = \sum_j k(x_j) \begin{pmatrix} l(\theta_1, \delta(x_j)) \\ \vdots \\ l(\theta_s, \delta(x_j)) \end{pmatrix},$$

which is a convex combination of the points (L_1,…, L_s), where $L_i = l(\theta_1, \delta(x_j))$, corresponding to the m actions $\delta(x_1),...,\delta(x_m)$. These losses can be achieved by the randomization $k(x_1),...,k(x_m)$ of pure actions $\delta(x_1),...,\delta(x_m)$, and so the data does not extend the convex set of loss points to risk points that are any better (Lindgren, 1971).

Section 3.4 Selecting a Decision Function

Section 2.6 introduced the concepts of dominance and admissibility which are extended to the decision function, by defining them in terms of the risk function. (Although this concept was addressed in the previous section, this section discusses more in depthly within the constructs of modeling statistical decision problems. So, a decision function δ^* dominates a decision function δ if and only if

$$R(\theta, \delta^*) \le R(\theta, \delta), \quad \text{for all } \theta.$$

The dominance is strict if the inequality is strict for at least one state of nature. A decision function is admissible if it is not dominated strictly by any other decision function. This was defined by Berger (1985) and stated in **Section 2.6**.

A decision function that involves an inadmissible action is inadmissible. If a is not admissible, it is dominated by some action a^* (Lindgren, 1971):

$$l(\theta_0, a^*) < l(\theta_0, a), \quad \text{for some } \theta_0.$$

If $\delta(x)$ assigns the action a to some possible value x_j of positive probability under θ_0, a new rule $\delta^*(x)$ is defined to be identical with $\delta(x)$ except that $\delta^*(x_j) = a^*$, and then

$$\begin{aligned} R(\theta_0, \delta^*) &= \sum_j l(\theta_0, \delta^*(x_i)) f(x_i; \theta_0) \\ &< \sum_j l(\theta_0, \delta(x_i)) f(x_i; \theta_0) = R(\theta_0, \delta) \end{aligned}$$

(The inequality follows because the losses in the sums are identical except for the term where $i = j$, in which case $\delta^* = a^*$ instead of a, yielding a smaller loss.) For states other than θ_0 the inequalities < are replaced by $\leq$, and the strict dominance of δ by δ^* follows.

The Minimax Approach…

The minimax principle provides a numerical measure of decision rules, namely, the maximum risk over the various states of nature:

$$M(\delta) = \max_{\theta} R(\theta, \delta).$$

The minimax decision function is the δ that minimizes this maximum risk.

The Bayes Approach…

Assigning prior probability weights $g(\theta)$ to the various states of nature determines the average risk over the states:

$$B(\delta) = E[R(\theta, \delta)] = \sum_{j} R(\theta_j, \delta) g(\theta_j).$$

This is the Bayes risk. The preferred decision rule is the decision function δ which minimizes this Bayes risk.

Recall the question posed at the last section:

> *Question: Would it make any difference, in studying a decision problem, if one used regret instead of loss?*

As in the no-data case, it can make a difference - at least, if one considesr a minimax approach. There are two ways to introduce the idea of regret - by applying it to the initial loss function, and by applying it to the risk, or expected loss. These result to the same thing and are shown below. The regret function was defined as

$$r(\theta, a_i) = l(\theta, a_i) - \min_{a} l(\theta, a),$$

so for each x,

$$r(\theta, \delta(x)) = l(\theta, \delta(x)) - \min_{a} l(\theta, a);$$

and the expected regret is then

$$E[r(\theta, \delta(X))] = E[l(\theta, \delta(X))] - \min_{a} l(\theta, a).$$

Now for any decision rule δ,

$$l(\theta, \delta(x)) \geq \min_a l(\theta, a), \quad \text{for all } x,$$

so that

$$E[l(\theta, \delta(X))] \geq \min_a l(\theta, a)$$

and

$$\min_\delta E[l(\theta, \delta(X))] = \min_\delta R(\theta, \delta) \geq \min_a l(\theta, a).$$

But if a^* is the action that gives the minimum $l(\theta, a)$, and one considers the decision rule $\delta^*(x) \equiv a^*$, then

$$\min_\delta R(\theta, \delta) \leq R(\theta, \delta^*) = E[l(\theta, a^*)] = l(\theta, a^*) = \min_a l(\theta, a) \text{ (Lindgren, 1971).}$$

Since the minimum regret over all rules δ is neither > nor < the minimum loss over all actions a, these must be equal. Therefore, the expected regret is the same as the "regretized" risk and is given as $E[r(\theta, \delta(X))] = R(\theta, \delta) - \min_\delta R(\theta, \delta)$.

Example 3-1: (cont'd)
(Lindgren, 1971)

Recall this example with the updated table of expected regrets. Shown also are the maximum risks and the Bayes risk for the corresponding prior probabilities $g(\theta)$.

$g(\theta)$	δ_1	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7	δ_8
.8	1	.9	.7	.4	.6	.3	.1	0
.2	0	.5	.4	.1	.9	.6	.5	1
$M(\delta)$	1	.9	.7	.4	.9	.6	.5	1
$B(\delta)$	.8	.82	.64	.34	.66	.36	.18	.2

Suppose we now from graphically analyzing this problem that the minimax mixed decision rule is a mixture decision rules 4 and 7 and can be computed as

$$p\begin{pmatrix}.4\\.1\end{pmatrix} + (1-p)\begin{pmatrix}.1\\.5\end{pmatrix}$$

which gives

$$.4p + .1 - .1p = .1p + .5 - .5p.$$

Thus, $p = 4/7$ and the minimum maximum risk is computed by substituting the value for p back in the equation above and obtaining the value 19/70. The Bayes approach works in much the same way, except that the prior distribution would be utilized.

Section 3.5: The Posterior Distribution

For completeness, a brief description of the posterior distribution is provided here within the decision model framework. Some of the concepts or relationships between conditional, marginal and joint were addressed in **Section 1** on probability measures.

Let X represent data with possible values $x_1,...,x_k$ and let Θ denote the state of nature, with possible values $\theta_1,...,\theta_m$. Assume that the distributions of X under the various states of nature are given:

$$f(x_i \mid \theta_j) = P(X = x_i \mid \Theta = \theta_j),$$

as well as the prior distribution for Θ:

$$g(\theta_j) = P(\Theta = \theta_j).$$

From these the joint probabilities $p(x_i, \theta_j)$:

$$p(x_i, \theta_j) = f(x_i \mid \theta_j) g(\theta_j),$$

and the marginal probabilities for X:

$$p(x_i) = \sum_j f(x_i \mid \theta_j) g(\theta_j)$$

can be constructed. The conditional probabilities for Θ given X=x_i are

$$h(\theta_j \mid x_i) = \frac{p(x_i, \theta_j)}{p(x_i)} = \frac{f(x_i \mid \theta_j) g(\theta_j)}{\sum_j f(x_i \mid \theta_j) g(\theta_j)}.$$

This relation is essentially Bayes theorem. The function $h(\theta_j \mid x_i)$ is the posterior probability function for Θ, corresponding to the given prior, $g(\theta)$.

There are 2 notes to make before closing this section (Lindgren, 1971).

1. When the distribution of the data X is unrelated to the state of nature (that is, is the same for all states), then the posterior probabilities are equal to the prior probabilities. Thus, if $f(x_i \mid \theta) \equiv k(x_i)$,

$$h(\theta \mid x_i) = \frac{f(x_i \mid \theta) g(\theta)}{k(x_i)} = g(\theta).$$

So the observation X does not alter the odds on the various states of nature from what they were before the observation.

2. If a prior belief is sufficiently strong, observations will not alter it. Thus if $g(\theta^*) = 1$ and $g(\theta) = 0$ for any other θ, then

$$k(x) = \sum_i f(x \mid \theta_i) g(\theta_i) = f(x \mid \theta^*);$$

and then provided $f(x \mid \theta^*) \neq 0$,

$$h(\theta \mid x) = \frac{f(x \mid \theta) g(\theta)}{f(x \mid \theta^*)} = \begin{cases} 1, & \text{if } 0 = \theta^*, \\ 0, & \text{for any other } \theta. \end{cases}$$

Section 3.6 Successive Observations

This is adapted from Lindgren (1971) and will be discussed further in the Part on Sequential Statistical Decision Theory.

The previous subsection has indicated that an observation can alter prior odds to posterior odds. The concern or question arises that if the first posterior distribution were considered as if it were a prior, should it result in yet a new posterior distribution? The interest holds as if a posterior distribution is used as a prior distribution with new data, the resulting posterior distribution is the same as if one had waited until all the data were at hand to use with the original prior distribution for a final posterior distribution.

Suppose that two observations are made, (X, Y), with probability function

$$f(x, y \mid \theta) = P(X = x, Y = y \mid \Theta = \theta).$$

Expressing conditional probabilities using the defining formulas in terms of joint probabilities gives

$$P(X = x, Y = y \mid \Theta) = \theta = P(X = x \mid \Theta = \theta) P(Y = y \mid X = x \text{ and } \Theta = \theta),$$

which can be re-written as

$$f(x, y \mid \theta) = f(x \mid \theta) f(y \mid x, \theta).$$

The posterior probabilities for θ given X=x, and given a prior $g(\theta)$ are

$$h_1(\theta \mid x) = \frac{f(x \mid \theta) g(\theta)}{\sum f(x \mid \theta_i) g(\theta_i)} = \frac{f(x \mid \theta) g(\theta)}{k(x)}.$$

Using this as a prior probability for $\Theta = \theta$ together with the new observation $Y=y$ yields the posterior probability

$$h_2(\theta \mid x, y) = \frac{f(y \mid x, \theta) h_1(\theta \mid x)}{\sum f(y \mid x, \theta_j) h_1(\theta_j \mid x)}$$
$$= \frac{f(y \mid x, \theta) f(x \mid \theta) g(\theta) / k(x)}{\sum f(y \mid x, \theta_j) f(x \mid \theta_j) g(\theta_j) / k(x)}$$
$$= \frac{f(x, y \mid \theta) g(\theta)}{\sum f(x, y \mid \theta_j) g(\theta_j)}.$$

This is precisely the posterior probability that would be computed based on data (X, Y) and the prior $g(\theta)$.

Section 3.7 Bayes Actions from Posterior Probabilities

The following discussion is found in Chapter 4 of Lindgren (1971), Chapter 4 of Berger (1985), Chapter 6 of French and Insua (2000) and Chapter 3 of Raiffa and Schlaifer (2000).

The Bayes decision rule for a given problem, corresponding to a certain prior distribution can be found by applying the Bayes principle to the risk table - weighting the various states of nature and choosing the decision function $\delta(\cdot)$ so as to minimize the Bayes risk, $B(\delta)$. This function of δ is the expected value of the risk

$$B(\delta) = \sum_i R(\theta_i, \delta) g(\theta_i).$$

Consider a different formalization.

Step 1: Recall the expression for $R(\theta, \delta)$ as it is computed from the loss function is:

$$R(\theta, \delta) = \sum_j l(\theta, \delta(x_j)) f(x_j \mid \theta).$$

Step 2: Substitute this into $B(\delta)$ yields

$$B(\delta) = \sum_i \sum_j \{l(\theta_i, \delta(x_j)) f(x_j \mid \theta_i)\} g(\theta_i).$$

Step 3: According to Bayes' theorem, the product $f(x_j \mid \theta_i) g(\theta_i)$ can be replaced by the product of the conditional probability in which the condition is the value of X, and the marginal probability $X=x_j$:

$$f(x_j \mid \theta_i)g(\theta_i) = h(\theta_i \mid x_j)p(x_j).$$

This is calculated by summing the joint probabilities for (Θ, X) over θ (which would be relevant in determining the posterior probabilities but is not at the moment):

$$p(x_j) = \sum_i P(\Theta = \theta_i, X = x_j)$$
$$= \sum_i f(x_j \mid \theta_i)g(\theta_i).$$

Step 4: The indicated double summation can be calculated equally well by summing first on *i* and then on *j*:

$$B(\delta) = \sum_j \left\{ \sum_i l(\theta_i, \delta(x_j)) f(x_j \mid \theta_i) g(\theta_i) \right\}$$
$$= \sum_j \left\{ \sum_i l(\theta_i, \delta(x_j)) h(\theta_i \mid x_j) \right\} p(x_j),$$

where the last expression is obtained by making the substitution from Bayes' theorem, as given above.

Note: The quantity in braces depends on the observed value x_i and on the action assigned to x_i by the decision function used:

$$L_h(\delta, x_j) = \sum_i l(\theta_i, \delta(x_j)) h(\theta_i \mid x_j).$$

It is the average of the losses $l(\theta, \delta(x_j))$ with respect to the posterior probability weight which is referred to as the expected posterior loss.

Step 5: Since $B(\delta)$ is a weighted sum of expected posterior losses with nonnegative weights:

$$B(\delta) = \sum_j L_h(\delta, x_j) p(x_j),$$

where L_h is made as small as possible. So if (for the observed to x_i) $\delta(x_j)$ is taken to be the action that minimizes the expected posterior loss L_h, the resulting rule is one that minimizes the Bayes risk and is a Bayes procedure (Lindgren, 1971).

To obtain the Bayes action given a particular *X*, it is not necessary to go through the computation of the whole decision rule where the number of decision functions is of a higher order than the number of actions.

Mathematically, the present approach is a simple process of minimizing a function of the actions, namely, the expected posterior loss, by selecting one of those actions; the earlier approach is a more complicated process of minimizing a function of the decision rules, over all possible decision rules. In summary, the earlier approach incorporates the data into the loss function to obtain the risk function where the original prior is then used on those risks to determine the Bayes rule. The present approach incorporates the data into the prior distribution to obtain the posterior distribution, which is used as an educated "prior" distribution on the original losses. The results are equivalent.

Part II: UTILITY THEORY - Concepts and Methods

Decision making under uncertainty places the evaluation of the consequences of possible actions at the forefront of the problems concerning decision problems. There are two main paradigms for modeling human uncertainty which has been applied to decision making under uncertainty. One involves probabilistic and statistical reasoning. Most common is the Bayesian perspective coupled with expected utility maximization. The other classification is rule-based deductive systems which are based on axiomatic foundations. Of particular interest is Savages expected utility which involves binary relations over functions defined on a measurable space. In contrast, von Neumann-Morgenstern's theory of expected utility maximization under risk derived a basic set of axioms from which was deduced the existence of mathematical functions which have useful properties in comparing alternative states and preferences.

The first section of this part presents the basic ideas and concepts of utility theory. It also introduces the utility within the context of monetary values. Sections 5 and 6 focus on the development of utility theory based on axiomatic foundations. The middle section focuses on the concept of ordinal utility and shifts to discuss the concept of subjective probability. The last section is a comparative assessment of the various theories that were developed, demonstrating the differences and similarities in their conception.

Section 4: Utility Theory – From Axioms to Functions

An analytic study of a decision problem seems to require the assumption of a mathematical model or structure of ordering among the various possible consequences of taking an action in the face of a particular state of nature. In evaluating these consequences of possible actions, two major problems are encountered. The first is that the values of the consequences may not have any scale of measurement; and the second is that even when there is a clear scale (usually monetary) by which consequences can be evaluated, the scale may not reflect the "true" value to the decision maker (Berger, 1985).

In speaking about what one is faced with as the result of making a decision when nature is in a certain state, the term consequence has been used, because the situation has been brought about in part by the making of the decision. In discussing utility, it is preferable to use the term prospect, without the implication that the future history has necessarily been brought upon by a decision-which happens to be the case in decision theory but is not generally so in the theory of utility.

A mathematical analysis is simpler, if it is possible to put prospects, not just in order, but on a numerical scale-to assign a numerical measure to each prospect. With reasonable assumptions about one's preferences among the various prospects, such an assignment, or function, can be achieved. The numbers are referred to as utilities and subsequently utility theory deals with the development of such numbers. Thus, utility is a function defined on the various prospects (or consequences) with which one is faced, measuring the relative desirability of these prospects on a numerical scale.

This section focuses on the basic axioms and general concept of utility while the proceeding sections discuss the development of the different "types" of utility within individual rational decision making.

Section 4.1: Preference Axioms

Every mathematical structure is based on a set of axioms. Acceptance of the axioms consolidates the acceptance of a whole string of consequences, or theorems, that follow from the axioms. This section presents the underpinning axioms, dealing with preferences, which imply the existence of a utility function, providing a numerical scale in terms of which the consequences of actions are assumed to be measured.

A notion of preference, or relative desirability, among prospects is assumed. The following notation[1] of the basic binary relation represents a weak preference taken to be at least as desirable:

$$P_1 \underset{\sim}{\succ} P_2 \text{ means: } \textit{prospect } P_1 \textit{ is at least as desirable as prospect } P_2.$$

[1] The symbols $\underset{\sim}{\succ}$ and $\succ$ will occasionally be used in the other direction, with the obvious meaning, namely, that the prospect on the small end of the symbol is the less desirable.

This relation is weak which assumes properties of completeness and transitivity (French and Insua, 2000). This latter trait – transitivity is stated below as axiom 2 and holds for the following two relations as well. This aforementioned weak order can be used to define the equal desirability of two prospects:

$$P_1 \sim P_2 \text{ means: } \textit{prospect } P_1 \textit{ and } P_2 \textit{ are equally desirable},$$

which is defined to mean that both $P_1 \underset{\sim}{\succ} P_2$ and $P_2 \underset{\sim}{\succ} P_1$ or simply that they are indifferent. Strict preference is expressed in the following manner:

$$P_1 \succ P_2 \text{ means: } \textit{prospect } P_1 \textit{ is preferred over prospect } P_2,$$

which is a condition that (by definition) exists when and only when $P_1 \underset{\sim}{\succ} P_2$, but P_1 and P_2 are not equally desirable. This latter relation $\sim$ is reflexive, that is, $P_1 \sim P_1$, if it is understood that a prospect is at least as good as itself. Moreover, by definition it is symmetric; that is $P_1 \sim P_2$ is equivalent to $P_2 \sim P_1$ (Lindgren, 1971).

These relations (along with axiom 2 – transitivity) result in the following consequences:

(i) If $P \sim Q$ and $Q \sim R$, then $P \sim R$.
(ii) If $P \underset{\sim}{\succ} Q$ and $Q \sim R$, then $P \underset{\sim}{\succ} R$.
(iii) If $P \succ Q$ and $Q \sim R$, then $P \succ R$.
(iv) If $P \succ Q$ and $Q \succ R$, then $P \succ R$.

Before clearly stating the key axioms, it is necessary to address the concept of a "mixture" of prospects (Lindgren, 1971). These are summarized into three aspects.

(1) If $\exists$ with probability p of a prospect P_1 and a probability (1-p) of a prospect P_2, this is called a random prospect and is considered to be a mixture of the prospects P_1 and P_2 and is denoted by $[P_1, P_2]_p$.

(2) More generally, if one faces P_1 with probability p_1, P_2 with probability p_2, and P_k with probability p_k, this is again a random prospect, called a mixture of $P_1,\ldots, P_k$ and denoted $[P_1,...,P_k]_{(p_1,...,p_k)}$.

(3) Similarly, the notation for considering random prospects composed of infinite prospects, P_1 with probability p_1, P_2 with probability p_2, and so forth, where $p_1+ p_2+\ldots=1$ is $[P_1, P_2,...]_{(p_1,p_2,...)}$.

These have been summarized into the adopted axioms as follows (Lindgren, 1971) & (Berger, 1985):

AXIOM 1: *Given any prospects P and Q, then either*

$$P \succ Q \text{, or } P \prec Q \text{, or } P \sim Q.$$

Axiom 1 states that, given any two prospects, either one is preferred over the other, or the other over the one, or they are equally desirable.

AXIOM 2: *If* $P \underset{\sim}{\succ} Q$ *and* $Q \underset{\sim}{\succ} R$ *then* $P \underset{\sim}{\succ} R$.

Axiom 2 is the axiom of transitivity of the preference relation. Its assumption precludes a type of inconsistency in the mathematical structure that may or may not accurately describe one's preferences.

AXIOM 3: *If* $P_1 \succ P_2$, *then for any probability p and prospect P, it follows that*

$$[P_1, P]_p \succ [P_2, P]_p.$$

AXIOM 3`: *If* P_1, P_2,… *and* Q_1, Q_2,… *are sequences of prospects such that* $P_i \succ Q_i$ *for i =1,2,..., then for any set of probabilities* $\alpha_1, \alpha_2, ...(\sum \alpha_i = 1)$

$$[P_1, P_2, ...]_{(\alpha_1, \alpha_2, ...)} \succ [Q_1, Q_2, ...]_{(\alpha_1, \alpha_2, ...)}.$$

Axiom 3 says that improving one of the prospects in a finite mixture improves the mixture, and Axiom 3` extends this notion to countably infinite mixtures.

AXIOM 4: *Given three prospects in the order* $P_1 \succ P_2 \succ P_3$, *there are mixtures* $[P_1, P_3]_p$ *and* $[P_1, P_3]_r$ *such that*

$$P_1 \succ [P_1, P_3]_p \succ P_2 \succ [P_1, P_3]_r \succ P_3.$$

Axiom 4 says when $P_1 \succ P_2 \succ P_3$ that there is no P_1 so wonderful that the slightest chance of encountering it instead of P_3 is better than any ordinary P_2; and that there is no P_3 so terrible that the slightest chance of encountering it instead of P_1 is worse than any ordinary P_2.

The following consequences are implied (Lindgren, 1971). For instance $P_1 \succ P_0$, it then follows that

(v) $\quad P_1 \succ [P_1, P_0]_p \succ P_0$, for $0 < p < 1$.

Since P_1 can be thought of as $[P_1, P_1]_p$ and $[P_0, P_0]_p$, so that (by Axiom 3)

$$P_1 = [P_1, P_1]_p \succ [P_1 \succ P_0]_p \succ [P_0, P_0]_p = P_0.^{2}$$

(vi) $[P_1, P_0]_p \succ [P_1, P_0]_q$, if $P_1 \succ P_0$ and $0 \le q < p \le 1$.

This states that one mixture of two prospects is preferred over another if there is a stronger probability of the more desirable prospect in the first mixture. That is as p increases from 0 to 1, the mixture $[P_1, P_0]_p$ becomes increasingly more desirable. A further consequence of the axioms is the existence of a utility function which is to follow.

Section 4.2 Coding Intermediate Prospects

The previous sub-section showed that given any two distinct prospects P_0 and P_1 with $P_0 \prec P_1$, all mixtures of these prospects lie between them in order of desirability:

$$P_0 \prec [P_1, P_0]_p \prec P_1.$$

Moreover, the larger p the more desirable is the mixture. Now, consider that if P is an arbitrary prospect between P_0 and P_1:

$$P_0 \prec P \prec P_1,$$

then there is some mixture of P_0 and P_1 equivalent to P.

THEOREM: *Given any two prospects P_0 and P_1 such that $P_0 \prec P_1$, and given any prospect P such that $P_0 \prec P \prec P_1$, then there is a unique number p between 0 and 1 such that $[P_1, P_0]_p$ is equivalent to P. (Lindgren, 1971)*

This result means that the prospects intermediate to given prospects can be thought of as lined up on the scale of numbers from 0 to 1, each being identified with a number on that scale. Equivalent prospects are located at the same point on the scale, and the ordering of prospects in desirability corresponds to ordering on the scale – from least desirable to more desirable prospects. A utility function is now easily constructible which provides a coding of utilities that gives a numerical representation of preferences-by taking the utility of a prospect to be just the number on the scale from 0 to 1. Thus, if $P \sim [P_1, P_0]_p$, define $u(P) = p$. For all prospects between P_0 and P_1, as well as for P_0 and P_1 themselves

$$u(P_0) = 0, \text{ and } u(P_1) = 1.$$

[2] The equality of two prospects means that they are precisely the same prospect.

There are certain properties satisfied by the utility function[3], defined for any P between given P_0 and P_1, as the value p in the mixture $[P_1, P_0]_p$ which is equivalent to P, as follows:

Utility Property A: If $P \succ Q$, then $u(P) \geq u(Q)$.

Property A, for any prospects P and Q between P_0 and P_1, follows from the fact that mixtures of P_0 and P_1 are ordered according to the proportion of P_1 in the mixture; thus, if $P = [P_1, P_0]_p$ and $Q = [P_1, P_0]_q$, and $P \succ Q$, then $p > q$. But $u(P) = p$ and $u(Q) = q$, which yields Property A.

Utility Property B: $u([P,Q]_r) = ru(P) + (1-r)u(Q).$

Property B, for these P and Q, observe that

$$[P,Q]_r = [[P_1, P_0]_p, [P_1, P_0]_q]_r = [P_1, P_0]_{pr+q(1-r)'}$$

which implies that the utility in $[P,Q]_r$ is

$$u([PQ]_r) = pr + q(1-r) = ru(P) + (1-r)u(Q).$$

Property B can be expressed in terms of expectation. For a mixed prospect $[P,Q]_r$, the utility is a random variable with possible values $u(P)$ and $u(Q)$ and corresponding probabilities r and 1-r; the expected utility is therefore $ru(P) + (1-r)u(Q)$, which is the utility assigned to the random prospect. The utility of a random prospect is the expected value of the utility (Berger, 1985).

Berger (1985) describes a 5 step procedure in constructing a utility function much of which is described above. He also includes in his construction of the utility function the fact that any linear function of such a utility would also satisfy **Utility Properties A** and **B**[4], and serve just as well as a utility function.

Defining a function $v(P)$ on prospects between P_0 and P_1 as follows:

$$v(P) = au(P) + b,$$

for $a > 0$, has the simple effect of shifting the origin and introducing a scale factor. The utility of P_0 and P_1 in this new scheme would be

$$v(P_0) = b \text{ and } v(P_1) = a + b.$$

[3] Berger, 1985.
[4] French and Insua, 2000.

Property A, for $v(P)$ is satisfied. To obtain property B compute

$$\begin{aligned} v([P,Q]_r) &= au([P,Q]_r) + b \\ &= a[ru(P) + (1-r)u(Q)] + b \\ &= r[au(P) = b] + (1-r)[au(Q) + b] \\ &= rv(P) + (1-r)v(Q). \end{aligned}$$

The matter of utilities for prospects that are not between two given prospects P_0 and P_1 has been neglected in this report. This concludes this sub-section.

Section 4.3 Utility for Money

In many practical situations, prospects are expressible in terms of amounts of money. In such cases, it is tempting to treat the number of dollars as utility, or as proportional to utility: $u(M) = M$ or $u(M) = kM$. Recall the introduction to this section which addressed that discounting the possibility that such a utility function overlooks other aspects of prospects addresses the issue that this scale of values is not usually a totally adequate basis in terms of which to analyze decision problems involving money. For one thing, the function $u(M) = M$ is not a bounded function of *M*, at least if *M* is allowed to range over all real numbers. The famous example of the "St. Petersburg Paradox" illustrates the difficulty (Lindgren, 1971) and (Berger, 1985):

Example 6-1:

You are offered, for a fee, the following random prospect. You will be given $\$2^N$ if in repeated tosses of a coin the first heads does not appear until the *N*th toss. What entry fee would you be willing to pay?

If money is utility, (that is $u(M) = M$), you should be willing to pay any fee up to the expected value of the payoff, because your utility would not be decreased by so doing. It can be shown that the probability of obtaining heads for the first time on the *N*th toss is $½^N$, and accepting this one obtains as the expected payoff:

$$E(payoff) = \sum_1^\infty 2^N \bullet \frac{1}{2^N} = 1 + 1 + 1 + \ldots = \infty.$$

(That is, the sum is not a finite number.) Despite the infinite expected payoff, it is found that people generally will not pay even a large but finite entry for the proposed game.

On the other hand, if it is assumed that one's utility function is

$$u(M) = \begin{cases} 2^M & \text{if } M \le 20, \\ 2^{20} = 1{,}048{,}576 & \text{if } M > 20, \end{cases}$$

the expected utility of the payoff is

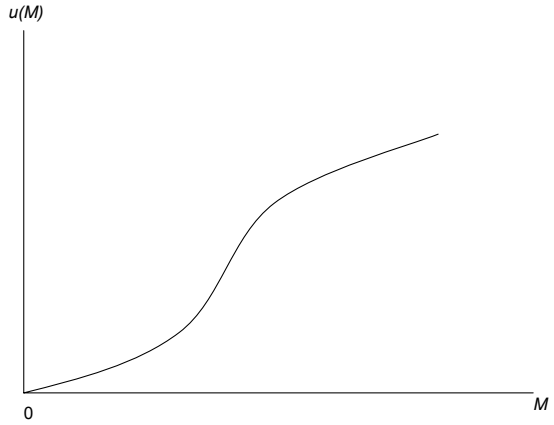
$$E[u(Payoff\,0] = \sum_{1}^{20} 2^N \bullet \frac{1}{2^N} + \sum_{21}^{\infty} (2^{20}) \frac{1}{2^N} = 20 + \left(\frac{1}{2} + \frac{1}{4} + ... \right) = 21.$$

The game then has the same utility as \$21.

Another aspect of the paradox is that no one could in fact offer the game honestly - no one has $\$2^{40}$ to pay out in the event of 39 tails in a row before the first heads, for example. On the other hand, if a gambling house with a capital of \$1,048,576 agrees to pay out the entire amount of N =20 or more, the expected payoff is \$21, as the computation above also shows.

Even without the paradox arising from permitting one's capital M to have any finite value, as in the example above, most people's utility for money is not strictly linear; although, it may be approximately linear over a restricted range of values of M. **Figure 4-1** shows a utility function $u(M)$ plotted against M that is something like most people's utility functions for money.

Figure 4-1:

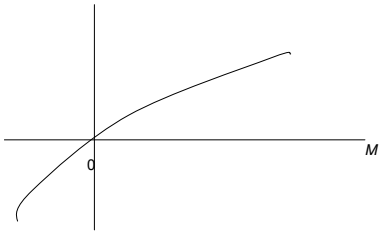


The function shown above is an increasing function of M, corresponding to the assumption that the greater the monetary value of a prospect the greater its desirability. It is unusual to encounter a situation in which a greater amount of capital is "worth" less than a smaller amount. The monotonic nature of the curve representing utility for money would show that that if one is offered a favorable chance prospect involving money repeatedly, the offer should be accepted. That is, if there is a net monetary gain each time, the decision maker's monetary assets increase - and therefore utility increases.

The expected utility is the appropriate guide for action in a single experiment, rather than expected monetary gain, but in a long sequence of experiments the monetary gain is relevant. In dealing with a given problem it is often convenient to focus attention on the

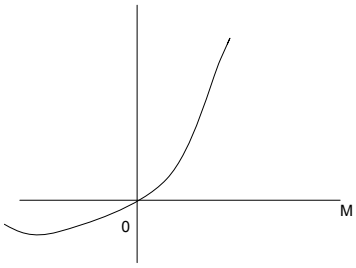
small portion of the curve representing the function $u(M)$. Label the money axis in terms of the amount of money to be lost or gained, rather than in terms of total assets. If M denotes an increase or decrease from an initial status M_0, $u(M)$ is really $u(M + M_0)$. It may also appear as though a person's utility function is different in different situations reflected as different points on the overall utility curve. Figure 4-1 showed that it is possible to have (locally) a utility for gain or loss that is concave down in some instances, and concave up in others, as shown in the figure below.

Figure 4-2(a):



This would be appropriate if the initial capital was near 0.

Figure 4-2(b):



This would be appropriate if the initial capital was large.

Section 4.4 Bets, Fair and Unfair

"A bet is a random or mixed prospect [of the type encountered at the beginning of the previous section]: One starts with an initial capital M_0 and according to the outcome an experiment of chance ends up either with M_L, an amount less than M_0, or with M_W, an amount greater than M_0. In the former case he loses, his capital being reduced by the amount $y = M_0 - M_L$, and in the latter case he wins, his capital being increased by the amount $x = M_W - M_0$. It is said he "puts up with" y and his opponent "puts up with" x, the winner taking both amounts." (Lindgren, 1971)

Definition:

The money odds in such a bet are said to be "y to x" that one wins the bet, or y/x. The probability odds in such a bet are said to be "p to 1-p" that one wins the bet, or $p/(1-p)$, where p is the probability that he wins. A bet is said to be a fair bet of the probability odds and the money odds are equal.

A fair bet can be thought of as the invariant principle addressed in **Section 2**. It is "fair" in the sense that neither side is favored. If the bet is fair, then

$$\frac{y}{x} = \frac{p}{1-p}, \text{ or } y(1-p) - px = 0.$$

Substitution of $y = M_0 - M_L$ and $x = M_W - M_0$ yields

$$(M_0 - M_L)(1-p) - (M_W - M_0)p = 0,$$

or

$$pM_W + (1-p)M_L = M_0.$$

A bet can be redefined as one in which the expected capital after the bet is equal to the capital of before the bet. There is no loss nor gain in capital on the part of the one bettor; however, if one player experienced a net expected loss, the bet would not be fair.

When utility is measured in terms of money, that is when $u(M) = kM$, the indifference relation stands – accepting or not accepting the bet are equally desirable:

$$u(bet) = pu(M_W) + (1-p)u(M_L) = k(pM_W + (1-p)M_L)$$
$$= kM_0 = u(M_0).$$

The equal desirability follows from the fact that taking the bet has the same utility as the initial capital and will be re-addressed in the sections concerning the development of utility theory.

Consider a general non-linear utility function $u(M)$, and a fair bet:

$$M_0 = pM_W + (1-p)M_L.$$

The utility of the bet is

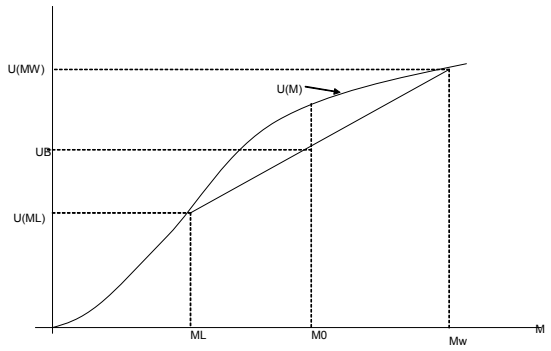
$$U_B = pu(M_W) + (1-p)u(M_L).$$

In vector notation

$$\begin{pmatrix} M_0 \\ U_B \end{pmatrix} = p\begin{pmatrix} M_W \\ u(M_W) \end{pmatrix} + (1-p)\begin{pmatrix} M_L \\ u(M_L) \end{pmatrix},$$

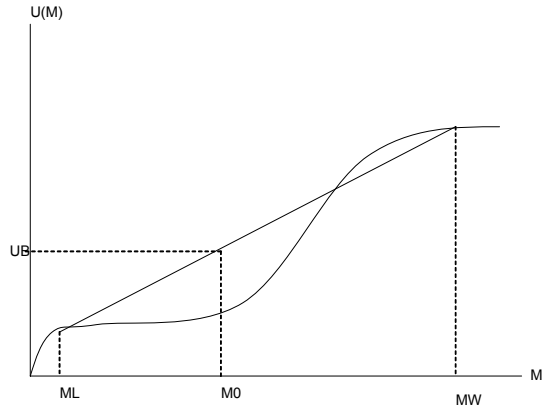
which is interpreted geometrically to mean that the point (M_0, U_B) is a convex combination of the two points $(M_W, u(M_W))$ and $(M_L, u(M_L))$. Thus, the point (M_0, U_B) lies on the line segment joining those two points, as shown in the figures below. Two cases are shown when the person is not indifferent to the fair bet.

Figure 4-3(a):



This represents a utility function and fair bet that is attractive.

Figure 4-3(b):



This represents a utility function and fair bet that is unattractive.

Consider a bet that is not fair, biased against the person whose interests are of concern and in terms of whose utility the situation is being studied; in such a case the expected capital after the bet is less than the initial capital:

$$p'M_W + (1-p')M_L = M' < M_0,$$

where p' is the probability of winning, and is smaller than the p corresponding to a fair bet for the same amounts of money. The above inequality can be restated in the form

$$\frac{p'}{1-p'} < \frac{M_0 - M_L}{M_W - M_0} = \frac{y}{x},$$

which says that the money odds are higher than the actual probability odds.

If U' denotes the utility in this unfair bet:

$$U' = p'u(M_W) + (1-p')u(M_L),$$

then the point (M', U') lies on the line segment joining the points $(M_W, u(M_W))$ and $(M_L, u(M_L))$:

$$\begin{pmatrix} M' \\ U' \end{pmatrix} = p' \begin{pmatrix} M_W \\ u(M_W) \end{pmatrix} + (1 - p') \begin{pmatrix} M_L \\ u(M_L) \end{pmatrix}.$$

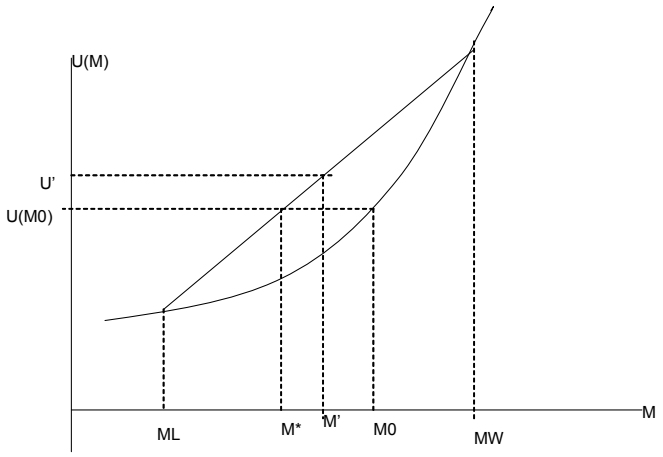
This is shown in ***figure 4-4***, in which is also plotted M_0, the initial capital, which is greater than the expected capital after the bet, M'. Decreasing M' amounts to decreasing p', that is, altering the probability odds, and there is a p^* corresponding to the smallest M^* for which the bet would be worth taking:

$$p^* M_W + (1 - p^*) M_L = M^*$$

or

$$p^* = \frac{M^* - M_L}{M_W - M_L}.$$

figure 4-4:



It was shown above a case in which the utility is such that a fair bet is advantageous. In much the same manner, it can be shown that a utility is such that a fair bet is not advantageous, which can be used to determine how unfair it is. This is omitted here. Nau (2000) discusses the issues surrounding fair bets and utility for money under equilibrium in depth. He characterizes those who are risk averse and those who are not in delimitating their actions under varying conditions.

Section 5: Development of Utility Theory I – Ordinal Utility

The development of rational choice[5] [decision] theory (in its modern form) dates back 50 years to a trio of publications: von Neumann and Morgenstern's *Theory of Games and Economic Behavior* (1944/1947), Kenneth Arrow's *Social Choice and Individual Values* (1951), and Savage's *Foundations of Statistics* (1954)[6]. These papers laid the foundation in the use of mathematical methods such as some of those discussed in **Section 6**. These include axiomatic methods, concepts of measurable utility and personal probability, and tools of general equilibrium theory and game theory. The focus of the previous section was to address some of these aspects while these next two sections discuss the development of utility theory.

For brevity purposes, this report only reviews models of individual rationality. These include **ordinal utility** (for choice under conditions of "certainty"), **expected utility** (for choice under conditions of "risk," where probabilities for events are objectively determined), **subjective expected utility** and **state-preference theory** (for choice under conditions of "uncertainty," where probabilities for events are subjectively determined—or perhaps undetermined), and several kinds of **non-expected utility theory**. The prior presentation of the underpinning preference axioms is revisited briefly in this section, providing some background on the concept of "ordinal" utility, prior to the discussions of "cardinal" expected utility presented in the next section.

Section 5.1: Ordinal Utility

Returning to the notation for preference relations and the stated axioms, suppose that there are only two preferences – for instance, apples and bananas[7]. A pair of numbers (a, b) then represents a possible bundle of commodities (namely a apples and b bananas) that a person might possess. Now suppose that this person has **preferences** among such groups of apples and bananas. If those preferences are complete, reflexive, transitive, and continuous (see **Section 6.1**), then there exists a utility function U which represents them, so that $(a, b) = (a', b')$ if and only if $U(a, b) = U(a', b')$. Recall the notion of indifference. Then it can be said that for any group or bundle (a, b) it is possible to find a unique number x such that the bundle (x, x), which has equal amounts of every preference, is precisely indifferent to (a, b). Let that number x be defined as $U(a, b)$. But this is not the only possible way to define a utility function representing the same preferences. For example, we could just as well define $U(a, b)$ to be equal to 2^x or $f(x)$ where f is any monotonic (i.e., strictly increasing) function. This was shown in **Section 4**.

Definition:

[5] Several economics and operations research papers interchange between the terms choice and decision theory.
[6] Nau, 2000.
[7] This could be anything. Lindgren uses beef and chicken in his example. The idea is to be able to assign a numeric value to something which may be numerically measured in terms of preference for a particular person. Much of this theory presented here is crucial to the understanding of consumer theory as found in microeconomics.

An ordinal value function $v : A \to \Re$ is a real-valued function which represents –or agrees with a weak preference in the following sense:

$$\forall a, b \in A, v(a) \geq v(b) \Leftrightarrow a \underset{\sim}{\succ} b .$$

The above definition was provided by French and Insua (2000) and simply states that an ordinal value function is a numerical assignment to a set of alternatives which represents the person or decision maker's preference ranking by a numerical ranking. Thus, the utility function representing a decision maker's preferences is said to be merely an **ordinal utility** because only the *ordering* property of the utility numbers is meaningful.

Example 5-1:

For example, suppose that $U(1, 1) = 1$, $U(2, 1) = 2$, $U(1, 2) = 3$, and $U(2, 2) = 4$. In other words, referring back to apples and bananas, one apple and one banana yield one unit of utility, two apples and one banana yield two units of utility, etc. Then this means *only* that the preference ordering is (2, 2) > (1, 2) > (2, 1) > (1, 1), i.e., two apples and two bananas are better than one apple and two bananas which are better than two apples and one banana which are better than one apple and one banana. However, it may not be concluded that two apples and one banana are "twice as good" as one apple and one banana or that one additional apple yields the same "increase in utility" regardless of whether you have one banana or two bananas to start with.

Recall that any monotonic transformation of U carries exactly the same preference information: if f is a monotonic function, and if V is another utility function defined by $V(a, b) = f(U(a, b))$, then V encodes exactly the same preferences as U. For example, letting $f(x) = 2^x$, we obtain a second utility function satisfying $V(1, 1) = 2$, $V(2, 1) = 4$, $V(1, 2) = 8$, and $V(2, 2) = 16$, which provides a completely equivalent representation of this hypothetical decision maker's preferences.

Ordinal value functions represent preference rankings and nothing more. There is no further information in the numbers used to represent the decision maker's preferences, such as strength of preference. More formally, the following theorem is given by French and Insua (2000).

THEOREM: *For any domain of objects A with a weak order* $\underset{\sim}{\succ}$*, there exists an ordinal value function if an only if there exists a countable order dense subset* $B \subset A$*, viz. B is such that* $\forall a_1, a_2 \in A, a_1 \underset{\sim}{\succ} a_2, \exists b \in B$ *such that* $a_1 \underset{\sim}{\succ} b \underset{\sim}{\succ} a_2$.

Thus, any strictly increasing transformation of an ordinal value function also agrees with the underlying weak preferences. Conversely, any two ordinal value functions which agree with the same preference relation are related by a strictly increasing transformation. Thus, the representation is said to be "unique up to a strictly increasing transformation" (French and Insua, 2000).

Section 5.2: Marginal Utility

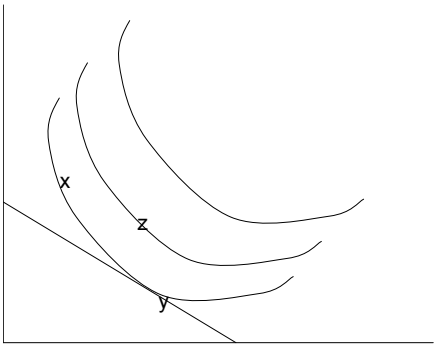
Suppose now that the non-unique ordinal utility function of concern is *differentiable*—i.e., it is a smooth function of the numbers of apples and bananas. Then **marginal utility (**of an additional apple or banana)[8] can be defined as the partial derivative of the utility function (with respect to apples or bananas), evaluated at the point of interest. Since it was previously stated that ordinal utilities are nearly-meaningless numbers, it might be plausible to believe that marginal utilities would be nearly-meaningless numbers as well; however, the *ratios* of marginal ordinal utilities are always uniquely defined. The ratios of marginal utilities are precisely the **marginal rates of substitution** between the prospects that leave total utility unchanged. Marginal rates of substitution are uniquely determined by preferences and as such they are *observable*, hence any two utility functions that represent the same preferences must yield the same ratio of marginal utilities between any two preference groups (Nau, 2000).

For example, if the ratio of the marginal utilities of apples to bananas is equal to 1/2 when the decision maker already has one apple and one banana means that the person would indifferently trade an ε fraction of an additional banana for a 2ε fraction of an additional apple. Then the same ratio must be obtained for any other utility function that is a monotonic transformation of the original one.

Section 5.3: Indifference Curves

Section 6 illustrated the geometrical perspective of a decision maker's utility curve. So, to recap, geometrically, an ordinal utility function determines the shape of **indifference curves** in the space of possible prospects or groups of prospects. An indifference curve is a set of points representing endowments that all yield the same numerical utility—i.e. that are equally preferred. A decision maker's indifference curves in $\Re^2$ space is illustrated in figure 5-1.

Figure 5-1:



Points ***x*** and ***y*** are on the same indifference curve, meaning that they yield exactly the same utility —i.e., the decision maker is precisely indifferent between them—whereas point ***z*** lies on a "higher" indifference curve, so it would be preferred to either ***x*** or ***y***. The *slope of a tangent line* to an indifference curve (such as the dotted line through point ***y***) is precisely the *marginal rate of substitution* between prospects at that point.

[8] The marginal utility of an apple is not necessarily the increase in utility yielded by an additional whole apple—it is the increase in utility yielded by an additional ε fraction of an apple, divided by ε, where ε is infinitesimally small.

Note 1: A person wishes to climb higher (to another indifference curve) if at all possible. Thus, if the person started at point $\boldsymbol{x}$ on the map, he or she would be indifferent to walking over to point $\boldsymbol{y}$, which is at the same elevation, but would prefer to climb higher up to point $\boldsymbol{z}$.

Note 2: While the shapes of the indifference curves are uniquely determined by preferences, the relative *heights* of the utility surface above different indifference curves are completely *arbitrary*.

Note 3: Another utility function that represented the same preferences would have to yield the same indifference curves, whose tangent lines would have exactly the same slopes at all points, but the corresponding utility surface might otherwise look very different.

Section 5.4: Assumptions - Marginal Utility and Multi-Attributed Theory

The assumption of diminishing marginal utility was briefly addressed in Section 6. This assumption means that if the holdings of all prospects are fixed except one, then the marginal utility of that prospect should decrease as the "endowment" of it increases (Nau, 2000). The term non-satiation refers to the fact that the marginal utility of every prospect always remains positive—i.e., it approaches zero only asymptotically, so that more is always better no matter how much you already have. (See **Section 4.3**.)

Decreasing marginal utility is actually not a strong enough condition for the decision maker's problem to always have a unique solution unless the utility function is also *additively separable*—i.e., of the form $U(a,b) = u_1(a) + u_2(b)$. If the number of prospects is increased to some n - i.e., there exists prospects a, b, c, d, …, etc., then the additive form of utilities refers to the consequences or alternatives to be multi-attributed in $\Re^n$.[9]

A stronger assumption of *strictly convex preferences* is usually made, which means that if $\boldsymbol{x} = \boldsymbol{z}$ and $\boldsymbol{y} = \boldsymbol{z}$, then $\alpha\boldsymbol{x}+(1-\alpha)\boldsymbol{y} > \boldsymbol{z}$ for any number α strictly between 0 and 1, where the vectors $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$ denote general multidimensional preferences. In other words, if $\boldsymbol{x}$ and $\boldsymbol{y}$ are both *weakly* preferred to $\boldsymbol{z}$, then everything "in between" $\boldsymbol{x}$ and $\boldsymbol{y}$ is *strictly* preferred to $\boldsymbol{z}$[10]. In particular, if $\boldsymbol{x}$ and $\boldsymbol{y}$ are indifferent to each other, then $\alpha\boldsymbol{x}+(1-\alpha)\boldsymbol{y}$ is strictly preferred to either of them—i.e., the decision maker would prefer to have an average of $\boldsymbol{x}$ and $\boldsymbol{y}$ rather than either extreme.

This property captures the intuition of diminishing marginal utility, namely that ½ $\boldsymbol{x}$ provides "more than half as much utility" as $\boldsymbol{x}$, but it does so by referring to observable preferences rather than unobservable utility. Lindgren (1971) discusses the case of arbitrary prospects more in depth in Chapter 2.3.

---

[9] This is explored in **Section 8**.
[10] See **Section 6.2**.

If preferences are strictly convex, then any utility function that represents those preferences is necessarily *strictly quasi-concave*, which means that its *level sets* are strictly convex sets, i.e., the set of all $\boldsymbol{y}$ such that $U(\boldsymbol{y}) = U(\boldsymbol{x})$ is a strictly convex set. The strict convexity of the level sets ensures that the decision maker's problem has a unique solution (Nau, 2000)[11].

Section 5.5 Subjective Probability

The classical concept of probability was initially introduced (in ***Part A***) as a repeatable experiment (of chance) that (at each trial) results in one of a given set of possible outcomes. The probability assigned to each outcome is thought of as an idealization of its relative frequency of occurrence in a long sequence of trials. In the case of a chance or random variable, these outcomes are ordinary numbers, characterized by non-numerical descriptions. When the outcomes are thought of as prospects with corresponding probabilities is termed a random prospect. However, this frequency concept does not always suffice when dealing with θ.

The theory of subjective probability was created to enable one to talk about probabilities when the frequency viewpoint does not apply (Berger, 1985). A quantification of "chances" involving one's beliefs or convictions concerning the occurrences of the event, and embodying the odds indicates that the matter of chance in this kind of experiment is actually personal. The term subjective probability is used to refer to probability that represents a person's degree of belief concerning the occurrence of an event. Although it can be assumed that the subjective probability of a given state exists, it is seldom easy to determine that probability (Lindgren, 1971).

Nau (2000) states that there are four "best-known" concepts of probability – (1) the classical interpretation, (2) the frequentist (or objective) interpretation, (3) the logical (or necessary) interpretation, and (4) the subjective interpretation. Although objective probability and subjective probability are different in conception and interpretation, they are not different as mathematical entities, in as much as the same basic axioms are assumed to hold for both. Any definitions or theorems concerning objective probabilities apply equally well to subjective probabilities. In particular, the expected value of a random variable, defined earlier (in Section 6) for the case of objective probabilities, would be defined in exactly the same way if the probabilities are subjective, namely, as the weighted sum of the possible values, where the weights are the corresponding (subjective probabilities).

This section discussed the role of ordinal utility. This closing sub-section has introduced the topic to be discussed in the following section which more formally develops the utility theory within its three comparable forms.

[11] Nau (2000) discusses this within the constructs of consumer theory as understood in microeconomic theory. He also provides a detail development of utility theory and equilibrium states as they formed in utility theory and impacted microeconomic theory.

Section 6: Development of Utility Theory II– Comparing Axiomatic Theories

In the last section, ordinal utility functions were discussed according to the marginalist view of modeling decision problems under conditions of certainty. This section compares the main elements of the theory of individual rational modeling under *risk and uncertainty*. These were developed in the mid-1900's: the theories of subjective probability (SP), expected utility (EU), and subjective expected utility (SEU). The primary sources of information for this section are Nau (2000) and French and Insua (2000) and were stated at the beginning of **Section 5**. Both provided detailed accounts of the developments of all three works and provide citations for the original publications and other perspectives on these developments.

Nau discusses these three theories in terms of four attributes or concepts – "axiomatic fever", "primitive preferences", "concepts of probability", and ""the independence condition for preferences → cardinal utility". Of these four concepts, the first two have been discussed thoroughly in the **Section 4**. The third topic of interest was addressed briefly in the closing sub-section of **Section 5**. However, this last concept has not been addressed this far. Before beginning a comparative look at the three theories of concern, a little insight into the importance of this last topic will be given first.

Section 6.1: The Independence Condition for Preferences → Cardinal Utility

Von Neumann and Morgenstern in their aims to establish a "new game-theoretic foundation for economics", decided it was necessary to first axiomatize a concept of *cardinal* measurable utility that could serve as a "single monetary commodity" whose expected value the players would seek to maximize through their strategy choices.

Recall from the previous section that decision making under conditions of certainty were modeled as ordinal utility functions. It was discussed that under such conditions, it is meaningless to say something like "the utility of y is exactly halfway between the utility of x and the utility of z." von Neumann and Morgenstern discovered that such a statement can be made if the decision maker has preferences not only among definite objects such as x, y, and z, but also among *probability distributions* over those objects.

Thus, it can be said that *by definition*, a 50-50 gamble between x and z has a utility that is exactly halfway between the utility of x and the utility of z, and if y is indifferent to such a gamble, then it too has a utility that is exactly halfway between the utility of x and the utility of z. There are several important things to note here (Nau, 2000) and (French and Insua, 2000).

Note 1: Additional "psychological" data is required to construct a cardinal utility function, namely data about preferences among probability distributions over rewards, not just data about preferences among the rewards themselves.

Note 2: Those preferences must satisfy an *independence condition*, namely that x is preferred to y if and only if an α chance of x and a $1-\alpha$ chance of z is preferred to an α chance of y and a $1-\alpha$ chance of z, regardless of the value of z—i.e., a preference x over y is independent of common consequences received in other events.[12]

Note 3: Once probabilities have been introduced into the definition and measurement of utilities, a cardinal meaning cannot be attached to the utility values *except* in a context that involves risky choices: if the utility of y is found to lie exactly halfway in between the utilities of x and z, then it cannot be conclude that the increase in the measure results from the exchange of x for y to be exactly the same as the increase that results from the exchange of y for z. All that can be said is that y is indifferent to a 50-50 gamble between x and z.

While von Neumann-Morgenstern utility function was being established other economists had discovered that a simple behavioral restriction on preferences-under-certainty implies that a consumer must have an additively-separable utility function, i.e., a utility function of the form $U(x_1, x_2, \ldots, x_3) = u_1(x_1) + u_2(x_2) + \ldots + u_n(x_n)$, which is unique up to an increasing linear transformation and is therefore cardinally measurable (Nau, 2000).

Section 6.2: Comparative Axioms of SP, EU, and SEU

Many authors contributed to the development of axiomatic theories of rational decision making under uncertainty; however this section concentrates on three main contributions: de Finetti's theory of subjective probability (SP), von Neumann and Morgenstern's theory of expected utility (EU), and Savage's theory of subjective expected utility (SEU). Savage's theory is essentially a merger of the other two and will be presented in conjunction with the works of Anscombe and Aumann. I have tried to be consistent with my notation; however various authors present various different aspects and thus, this is an amalgamation of work published by (Nau, 2000) and (French and Insua, 2000).

Table 6-1 (appended to the end of section 6) presents the axioms at a glance as a comparative means and was adapted from Nau (2000). This is followed by a discussion on some of the key elements of comparison for the three theories of concern.

In all three theories, there is a primitive ordering relation, which I will denote by $\succsim$. In SP theory, the objects of comparison are *events* A, B, etc., which are subsets of some grand set of possible *states of nature*. The ordering relation between events is one of *comparative likelihood*: $A \succsim B$ means that event A is considered "at least as likely" as event B. In EU theory, the objects of comparison are *probability distributions* f, g, etc., over some set of possible *consequences*, which could consist of amounts of money or

[12] The independence condition is implicit in von Neumann and Morgenstern's axiomatization.

other tangible rewards or prizes, or they could simply be elements of some abstract set. The ordering relation is one of *preference*: $f = g$ means that distribution f is "at least as preferred" as distribution g. SEU theory merges the main elements of these two frameworks: the objects of comparison are "acts," which are real or imaginary mappings from states of nature to consequences. In other words, an act is an assignment of some definite consequence to each state of nature that might obtain. The ordering relation is *preference*, as in EU theory.

Similarity of Axioms Across Theories

1. Assumption that the relation is *complete*: for any A and B, either $A \underset{\sim}{\succ} B$ or $B \underset{\sim}{\succ} A$ or both. Hence, it is always possible to determine at least a weak direction of ordering between two alternatives.

2. Assumption that the ordering relation is *transitive*: i.e., $A \underset{\sim}{\succ} B$ and $B \underset{\sim}{\succ} C$ implies $A \underset{\sim}{\succ} C$.

3. Assumption that the relation has a property called *independence* or *cancellation*, which means that in determining the order of two elements (such as $A \underset{\sim}{\succ} B$), there is some sense in which *common events or consequences* can be ignored on both sides.

A Closer Look at the Independence Condition Across Probability Theories

1. The **independence condition for subjective probability** states:

$$A \underset{\sim}{\succ} B \Leftrightarrow A \cup C \underset{\sim}{\succ} B \cup C$$

for any C that is disjoint from both A and B. Simply stated, if the same non-overlapping event is joined to two events whose likelihoods are being compared, the balance of likelihood is not "tipped". This condition also implies:

$$A \cup C \underset{\sim}{\succ} B \cup C \Leftrightarrow A \cup C^* \underset{\sim}{\succ} B \cup C^*$$

for any C, C^* that are disjoint from both A and B. Thus, a common disjoint event on both sides of a likelihood comparison can be replaced by another common disjoint event without tipping the balance.

2. The **independence condition for expected utility** is:

$$f \underset{\sim}{\succ} g \Leftrightarrow \alpha f + (1-\alpha)h \underset{\sim}{\succ} \alpha g + (1-\alpha)h.$$

The expression $\alpha f + (1\text{-}\alpha)h$ is an "objective mixture" of the distributions f and h—i.e., an α chance of receiving f and a $1\text{-}\alpha$ chance of receiving h, where the chances are objectively determined. Thus, if f and g are objectively mixed with the *same* other distribution h, in exactly the *same* proportions, the balance of preference is not tipped between them. The a decision-tree illustration of this condition is and its interpretation is explained in **Section 9**.

The condition also implies:

$$\alpha f + (1-\alpha)h \underset{\sim}{\succ} \alpha g + (1-\alpha)h \Leftrightarrow \alpha f + (1-\alpha)h^* \underset{\sim}{\succ} \alpha g + (1-\alpha)h^*,$$

which says that if f is preferred to g when they are both mixed with a common third distribution h, then the direction of preference isn't changed if they are mixed with a different common distribution h^* instead. In other words, the "common element" h can be replaced with any other common element.

3a. The **independence condition for subjective expected utility** (Savage's axiom P2, otherwise known as the "sure thing principle" as noted by Nau (2000)) states:

$$Af + (1-A)h \underset{\sim}{\succ} Ag + (1-A)h \Leftrightarrow Af + (1-A)h^* \underset{\sim}{\succ} Ag + (1-A)h^*,$$

for any non-null event A[13]. The expression $Af + (1-A)h$ is a "subjective mixture" of f and h: it means the consequence specified by act f in every state where event A is true is achieved, and the consequence specified by act h in all other states is achieved. Simply, this condition says that comparing two acts which happen to agree with each other in some states (namely the states in which A is not true), then different agreeing consequences in those states can be substituted without tipping the balance. This property is necessary to define the concept of a *conditional preference* between two acts: the relation $Af + (1-A)h \underset{\sim}{\succ} Ag + (1-A)h$ means that f is preferred to g *conditional on event A*.

3b. There is also **a second kind of independence condition** that also appears in SEU theory, namely an assumption about the *state-independence* of preferences for particular consequences. This assumption (Savage's P3) states that if x, y, and z are *constant acts* (acts that yield the same consequence in every state of nature) and A is any non-null event, then:

$$x \underset{\sim}{\succ} y \Leftrightarrow Ax + (1-A)y \underset{\sim}{\succ} Ay + (1-A)z.$$

[13] A non-null event is an event that has non-negligible probability in the sense that at least some preferences among acts are affected by the consequences received in that event.

Thus, if the consequence x is preferred to the consequence y "for sure," then it is also preferred to y conditional on any event, other things being equal. This assumption is needed to help separate utilities from probabilities—i.e., to ensure that relative utilities for consequences don't depend on the states in which they are received.

The important thing to note is that, in all these conditions, the common additive factors can be ignored or canceled *(e.g., C, h, z, etc.)* on opposite sides of the ordering relation. This property is necessary in order for the relation to be represented by a probability distribution that is *additive* across mutually exclusive events and/or a utility function that is a *linear in probability*. The additivity or linearity of the preference representation, in turn, means that the direction of preference or comparative likelihood between two alternatives depends only on the *difference* between them.

Section 6.3: Subjective Probability versus Expected Utility theory

In de Finetti 's 1937 paper, he pointed out that there are actually two equivalent ways to axiomatize the theory of subjective probability. One way is in terms of a binary relation of comparative likelihood (as illustrated in the table above). The other way is in terms of the acceptance of monetary *gambles*. The two theories are "dual" to each other—i.e., they are really the *same* theory but their working parts are merely labeled and interpreted in different ways (Nau, 2000).

Suppose that there is some set S of mutually exclusive, collectively exhaustive states of nature. Consider possible distributions of monetary wealth over those states. Thus, if f is a vector representing a wealth distribution, then $f(s)$ is the amount of money you get if state $s \in S$ occurs.

A0: (reflexivity) $f \succsim f$ **for all** f**.**

A1: (completeness) for any f **and** g**, either** $f \succsim g$ **or** $g \succsim f$ **or both.**

A2: (transitivity) if $f \succsim g$ **and** $g \succsim h$**, then** $f \succsim h$**.**

A3: (strict monotonicity) if $f(s) > g(s)$ **for all** s**, then** $f \succ g$ **(i.e., NOT** $g \succsim f$**).**

A4: (independence) $f \succsim g \Leftrightarrow \alpha f + (1-\alpha)h \succsim \alpha g + (1-\alpha)h$ **for all** h **and** $0 < \alpha < 1$**.**

In the presence of the ***completeness axiom***, the independence axiom also works in reverse, i.e., $f \succsim g \Leftrightarrow \alpha f + (1-\alpha)h \succsim \alpha g + (1-\alpha)h.$ for all $0 < \alpha < 1$. This is so because the completeness axiom requires the decision maker to have a definite direction of preference between f and g, and if it were not the same as the direction of preference between $\alpha f + (1\text{-}\alpha)h$ and $\alpha g + (1\text{-}\alpha)h$, the independence axiom would produce a contradiction.

This means the decision maker effectively has *linear utility for money*, because scaling both f and g up or down by the same positive factor and/or increasing the decision maker's initial wealth by a constant amount h does not lead to a reversal of preference.

There are many results and important conclusions that follow from this line of thought which are discussed in some depth in Nau (2000). Two of prominence are: The first of which is that the direction of preference between any two wealth distributions depends only on the difference between them. And the second of which states that the inequalities between different pairs of wealth distributions can be added.

Now consider the following set of axioms. Consider now the differences between wealth distributions as gambles – of which there are two types: acceptable and unacceptable.

B0: (reflexivity) $\mathbf{0 \in G}$
B1: (completeness) for any $\boldsymbol{x}$**, either** $\boldsymbol{x \in G}$ **or** $\boldsymbol{-x \in G}$ **or both**
B2: (linearity) if $\boldsymbol{x \in G}$**, then** $\boldsymbol{\alpha x \in G}$ **for any** $\boldsymbol{\alpha > 0}$
B3: (additivity) if $\boldsymbol{x \in G}$ **and** $\boldsymbol{y \in G}$**, then** $\boldsymbol{x + y \in G}$
B4: (coherence) $\boldsymbol{G}$ **contains no strictly negative vectors**

These are the assumptions used by de Finetti in his gamble-based method of axiomatizing subjective probability, and (by the arguments above) they are logically equivalent to A0–A4 (Nau, 2000).

SP THEOREM: *If preferences among wealth distributions over states satisfy A0-A1-A2-A3- A4 (or equivalently, if acceptable gambles satisfy B0-B1-B2-B3-B4), then there exists a unique subjective probability distribution* π *such that* $f = g$ *if and only if the expectation of the wealth distribution f is greater than or equal to the expectation of distribution g according to* π *(or, equivalently, x is an acceptable $-gamble if and only if its expectation is non-negative according to* π*)* .

This is **de Finetti's theorem on subjective probability**.

Consider reversing the role of money and probabilities. Let C be some set of consequences or prizes, and let f, g, etc., denote probability distributions over those prizes. Thus, $f(c)$ now represents the *probability* of receiving *prize c* rather than an amount of money. The axioms are as follows:

A0: (reflexivity) $f \succsim f$ **for all** f

A1: (completeness) for any f **and** g**, either** $f \succsim g$ **or** $g \succsim f$ **or both**

A2: (transitivity) if $f \succsim g$ **and** $g \succsim h$**, then** $f \succsim h$

A3: (non-triviality) there exist f^* **and** g^* **such that** $f^* \succ g^*$ **(i.e., NOT** $g^* \succsim f^*$**)**

A4: (independence) $f \succsim g \Leftrightarrow f + (1-\alpha)h \succsim g + (1-\alpha)h$ **where** $0 < \alpha < 1$.

A5: (continuity): for fixed f**, the set of** g **such that** $f \succsim g$ **is closed, and vice versa**

There is a small difference in Axiom 3 which is the strict monotonicity condition is weakened to a non-triviality condition.

The direction of preference between any two probability distributions depends only on the *difference* between those two distributions. Note that a difference between two probability distributions also a kind of "gamble" in the sense that it changes your *risk profile*—it just does so by shifting the probabilities attached to fixed prizes rather than by shifting the amounts of money attached to events with presumably-fixed subjective probabilities. Then the followings axioms are developed:

B0: (reflexivity) $0 \in G$.
B1: (completeness) for any x**, either** $x \in G$ **or** $-x \in G$ **or both.**
B2: (linearity) if $x \in G$**, then** $\alpha x \in G$ **for any** $\alpha > 0$.
B3: (additivity) if $x \in G$ **and** $y \in G$**, then** $x + y \in G$.
B4: (non-triviality) there is at least one non-zero p**-gamble *not* in** G.
B5: (closure): G **is closed.**

The continuity/closure (**A5**/**B5**) requirement is needed to rule out unbounded utilities as well as lexicographic preferences (Nau, 2000) and French and Insua (2000).

Let u denote a *utility function for prizes*, because $f = g$ if and only if $f'u = g'u$, and the quantities on the left and right of this expression are just the expectations of u under the distributions f and g, respectively. The function u is subject to arbitrary positive scaling, and a constant can also be added to it without loss of generality (wlog), because this will just add the same constant to the expected utility of every probability distribution.

EU THEOREM: The results are summarized as the expected utility theorem. If preferences among probability distributions over consequences satisfy A0-A1-A2-A3-A4 (or equivalently, acceptable p-gambles satisfy B0-B1-B2-B3-B4), there exists a utility function u, unique up to positive affine transformations, such that $f = g$ if and only if the expected utility of f is greater than or equal to the expected utility of g (or, equivalently, x

is an acceptable *p*-gamble if and only if it yields a non-negative change in expected utility according to *u*).

This is **von Neumann-Morgenstern's theorem on expected utility**, and it is isomorphic to de Finetti's theorem on subjective probability, merely with the roles of probabilities and payoffs reversed (French and Insua, 2000).

The axioms of subjective probability or expected utility are satisfied if and only if (a) your set of acceptable directions of change in (stochastic) wealth is a convex cone, (b) it excludes at least some directions—especially those that lead to a sure loss—and (c) it is always the same set of directions, no matter what your current distribution.

Section 6.4: Subjective Expected Utility (Anscombe-Aumann & Savage)

Savage combined the features of subjective probability and expected utility theory in order to model situations in which the decision maker may attach their own subjective probabilities to events and their own subjective utilities to consequences. Anscombe & Aumann followed in much the same way, but had a simpler approach by simply merging Savage's concept of an "act" (a mapping of states to consequences) with von Neumann and Morgenstern's concept (an objective probability distribution over consequences), (French and Insua (2000).

Anscombe and Aumann proposed that the objects of choice should be *mappings from states to objective probability distributions over consequences*. Such objects are known as "horse lotteries" in the literature. This is supposed to conjure up an image of a horse race in which each horse carries a different lottery ticket offering objective probabilities of receiving various prizes. By applying the von Neumann-Morgenstern axioms of expected utility to horse lotteries, together with one more assumption, you get probabilities and utilities together. Henceforth, let *f*, *g*, and *h* denote horse lotteries.
Think of them as vectors with doubly-subscripted elements, where $f(c,\theta)$ denotes the objective probability that the horse lottery *f* assigns to consequence *c* when state θ occurs (i.e., when horse *s* wins the race). Furthermore, let *x*, *y*, and *z*, denote horse lotteries that are "constant" in the sense that they yield the same objective probability distribution over consequences in every state of the world (i.e., no matter which horse wins).

Let the usual axioms A0-A5 apply to preferences among horse lotteries, together with one additional axiom, which is essentially the same as Savage's P3:

A6: (state-independence): $x \underset{\sim}{\succ} y \Leftrightarrow Ax + (1-A)z \underset{\sim}{\succ} Ay + (1-A)z$

for all constant *x*, *y*, *z* and every event *A*

Here, $Ax + (1-A)z$ denotes the horse lottery that agrees with *x* if *A* occurs and agrees with *z* otherwise, where *A* is an event (a subset of the states of the world).

Note1: From axioms A0-A5 and some work, the existence of an *expected-utility* vector v such that $f = g$ if and only if $f'v = g'v$. The elements of the vector v are subscripted the same as those of f and g, with $v(c, \theta)$ representing the expected utility of consequence c when it is received in state θ. Intuitively, $v(c, \theta)$ combines the *probability* of the state with the *utility* of the consequence.

Note 2: Axiom A6 then implies the further restriction that the expected utilities assigned to different consequences in state θ must be *proportional* to the expected utilities assigned to the same consequences in any other state, because it requires *conditional preferences* among objective probability distributions over consequences to be the same in all states of the world. It follows that v can be decomposed as $v(c, s) = p(s)u(c)$, where $p(s)$ is a (unique!) non-negative probability assigned to state s and $u(c)$ is a state-independent utility assigned to consequence c.

Note 3: By convention, the *utilities* attached to particular consequences are assumed to be identical—not merely proportional—in different states, in which case the ratio of $v(c, \theta)$ to $v(c, \theta')$ can be interpreted as the ratio of the subjective *probabilities* of states θ and θ'.

Note 4: Preferences among Savage acts are represented by a unique probability distribution p over states and a state-independent utility function u over consequences, with act f preferred to act g if and only if the expected utility of f is greater than or equal to the expected utility of g. This is summarized as the subjective expected utility theorem. However, before stating the theorem, one further remark is made which addresses a weakness in this point.

Remark: It should be noted that the uniqueness of the derived probability distribution depends on the unstated assumption that consequences are not only similarly ordered by preference in every state of the world (which is the content of axiom A6 above), but they also have the *very same numerical utilities* in every state.

SEU THEOREM: If preferences among horse lotteries satisfy A0-A1-A2-A3-A4-A5-A6, then there exists a unique probability distribution p and a state-independent utility function u that is unique up to positive affine transformations such that $f = g$ if and only if the expected utility of f is greater than or equal to the expected utility of g.

Section 6.5: Incomplete Preferences

The completeness axiom causes some distress. It forces the notion that a "gamble" must be accepted. So, although, it may be unacceptable according to some axioms – it is still acceptable according to the completeness axiom. The three theorems staed above with respect to each theory is restated generally with incomplete preferences.

SP THEOREM with incomplete preferences: If preferences among wealth distributions over states satisfy A0-A2-A3-A4-A5, then there exists a convex set P of subjective probability distributions such that $f = g$ if and only if the expectation of the

wealth distribution f is greater than or equal to the expectation of distribution g for every distribution π in P (or, equivalently, x is an acceptable gamble if and only if its expectation is non-negative for every distribution π in P) .

Similar results are obtained when you drop the completeness axiom from EU or SEU theory: you end up with convex sets of utility functions or expected-utility functions.

EU THEOREM with incomplete preferences: If preferences among probability distributions over consequences satisfy A0-A2-A3-A4-A5 (or equivalently, acceptable *p*-gambles satisfy B0-B2-B3-B4-B5), there exists a convex set U of utility functions, unique up to *positive affine transformations*,[14] such that $f \succ\!\!\!\sim g$ if and only if the expected utility of f is greater than or equal to the expected utility of g for every utility function u in U (or, equivalently, x is an acceptable *p*-gamble if and only if it yields a non-negative change in expected utility for every utility function u in U). See French and Insua (2000).

SEU THEOREM with incomplete preferences: If preferences among horse lotteries satisfy A0-A2-A3-A4-A5-A6, then there exists a convex set V of expected-utility functions, at least one of which can be decomposed as the product of a probability distribution p and a state-independent utility function u, such that $f \succ\!\!\!\sim g$ if and only if the expected utility of f is greater than or equal to the expected utility of g for every expected-utility function v in V.

[14] Note: Theorem 3 addresses positive affine transformations in Chapter 2 of French and Insua, (2000).

Table 6.1: Comparing Axioms: SP, EU & SEU

	de Finetti	**Von Neumann-Morgenstern**	**Savage**
Theory of…	Subjective Probability	Expected utility	Subjective expected utility
Features	States	Abstract set closed under probability mixtures	States, consequences
Objects of comparison	Events A,B,C (sets of states)	Probability distributions f,g,h	"Acts" f,g,h, (mappings from states to consequences); constant acts x,y,z
Primitive relation $\succ$ $\underset{\sim}{}$	Comparative Likelihood	Preference	Preference
Reflexivity axiom	$A \underset{\sim}{\succ} A$	$f \underset{\sim}{\succ} f$	$f \underset{\sim}{\succ} f$
Completeness axiom	For all A,B: if $A \underset{\sim}{\succ} B$ or $B \underset{\sim}{\succ} A$ or both	For all f,g: if $f \underset{\sim}{\succ} g$ or $g \underset{\sim}{\succ} f$ or both	P1a:For all f,g: if $f \underset{\sim}{\succ} g$ or $g \underset{\sim}{\succ} f$ or both
Transitivity axiom	For all A,B,C: if $A \underset{\sim}{\succ} B$ or $B \underset{\sim}{\succ} C$ then $A \underset{\sim}{\succ} C$	For all f,g,h: if $f \underset{\sim}{\succ} g$ or $g \underset{\sim}{\succ} h$ then $f \underset{\sim}{\succ} h$	P1b: For all f,g,h: if $f \underset{\sim}{\succ} g$ or $g \underset{\sim}{\succ} h$ then $f \underset{\sim}{\succ} h$
Independence axiom (a.k.a. substitution, cancellation, separability, sure-thing principle)	If $A \cap C = B \cap C = \varnothing$ then $A \underset{\sim}{\succ} B$ iff $A \cup C \underset{\sim}{\succ} B \cup C$	If $f \underset{\sim}{\succ} g$ then $\alpha f + (1-\alpha)h \underset{\sim}{\succ} \alpha g + (1-\alpha)h$ for all α strictly between 0 and 1. the converse is also true as a theorem, whence: $\alpha f + (1-\alpha)h \underset{\sim}{\succ} \alpha g + (1-\alpha)h$ iff $\alpha f + (1-\alpha)h^* \underset{\sim}{\succ} \alpha g + (1-\alpha)h^*$	P2: For all f,g,h,h^*, if event A is non-null, then $\mathrm{A}f + (1-\mathrm{A})h \underset{\sim}{\succ} \mathrm{A}g + (1-\mathrm{A})h$ iff $\mathrm{A}f + (1-\mathrm{A})h^* \underset{\sim}{\succ} \mathrm{A}g + (1-\mathrm{A})h^*$
State-independent utility axiom ("value can be purged of belief")			P3: For all constant acts x,y,z, if event A is non-null, then $x \underset{\sim}{\succ} y$ iff $\mathrm{A}x + (1-\mathrm{A})z \underset{\sim}{\succ} \mathrm{A}y + (1-\mathrm{A})z$

			$\mathrm{A}x + (1-\mathrm{A})z \underset{\sim}{\succ} \mathrm{A}y + (1-\mathrm{A})z$
Qualitative probability axiom ("belief can be discovered from preference")			P4: For all events A & B and constant acts x,y,z^*,y^* such that $x \succ y$ and $x^* \succ y^*$: $Ax + (1-A)y \succ Bx + (1-B)y$ iff $Ax^* + (1-A)y^* \succ Bx^* + (1-B)y^*$
Non-triviality axiom	$\Omega \succ A \succ \varnothing$ for any A that is "uncertain"	$\exists f$ and g such that $f \succ g$	P5: $\exists f$ and g such that $f \succ g$
Continuity axiom	There exists a fair coin	If $f \succ g \succ h$, then $\alpha f + (1-\alpha)h \succ \beta g + (1-\beta)h$ for some α and β strictly between 0 and 1.	P6: If $f \succ g$, there exists a finite partition of the set of states into events $\{\mathrm{A}_i\}$ such that $\mathrm{A}_i f + (1-\mathrm{A}_i)h \succ (1-\mathrm{A}_i)g + \mathrm{A}_i h$ for any event A_i in the partition and h.
Dominance axiom			P7: If $f(\theta) \succ g(\theta)$ one every state θ in event A, then $\mathrm{A}f + (1-\mathrm{A})h \underset{\sim}{\succ} \mathrm{A}g + (1-\mathrm{A})h$
Conditional probability axiom	If $A \subseteq C$ and $B \subseteq C$ then $A \mid C \underset{\sim}{\succ} B \mid C$ iff $A \underset{\sim}{\succ} B$		

Part III: SEQUENTIAL STATISTICAL DECISION THEORY - Introduction to Markov Decision Processes

Concepts of decision processes (as those presented in prior sections) have assumed single decision making problem. However, this is rarely the case. In most instances, decisions are "nested in a series of related decisions" (French and Insua, 2000). This iterates the concept that a decision is contingent upon the outcome of previous decisions made. The formulation of sequential problems is most naturally expressed as a recursive equation and its solution is credited to Bellman. Subsequent developments in sequential methods have instigated growth in sequential probability leading to a large body of theory on Markov decision processes. This sequential pattern of decision processes is a concept that has wide applications to all areas of the sciences, engineering, operations research, finance, etc. Consequently, these developments and their applicability to a wide range of problems have propagated the use of sequential methods in subsequent fields.

This final theoretical part of the report focuses on sequential decision processes. An examination of sequential sampling leads to the formulations of decision rules and stopping rules in statistical decision theory. These concepts establish the goal for pre-posterior and sequential analysis. Since many problems can be formulated in Markovian ways, the latter section focuses on Markov chains and Markov decision processes as an extension to the developments in sequential probability. Emphasis is placed on abstract and technical formulations of the concepts and methodology. This provides a framework for formulating sequential decision problems.

Section 7: Preposterior and Sequential Methods

The framework for decision problems as introduced in Part I assumed a fixed sample of observations for a single decision. Noted in the framework, decisions lead to consequences which may subsequently lead to further decision making. A natural extension of the framework to encompass sequential decision processes establishes new concepts for statistical decision theory. That is the option to experiment or make new observations before making a decision – a term which is coined pre-posterior analysis. With a goal to minimize cost, sampling is the principle issue in sequential statistical decision theory and establishes the concepts of optimal stopping rule and decision rules. This instigated the expansion of hypothesis testing to encompass sequential methods.

Section 7.1: A Framework for Sequential Decision Problems

The framework for decision problems thus far, has centered on decision making either with or without experimentation. However, the choice or the decision to experiment is also a very important aspect of statistical decision theory. This decision making process is frequently termed pre-posterior analysis. The goal is to minimize the overall cost which incorporates both the decision loss as well as the cost of conducting the experiment.

The first introduction to statistical decision problems introduced the concept of having observations upon which a decision could be based. Those statistical procedures assumed a fixed sample size, although it was not explicitly stated. Other strategies suggest sequential sampling and moreover, multi-stage procedures. Extending the framework for decision problems, denote the sample space for independent random variables $X_1, X_2, \ldots$ to be X_i and define $\mathbf{X}^j = (X_1, X_2, \ldots, X_j)$ where $\mathbf{X}^j$ is assumed to have density $f_j(\mathbf{x}^j \mid \theta)$. Now suppose that $X_1, X_2, \ldots$ is a sequential sample from the density $f(x \mid \theta)$; then

$$f_j(\mathbf{x}^j \mid \theta) = \prod_{i=1}^{j} f(x_i \mid \theta).$$

The cost of the experiment is attributed to two factors: n – the number of observations taken and s – the way in which the observations are taken (Berger, 1985). Thus, $l(\theta, a, n, s)$ denotes the overall loss when θ is the true state of nature and a is the action taken. Note that $l(\theta, a, n, s)$ can be considered to be a sum of the general loss function $l(\theta, a)$ and the sampling cost $c(n, s)$. Berger(1985) states that these functions are additive when the decision maker has a linear utility function which can be stated as

$$l(\theta, a, n) = l(\theta, a) + \sum_{i=1}^{n} c_i,$$

where $l(\theta, a, n)$ is the shortened form of the overall loss function for a given sampling approach and c_i represents the *i*th observation cost.

Recall that the loss function can be defined as $l(\theta, a) = -u(\theta, a)$. Thus, if the utility function is nonlinear and there is a monetary gain or loss, $g(\theta, a)$, then $l(\theta, a, n, s) = -u(g(\theta, a) - c(n, s))$.

The difficulty in sequential decision problems resides in all the possible ways in which the observations are taken. The two most common approaches are (a) the fixed sample method and (b) the sequential sampling approach. Although the latter approach is of primary interest, this section begins by describing the methodology of determining the optimal fixed sample size, leaving the latter approach for the remainder of this section.

Recall that pre-posterior analysis was coined as such as the choice is made before the data is obtained. Thus, the "non-statistical" decision problem – that is the no-observation problem first. Within the Bayesian context, let $\pi = P_\theta(\cdot)$ and $r_n(\pi)$ be the Bayes risk of the optimal policy with at most *n* stages with knowledge π. Then the situation in which the decision maker has no option to make an observation and must choose an action can be expressed as:

$$r_0(\pi) = \min_{a \in A} E_\theta[l(a, \theta)]$$

where the expectation over θ is taken with respect to π. Another way of expressing this is to let δ_n^π denote a Bayes decision rule for this problem and incorporating the loss in observing $\mathrm{X}^n = (X_1, ..., X_n)$. The equivalent expression is then

$$r_0(\pi) = E^\pi E_\theta^{\mathrm{X}^n}[l(\theta, \delta_n^\pi(\mathrm{X}^n), n)].$$

To determine the optimal fixed sample size in a decision problem requires resolving the *n* which minimizes $r_n(\pi)$. In simplicity, this can be resolved by differentiating $r_n(\pi)$ with respect to *n* and setting the equation equal to zero.

If it assumed that the overall loss function has the aforementioned additive structure, then δ_n^π is the Bayes decision rule for the loss function. Furthermore, if the fixed sample size loss is equal to the sequential loss, a sequential analysis of the problem is considered cheaper then a fixed sample size analysis (Berger, 1985). The advantages of sequential analysis are somewhat obvious and at the forefront lies the decision maker choice to gather the required amount of data needed to make a decision for a specified degree of precision. The remainder of this section of the report focuses on the foundations, concepts and methods of sequential analysis.

Section 7.2: Sequential Analysis – The Basics

The distinguishing feature of sequential analysis is the ability to take observations one at a time with the decision maker or experimenter having the option of stopping the experiment and making a decision. Keeping within the sequential decision problem framework presented earlier, let's assume that the cost is proportional to the number of observations; and moreover that the loss function is increasing in *n*. Thus, it can then be reiterated that if *n* observations are taken sequentially, at which point the decision maker chooses to take action $a \in A$, then the loss function when θ is the true state of nature can be expressed as $l(\theta, a, n)$.

A nonrandomized sequential decision procedure is denoted $d = (\varphi, \delta)$ and consists of two components: φ – a stopping rule; and δ – a (terminal) decision rule. A stopping rule is characterized by a family of functions of the form

$$\varphi = \{\varphi_0, \varphi_1(X_1), \varphi_2(X_2)...\}.$$

In the nonrandomized case,

$$\varphi_0 = \begin{cases} 0, \text{ iff the rule says to take the first observation} \\ 1, \text{ iff the rule says to take the no observations} \end{cases}$$

and

$$\varphi_n(X_1, X_2, ...) = \begin{cases} 0, \text{ iff the rule says to take another observation,} \\ \quad \text{given that } X_1, ..., X_n \text{ are at hand} \\ 1, \text{ iff the rule says to take no more observations,} \\ \quad \text{given that } X_1, ..., X_n \text{ are at hand.} \end{cases}$$

That is to say that φ_0 is the probability of making an immediate decision without sampling while φ_n is the probability (zero or one) of stopping sampling and making a decision after X_n is observed. Thus, given the functions φ_n one can determine when to stop; If the rule is given in other terms, one can determine the functions φ_n from the rule.

Alternatively, the stopping rule can be characterized by the following family of functions

$$\tau = \{\tau_0, \tau_1(X_1), \tau_2(X_2), ..., \}$$

where $\tau_0 = \varphi_0$, and

$$\tau_n(X_1,...,X_n) = (1-\varphi_0)[1-\varphi_1(X_1)]...[1-\varphi_{n-1}(X_1,...,X_{n-1})]\varphi_n(X_1,...,X_n)$$
$$= \begin{cases} 0, & \text{if the rule either stopped the sampling prior to the } n\text{th} \\ & \text{observation or says to go on after it} \\ 1, & \text{if the rule calls for stopping sampling for the first time at} \\ & X_n, \text{given the stream } X_1, X_2,... \end{cases}$$

Similar to the φ_n functions, given the τ_n functions one can determine when to stop, or given the stopping rule, one can determine the τ_n functions.

The (terminal) decision rule is simply defined as a collection of decision rules of the type used in problems with fixed sample size (Lindgren, 1971). Thus, a terminal decision rule consists of a series of decision functions $\delta_0, \delta_1(X_1), \delta_2(X_2),...$ where $\delta_i(X_i)$ is the action taken after sampling has stopped after observing X_i. Thus, δ_0 is the action taken when no data is available and a decision is taken immediately.

Since the overall problem at hand, is to determine the final sample size at which φ says to stop and make a decision, it is favorable to discuss this issue in terms of the stopping time *N*. Thus, given a sequence of observations $X_1, X_2,....,$ the function τ_n gives the conditional probability that precisely *n* observations are called for:

$$P(N = n \mid X_1, X_2,...) = \tau_n(X_1,...., X_n).$$

Further, the probability of stopping at the *n*th observation, is simply the expected value assuming a particular decision and stopping rule:

$$P_{\theta=\{\tau,\delta\}} = E\{\tau_n(X_1,...., X_n)\}, \text{ (Lindgren, 1971).}$$

More formally, the stopping time is the random function of $\boldsymbol{X}$ given (for a nonrandomized sequential procedure) by

$$N(\mathbf{X}) = \min\{n \geq 0 : \tau_n(\mathbf{X}^n) = 1\}, \text{ (Berger, 1985).}$$

Thus, $\{N = n\}$ is the set of observations for which the sequential procedure stops at time *n*; and hence there is nothing to prevent $N = \infty$, as shown in Berger (1985):

$$P(N < \infty) = P(N = 0) + \sum_{i=1}^{\infty} P(N = n)$$
$$= P(N = 0) + \sum_{i=1}^{\infty} \int_{N=n} dF_n(\mathbf{x}^n \mid \theta)$$

But, in applying sequential procedures it is favorable to have a rule for which $P(N<\infty)=1$.

Section 7.3: Bayesian Sequential Procedures

Recall in section 7.1 that a pre-posterior analysis includes both the decision loss and the sampling or experiment cost when a particular action is taken for a particular state of nature. If a sequence of observations is taken at a cost of $c(N)$, for *n* observations and a sequential procedure $d=(\varphi,\delta)$ is used, the loss incurred is

$$l(\theta,d)+c(N)$$

where *N* is the random variable whose value is the number of observations actually used in reaching a decision. And the expected loss, or risk function, is subsequently expressed

$$r(\theta,d)=E[l(\theta,d)+c(N)].$$

The Bayes risk of a sequential procedure $d=(\varphi,\delta)$ is defined to be

$$r(\pi,d)=E^{\pi}[r(\theta,d)].$$

Bayesian analysis with fixed sample size problems is straightforward; Bayesian sequential analysis is quite cumbersome. So the objective of Bayesian sequential analysis is to find the sequential procedure which minimizes $r(\pi,d)$. This is defined as

$$r(\pi)=\inf_{\mathbf{d}} r(\pi,d).$$

Note that the idea here is that every stage of the procedure, one should compare the (posterior) Bayes risk of making an immediate decision with the "expected" (posterior) Bayes risk (Berger, 1985). Now suppose that there are potentially *n* stages, corresponding to observations $X_1,X_2,....,X_n$. If the Bayes decision rule δ^{π} calls for taking all *n* observations in a particular application, the terminal decision is made according to the Bayes criterion. That is, the posterior distribution is applied to the given loss function to obtain averages over which the possible actions are ordered and the optimum chosen (Lindgren, 1971). Assuming that $l(\theta,a,n)\geq -K$, it then follows that

$$
\begin{aligned}
r(\pi, \mathbf{d}) &= E^{\pi}[r(\theta, \mathbf{d})] \\
&= P(N=0)E^{\pi}[l(\theta, \delta_0, 0)] + \sum_{n=1}^{\infty} \int_{\Theta} \int_{\{N=n\}} l(\theta, \delta_n(\mathbf{x}^n), n) dF_n(\mathbf{x}^n \mid \theta) dF^{\pi}(\theta) \\
&= P(N=0)E^{\pi}[l(\theta, \delta_0, 0)] + \sum_{n=1}^{\infty} \int_{\{N=n\}} \int_{\Theta} l(\theta, \delta_n(\mathbf{x}^n), n) dF^{\pi^n}(\theta) dF^{m_n}(\mathbf{x}^n) \\
&= P(N=0)r_0(\pi, \delta_0, 0) + \sum_{n=1}^{\infty} \int_{\{N=n\}} r_0(\pi^n, \delta_n(\mathbf{x}^n), n) dF^{m_n}(\mathbf{x}^n).
\end{aligned}
$$

(Note that the subscript *m* refers to the marginal density.) Thus, $r(\pi, d)$ is minimized if δ_0 and δ_n are chosen to minimize the posterior expected loss. This is equivalent to the fixed sample size issue (Berger, 1985). Subsequently, if the stopping rule τ suggests at least $n - k$ observations, then the problem of whether to stop (using the Bayes terminal decision) or obtain more data is solved by comparing two conditional expected losses (Lindgren, 1971). One of these conditional expressions is based on those $n - k$ observations, while the other is to take more observations. This is summarized by the following theorem found in Lindgren (1971).

> *If τ is a given stopping rule and π a given prior, then the Bayes risk $r(\pi, d)$ is minimized by that τ by the decision rule* **d** *where d_i is the Bayes decision rule based on the first i observations considered as a sample of fixed sample size i.*

Thus, half the problem of determining the Bayes sequential problem is deciphered as regardless of the stopping rule used, the optimal action, once one has stopped, is simply the Bayes action for the given observations (Berger, 1985). Furthermore, if the posterior Bayes risk is at time *n* is a constant, independent of the actual observation then the optimal stopping rule corresponds to the choice, before an experiment, of an optimal fixed sample size *n*. However, this situation is rare. Yet when they are independent, the problem of determining the optimal stopping time is still a question and is the focus of the next section.

Section 7.4: Optimal Stopping Time

The difficulty of determining a Bayes sequential stopping rule resides in the fact that there is an infinite "stream of identically distributed observations" upon which to base a sequential test. However, this is made easier if the Bayes stopping rule can be determined stage by stage. This simplification arises because at every stage the future looks exactly the same (Berger, 1985). The conditional expected loss (given the observations) is the function corresponding to the posterior probabilities at each stage. Hence, those posterior probabilities depend on the observations (at hand) which enter the decision process through the posterior distribution (Lindgren, 1971).

There are many situations which may arise when trying to determine the optimal stopping rule. Before considering these situations it is important to develop the framework for such a test. So, for a given procedure $d = (\tau, \delta)$ and given losses l_a and l_b corresponding to the erroneous acceptance and rejection of the null hypothesis, respectively, the expected loss is

$$r(\theta; \tau, \delta) = \begin{cases} l_a \alpha + cE(N \mid H_0) \equiv R_0, \text{ if } \theta = H_0 \\ l_b \beta + cE(N \mid H_1) \equiv R_1, \text{ if } \theta = H_1, \end{cases}$$

where α and β are the usual error sizes, c is the cost per observation and N is the number of observations used in making a decision. Then assuming a prior distribution π assigned to H_0, the Bayes risk is

$$r(\pi, d) = \pi R_0 + (1 - \pi) R_1,$$

and the Bayes procedure which minimizes $r(\pi, d)$. Lindgren (1971) demonstrates that if no observations are permitted by the stopping rule τ, the minimum Bayes risk is just its value when the no-data Bayes procedure is used (as mentioned previously):

$$r_0(\pi) = \min r(\pi, d) = \begin{cases} l_a \pi, \text{ if } \pi \le \dfrac{l_a}{l_a + l_b}, \\ l_a (1 - \pi), \text{ if } \pi > \dfrac{l_a}{l_a + l_b}. \end{cases}$$

This is shown graphically in figure 7-1 below. Recall that this was shown in section 2 as it is essentially the no data problem or "non-statistical" decision problem.

Figure 7-1:

Now consider a set of rules, C_1, which suggest obtaining at least one observation. The minimum Bayes risk for this set of rules is denoted:

$$r_1(\pi) = \min_{C_1} [\pi R_0 + (1 - \pi) R_1].$$

Figure 7-2:

The figure above represents a general graphical view of this function of π and has the following properties (Lindgren, 1971):

(a) $r_1(\pi) \ge c$,
(b) $r_1(0) = r_1(1) = c$,
(c) the graph of $r_1(\pi)$ is concave down.

The first property suggests that at least one observation should be taken at cost *c*. The second property suggests that if the prior is either 0 or 1, then no amount of sampling will alter the distribution, so the cost is just *c* plus the 0 of the no-data curve. The last property simply implies that the curve $r_1(\pi)$ is made up of a "family of line segments joining pairs of points $(0, R_0)$ and $(1, R_1)$. Berger (1985) and Lindgren (1971) both establish more formally the concavity. With these properties, it is now possible to graphically compare the curves. For instance, if the graph of $r_1(\pi)$ lies above $r_0(\pi)$ then this suggests not to take any observations. This visual analysis can be carried out for the given hypotheses.

Lindgren (1971) provides three steps for determining the appropriate Bayes sequential procedure. This can be summarized as follows:

1. Given losses l_a and l_b, and cost *c* per observation, the values of *C* and *D* as shown in ***figure 7-2*** can be determined from the graphs of $r_0(\pi)$ and $r_1(\pi)$.
2. If $\pi \le C$, reject H_0, and if $\pi \ge D$, do not reject H_0; otherwise sample and determine *A* and *B* (from *C* and *D* and π) where

$$A = \frac{1-\pi}{\pi}\frac{C}{1-C}, \text{ and } B = \frac{1-\pi}{\pi}\frac{D}{1-D}.$$

   This is established since the condition for continuing sampling, can be expressed in terms of the likelihood ratio for the first *n* observations such that $A < \Lambda_n < B$.
3. After each observation X_n determine the corresponding likelihood ratio Λ_n for $X_1, ..., X_n$; if $\Lambda_n \le A$, reject H_0, and if $\Lambda_n \ge B$, do not reject H_0. If $A < \Lambda_n < B$, take any other observation.

The draw back to this procedure is that $r_1(\pi)$ is rarely known; however, given the properties of $r_1(\pi)$, it is possible to determine the form of the Bayes sequential procedure.

Section 7.5: Sequential Probability Ratio Test

The sequential probability ratio test is provides one of the first applications of sequential analysis within statistical analysis (French and Insua, 2000). It tests a simple hypothesis H_0 against a simple alternative H_1. As constructed in the previous sub-section, the following constants are defined where α and β are the respective error sizes:

$$A = \frac{\alpha}{1-\beta}, \quad B = \frac{1-\alpha}{\beta}$$

These constants form the limits for the likelihood ratio Λ_n computed after each observation is taken. Then if $\Lambda_n \leq A$, the sampling stops and the null hypothesis is rejected. If $\Lambda_n \geq B$, the sampling stops and the hypothesis is not rejected. And subsequently, if $A < \Lambda_n < B$, then another observation is taken.

Notice that this is essentially the Bayes sequential procedure described earlier. The essential difference is that there is no formal loss structure in this instance; however given any sequential probability likelihood ratio test essentially means that there exist losses and a sampling cost per observation such that the Bayes sequential test for some prior distribution is the given sequential probability ratio test. Thus, the analysis both statistically and graphically proceeds in the same manner.

The decision theoretic approach of balancing the competing factors of expected losses and sampling costs resides in both the Bayesian and non-Bayesian approaches; and thus makes it difficult to determine an optimal policy or rule. However, French and Insua (2000) state that there is an intuitive structure for the optimal policy. That it is to say, that if the likelihood ratio of the observations is not sufficiently small or large, then keep sampling – otherwise, stop and take action. Even the constant values of *A* and *B* remain undefined, knowing the general structure of the optimal policy allows for finding possible good values. Berger (1985), French and Insua (2000), and Lindgren (1971) provide further details on the development, theory and other methodological issues (constraints) with regards to the sequential probability ratio test.

Section 8: Markov Decision Processes

The previous section provided a framework for sequential decision problems. Such problems can be formulated in Markovian ways which has led to large body of theory on Markov decision processes. Decision processes which make a sequence of probabilistic transitions and in which the transition probabilities depend only on the current state of the system are said to undergo a Markov process. Subsequently, if there are alternative actions amongst which a choice can be made then this is referred to as a Markov decision processes. Since such processes are comprised of Markov chains they are also attributed with the Markovian properties which provide favorable analytic abilities. This section begins with introducing the underlying concepts of Markov chains and their properties as well as other attributes of both finite and infinite chains. This section concludes with the framework and computational aspects of Markov decision processes for determining optimal policies.

Section 8.1: Sequential Decision Problems Formulated in Markovian Ways

The previous section introduced a framework for sequential decision making under uncertainty. Intervals of time separate the stages at which decisions must be made, and the effect of a decision at any stage is to influence the transition from current to succeeding state. If the transition from state to state is a probabilistic sequence and the transition probability from current to succeeding state is only dependent on the current state, then the decision process is said to have the Markovian property. This section begins by providing the basic notation and framework for sequential decision making processes which can be formulated in Markovian ways.

A random process when observed over time will be in different states at different times; and this change in state is referred to as state transition. Suppose a process has N states and its behavior over time is specified by the state the "system" is in at each stage in time. In this report, the stage structure for the sequential problems considered (whether they are time based or not) will be discrete. More formally, a typical random process X is a family $\{X_t : t \in T\}$ of random variables indexed by some set T. Further, if we define $T = \{0,1,2,...\}$, the system is a discrete-time process (Grimmett and Stirzaker, 1982). In sequential decision problems, we may define $\{X_0, X_1, X_2, ...\}$ to be a sequence of random variables which take values in some countable set S, the state space. Thus each X_n is a discrete random variable taking on one of N possible values, where $N = |S|$. It is evident that this system is not deterministic, and hence can be neither prophesied nor achieved. It is possible, however, to deduce some statistical description about the behavior of the process (Buchanan, 1982).

Definition: The process X is a Markov chain if it satisfies the Markov property

$$P(X_n = s \mid X_0, X_1, ..., X_{n-1}) = P(X_n = s \mid X_{n-1})$$

for all $n \geq 1$ and $s \in S$. (Grimmett and Stirzaker, 1982)

For simplicity (of notation), let $p_{ij}(n)$ denote the transition probability from *i* to *j* if the transition takes place at the *n*th step. The Markovian assumption that only the current stage (*n*) and the state (*i*) are relevant to the determination of the future behavior of the process is described as being memoryless. This assumption is attributable to the attractive use of Markov chains in sequential decision problems.

Much of theory of Markov chains is simplified by the condition that *S* is finite (Grimmett and Stirzaker, 1982). A finite Markov chain is said to be time-homogenous if for every pair of states *i* and *j*, $p_{ij}(n) = p_{ij}$ for all *n*. Such a chain is also described as having stationary transition probabilities. The transition matrix for a Markov chain is the $N \times N$ matrix $P = [p_{ij}]$, where *N* is the number of states. The transition probability for a Markov chain is a stochastic matrix.

Theorem: *P* is a stochastic matrix, which is to say that
(a) *P* has non-negative entries, or $p_{ij} \geq 0$
(b) *P* has row sums equal to one, $\sum_j p_{ij} = 1$.

This summarizes the basic elements of Markov chains. The remainder of this section presents the favorable attributes of Markov chains which demonstrate the potential applicability of Markov chains, especially in infinite stage Markov decision processes.

Section 8.2: Multi-step Transition and State Probabilities

The previous subsection focused on finite stage Markov decision problems. This part now shifts to work with infinite stage Markovian decision problems. The transition probability matrix *P* describes the likelihood of transitions from state to state; and since our interest lies in the evolution of *X*, it becomes important to determine the multi-step transition probability values.

Define $\varphi_{ij}(n) = p(X_n = s \mid X_0)$ for $1 \leq i, j \leq N$ and $n = 0,1,2...$ as the n-step transition probabilities for the Markov chain (defined by *P*). If the system starts in state *i* at time 0 and is in state *j* at time (*n*+1), the result is achieved by transition in *n* steps from state *i* to some state *k* (for which the probability is $\varphi_{ik}(n)$) followed by transition from *k* to *j* (for which the probability is p_{kj}). Thus, given that the state at the *n*th step is *k*, the probability of transition to *j* in (*n* + 1) steps is $\varphi_{ik}(n)p_{kj}$. These *N* stages for the transitions from *i* to *j* given the *n*th state are mutually exclusive and exhaustive, hence

$$\varphi_{ij}(n+1) = \sum_{k=1}^{N} \varphi_{ik}(n)p_{kj}$$, (Buchanan, 1982).

Buchanan (1982) provides the following justification.

$$\begin{aligned}
\varphi_{ij}(n+1) &= P\{X_{n+1} = s \mid X_0\} = \sum_{k=1}^{N} P\{X_{n+1} = s, X_n \mid X_0\} \\
&= \sum_{k=1}^{N} P\{X_n = s \mid X_0\} P\{X_{n+1} = s \mid X_n, X_0\} \\
&\qquad \text{from the defintion of conditional probability} \\
&= \sum_{k=1}^{N} P\{X_n = s \mid X_0\} P\{X_{n+1} = s \mid X_n\} \\
&\qquad \text{by the Markovian assumption} \\
&= \sum_{k} \varphi_{ik}(n) p_{kj}.
\end{aligned}$$

Since multi-step transition probabilities are probabilities, they must also adhere to the regular conditions of probability. That is they must satisfy the relationship that $0 \le \varphi_{ij} \le 1$ and $\sum_{j=1}^{N} \varphi_{ij}(n) = 1$. The next step is to formulate this into a matrix form which in turn develops the Chapman-Kolmogorov equations (Grimmett and Stirzaker, 1982).

The equation for the multi-step transition probabilities can be represented in (stochastic) matrix form (quite simply). Define $\Phi(n) = [\varphi_{ij}(n)]$ as the *n*-step transition probability matrix for all *n*. From the justification above, we can rewrite the final statement as

$$\begin{aligned}
&\Phi(n+1) = \Phi(n)P, \text{ for } n = 0,1,..., \\
&\Phi(0) = I,
\end{aligned}$$

where *I* is the $N \times N$ identity matrix. So,

$$\begin{aligned}
\Phi(1) &= \Phi(0)P = P, \\
\Phi(2) &= \Phi(1)P = P^2, \\
\Phi(3) &= \Phi(2)P = P^3,
\end{aligned}$$

and thus more generally $\Phi(n) = \Phi(n-1)P = P^{n-1}P = P^n$. This can then be deduced to the special case of the Chapman-Kolmogorov equations as stated in Grimmett and Stirzaker (1982):

$$\varphi(m+n) = \sum_{k} \varphi_{ik}(m) p_{kj}(n).$$

Hence, $P^{m+n} = P^m P^n$ and so P^n is the *n*th power.

Now define $\Phi = \lim_{n\to\infty} P^n$ (if it exists) and call Φ the limiting multi-step transition probability matrix or summarized as the limiting distribution for the chain. Note as the number of transitions increases, the influence of the original state diminishes. This is expected since the Markovian property assumes no memory to the process. It is important to keep clear that the probabilities of transitions from state to state are fully described by *P*, while Φ summarizes the long-term behavior of the process. Thus, the above theorem relates the long-term development to short-term development, and informs us how X_n depends on X_0.

Section 8.3: Some Classes of Markov Chains

This next sub-section considers the way in which the states of a Markov chain are related to each other. This will lead to a classification of the states. First we begin with some basic terminology which essentially focuses on the term communication (which exists between states).

> ***Definition:*** It is said that *i* communicates with *j*, written $i \to j$, if the chain may ever visit *j* with positive probability, starting from *i*. That is, $i \to j$ if $\varphi_{ij}(m) > 0$ for some $m \geq 0$. It is also said that *i* and *j* intercommunicate if $i \to j$ and $j \to i$, in which case we write $i \leftrightarrow j$.

Thus, two states intercommunicate if transition is possible in either direction in some finite number of steps. The number need not be the same for the directions of communication (Buchanan, 1982). Grimmett and Stirzaker (1982) proceed by stating the following theorem.

> ***Theorem:*** If $i \leftrightarrow j$ then
> (a) *i* and *j* have the same period
> (b) *i* is transient if and only if *j* is transient
> (c) *i* is null persistent if and only if *j* is null persistent

The first property, period is defined as the greatest common divisor of the epoch at which return is possible (Grimmett and Stirzaker, 1982). Further, the state is said to be periodic if the divisor is > 1 and aperiodic if the divisor is equal to one. The second is probably the most important property of the relationship of communication. Buchanan (1982) demonstrates this property by claiming that since $i \leftrightarrow j$ and $j \leftrightarrow k$, then there exists numbers n_1 and n_2 such that $\varphi_{ij}(n_1)$ and $\varphi_{jk}(n_2)$ are both positive. From the Chapman-Kolmogorov equation, we get

$$\varphi(n_1 + n_2) = \sum_{l=1}^{N} \varphi_{il}(n_1)\varphi_{lk}(n_2) \geq \varphi_{ij}(n_1)\varphi_{jk}(n_2) > 0.$$

This establishes the communication from $k \leftrightarrow i$. The last property which draws upon the term persistent simply claims that state *i* is persistent if the probability of eventual

return to *i*, having started from *i*, is 1. Subsequently, if this probability is strictly less than 1, *i* is transient. In any Markov chain, sets of states (or at least one set of states) where all states which are members of the same set can be found which communicate with one another (Buchanan, 1982).

There are many classifications of states besides the ones already mentioned. We now define an ergodic set of states which is said to be a set of in which every state communicates with every other state of the set. Grimmett and Stirzaker (1982) define a state to be ergodic if it is persistent, non-null, and aperiodic. Thus, an ergodic set of states is a set of states in which every state can be reached from every other state of the set, and once the set is entered there can be no transition out of the set. Of course, an ergodic state is an element of an ergodic set. Buchanan (1982) suggests considering ergodic states to be "collective" trapping states, whereby every Markov chain contains at least one ergodic state. It should also be noted that there may be states which do not belong to an ergodic state; these are referred to as transient states.

We now provide another definition or classification of states which leads to a classification of chains.

Definition: A set of *C* of states is called
(a) closed if $\varphi_{ij} = 0$ for all $i \in C, j \notin C$
(b) irreducible if $i \leftrightarrow j$ for all $i, j \in C$

Once a chain has taken a value in a closed set *C* of states, it never leaves. States of this type are absorbing states. The equivalence class $i \leftrightarrow j$ simply iterates that an irreducible set *C* is aperiodic if all states in the set are aperiodic. (Grimmett and Stirzaker, 1982)

The final remark to be made here is with regard to defining a regular Markov chain. That is a Markov chain whose states form a single ergodic set is an ergodic chain. If such an ergodic set has all entries of the multi-step transition probability matrix $\Phi(n) = P^n$ greater than zero, then it is said to be regular. Since $\Phi(n+1) = P\Phi(n)$, then P^{n+1} will have no zero entries if P^n does not have any (Buchanan, 1982). Thus, reaching a sufficient number of transitions at which all state to state transitions are possible, any additional transitions will not lead to any such transitions having zero probability again. And, such an ergodic chain which is regular in subsequently termed a regular Markov chain.

It is a property of regular Markov chains that the limiting multi-step transition probability matrix $\Phi = [\varphi_{ij}]$ will have identical rows. Denote the state probability as $\pi_i(n)$. Buchanan (1982) shows that the state probability vector $\pi(n) = \pi(0)\Phi(n) = \pi(0)P^n$ at time *n*. From this, we get $\pi = \pi(0)\Phi$ by taking the limit of both sides as *n* tends to infinity. Buchanan (1982) uses this to show that for a regular Markov chain that the limiting state probability vector π is independent of the initial state probability vector $\pi(0)$ and is the same as the identical rows of the limiting multi-step transition probability matrix Φ. This concludes the introduction to Markov chains and their properties.

Section 8.4: Markov Decision Processes

The section now extends the properties and method of Markov chains to develop the body of knowledge known as Markov decision processes. In a Markov decision process there is a system with states labeled $i = 1,...,N$. In state *i* there are a number of actions available. If action *k* is chosen a return $r(k,i)$ is generated and the system goes to state *j* with probability $p(i,j,k)$. A set of actions, one for each state, constitutes a policy. Under a given policy δ the systems is a Markov process with returns and has steady state gain $g(\delta)$. A policy which has the highest gain (or the lowest in a minimization problem) is an optimal policy. A policy whose gain differs from the optimal by not more than a known amount is called a tolerance optimal policy. The formulation and solution of Markov decision problems will be illustrated by extending the decision structure.

Suppose a system has a finite number of states, $\theta_i \in \Theta = \{\theta_1, \theta_2, ..., \theta_m\}$. The system or process may move between states at each infinite series of equal stages (or time steps) $t = 1,2,...$. At each time step *t*, the decision maker is to select an action $a^t = a_k \in A$ where *A* is finite. Depending on the action a_k chosen by the decision maker at time *t* and its current state θ^t, the system evolves to its next state θ^{t+1}. Then the probabilities

$$p_{ij}^k = p_{\theta^{t+1}}(\theta^{t+1} = \theta_j \mid \theta^t = \theta_i, a_k)$$

describe the evolution of the system. Thus, it is assumed that the probabilities hold the Markovian property and subsequently, do not depend on either the earlier states of the process nor any of the earlier actions. Recall that in a decision framework, every action taken with the current state results in a consequence. Similarly, we define $c^t(a_k, \theta_i)$ which is the consequence a decision maker receives at each stage. Further, this can be extended to show the overall "timestream" of consequences (French and Insua, 2000).

Now that Markov decision processes have been placed within a decision theoretic framework, we now extend this to include the concepts of utility theory. That is, we now place things within the context of rewards rather than losses. (Note that reward includes the cost of taking the action.) Assume that the reward received at stage *t* is $u^t(c^t(a_k, \theta_i)) = r_t(k,i)$. It follows that all $r_t(k,i)$ are bounded. Since the aim is to find the optimal policy with minimal long-term average costs, it is necessary to consider the increase in the value of state *i* from stage to stage. Hastings and Mello (1978) provide the formulation for the bounds and convergence test.

There are two common ways of evaluating the overall timestream of rewards in Markov decision processes. The first includes a discount factor ρ and a discounted additive utility function such that

$$u(c^1(a^1, \theta^1), c^2(a^2, \theta^2), ...) = \sum_{t=1}^{\infty} \rho^{t-1} r_t(k,i).$$

The other common approach is to choose a series of actions to maximize her average reward:

$$u(c^1(a^1,\theta^1),c^2(a^2,\theta^2),...) = \lim_{T\to\infty} \frac{\sum_{t=1}^{T} r_t(k,i)}{T}.$$

In the first approach presented, the decision maker seeks a policy for choosing actions which maximize the total expected discounted reward. The latter approach gives the long-run average monetary return or the expected monetary value of the policy.

Finite Markovian decision processes decide upon a policy to be applied at a (previously known) finite number of decision epochs with the objective of optimizing some measure of performance. However, in certain instances, the finite stage aspect may be too limiting a prospect which then "unbounds" the criterion presented above. Having an infinite Markovian decision process has the problem of approximation itself, which due to the attributes of Markov chains, the infinite approximation is a feasible proposition. For a finite stage problem, the criterion for discriminating between policies was the maximization of the expected reward associated with a policy. In the infinite stage problem, the cumulative expected reward from using a given policy will not be finite and hence, will increase without bound. Discounting the rewards by some factor as stated above, obviates this problem and consequently gives an expected discounted reward for each policy which is finite. There are other propositions for handling infinite Markov decision process. Some of these alternatives are presented in Hastings and Mello (1978) and Buchanan (1982). French and Insua (2000) also provide a discussion on the merits of some of alternative functions of utility to solve Markov decision processes.

Now a decision policy $(\delta_1,\delta_2,...)$ for a Markovian decision process is a sequence of decision rules which prescribe for each stage *t*, the action the decision maker should take if the current state of the process is θ^t (French and Insua, 2000). If the decision rule selected does not depend on the stage *t*, then the policies are said to be stationary. A stationary policy applied to a Markov decision process determines the transition probabilities from state to state and thus, induces a Markov chain. If it is assumed that this Markov chain is single and ergodic then the assumption can be made that for each policy there exists a state which can be reached from any other state under the policy. It can then be further delineated to use the attributes of Markov chains to solve such decision processes. (Computational issues are expounded in ***section 12***.)

There are several algorithms (or classes of algorithms) which can utilized for solving Markov decision processes. Linear programming is one approach used. However, value iteration and policy iteration are the most common. The first of these, value iteration, considers that if the decision maker takes action a_k and receives $r_t(k,i)$ immediately and then passes to state θ_j with probability p_{ij}^k, and continues optimally receiving the expected total discounted reward, the sequence will eventually converge for any starting point. There are many variants of this algorithmic procedure. French and Insua (2000)

illustrate a couple of these and Mello and Hastings (1978) present another variant. The latter algorithmic procedure, policy iteration has a similar objective. It begins with an arbitrary policy, instead of an arbitrary function as defined for the value iteration algorithm. This procedure begins with an arbitrary (stationary) policy, evaluates the function which defines the total expected discounted reward and seeks for an improved policy and until convergence is reached. This has described the possible approaches for solving infinite Markov decision processes. Finite Markov decision processes are simpler and can be solved by backward dynamic programming or backward induction which is presented in ***section 13***.

Part IV: MODEL BUILDING FOR DECISION ANLYSIS: Topics in Probabilistic Graphical Models

Methodologies for decision processes (as those presented in prior sections) are based upon probability and utility or different ways of handling uncertainty. Subsequently, certain criteria should be satisfied as a provision for guiding decision-based problems. For instance, French and Insua, list axiomatic basis, lack of counter examples, feasibility, robustness, "transparency to users", and "comparability with a wider philosophy" as criteria which should be satisfactorily determined. The first two provide the theoretical merit; the second two, feasibility and robustness, relate to the practicality of performing the analysis while the latter two entail the implementation of the analysis results. These latter parts of the report focus on the practicality and implications of decision processes and attempt to bridge the gap between the conceptual and practical components of decision theory. However, none of this is applicable without a decision model upon which the criteria is built and subsequently assessed.

This part of the report focuses on building models to illustrate decision problems under various assumptions. Emphasis is placed on probabilistic graphical models. These probability models are extensions of the procedures and methodology presented in earlier sections of this report. Concepts of problem structuring, parameters and attributes lay the foundation for model building and are introduced in the description of each inception of the model building process. Each section presented in Part IV describes a modeling approach for a particular formulation and framework for a given decision problem.

Section 9: Graphical Representation of Decision Problems

Section 1 introduced the framework for decision problems. The components of a decision problem provide the construction for a graphical representation. An introduction to graphical models for describing decision processes is illustrated through the conception of decision trees and influence diagrams. A comparative assessment of these two formulations demonstrates the strengths and weakness of the approaches in both structural and practical settings. In this section, a formulation of a general decision problem is revisited in a concise manner in order to introduce the assumptions and criteria required for model building.

Section 9.1: Modeling Decision Problems

Sections 1 through **3** demonstrated the key concepts and ideas pertaining to decision modeling. In essence, decision problems consist of actions, states of nature and consequences. Without forgoing the complexities of modeling such problems, a decision model can be concisely formulated as

$$a \oplus \theta \to c\,.$$

That is, the interaction between an action and a state of nature leads to a consequence.

Recall that the choice of action or decision is under the control of the decision maker; but the state which pertains is beyond his/her control. When both the action space A and the state space are finite, the model can be presented in a decision table as shown below.

Actions	**States of Nature**			
	θ_1	θ_2	$\cdots$	θ_n
a_1	c_{11}	c_{12}	$\cdots$	c_{1n}
a_2	c_{21}	c_{22}	$\cdots$	c_{2n}
.	.	.	$\cdots$	.
.	.	.	$\cdots$	.
.	.	.	$\cdots$	.
a_m	c_{m1}	c_{m2}	$\cdots$	c_{mn}

The simplest discussions of decision theory assume that the decision maker chooses an action (row) in the table, whereupon nature will 'choose' a state (column), leading to the consequence. This tabular representation of decision problems has very close conceptual and historical connections with game theory (which is not explored in this report).

With respect to the three elements – actions, states and consequences, the terminology may vary; however, the distinction between these elements is common. This separation is found throughout statistical decision theory, and leads to a model of decision making in which there is some representation of uncertainty relating to the unknown state, some

representation of preference between the possible consequences, and a mechanism for bringing these two representations together to guide choice. This statement merely emphasizes the adoption of the model $a \oplus \theta \to c$.

Conceptually, the "consequences" considered thus far, need not have been considered to be numerical. That it is to say, that they may be thought of as descriptions of the outcome of actions chosen under various possible "states of nature". Similarly, neither the actions nor the states of nature need to be thought of quantitatively. Then the previously defined sets C and A can be taken in a general sense without any further mathematical structure other than the existence of suitable (σ-) fields required for the introduction of probability measures, (French and Insua, 2000).

The tabular representation (shown above) is replaced by a functional one when the action and state spaces are infinite. The notation $c : A \times \Theta \to C$ is commonly used however, other authors (such as Savage) employ an equivalent but more succinct notation which identifies the actions from the state space to the consequence space: $a : \Theta \to C$. Notably, neither notation recognizes the possibility that the set of possible states may depend upon the action chosen (French and Insua, 2000).

Statistical decision theory, on the other hand, assumes that the observation $X = x$ from an observation space X is observed according to the conditional distribution $P_X(\cdot \mid \theta)$. Thus, decision problems can be formulated as $D = A^X = \{d \mid d : X \to A\}$ which relates the decision maker's actions with each possible observation and presents the set of all possible choices. Thus, a choice $d \in D$ induces a probability distribution for each pair $(d, \theta) \in D \times \Theta$ as follows:

$$P_C(c' \mid \theta) = \int_{c(d(x),\theta)=c'} dP(x \mid \theta).$$

Consequently, this structure of statistical decision theory reduces to the general form of a decision problem. Thus, three representations of decision modeling have been presented each shown to have the same general structure.

Section 9.2: Model Building – Fundamental and Means Objectives

Modeling a decision problem (in practicality) requires three fundamental steps, as stated by Robert T. Clemen in *Making Hard Decisions*. These can be summarized concisely as (1) identifying and structuring the values and objectives, (2) structuring the elements of the decision situation into a logical framework, and (3) refining the precise definition of all of the elements of the decision model. The aforementioned steps not only provide the foundation for the decision making process but also provide the construction process for a graphical representation of decision problems and methodology.

Much of the work presented in this report has focused on single objective problems; however, often, multiple objectives or multi-attribute problems surface which encompass conflicting goals. For instance, the investor may wish to maximize the financial return of an investment portfolio but also minimize the volatility of the portfolio's value. Thus, it

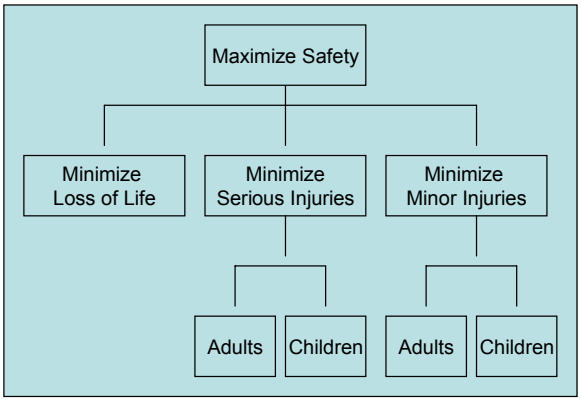
is important to address the issue of value structuring which is necessary in order to graphically display complex decision processes.

The consequence space *C*, thus far, has simply been stated as a set of objects which can be defined on an interval of the real line. When considering multiple objectives, the consequence space becomes complex and presents a multi-attributed structure. To clarify, by an attribute refers to a factor to be considered and when given a "direction" of preference, say minimize or maximize, it is termed objective. Thus, the problem arises of structuring these objectives in a consequential manner.

Once a set of objectives, consistent with the decision context, has been established, the objectives should be separated into means and fundamental objectives. This step provides a value structuring process whereby distinguishing between those objectives that are important because they help achieve other objectives and those that are important because they reflect the goal – what we really want to accomplish. This process of arranging attributes holds cognitive advantages through graphical modeling and representation.

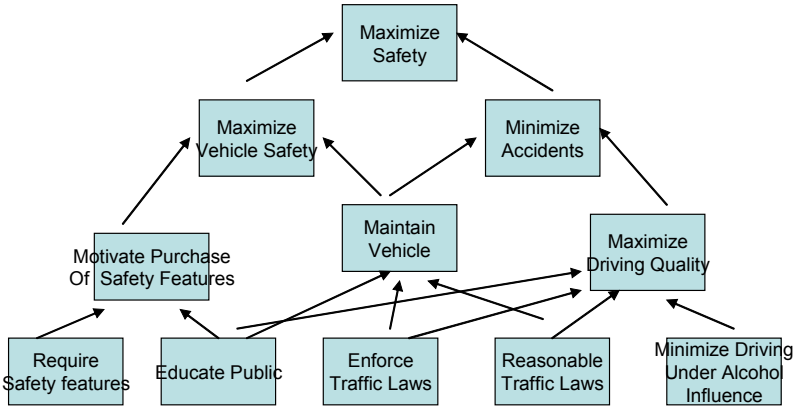
For example, fundamental objectives are organized into hierarchies. The upper levels in a hierarchy represent more general objectives, and the lower levels explain or describe important elements of the more general levels. The figure below illustrates a possible hierarchy that might arise in the context of defining vehicle regulations. A higher level fundamental objective might be "Maximize safety" where lower level fundamental objectives might be "Minimize Loss of Life," "Maximize Serious Injuries," and "Minimize Minor Injuries." Furthermore, these can be separated in order to pertain to particular issues for a particular group. As seen in this example, these objectives are parsed out separately for adults and children.

Figure 9-1:



In contrast to a hierarchy organization, means objectives are organized into networks. For the same example, means objectives may include "Minimize accidents" and "Maximize Use of Vehicle Safety Features". There are several other possible means objectives which may work together (network) in order to accomplish the maximization of safety. That is means objectives can be connected to several objectives in order to accomplish the task at hand.

Figure 9-2:



Structuring the fundamental objectives into a hierarchy or tree is crucial for a multiple decision process; however, the means objective network provides an important basis for generating creative new alternatives. In some instances, a well measured means objective can sometimes substitute for a more difficult to measure fundamental objective (Clemen, 1996). Brownlow and Watson (1987) confer that structuring attributes assist the "cognitive overload" brought about by the complexities of a decision making process. In reality, problems are much complex in structure than the one presented here. However, the graphical interface that induces this methodology still requires the building of the model or decision problem.

There are several ways in which a decision maker can choose to build such a model. Although, much of this discussion is intuitive, there are methods or techniques which facilitate this thinking process. French and Insua (2000) refer to the "top-down" questions; subsequently, Clemen (1996) presents table 9-1 which summarizes the questions or four techniques for organizing the two types of objectives which are useful in constructing means objectives networks and fundamental objectives hierarchies.

Table 9-1: Constructing Mean-Objectives Networks and Fundamental-Objectives Hierarchies

	Fundamental Objectives	**Mean Objectives**
To Move:	*Downward in the hierarchy*	*Away from Fundamental Objectives*
To Ask:	*What do you mean by that?*	*How could you achieve this?*
To Move:	*Upward in the hierarchy*	*Towards Fundamental Objectives*
To Ask:	*Of what more general objective is this an aspect?*	*Why is that important?*

It is obvious that attributes must hold certain properties or meet certain requirements in order to be useful. Recall, that the intention is to quantify such attributes. Hence, they must be measurable on some scale for each consequence. French and Insua (2002) distinguish three types of scale, noting that they may be objective or subjective measures. The first, a natural scale, provides a direct measure such as cost. The second, a constructed or subjective scale, defines the objective as well as indicates the impact of a specific consequence. That is they are created specifically within the context of the decision problem, e.g. minimizing caregiver burden. The third scale, a proxy, uses an attribute which is perceived to describe the objective and is measurable. Keeney and Raiffa (1976) affirm that no two attributes should measure the same aspect and yet should distinguish between consequences. Other practical concerns such as decision ownership and feasibility, although important in the practical setting, are not explored in this report.

Section 9.3: Model Building – A Graphical Approach

The process of specifying, structuring, and sorting out the means objectives from the fundamental objectives is the initial step of building a graphical model of the decision problem. The focus now turns on the graphical representation of the elements of a decision problem – decisions or alternatives, uncertain events or outcomes, and consequences. The most common graphical representation of decision problems is a decision tree. It has been superceded, in recent years, by influence diagrams.

Decision trees and influence diagrams, both, provide a graphical approach to modeling decision problems. This (sub-)section will focus on the common elements used is such model building processes. The latter section delved into the specifics for each modeling approach and concludes with a comparative assessment. Nonetheless, both approaches require the elements of a decision problem. Although notation varies from source to source, there are three types of nodes which are used both in influence diagrams and decision trees and can be stated as follows:

- Square nodes represent decisions (or actions) and mathematically are associated with a decision set D.
- Circular nodes represent random quantities or events which are associated with a random outcome X.
- Diamond or rounded square nodes represent consequences (value) of the decision process. These are represented by the set of all possible consequences C.

These shapes are generally referred to as decision nodes, chance nodes and consequence nodes. Nodes are integrated into graphs or networks, connected by arrows/branches or (directed) arcs. These nodes are assimilated such that they coincide with the progression of time. These are useful aids in model building. The following two sections describe a general procedure for constructing and modifying influence diagrams and decision trees to model the structure of a decision problem.

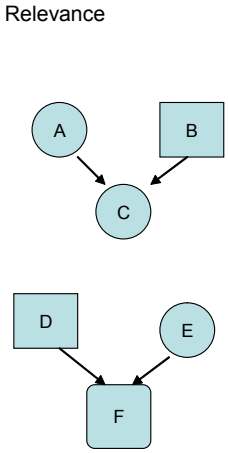
Section 9.4: Constructing Influence Diagrams

Influence diagrams provide a visual perspective of the decisions, uncertain events (states of nature), and consequences and the subsequent interrelationships that exist within a decision problem. An appropriate model ensures the probabilistic structure of a decision problem, the timing of available information and interdependence of decisions that can be taken and which may arise under certain states of nature, in a compact form (Marshall and Oliver, 1995).

Model building of decision problems as seen can become quite complex; however, influence diagrams hold many advantages. First, they provide a framework by which a decision maker can reject or confirm assumptions and accurately model dependencies through a graphical display. Secondly, complex decision problems have the innate ability to become "messy". Influence diagrams provide a format whereby the large volume of information can be summarized into relevant and sufficient matters. From a practical viewpoint, they provide alternative situations and are easily interpretable. The use of algorithms and numerical techniques also enable an efficient and simple analysis.

The design and construction of influence diagrams are based on the aforementioned three basic elements – decision nodes, chance (states of nature) nodes, and consequence nodes, connected by arcs. A node at the beginning of an arc is qualified as a predecessor; and consequently, a node at the end of an arc is referred to as a successor. The rules for using arcs to represent relationships among the nodes are shown in ***figures 9-3*** and ***9-4***. In general, an arc can represent either relevance or sequence (Clemen, 1996). The direction of the arc indicates the meaning and brings into the relevance or sequence context.

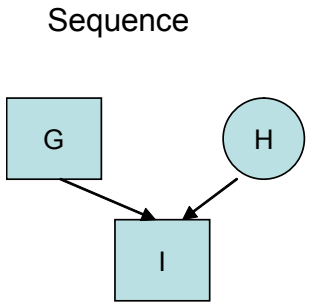
Figure 9-3:



For example, the figure above illustrates an arrow pointing into a chance node, indicating relevance. This demonstrates that the predecessor is relevant for assessing the chances associated with the uncertain event. In the diagram above, the first cluster of nodes shows an arrow (arc) from A to C. This suggests that the chances (states of nature) associated with C may be different for different outcomes of A. Likewise, an arrow pointing from a decision node to a chance node means that the chosen decision is relevant for assessing the chances associated with the succeeding uncertain event. For instance,

the choice taken in decision B is relevant for assessing the chances associated with C's possible outcomes. Relevance arcs can also point into consequence or calculation nodes, indicating that the consequence or calculation depends on the specific outcome of the predecessor node. The second cluster of nodes shows that consequence node F depends both on decision D and event E.

Figure 9-4:



When the decision maker has a choice to make, the choice would normally be made on the basis of information available at the time. Arrows that point to decisions represent information available at the time of the decision and hence represent sequence (figure 9-4). Such an arrow indicates that the decision is made knowing the outcome of the predecessor node. An arrow from a chance node to a decision means that from a decision maker's point of view, all uncertainty associated with a chance event is resolved and the outcome is known when the decision is made. Thus, information is available to the decision maker regarding the event's outcome. This is the case, as illustrated in the figure above, with chance node H and decision node I. The decision maker waits to learn the outcome of H before making decision I. An arrow from one decision to another decision simply means that the first decision is made before the second, such as decision nodes G and I. Thus, the sequential ordering of decisions is shown in an influence diagram by the path of arcs through the decision nodes.

A simple procedure for constructing an influence diagram to model the structure of a decision problem is stated in *Decision Making and Forecasting* by Marshall and Oliver (1995). These sequential steps are summarized below.

1. Create a preliminary list of decisions and random events (or quantities of interest) whose outcomes are believed to be important in the formulation of the problem. Identify the attributes and objectives that are to be used to measure the consequence of the decisions and outcomes.
2. Name each random quantity and decision. Represent each random quantity with a circular node and a decision with a square node. Draw them in order of occurrence from left to right.
3. Identify any influences or dependencies between random quantities and decision. Insert directed arcs between nodes that influence one another with the direction corresponding to the natural influence believed.

4. Determine any conditional independencies and represent them accurately.
5. Check to see that there are no directed cycles – i.e. a connected set of arcs in a directed path that leads out of one node and back itself.
6. Check that there any decision node that occurs before a later decision node has a directed arc from the former to the latter, Similarly, a chance node known to a given decision mode must be known to a later decision node. This is a requirement of the principle of coherence (Marshall and Oliver, 1995).

This process is fairly simple and thus alternative diagrams can be easily drawn. Clemen (1996) offers some other remarks on the construction of influence diagrams. First, the nature of the arc-relevance or sequence can be ascertained by the context of the arc within the diagram. Secondly, as stated in 5 of the procedure above, an influence diagram should not contain any cycles. Furthermore, although the construction of an influence diagram may be technically correct, there is no clear cut supposition to suggest that the influence diagram constructed is the only correct one. Thus, it cannot be stated that there is a unique correct diagram but that there are many ways in which a diagram can appropriately represent a decision problem. The representation that is the most appropriate is the one that is requisite for the decision maker. That is, a requisite model contains everything that the decision maker considers important in making the decision (Raiffa and Schlaifer, 2000). Incorporating all of the important concerns[1] of the decision is the only way to get a requisite model and adequate representation of the problem.

Section 9.5: Constructing Decision Trees

Influence diagrams are an excellent display of a decision problem's basic structure; however, they hide much of the detail which may be pertinent to a decision maker. To display more of these details, a decision tree can be constructed. Yet, they hold the drawback that the size of the problem cannot always be practically represented.

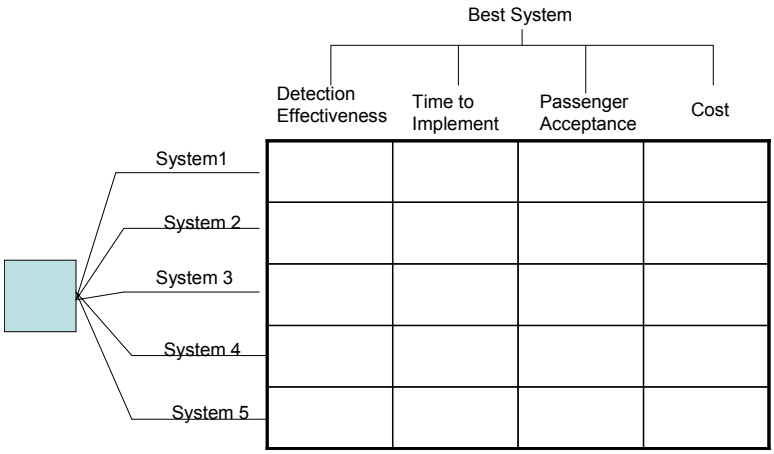
The design of a decision tree holds much of the character of an influence diagram. That is, squares represent decisions to be made, circles represent chance nodes and consequence nodes are found at the ends of the branches which connect the nodes. The interpretation of decision trees follows much the same path as explained in ***section 9.4*** on influence diagrams.

In designing a decision tree, the options represented by branches from a decision node must be such that the decision maker can choose only one option. However, in some instances, combination strategies are possible. Also, each chance node must have branches that correspond to a set of mutually exclusive and collectively exclusive outcomes. Mutually exclusive means that only one of the outcomes can happen. Collectively exclusive means that no other possibilities exist; one of the specified outcomes has to occur. Putting these two specifications together means that when the uncertainty is resolved, one and only one of the outcomes occurs.

[1] Sensitivity analysis is helpful in determining which elements are important.

A decision tree represents all possible paths that the decision maker might follow through time, including all possible decision alternatives and outcomes of chance events. In a complicated decision situation with many sequential decisions or sources of uncertainty, the number of potential paths may be very large. Each path, consequently, represents a particular scenario that could unfold over time. It is sometimes useful to think of the nodes as occurring in a time sequence.

Figure 9-5:



Including multiple objectives in a decision tree is straightforward; at the end of each branch, simply list all of the relevant consequences. An easy way to accomplish this systematically is with a consequence matrix as shown in the figure above. This is an example of a decision tree representation of FAA's multiple objective bomb detection decision (Clemen, 1996). Each column of the matrix represents a fundamental objective, and each row represents an alternative. Evaluating the alternatives requires "filling in the boxes" in the matrix; each alternative must be measured on every objective. Some basic decision trees will be described later in this section in conjunction with influence diagrams more illustratively.

Marshall and Oliver (1995) provide a procedure for constructing decision trees, once the set of random events and actions have been defined together with their sets of probabilities and respective utilities (cost/payoff). Much like (sequence) influence diagrams, decision trees are constructed in chronological order and nodes are connected by branches. The steps include:

1. Draw a branch for each possible action from the first (leftmost) decision node. If it is a chance node, do the same drawing a branch for each possible consequence.
2. Label each branch emanating from a chance node *i* with a unique element from set $\theta_i \in \Theta$, with the corresponding probability and consequence.
3. Label each branch emanating from a decision node and its respective loss or consequence.
4. Place the consequence values from the set on the terminal nodes at the end of each branch.

The designation between decisions and chance nodes is critical. Placing a chance event before a decision means that the decision is made conditional on the specific chance outcome having occurred. Conversely, if a chance node is to the right of a decision node, the decision must be made in anticipation of the chance event. The sequence of decisions is shown in a decision tree by order in the tree from left to right. If chance events have a logical time sequence between decisions, they may be appropriately ordered. If no natural sequence exists, then the order in which they appear in the decision tree is not critical, although the order used does suggest the conditioning sequence for modeling uncertainty (Clemen, 1996).

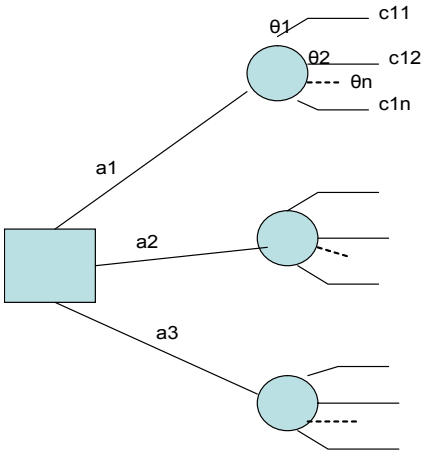
The key to constructing a decision tree from an influence diagram is the specification of every possible outcome of every node: chance outcomes from each chance node, actions from each decision node, and all possible consequences defined. Thus, a path in a decision tree is a series of decisions and random outcomes that lead from a single initial chance or decision node to a distinct end point or terminal node. The one fundamental problem with decision trees, as stated in Marshall and Oliver (1995), is that the dependencies among branches on each path are not made evident as apparent by influence diagrams.

Section 9.6: Comparing Problem Structuring

Modeling decision problems can be formulated by use of graphical methods, such as those presented above. Thus, for analysis purposes, it is beneficial to construct either a decision tree or influence diagram and bring the structure of the decision problem to the forefront. In this section, a general formulation of decision problem is presented in both graphical forms and provides the basis for a comparative assessment of these approaches.

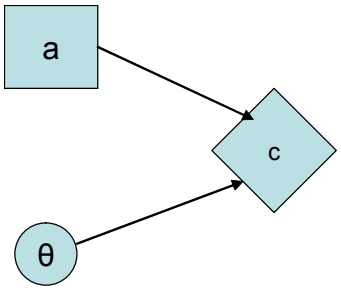
Figure 9-6 below provides a decision tree representation of the general problem presented in the table in ***section 9.1***. The decision tree displays the different possible scenarios, but does not provide a clear picture of the interrelation and influences between the uncertainties and decisions. Thus, it is not immediately clear whether the *m* chance nodes at the end of the action branches represent the same uncertainty about θ, i.e. the probability distribution of θ is not conditioned on the choice of action. The influence diagram in ***Figure 9-7*** makes this clear.

Figure 9-6:



The *mn* possible scenarios (histories) of the decision tree are represented by the *mn* possible paths in the tree.

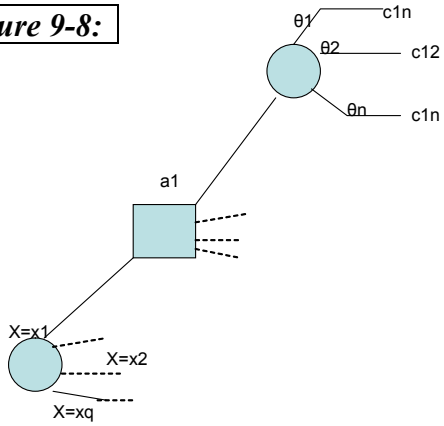
Figure 9-7:



The arrows indicate that the consequence depends on the choice of action and the unknown state; but the absence of an arrow from the action node to the state nodes indicates that the uncertainty about the state is not conditioned on the choice of action.

Figure 9-8 and ***9-9*** provide a decision tree and influence diagram for a statistical decision problem. These emphasize a further difference between the two representations. From the tree, it is clear that the observation *X* is made before an action is decided upon. So, in this case, the uncertainty concerning θ differs between chance nodes, being conditional upon the observed value of *X*. Within the influence diagram, it might appear that the arrow from *X* to *a* indicates that *X* becomes known to the decision maker before the action *a* is chosen. But if this interpretation is made for all influence arrows, it may be assumed that the uncertainty about θ is resolved and becomes known to the decision maker before *X* is observed, which it does not.

Figure 9-8:



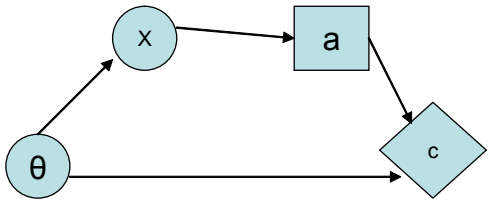


Figure 9-9:

Generally in influence diagrams, an arrow entering a chance node indicates that the probability distribution represented by that node is conditional on the quantity (random variable or decision) at the other end of the arrow. An arrow entering a decision node indicates that the quantity at the other end of the arrow is known to the decision maker at the time of the decision. An arrow entering a value node indicates that the quantity in the value node is partially determined by the quantity at the other end of the arrow. In summary, a decision tree representation displays two facets of a problem particularly clearly: contingency and temporal relationships between decisions and realizations of random events. An influence diagram[2] representation hides these facets but does bring influences-or, in probabilistic terms, conditionality-to the forefront. Together they bring complementary perspectives on the issues facing a decision maker.

[2] An influence diagram can be used to represent the structure of a decision maker's knowledge, i.e. beliefs, simply by avoiding the use of decision or value nodes. Such a diagram is referred to as a belief net and its use is most common in the field of artificial intelligence.

Decision trees have a difficulty in that for many problems they rapidly become very large: too large for the eye to comprehend as one. As such they have been described as a "bushy mess" (French and Insua, 2000). Thus, decision trees are often displayed as a series of sub-trees. Influence diagrams are a much more compact representation. However, their advantage in this respect is, in a sense, illusory. Decision trees can represent asymmetric decision problems, i.e. problems in which a particular choice of action at a decision node makes available different choices of action at subsequent decision nodes that those available after an alternative choice (Raiffa & Schlaifer, 2000). Such asymmetric problems are rather the rule than the exception in decision analysis. Within the classes of problems commonly considered in statistical decision theory, asymmetry is less common. It is the asymmetry within problems that causes the decision trees to grow into bushy messes; and it is unfortunate that, as commonly formulated, influence diagrams cannot model asymmetric problems.

The discussion and the examples shown demonstrate, on the surface at least, that decision trees display considerably more information than do influence diagrams. And is previously commented, decision trees get "messy" much faster than do influence diagrams as decision problems become more complicated. Even one of the most complicated decision trees may not be capable of showing all of the intricate details contained in an influence diagram depicting the same decision problem. In practice, the understanding of the graphical representation of influence diagrams is optimal as it is regarded as especially easy to understand regardless of one's mathematical training (Clemen, 1996).

Both methods are worthwhile, and they complement each other well. Influence diagrams are particularly valuable for the structuring phase of problem solving and for representing large problems. Decision trees display the details of a problem. Yet, the ultimate decision need not depend on the representation, as influence diagrams and decision trees are isomorphic. That is a properly built influence diagram can be converted into a decision tree, and vice verse; although the conversion may not be easy. One strategy is to start by using an influence diagram to help understand the major elements of the situation and then convert to a decision tree to fill in details.

Influence diagrams and decision trees provide two approaches for modeling a decision. Because the two approaches have different advantages, one may be more appropriate than the other, depending on the modeling requirements of the particular situation. For example, if it is important to communicate the overall structure of a model to other people, an influence diagram may be more appropriate. Careful reflection and sensitivity analysis on specific probability and value inputs may work better in the context of a decision tree. Using both approaches together may prove useful; the goal, after all, is to make sure that the model accurately represents the decision situation. Because the two approaches have different strengths, they should be regarded as complementary techniques rather than as competitors in the decision modeling process.

Section 10: Probabilistic Graphical Models

The axiomatic developments presented in Part II justify the subjective expected utility representation of a rational decision maker's preferences in a decision problem; but ignore the structure in the state, observation or consequence spaces. This section focuses on the process of structuring the "unknowns" into parameters and observations thereby facilitating the probability model building. This sections begins be re-visiting a decision model framework with emphasis on the statistical elements of parameters and observations. This leads to probability modeling and the representation of probabilistic graphical models from a Bayesian perspective. The development of models such as Bayesian networks (belief nets), multi-level or hierarchical models and exchangeability are the emphasis in the later portion of this section.

Section 10.1: Subjective Expected Utility Model – Rebuilding the Decision Framework

The modeling of a decision maker's preferences or beliefs is justifiably represented through the established subjective expected utility axioms. Such a model has been described as having three essentials: the action $\{A\}$, states $\{\Theta\}$, and consequence $\{C\}$ spaces. Drawing upon the scientific process of modeling, French and Insua (2000) restate the previously defined decision model in a more general framework containing parameters and observable variables. Thus, let's consider a model with parameters θ and input variables α of the form $M(\alpha,\theta)$. French and Insua (2000) proceed to develop comments of such a modeling framework.

The model $M(\cdot,\cdot)$ used to predict a set of observed values $y = M(\alpha,\theta)$, consists of α values which are observed while the parameters θ contain information obtained from previous studies. That is the uncertainty lies within θ whereas α is known. It is proposed that in decision modeling, partitioning α into $(\alpha_a, \overline{\alpha}_a)$ relates back to the model previously defined. If α_a is defined to be the set of actions such that $a \in A$ and $\overline{\alpha}_a$ is defined to be those random quantities which are unknown, then re-labeling $\overline{\alpha}_a$ as θ gives $M(\alpha,\theta)$. As stated in ***section 9***, there is no correct model. Such models simply provide a means of encoding the current assumptions and stating the understood relationships. These models are descriptive in nature and further taken to be deterministic accept when allowing a prescriptive interpretation of probability.

The decision model presented in Part I and revisited in the previous section concedes that the consequences of a decision are assumed to be determined by the action chosen and the unknown state:

$$a \oplus \theta \to c\,.$$

The decision model can be restated as

$$c(\alpha,\theta) = M(\alpha_a,\theta).$$

Based on the axiomatic developments in Part II, the subjective expected utility model takes the following form:

$$\text{Choose } a = \arg\max_a E_\theta[u(c(a,\theta))].$$

Thus, encoded in the model is the decision maker's uncertainty and preferences. The observed value X is subtly introduced into the model by redefining the parameter θ to include the observations: $\theta = (\theta, X)$. The justification for the use of conditional probabilities is provided by the following axiom (French and Insua, 2000):

$$\forall R,S,T \in Q, (R \mid T) \gtrsim (S \mid T) \Leftrightarrow R \cap T \gtrsim S \cap T.$$

DeGroot (1970) proceeds with the following theorem based on the axioms stated in Part II of this report by asserting that $\exists P_\theta(\cdot)$, a probability distribution on Θ, such that

(a) $\forall R,S \in Q : R \gtrsim S \Leftrightarrow P_\theta(R) \geq P_\theta(S)$;

(b) $\forall R,S,T \in Q \text{ with } T \succ \emptyset : (R \mid T) \gtrsim (S \mid T) \Leftrightarrow P_\theta(R \mid T) \geq P_\theta(S \mid T)$.

DeGroot's next step is to construct the utility function. Including the observed value X=x, gives the subjective expected utility model to be

$$\text{Choose } a = \arg\max_a E_\theta[u(c(a,\theta)) \mid x].$$

Note that the loss function $l(a,\theta) = -u(c(a,\theta))$ and $\delta(x) = a$ which is the action taken after observing x. This brings us back to the development of the statistical decision problem presented in Part I within the context of the developments found in Part II. This leads directly into the building of probability models.

Section 10.2: Building the Probability Framework

The decision model elements: A, Θ, and C (subsequently defined as the action, state, and consequence spaces) which constitute the parameters and random and/or observed variables of the model have been prescribed as vectors of real or integer numbers or categorical variables. In the previous section, both the general form of the model and the consequence model $c(\alpha,\theta) = M(\cdot,\cdot)$ are given to be deterministic; unless accentuated by a "propensity interpretation of probability" (French and Insua, 2000). Decision making under uncertainty is quantified or assessed through the development of probability distribution and epitomizes the intrinsic nature of statistical decision theory.

The joint distribution of the unknown states of nature θ and the observed random variable X are expressed through the joint distribution $P_{\theta,X}(\theta,X)$. The relationship between the deterministic model $M_X(\cdot,\cdot)$ and the ascribed joint probability distribution $P_{\theta,X}(\theta,X)$

accounts for the decision maker's uncertainties and the confounding effects within the modeling framework.

As opposed to the descriptive structure of decision modeling construed in earlier sections, the prescriptive approach discussed here is indicative of the structure of a decision maker's thoughts or belief process (French and Insua, 2000) and resides in the relationship between the construction of probability distributions and utility functions. A review of the development of the form or structure of the joint probability distribution (the Bayesian perspective) leads to the construction of probability models and an understanding of the graphical approach.

The development of the distributions $P_{\theta,X}(\theta, X)$, $P_\theta(\theta \mid X)$, and $P_X(X \mid \theta)$ is intrinsic in determining the confounding effects of both observations and parameter estimates. The construction of the joint probability distribution elicits the Bayes' theorem. Furthermore, the Bayes' theorem relates the joint distribution to the marginal and conditional distributions such that joint probability of two events *A* and *B* are given as:

$$P_\theta(A, B) = P_\theta(A \mid B) P_\theta(B).$$

The Bayes' theorem re-arranges this form to demonstrate that

$$P_\theta(A \mid B) = \frac{P_\theta(B \mid A) P_\theta(A)}{P_\theta(B)}.$$

The events *A* and *B* are independent, denoted $A \perp B$. Thus, the symmetry of the condition given above is immediate. This concept of conditional probability within the utility context is introduced by DeGroot (1970) in his presentation of the subjective expected utility axioms as shown in s***ection 10.1***.

The concept of conditional probabilistic independence extends further the modeling structure of a decision maker's beliefs (French and Insua, 2000). And the Bayes theorem tells us that "our revised belief for A", the posterior probability $P_\theta(A \mid B)$, can be obtained by the product of the prior probability $P_\theta(A)$ and the ratio $P_\theta(B \mid A) / P_\theta(B)$ where the numerator of the ratio is the likelihood of event *A*. Thus, the posterior probability can be expressed in the following form:

$$\textbf{posterior} \propto \textbf{prior} \times \textbf{likelihood}$$
$$P_\theta(A \mid B) \propto P_\theta(A) \times P_\theta(B \mid A).$$

This process of identifying conditional independence allows for the beliefs of a decision maker to be represented in the form of a probability distribution. Furthermore, they aid in developing the structural form of the utilities and in the assessment of the problem.

Section 10.3: Directed Acyclic Graphs (DAGs) and Probability

As illustrated in ***section 9*** decision models can be represented graphically and provide insight into the structure of the problem. It was stated that influence diagrams constitute the formation of a network through a graphical framework. A formal description of directed acyclic graphs (DAGs) is given in this section and offers a graphical approach to Bayesian inference. This section begins with a description of directed graphs as stated in graph theory and leads to its abilities to illustrate the Bayesian inference process.

A directed graph consists of nodes and directed arcs, as stipulated in the section on influence diagrams. Formally, we may consider a network with nodes $i, j = 1,2,...,n$ and a subset of the possible ordered pairs (i, j) which specifies the directed arcs. A path consists of a sequence:

$$i_1 \to i_2 \to i_3 \to ... \to i_k .$$

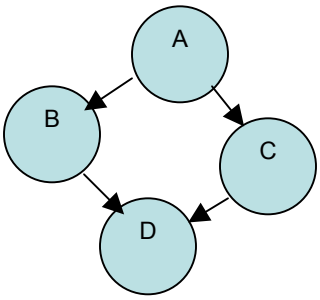
Such a directed graph is said to be acyclic when there exists no path with $i_1 = i_k$. The order of the nodes can be rearranged so that every directed arc $i \to j$ has $i < j$. Bather (2000) gives the following proposition:

If a directed network is acyclic, then the vertices can be renumbered in such a way that i < j for each directed arc $i \to j$.

Bather (2000) goes on to prove this proposition which is omitted from this report. It should also be noted that a path only exists if there is a link from one node to another; else they are disconnected. This representation is useful when discussing causation. For completeness purposes, a few terms in relation to graph theory are provided below.

Much of the terminology surrounding graph theory is centered on kinship relationships such as parents, child, ancestors, etc. A node in a directed graph is said to be a root if it has no parents and a sink if it has no children (Pearl, 1988). Every DAG has at least one root and one sink. The term tree is used to qualify a connected DAG in which every node has at least one parent. Similarly, if every node has a child then it is referred to as a chain. A graph is said to be complete if every pair of nodes is connected by an arc. The figure below illustrates an example of a DAG containing four nodes.

Figure 10-1:



This mathematical structure, a directed graph, represents a causal network consisting of variables and directed links. This causal framework suggests the way influence may run within a causal network. Two nodes (variables) are said to be separated if "new evidence on one of them has no impact on our belief of the other" (Jensen, 2001). There are three types of connections between triplets of variables *A*, *B*, and *C*.

Figure 10-2a:

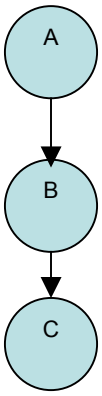
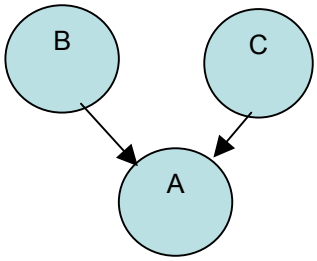
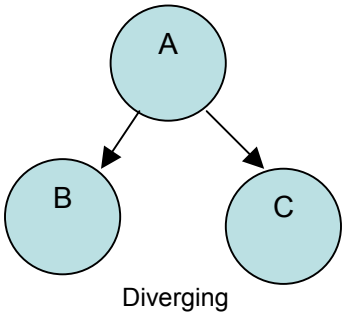


Figure 10-2b:



Converging

Figure10-2c:



Diverging

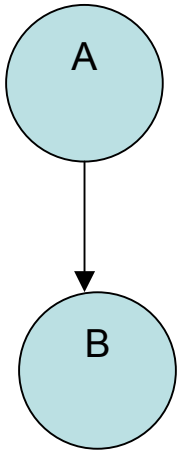
In both the linear and diverging cases, we have $(A \perp B \mid C)$ while in the converging case we have $(A \not\perp B \mid C)$.

Definition: (Jensen, 2001)
Two variables A and B in a casual network are said to be d-separated if for all trails between A and B there is an intermediate variable V such that either

i. *the connection is serial (linear) or diverging and the state of V is known; or*
ii. *the connection is converging and neither V nor any of V's descendents have received evidence.*

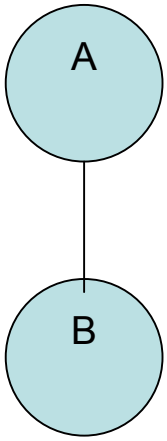
This follows the properties of conditional independence (described in ***Section 10.2***) as the quantification of uncertainty in causal structures must obey the principle that whenever *A* and *B* are d-separated then new information on one of them does not change the certainty of the other.

Figure 10-3(a):



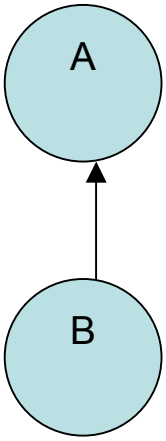
P(A)P(B|A)

Figure 10-3(b):



P(A,B)

Figure 10-3(c):



P(B)P(A|B)

For Bayesian inference, DAGs illustrate the prior to posterior inference process. Each diagram above represents the joint distribution of two random variables *A* and *B*. A causal interpretation follows from ***figure 10-3(a)*** which describes the prior beliefs. The posterior beliefs are represented by ***figure 10-3(b)*** and elicits an inferential process. Such a representation forms the basis of Bayesian networks or belief nets.

Section 10.4: Directed Acyclic Graphs (DAGs) for Inference on Bayesian Networks

The use of DAGs for Bayesian inference illustrates the prior to posterior inference process as demonstrated in the previous section. Bayesian networks, on the other hand, are generally more complicated than those presented above. Robert Cowell in his article *Introduction to Inference for Bayesian Networks*, defines Bayesian networks as a

"*model representation for the joint distribution of a set of variables in terms of conditional and prior probabilities, in which the orientations of the arrows represent*

influence, usually though not always of a causal nature, such that these conditional probabilities for these particular orientations are relatively straightforward to specify (from data or eliciting from an expert)".

Hence, in this section, the focus falls on the inferential procedure which involves the calculation of marginal probabilities conditional on the observed data using Bayes' theorem. This is equivalent to (diagrammatically) to reversing one or more of the Bayesian network arrows. This section defines a Bayesian network and shows how one can be constructed from prior knowledge.

Let's first consider a simple causal network where A is a parent of B. Then it is simple to calculate the $P_\theta(B \mid A)$. However, if C is also a parent of B then the individual conditional probabilities $P_\theta(B \mid A)$ and $P_\theta(B \mid C)$ are not sufficient in providing any information on the interaction between A and B. Hence, a specification of $P_\theta(B \mid A, C)$ is required. This ideology can be expanded to consider n random variables. Typically, the interest lies in looking for relationships among a large number of variables for which the Bayesian network is suitable for.

A Bayesian network for a set of variables $\mathbf{X} = \{X_1, X_2, ..., X_n\}$ consists of a set of directed arcs which in turn provide (a) a network structure S that encodes a set of conditional independence assertions about variables in **X**, and (b) a set P of local probability distributions associated with each variable (Heckerman, 1998; Jensen, 2001). Thus, the variables together with the directed arcs form a directed acyclic graph. Together these components:

i. define the joint probability distribution for **X**; and
ii. provide a one-to-one correspondence between S and **X**.

Thus, a Bayesian network is a DAG whose structure defines a set of conditional independence properties. These properties can be found through graphical manipulations such as those presented in ***section 10.3*** (Pearl, 1988). It has been contended that any uncertainty must obey the definition of d-separation (Jensen, 2001). This is to say, that a conditional probability distribution is associated with each node where the conditioning is on the parents of the node $P(X_i \mid par(X_i))$. Further, d-separation can be used to read off conditional independencies from DAG representation of a Bayesian network.

Given a network structure S, the joint probability distribution over the set of all variables U is given by

$$P(U) = \prod_i P(X_i \mid par(X_i)),$$

where the $par(X_i)$ is the parent set of X_i. Thus, the local probability distributions P are simply the distributions corresponding to the terms in the product of the equation given above (Heckerman, 1999). Jensen (2001) defines this as the chain rule.

Next let's consider the process of building a Bayesian network. Heckerman (1999) illustrates this process through an example. Here, we generalize his "learning by example" approach. The first task is much like that of the one described in developing a decision tree or influence diagram. That is to say this initial task entails the more logical structuring of the parameters and/or variables of interest. This can be characterized to include: (1) the correct identification of the goals of modeling (i.e. prediction versus explanation versus exploration); (2) the identification of all possible observations that may be relevant to the problem; (3) the identification of a subset of relevant observations to model; and (4) the classification of observations into variables having mutually exclusive and collectively exhaustive states. As seen earlier, this is a task embedded in decision analysis and hence is not exclusive to Bayesian modeling. This concludes the more logical framework development in the construction of a Bayesian network.

The next step in the construction of a Bayesian network entails the more mathematical or statistical framework. It is at this stage a DAG encoding the "assertions of conditional independence" are built. This approach is based on the chain rule of probability which is equivalent to the $P(U)$ shown above. In general the chain rule of probability is defined as

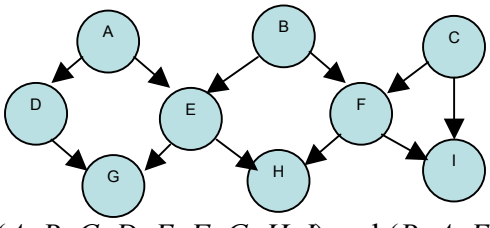
$$P(X) = \prod_{i=1}^{n} P(x_i \mid x_1,...,x_{i-1}) = \prod_{i=1}^{n} P(x_i \mid \pi_i).$$

Then for every X_i there exists some subset $\Pi_i \subseteq \{X_1,...,X_{i-1}\}$ such that X_i and $\{X_1,...,X_{i-1}\} \setminus \Pi_i$ are conditionally independent given Π_i. Thus, it is clear that the variables sets $(\Pi_1,...,\Pi_n)$ correspond to the parent nodes of a Bayesian network and in turn specifying the arcs in the network structure *S*.

It follows that in determining the structure of the Bayesian network requires ordering the variables and determining the most appropriate subset of variables. DAGs can always have their nodes linearly ordered so that for each node *X* all of its parents precedes it in the ordering. Such an ordering is referred to as a topological ordering (Cowell, 1999).

Consider the graph shown below with nine nodes. This has been taken from Cowell (1999).

Figure 10-4:



This example shows that (*A*, *B*, *C*, *D*, *E*, *F*, *G*, *H*, *I*) and (*B*, *A*, *E*, *D*, *G*, *C*, *F*, *I*, *H*) are two possible topological orderings. Thus, this task (of ordering the variables) may not be the most feasible or reasonable process to be used unless done so under a more logical

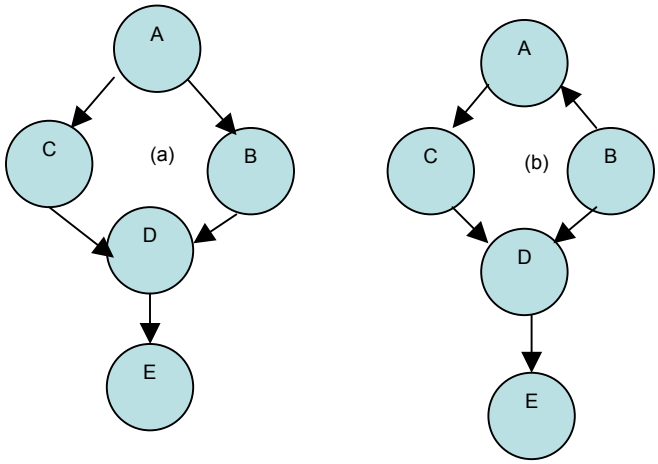
framework. For instance, if the variable ordering chosen is not probable, the resulting network structure may fail to reveal some of the important conditional independencies or in the worst case $n!$ variable orderings may need to be explored. Thus, applying the semantics of causal relationships (as observed naturally) readily asserts the corresponding conditional dependencies.

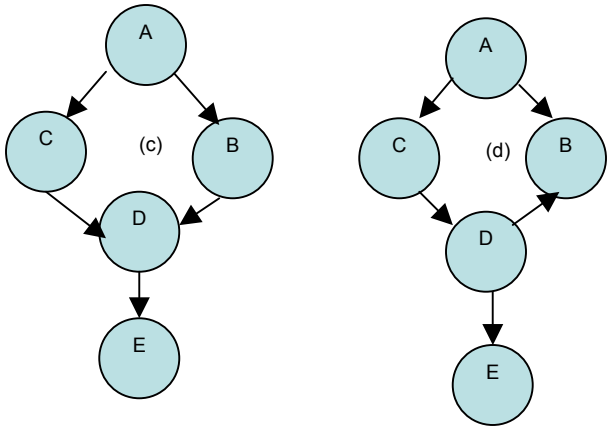
Now, the task at hand is to compute the joint density over the set of all variables U as defined earlier. According to the DAG, this is referred to as recursive factorization or "the distribution being graphical over the DAG" (Cowell, 1999). It follows a similar procedure illustrated above by simplifying the individual terms (or nodes) with respect to their structural parents. Then, the final step in the construction of Bayesian networks "simply" entails the assessment of the local probabilities, defined as $P(X_i \mid par(X_i))$ for each i. This concludes the systematic approach to the construction of Bayesian networks.

It has been illustrated that each of these models determines a set of conditional constraints, represented implicitly in the DAG. However, the implied criteria or assumptions have not been discussed explicitly, and hence is made mention of here which Pearl (1988) refers to as stability. The stability condition requires that all of the probabilistic independence relations implied by the model should be invariant across (small) perturbations to the parameters of the model. That is the parameters of the model should not be functionally related to each other.

Furthermore, Verma & Pearl (1990) proved that two DAGs are observationally equivalent if they have (i) the same skeleton; and (ii) the same sets of nodal structure – i.e. two converging arcs whose tails are not connected by an arc. Using this criterion, it can be seen that ***figure 10-5(a) and (b)*** that probabilistically they are indistinguishable while that ***figure 10-5(c) and (d)*** are not. Thus, a variety of belief nets can represent the same conditional independencies.

Figure 10-5:





So far the emphasis has been much on the construction of a Bayesian network (from prior knowledge, data or a combination). And the need is usually to determine various probabilities of interest from the model. Thus, we now focus on probabilistic inference in Bayesian networks. Because a Bayesian network for U determines a joint probability distribution for U (a set of all variables), in principle, the Bayesian network can be used to compute any probability of interest as described above. However, the task to refine the structure and local probability distributions of a Bayesian network given data results in a set of techniques for data analysis that combines prior knowledge with data to produce improved knowledge (Heckerman, 1999). This is process referred to as "belief updating in Bayesian networks". Jensen (2001) summarizes the mathematical process involved:

1. Use the chain rule to calculate $P(U)$.
2. Determine the $P(U \mid \theta_s, S^h) = \prod_{i=1}^{n} P(X_i \mid par(X_i), \theta_s, S^h)$ where θ_i is the vector of parameters for the distribution $P(X_i \mid par(X_i), \theta_s, S^h)$, θ_s is the vector of parameters $(\theta_1, ..., \theta_n)$, and S^h denotes the configuration of the network.
3. Marginalize $P(U \mid \theta_s, S^h) = \prod_{i=1}^{n} P(X_i \mid par(X_i), \theta_s, S^h)$ to $P(X_i \mid par(X_i), \theta_s, S^h)$ for each X_i in U.
4. Normalize $P(X_i, par(X_i), \theta_s, S^h)$ results in $P(X_i \mid par(X_i), \theta_s, S^h)$.

This illustrates the basic ideas for learning probabilities (and structure) within the Bayesian or belief net framework. This section has described the construction steps in a sequential manner; however, it is important to assert that many of these "steps" are intermingled in practice. Both the problem of structuring and the assessments of probability can lead to changes in the network structure. The section following presents an extension of both probability modeling and belief nets within a hierarchical form.

Section 10.5: Hierarchical Structures in Belief Nets

The natural occurrence of hierarchical structures in physical phenomenon is quite common. Moreover, these structures find themselves in belief net representations as hierarchical models. Hierarchical models were introduced in statistical modeling in 1972 by Lindley and Smith (Gelman et. al, 1995). Such models hold several advantages.

The simplest form of a hierarchical model can be described as a representation of beliefs which consists of a layered structure. For instance, consider a set of observations $Y_1, Y_2 ..., Y_n$ each related independently to a set of parameters $\theta_1, \theta_2, ..., \theta_n$ which share some similarity or characteristic. The assumption about θ_i having a common distribution $P_\theta(\cdot \mid \gamma)$ represents this "similarity". A third level suggests that γ may be represented by a prior distribution with parameters ξ. Thus the model described is a three level model:

$$\begin{aligned}
&\text{Level 1}: \; Y_i \sim P_{Y_i}(\cdot \mid \theta_i), i = 1, ..., n. \\
&\text{Level 2}: \; \theta_i \sim P_\theta(\cdot \mid \gamma), i = 1, ..., n. \\
&\text{Level 3}: \; \theta \sim P_\gamma(\cdot \mid \xi).
\end{aligned}$$

Thus, the concept here is to simply extend a common parameter θ to model the distribution of many observable quantities X_i. If the parameter has many components then it may be useful to specify their joint prior distribution using a common hyper-parameter, say γ (Neal, 1996) which is given its own prior. This hierarchy of a structure is referred to as hierarchical modeling.

If the θ_i are independent given γ, then

$$P(\theta) = P(\theta_1, ..., \theta_n) = \int P(\gamma) \prod_{i=1}^{n} P(\theta_i \mid \gamma) d\gamma \, .$$

Although the γ could have been dispensed (with θ given a direct prior); however, Neal (1996) states that using a hyper-parameter "may be much more intelligible". This could be extended to the level 1 data which would then require the decision maker to integrate over θ.

In the hypothetical model illustrated above, the distribution $P_\theta(\cdot \mid \gamma)$ is common to all θ_i, while the distributions $P_{Y_i}(\cdot \mid \theta_i)$ may vary. This allows for the articulation of dependence or independence in the beliefs about the components, if the Y_i are multi-dimensional (French and Insua, 2000). Further, a variety of experimental conditions can be ascribed.

A simple and natural occurring example of a hierarchy structure is in determining the school to school variability of student achievement. Suppose the response is math achievement on a standardized test. The innate structural formation describes students to

be nested within classrooms, classrooms nested within schools, and this could go on and on. This would mean that θ_i is the mean math score for classroom *i*, specified by a normal distribution for that particular classroom. Thus, this allows us to update or modify our belief of the math ability of students in general by adjusting the likely value of the hyper-parameter γ. Gelman et al. (1995) presents the advantages of hierarchical models in terms of statistical inference. Such strengths draw upon the ability to learn about the hyper-parameter γ as well as drawing upon some information about the particular effects of θ_i.

These three level hierarchical models are not restricted to the structural formation illustrated above but have also been used to structure statistical inference in terms of observational and modeling error and prior information on the parameters. Furthermore, hierarchical models are not restricted to three levels; more may be appropriate. That judgment is somewhat left to the decision maker.

Section 10.6: Exchangeability and Other Forms of Probability Models

One of the most fundamental concepts inlaid within subjective probability modeling is the concept of exchangeability introduced by De Finetti's. Conditional independence, Bayesian networks, and hierarchical models give the decision analysts the tools for exploring and constructing the form of the joint distribution through a decomposition into its marginal and conditional components; However, the independence conditions to do not suggest the parametric form of any of the conditional or marginal distributions whereas exchangeability can (French and Insua, 2000).

De Finetti's "twist of hypothetical permutation suggests that rather than "judging" the resemblance between two groups, the decision maker should "imagine" a hypothetical exchange of two groups and then decide whether the observed data under the swap would be distinguishable from the actual data (Pearl, 1999). Let *A* and *B* denote two groups, treated and untreated, and let $P_{A1}(y)$ and $P_{A0}(y)$ denote the distribution of group *A* under two hypothetical conditions, treatment and no treatment. Pearl explains that if the interest lies in some parameter μ of the response distribution then μ_{A1} and μ_{A0} denote the values of the parameters in the corresponding distribution $P_{A1}(y)$ and $P_{A0}(y)$, (with μ_{B1} and μ_{B0} defined similarly for group *B*. Thus, in measuring the pair (μ_{A1}, μ_{B0}), the hypothetical swap then suggests measuring (μ_{B1}, μ_{A0}).

Pearl (1999) defines two groups to be exchangeable relative to parameter μ if the two pairs are in indistinguishable, that is if

$$(\mu_{A1}, \mu_{B0}) = (\mu_{B1}, \mu_{A0}).$$

French and Insua (2000) extend this concept and state the following proposition:

A collection of random variables is exchangeable if and only if any n-tuple, with n less than or equal to the sixe of the collection, has the same distribution as any other n-tuple.

Thus, "exchangeability develops the parametric form of a distribution from relatively weak ideas and of symmetry between future observables rather than from modeling ideas derived from scientific understanding" (French and Insua, 2000). Literature also contends that these sets of symmetry conditions are not easily identifiable, and thus, sometimes the parametric structure of the distribution is based more on intuition.

The methods illustrated in ***Section 10*** provide a framework for the structuring of the functional form of the decision maker's probability distributions. In practice, many of the methods are mixed and matched to accomplish the ultimate goal. Hence, a decision maker may choose to build a Bayesian network and then focus on some nodes and borrow conditions of exchangeability and modify the structure. Either way, there are many choices in the structuring of decision problems; and this section introduces some of these concepts.

Section 11: Probabilistic Graphical Models with Markovian Properties

The framework for constructing probabilistic graphical models as DAGS is an initial step in modeling complex decision problems; however, the formalism of hidden components or sequential decision making processes has been ignored, thus far. This section does not survey sequential decision making but does strive to present some of the extensions of the graphical models presented earlier. It focuses primarily on sequential problems which can be formulated in Markovian ways. Such applications stem from the body of theory presented in PART III. Thus, we begin with exact inference when countering hidden nodes and then briefly introduce dynamic linear modeling which underpins many Bayesian forecasting techniques. This extends both probability modeling and the representation of probabilistic graphical models described in the previous section. The development of models such as Bayesian networks (belief nets) or hierarchical models are extended to the development of neural networks, hidden Markov models, and dynamic linear models.

Section 11.1: Inference in Probabilistic Graphical Models

The modeling of probabilistic graphs was introduced in ***section 10***. This section extends such structures to discuss probabilistic inference in graphical models which hold hidden components or nodes. The problem in such inference making resides in the computation of the conditional probability distribution over the unobserved or hidden nodes given the observed information.

Let H denote the set of hidden nodes and E denote the set of "evidence" or observed nodes. Thus the goal is to calculate the

$$P(H \mid E) = \frac{P(H, E)}{P(E)}.$$

Further, another aim is to calculate the marginal probabilities in graphical models, in particular the $P(E)$. Notably this is the likelihood function which of course has a direct relationship with the conditional probability $P(H \mid E)$. Thus, these distributions are intertwined and hence, any inference or algorithmic approaches to computing such probabilities coincide.

As shown in ***section 10***, when belief updating in Bayesian networks, the joint probability distribution $P(U \mid \theta_s, S^h)$ is sufficient for the calculations required above. However because the joint probability and corresponding calculations (based on its decomposition) increase exponentially with the number of variables or nodes, more efficient methods are desired (Jensen, 2001). Jordan (1999) presents an array of such methods encompassing both exact and approximation algorithms. The junction tree algorithm, an exact approach is introduced through the concepts of "moralization" and "triangulation".

The focus in this part of the report has been on graphical models – more specifically on DAGs. To recap, DAGs (directed acyclic graphs) can be numerically assessed determining the local conditional probabilities associated with each mode. That is these conditional probabilities give the $P(X_i \mid par(X_i))$. To determine the joint probability distribution of all of the *N* nodes, the product is taken over all possible nodes:

$$P(U) = \prod_i P(X_i \mid par(X_i)).$$

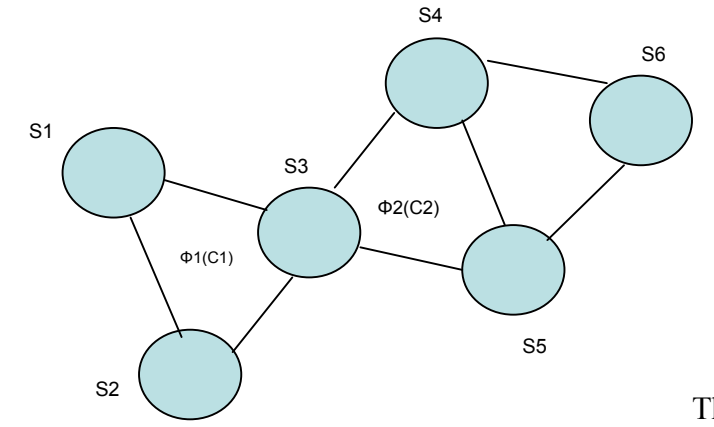
Thus, the probability inference in DAGs involves the computation of conditional probabilities under this joint distribution.

This report has neglected undirected graphical models, thus far, as the introduction to this subject area was directed by Bayesian networks in a very systematic approach. However, we introduce them here. First off, such models are also referred to as Markov random fields. They are numerically assessed by "associating potentials with the cliques of the graph" (Jordan, 1999). A potential is defined as a function on the set of configurations of a clique that associates a positive number with each configuration. And a clique is a subset of nodes which are fully connected and maximal. To compute the joint probability distribution for all of the nodes, the product is taken over all the clique potentials:

$$P(U) = \frac{\prod_{i=1}^{M} \phi_i(C_i)}{\sum_U \left\{ \prod_{i=1}^{M} \phi_i(C_i) \right\}},$$

where *M* is the total number of cliques and the normalization factor (the denominator) sums the numerator over all configurations. ***Figure 11-1*** is an example of an undirected graph adapted from Jordan (1999).

Figure 11-1:



The cliques are $C_1\{X_1, X_2, X_3\}$, $C_2\{X_3, X_4, X_5\}$ and $C_3\{X_4, X_5, X_6\}$.

The junction tree algorithm is commonly applied as a means of making exact inferences and compiles directed graphical models into undirected graphical models. Jensen (2001) provides a complete description and sequential steps involved in this graph theoretic representation of this approach. Recall, that it was stated that this approach required to main steps, "moralization" and "triangulation". These are described next.

The first stage, moralization, translates the directed graph into an undirected graph. Both graphs use the produce of the local functions to obtain the joint probability distributions of the nodes. The directed form, $P(X_i \mid par(X_i))$, does hold the property of being a real-valued function over the configuration; however, these nodes are not always situated together within a clique. Thus, the objective is to "marry" the parents of all of the undirected edges in the graph. This is done by simply dropping the arcs on the other edges in the graph (Jordan, 1999). The resulting graph is called a moral graph and can be used to represent the probability distribution on the original directed graph within the undirected structure.

The second stage of the algorithm is referred to as triangulation. This processes the moral graph (taking it as input) and produces an undirected graph (perhaps with additional edges). This graph holds a property which allows recursive calculations of probabilities (Jordan, 1999). In a triangulated graph, joint probabilities can be built up sequentially through the graph. ***Figure 11-2*** illustrates two graphs – (a) represents a non-triangulated graph whereas (b) demonstrates its triangulated form. Thus, if there are 4 cycles which do not have an edge, it is considered not triangulated but can be simply formulated to be by simply adding a chord as seen below.

Figure 11-2(a):

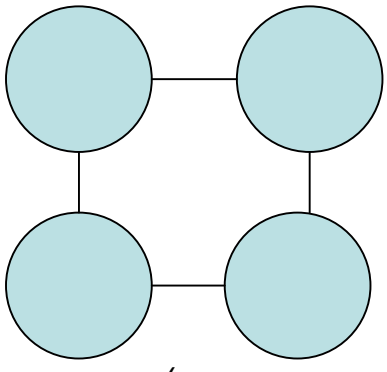
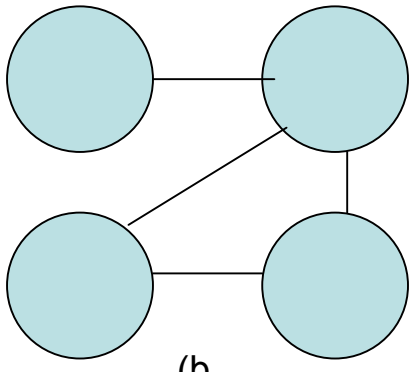


Figure 11-2(b):

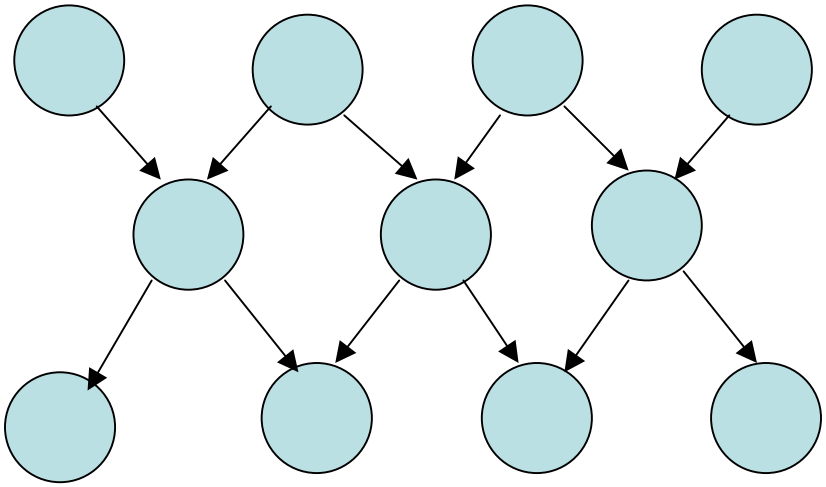


Once a graph has been triangulated, the cliques of the graph can be arranged into a data structure known as a junction tree. The aforementioned algorithmic structure is made possible for trees and not networks because of the independent properties. That is, if X is instantiated then all its neighbors are independent (Jensen, 2001). To summarize, through graph theoretic methods, the independence properties in a network are converted and organized into a set of sets of variables (cliques) and to construct a tree over the cliques, resulting in a junction tree.

Section 11.2: Neural Networks

Neural Networks are Layered graphs endowed with a nonlinear "activation" function at each node (Jordan, 1999). A neural network is characterized by its: (1) architecture – its pattern of connections between the nodes; (2) learning algorithm – its method of determining the weights on the connections (probability distributions); and (3) activation function – which determines the output (Ripley, 1996). ***Figure 11-3*** below illustrates the layered graphical structure of a neural network.

Figure 11-3:



Typically, a neural network consists of a large number of processing elements (nodes), or neurons (when speaking of artificial neural networks). Each node has an internal state, referred to as its activity level or activation which is a function of the inputs it has received. Thus, a node sends its activation as a signal to several other nodes; however, it can only send one "message" at a time.

Let's first consider a typical example of a simple neural network. Actually, for simplicity, let's consider a single output node *Y* which receives input (activation) from *n* nodes $X_1,...,X_n$. From each input node to the output node *Y*, there is an associated weight θ_{ij}. Then to determine the output of the node *Y*, simply sum the inputs with their respective weights (consider multiple regression with θ_{ij} as the regression coefficients):

$$y_{ij} = \theta_{i0} + \theta_{1j}x_{1j} + \theta_{2j}x_{2j} + ... + \theta_{nj}x_{nj}$$

The θ_{i0} is the "bias" parameter which allows one to change (or bias) the output independently of the inputs. The next step is to apply an activation function, of which a number of them could be used. However, a common activation function is the logistic sigmoid function (an *S*-shape curve).

Let's consider an activation function that is bounded between 0 and 1, such as that obtained through the logistic function:

$$f(y) = 1/(1+e^{-y}).$$

Thus, a binary variable X_i is associated with each node in a graphical model; and the interpretation of the activation of the node is the probability that the associated binary variable takes one of its two values. Thus, the corresponding logistic function is:

$$P(X_i = 1 \mid par(X_i)) = \frac{1}{1+e^{-\sum_{j \in par(X_i)} \theta_{ij} X_j - \theta_{i0}}}.$$

Thus, the θ_{ij} are the parameters associated with the edges between the parent nodes *j* and node *i*, and θ_{i0} is the bias parameter associated with node *i*. This is the sigmoid belief network introduced by Radford Neal (1992). Such treatments of neural networks hold many advantages including the ability to perform diagnostic calculations and to handle missing data (Jordan, 1999).

From ***figure 11-3*** above, it is quite clear that a node in a neural network, generally, has the preceding layer as all of its parent nodes. Thus, when applying the junction tree algorithm, the moral graph links between all of the nodes in this layer, as illustrated below. Further note that the by definition (***Section 11.1***) the output nodes are the "observed" or "evidence nodes and thus the hidden nodes are probabilistically dependent and so are any of its ancestors (Jordan, 1999). Computationally this can be quite intensive as the clique size due to the triangulation procedure will grow exponentially.

Section 11.3: Hidden Markov Models

This section introduces hidden Markov models which is one of the simplest kinds of a dynamic Bayesian networks. As suggested by the model's name, hidden Markov models hold Markovian properties for which we refer to PART III. Thus, instead of regurgitating such properties, we simply relay the directed Markovian property in graphical models.

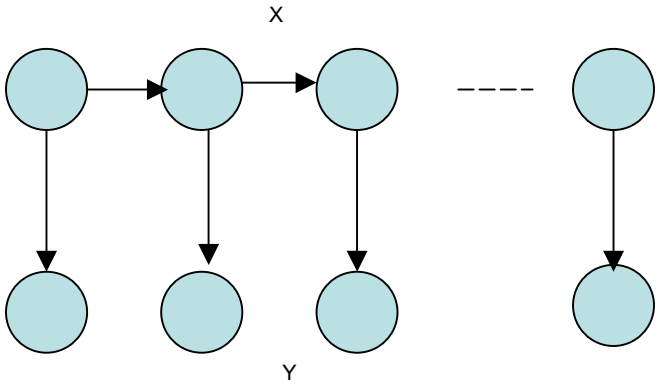
This directed Markovian property is a conditional independence property which states that a variable is conditionally independent of its non-descendents given its parents:

$$X \perp ND(X) \mid par(X)$$

where $ND(X)$ denotes the non-descendents of *X*. From ***section 10***, it was mentioned that the conditional probability $P(X_i \mid par(X_i))$ did not necessarily mean that if the $par(X) = \pi^*$ then the $P(X = x) = P(x \mid \pi^*)$. That is the suggestion was that any other information is irrelevant for this condition to hold. In DAGs, this information refers to the knowledge about the node itself or any of its descendents. So, we can summarize this

by saying that having information about a non-descendent does not tell us anything more about *X*, because either it cannot influence or be influenced by *X* either directly or indirectly (Cowell, 1999). (Or if it can be influenced indirectly, then it can only be done though influencing the parents which are all known anyway.)

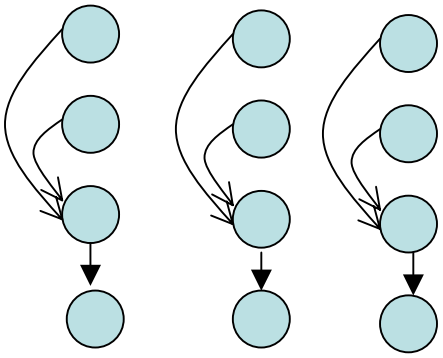
Figure 11-4:



A hidden Markov model (HMM) is a graphical model in the form of a chain as seen in the figure above. Applying the same property, consider a sequence of multinomial "state" nodes X_i and assume that the conditional probability of X_i is independent of its immediate predecessor X_{i-1}. This forms a matrix of transition probabilities $A = P(X_i \mid X_{i-1})$ which is time invariant. The output nodes Y_i have an associated "emission" probability law $B = P(Y_i \mid X_i)$ which is also time invariant.

In determining probabilistic inferences in such graphical models, the output nodes are treated as evidence nodes and the states as hidden nodes. The EM (expectation-maximization) algorithm is generally applied to update the parameters A, B, π. This two step iterative procedure first computes the necessary conditional probabilities and then updates the parameters via weighted maximum likelihood. Thus, the moralization and triangulation procedures of the junction tree algorithm are innocuous or useless for such graphical models as hidden Markov models (in this form).

Figure 11-5:



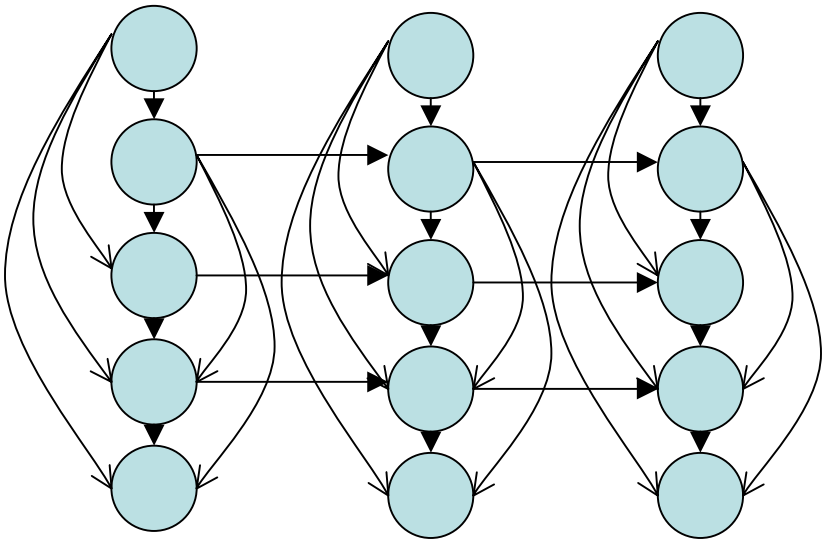
Now consider a variation on the aforementioned hidden Markov model with a number of chains. Such a graphical model is called a factorial HMM and is shown above. In general, the structure is composed of *m* chains. Hence, the state node for the *m*th chain at time *i* is denoted $X_i^{(m)}$ and corresponding transition matrix for the *m*th chain is denoted $A^{(m)}$. Thus, the overall transitional probability for the full model is computed by taking the product across the intra-chain transition probabilities:

$$P(X_i \mid X_{i-1}) = \prod_{i=1}^{m} A^{(m)}(X_i^{(m)} \mid X_{i-1}^{(m)}).$$

In this situation, the triangulation process will result in an exponential growth of cliques as there are multiple chains to be considered. Thus, a factorial hidden Markov model does not hold the tractable properties held by its more general formalism.

Another variation or amalgamation of two previously described models is hidden Markov decision trees (HMDT). As the name suggests, this is a model which is essentially a decision tree endowed with Markovian properties (Jordan, 1999). In an HMDT, the decisions in the decision tree are not only conditional on the current data point, but are also conditional on the decision at the previous moment in time.

Figure 11-6:



The figure above illustrates such a model where the horizontal edges represent the Markovian temporal dependence. Note that the term temporal is used as it is assumed that the model structure does not change (Murphy, 2003). Thus, given a sequence of input vectors U_i and a corresponding output vectors Y_i, the task at hand is to compute the conditional probability distribution over the hidden states again. This is much like the case of the FHMM and is shown to be intractable when making inferences (numerically) (Jordan, 1999).

Several other variations of the HMM exist including those of higher orders where each state depends on the previous *k* states as opposed to the single previous state or those with a mixture of both continuous or discrete nodes. Either way, the number of cliques for such models grows exponentially. Thus, although these models provide a graphical representation of complex problems, they are not easily tractable especially when concerned with exact inference algorithms. However, many of these can still be assessed using approximation methods which will be discussed generally in ***section 11.5***.

Section 11.4: Linear Dynamic Models

Another important type of dynamic Bayesian networks is the class of decision problems which has arisen out of the area of control of stochastic systems within the engineering field. Essentially, the objective is to keep systems running over a number of stages along target trajectories of states (French and Insua, 2000). They have also been developed for the problem without controls, known as dynamic linear models (DLM). Their importance resides in their formulation as a tool for Bayesian forecasting and sequential procedures.

Linear dynamic models have the same topology as HMMs, but all the nodes are assumed to have linear-Gaussian distributions. Let's consider the basic structure:

$$y_t \mid \theta_t \sim (F_t^T \theta_t, V_t)$$
$$\theta_t \mid \theta_{t-1} \sim (G_t \theta_{t-1}, W_t)$$

where y_t is scalar (data) and θ_t is the state vector. The model is defined by four quantities $\{F_t, G_t, V_t, W_t\}$ which are assumed to be known at time *t*: F_t is a *n* vector, G_t is a $n \times n$ matrix, V_t is a non-negative scalar, and W_t is a $n \times n$ symmetric and positive (semi)-definite variance matrix. A more general way of stating this model is:

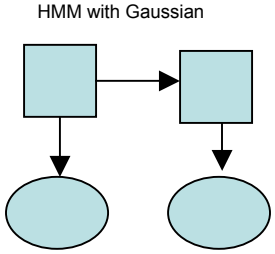
Observation equation:

$$y_t = F_t^T \theta_t + v_t, \quad v_t \sim (0, V_t)$$

System equation:

$$\theta_t = G_t \theta_{t-1} + w_t, \quad w_t \sim (0, W_t)$$

Figure 11-7:



The best way to think of a DLM is through is graphical structure as shown above. This graph has two noticeable features. The first, is that the state vector θ_t follows a Markov process, so that $\theta_k \perp \theta_j \mid \theta_{k-1}$, for all $1 \le j \le k \le t$. This implies that $w_j \perp w_k$. The second is that θ_k separates y_k from everything else, so that $y_k \perp y_j, \theta_j \mid \theta_k$. This further implies that $v_k \perp w_k$ and $v_k \perp v_j, w_j$ (West et. al., 1997).

Let $D_k = \{y_1, ..., y_k\}$ represent the data where y_k takes on either a numeric or missing value. Then the prior beliefs about the state vector can be written as:

Prior information:

$$\theta_0 \mid D_0 \sim (m_0, C_0).$$

The DLM allows us to compute m_k and C_k from m_{k-1}, C_{k-1}, and y_k. In this way, the information sets D_k are defined recursively and can be computed through stochastic dynamic programming algorithms; however, can be manageably daunting. Nonetheless, these models do provide highly effective means of modeling univariate time series (French and Insua, 2000).

Section 11.5: Variational Methods for Dynamic Bayesian Networks

The junction tree algorithm and dynamic programming algorithms provide solutions for probabilistic inferences in graphical models. However, as we have seen in this section, a need for other approaches is crucial for many forms of dynamic Bayesian networks as they grow in complexity. Even in cases where exact algorithms are manageable, it is still important to consider approximation methods.
A graphical model specifies a complete joint probability distribution over all the nodes. Given the joint probability distribution, all possible inferences queries can be determined by their marginal distributions. Computing several marginal distributions at the same time can be done using dynamic programming which avoids the redundant computation required by other marginalization procedures. However, much like the junction tree algorithm, this grows exponentially pending the size of its largest subset of nodes.

Approximation algorithms provide an alternate solution. Jordan (1999) provides an overview of some of the valuable and existing approximation methods. Variational methods and sampling methods are two approximation approaches readily used.

Variational methods generally provide bounds on probabilities of interests (Jordan, 1999). The simplest example is the mean-field approximation which exploits the law of large numbers to approximate large sums of random variables by their means (Murphy, 2003). This procedure requires the decoupling of all the nodes and updating of the variation parameters, formulating a lower bound on the likelihood. The objective of such methods is to convert a complex problem into a simpler one.

Since the basic idea (stemming from above) is to simplify the joint probability distribution by transforming the local probability functions, an appropriate choice can lead to a simplified inference problem. Transforming some of the nodes may allow for some of the original graphical structure to remain intact or introduce a new graphical structure to which exact methods may be applied. The sequential approach makes use of variational transformations and aims to transform the network until the resulting structure is amenable to exact methods (Jordan, 1999; French and Insua, 2000). Other similar approaches exist and have shown to be fruitful in accomplishing the inference task. This has been touched on briefly through some of the examples in this section.

Another approach to the design of approximation algorithms involves sampling (Monte Carlo) methods. Markov chain Monte Carlo (MCMC) methods are a more efficient approach and many of these algorithms have been applied to the inference problem in graphical models. Their simplicity in implementation and theoretical overtures of convergence make them advantageous; however, they can be slow to converge and it can be hard to diagnose their convergence (Jordan, 1999).

The general inference problem is the computation of the conditional distribution over the hidden values given the observed information. There is a lot of literature in this area and especially on algorithms to accomplish this task. Some of these have been briefly discussed in this section. These various approaches to inference are not mutually exclusive, and in many instances, an amalgamation of such methods may prove to be the best solution for a given graphical model. Some of these more common methods are described in more detail in the PART V to illustrate analysis procedures.

Part V: ANALYTIC & COMPUTATIONAL METHODS FOR DECISION PROBLEMS

The complexity of decision problems has led to the rapid development of methodological solutions to these problems. The graphical structure of probabilistic models has also added a more intuitive approach to the problem at hand. These developments have also led to the advancement of computational methods to deal with the complexities of the numerical assessment of these problems which arises. The task at hand is probabilistic inference, whether it lies in graphical models or in expected losses. Recall that French and Insua (2000) commented on the feasibility and robustness which relate to the practicality and implications of decision processes. Thus, the need for computational methods which meet the criteria is essential to decision modeling.

Earlier parts of this report introduced the framework for modeling decision problems in introduced the decision theoretic aspects. PART IV of this report focused on building models to illustrate decision problems with an emphasis on probabilistic graphical models. This part of the report concludes this paper by discussing some of the computational issues which arise in decision analysis. Computational methods including Monte Carlo methods and dynamic programming are introduced as both feasible and efficient approaches for the analytic assessment of decision problems. These analytic concepts are the focus of the beginning sections. The last section is an application of the typical methodology used in a prototypical example. This concludes the methodological issues which arise in decision problems.

Section 12: Decision Theoretic Computations

Part I and Part IV introduced the framework for decision models and the construction of graphical models, respectively. Both approaches introduced the computational issues which arise. Much of these issues reside in the calculation and or minimization of probability distributions. Some of the methodological approaches were introduced; however were never developed. This section re-introduces the problem and begins to address algorithms which aim to solve some of these computational constraints. A review of analytic approaches and numerical integration are presented but demonstrate some limitations which are resolved though Monte Carlo methods. Although this section concentrates on Bayesian decision theoretic computations, other statistical decision theoretic approaches face the same computational problem. Therefore, they provide insight into solving the computational issues which arise through the classical approach.

Section 12.1: Computational Issues in Decision Problems

The first half of this report focused on the construction and development of decision models. Recall that the loss function $l(\theta,a)$ is one of the key attributes of decision theory. Moreover, the process of calculating and minimizing expected losses is the key objective. This section addresses the computation issues which arise in the modeling of decision problems and specifically these objectives.

French and Insua (2000) state that there are two objectives which need to be solved:

$$\min_{a\in A} E_\theta[l(a,\theta)\mid x] = \min_{a\in A}\int l(a,\theta)\,p_\theta(\theta\mid x)d\theta$$

and

$$\min_{a\in A}\int l(a,\theta)\,p_\theta(\theta\mid x,a)d\theta\,.$$

The former is simply the calculation of an expectation. In this latter case, the probability model involves dependencies on a which occurs in influence diagrams and decision trees formulations. Assuming the optimal action is known means that the explicit computation of the posterior probability is only required. Thus the solution of these decision problems requires the computation of a posterior expectation.

In most instances, the integrals involved are difficult to evaluate analytically; and hence more powerful integration methods need to be implemented. This section reviews analytical and numerical approaches before introducing simulation or Monte Carlo methods. The limitation of the analytic and numeric approaches draws the importance for more computationally efficient and plausible methods. However, these methods are not without their drawbacks and are addressed. The latter part of this section attends to combining the issues of optimization and integration. Although, there is a focus on the

Bayesian decision theoretic computations, these issues are synonymous to other statistical decision theoretic issues; and hence provide insightful possibilities.

Section 12.2: The Bayesian Framework

In addressing the computational issues within a Bayesian framework, it deems necessary to describe the Bayesian perspective in order to fully understand the issues which arise is such methodology. Thus, we begin with an overview of Bayesian data analysis.

The essential characteristic of Bayesian methods is their explicit use of probability models for quantifying uncertainty in inferences based on statistical data analysis. It can be conceptualized by dividing it into the following three steps: (1) setting up a full probability model; (2) conditioning on the observed data; and (3) evaluating the fit of the model and the implications of the resulting posterior distribution (Gelman et al. 1995)

Setting up a full probability model entails the construction of a joint probability distribution for all observable and unobservable quantities in a problem. The model should be consistent with knowledge about the underlying scientific problem. The second step involves calculating and interpreting the appropriate posterior distribution – the conditional probability distribution of the unobserved quantities of ultimate interest, given the observed data. Lastly, questions addressing "does the model fit the data, are the substantive conclusions reasonable, and how sensitive are the results to the modeling assumptions in the first step" (Gelman et al. 1995) should be tended to.

Contextually, it is often easier (with the Bayesian approach) to build highly complex models when such complexity is realistic. The central feature to Bayesian inference is the direct quantification of uncertainty which in principle allows models with multiple parameters and complex layers to be fit (Gelman et al. 1995) with a conceptually simple method. The purpose of this report is not to glorify Bayesian inference over the frequentist approach as there is no need to choose exclusively between the two perspectives nor is it without its constraints. However, it is the purpose of this section to introduce and outline the general concept of the Bayesian paradigm. Before launching into a discussion on computational advances in Bayesian methods, a review of some terminology and standard notation and a formal presentation of the problem are provided.

The task in decision problems has been to make probabilistic inferences. Statistical inference refers to the process of drawing conclusions about a population on the basis of measurement or observations made on a sample of individuals from the population (Everitt, 2002). From a Bayesian perspective, there is no fundamental distinction between observables and parameters of a statistical model as they are all considered random quantities. Thus, here in lies the task – the quantification of such uncertainty.

Recall that Bayesian data analysis involves (first) setting up the full (joint) probability model. Let D denote the observed data, and θ denote model parameters and missing data. The joint distribution is comprised of two parts: a likelihood $P(D \mid \theta)$ and a prior

distribution $P(\theta)$. To summarize, obtaining the likelihood $P(D \mid \theta)$ describes the process which gives rise to the data D in terms of the unknown parameters θ while the prior distribution $P(\theta)$ expresses what is known about θ, prior to observing the data. Specifying $P(D \mid \theta)$ and $P(\theta)$ gives a full probability model, in which

$$P(D, \theta) = P(D \mid \theta)P(\theta).$$

The second step in Bayesian data analysis commemorates the Bayes Theorem is used to determine the distribution of θ conditional on D, after observing the data:

$$P(\theta \mid D) = \frac{P(\theta)P(D \mid \theta)}{\int P(\theta)P(D \mid \theta)d(\theta)}.$$

Applying the Bayes Theorem derives the posterior distribution of θ and is the object of all Bayesian inference (Gilks et al. 1996).

Moments, quantiles and other features of the posterior distributions can be expressed in terms of posterior expectations of functions of θ. The posterior expectation of a function $f(\theta)$ is

$$E[f(\theta) \mid D] = \frac{\int f(\theta)P(\theta)P(D \mid \theta)d\theta}{\int P(\theta)P(D \mid \theta)d\theta}.$$

The integration of this expression has until recently been the source of most of the practical difficulties in Bayesian inference. The problem of calculating expectations in high-dimensional distributions adheres both to Bayesian inference and some areas of frequentist inference (Geyer, 1995). To avoid an unnecessarily Bayesian "flavor" to this discussion, the problem is restated in more general terms.

Let X be a vector of k random variables, with distribution $\pi(\cdot)$. In Bayesian applications, X will comprise of both model parameters and missing data; in frequentist applications, it may comprise of data or random effects (Gilks et al. 1996) and Draper (1997). For Bayesians, $\pi(\cdot)$ will be a posterior distribution (as described above) while for frequentists it will refer to a likelihood. However, either way, the task is to evaluate the expectation

$$E[f(\theta)] = \frac{\int f(x)\pi(x)dx}{\int \pi(x)dx}$$

for some function of interest $f(\cdot)$.

A common situation arises in such that the normalizing constant (the denominator of the above equation) is unknown. In Bayesian practice, $P(\theta \mid D) \propto P(\theta)P(D \mid \theta)$; however the

normalizing constant $\int P(\theta)P(D \mid \theta)d(\theta)$ is not easily evaluated and is the focus of the next section. Here, it has been assumed that X takes values in k-dimensional space, i.e. it is comprised of k vectors of continuous random variables. Although, this problem has been described generally, X could consist of discrete random variables (in which case the integral would be replaced by a summation) or by a mixture of both continuous and discrete random variables.

Section 12.3: Integration Methods

As discussed in the previous section, the difficulty and the interest in Bayesian inference is in evaluating expectations $\int f(x)\pi(x)dx$ for various functions $f(x)$. Integration is usually simple in conjugate models – priors and posteriors are members of the same parametric family of distributions and thus posteriors are easily determined (French and Insua, 2002). Even for conjugate models, there may be cases (moments or probabilities) for which the solution may not be computed explicitly.

In the last decade alone, there have been a number of advances in numerical (Tierney and Kadane, 1986) and analytic (Smith et al. 1987) approximations for such calculations, largely in part due to increased computing power. Some of these alternatives such as numerical evaluation are difficult and inaccurate in $k > 20$ dimensions; analytic approximations such as the Laplace approximation which is sometimes appropriate (Gilks et al. 1996). Numerical quadrature methods for approximating integrals such as the common trapezoid and Simpson's rules (CSC260H, 2001) are commonly applied. Better rules are obtained when the form of the integrand is taken into account such as Gaussian quadrature methods which approximates well when the density function is a normal kernel (French and Insua, 2002). Such procedures also extend into k dimensions. Asymptotic methods based on asymptotic normality arguments also exist. However, integration strategies based on these arguments are only valid if samples sizes are large and other necessary conditions apply. These methods are efficient in problems with specific structure and/or low dimensionality.

In more complex problems, simulation or Monte Carlo integration is required. Monte Carlo methods are often the method of choice involving high-dimensional problems, but it is important to note that many of these computational methods complement each other rather than compete (Tierney and Mira, 1999).

Monte Carlo Integration

The basic objective of the Monte Carlo approach is to draw an approximate i.i.d. sample from $\pi(\cdot)$ which can then be used to compute sample averages as approximations to population averages (STA4276H, 2003). Although this is an oversimplification of the approach it introduces the concept concisely. More formally, Monte Carlo integration evaluates $E[f(X)]$ by drawing samples $\{X_t, t = 1,...,n\}$ from $\pi(\cdot)$ and then approximating

$$E[f(X)] \approx \frac{1}{n}\sum_{t=1}^{n} f(X_t).$$

So, the population mean of $f(X)$ is estimated by a sample mean; the average value has some basic properties which are very important. When the samples $\{X_t\}$ are independent, the Law of Large Numbers ensures that the approximation can be made as accurate as desired by increasing the sample size *n* (Gilks et al. 1996). Thus, *n* is under the control of the analyst; it is not the size of a fixed data sample.

The problem at hand is to simulate observations from a posterior distribution, obtained via Bayes' Rule as

$$P(\theta \mid X) = \frac{P(\theta)P(X \mid \theta)}{\int P(\theta)P(X \mid \theta)d(\theta)}.$$

In general, drawing samples $\{X_t\}$ from $\pi(\cdot) = P(\theta \mid X)$ is not feasible, since $\pi(\cdot)$ can be quite complex and non-standard. However the samples $\{X_t\}$ need not be independent; it can be generated by any process which (loosely) draws samples throughout the support of $\pi(\cdot)$ in the correct proportions (Gilks et al. 1996). One way this is accomplished is through a Markov chain having $\pi(\cdot)$ as its stationary distribution.

Let $X_1, X_2, ..., X_t, ...$ be a sequence of random variables which have outcomes from some set *S*. Also, assume that the conditional distribution of X_t given all previous observations $X_{t-1}, X_{t-2}, ...$ only depends on the last observation. That is, if the current value is known then the prediction of the next value does not require the previous values (CHL5223H, 2003). This property is known as the Markovian property and such a sequence of random variables is known as a Markov chain. More formally this is stated as suppose a sequence of random variables $\{X_t, t = 1, ..., n\}$ are generated such that at each time $t \geq 1$, the next state X_{t+1} is sampled from a distribution $P(X_{t+1} \mid X_t)$ which depends only on the current state of the chain, X_t. Thus, to summarize, given X_t, the next state X_{t+1} does not depend further on the history of the chain $\{X_1, X_2, ..., X_{t-1}\}$.

An important issue in basic Markov chain properties is the affect of the starting value of the chain or state X_1 affect X_t. The trait of the distribution of $X_t \mid X_1$, denoted $P^{(t)}(X_t \mid X_1)$, without the intervening variables $\{X_2, X_3, ..., X_{t-1}\}$ suggests that X_t depends directly on X_1. Subject to some conditions, the chain will gradually 'forget' its initial state and $P^{(t)}(\cdot \mid X_1)$ will eventually converge to a unique limiting distribution (Gilks et al, 1996). That is, X_t settles down to a limiting distribution no matter what the value of the starting value of X_1 is.

There are some basic properties of Markov chains which should be stated for completeness. The first is that if the set of states, *S*, is finite then the Markov chain holds two properties: aperiodicity and irreducibility; if the set of states is infinite then an additional property of an invariant distribution exists (CHL5223H, 2003). Thus, for a finite Markov chain, the limiting distribution exists of it is aperiodic (the chain is not cyclic) and irreducible (possible to move from one state to another) while infinite Markov chains hold the additional property of having an existing invariant or stationary distribution. The stationarity of a distribution simply refers to the limit as $n \to \infty$ is $\pi(\cdot)$. Thus, as discussed earlier, as t increases, the sample points $\{X_t\}$ look increasingly like dependent samples from the stationary distribution denoted as $\phi(\cdot)$ based on a long burn-in of m iterations, where $t = m+1,...,n$. The output from the Markov chain is used to estimate the expectation $E[f(X)]$, where X has distribution $\phi(\cdot)$. The important by-product is the "path averages" to the expected value under the limiting distribution where burn-in samples are usually discarded for the calculation of the estimator:

$$\bar{f} = \frac{1}{n-m}\sum_{t=m+1}^{n} f(X_t) \to E_{\infty}[f(X)].$$

The convergence to the expectation is confirmed by the ergodic theorem whereby the above estimator is referred to as the ergodic average (Gilks et al. 1996). Therefore these path averages can be used to estimates expected values of the limiting distribution.

(See Part III for a discussion on Markov chains and their properties.)

Section 12.4: Markov Chain Monte Carlo Methods

Recall that is was stated that the basic objective of the Monte Carlo approach was to draw an approximate i.i.d. sample from $\pi(\cdot)$ which could then be used to compute sample averages as approximations to population averages. Since direct i.i.d. sampling is rarely possible, a sample from a similar distribution is obtained and a Markov chain as its unique invariant distribution is constructed (Tierney and Mira, 1999). These are the concepts which formulate the Markov chain Monte Carlo (MCMC) method.

As explained the Monte Carlo techniques generate random variables having certain discrete distributions. However, once it becomes too complex, Markov chains can be used. If a Markov chain with a stationary distribution is the same as the desired probability distribution $\pi(\cdot)$ (the target distribution), then the Markov chain can be run for a long time, say m iterations. The probability that the chain is in state i will be approximately the same as the probability that the discrete random variable equals i. The idea behind the method is simple. Simply, a Markov chain with invariant distribution $\pi(\cdot)$ is run and sample path averages are used to approximate expectations under $\pi(\cdot)$. If the chain is irreducible and the expectations exist, then the sample path averages converge to the expectations. The problem of finding the Markov chain with the desired stationary distribution still exists.

The ergodic average shows how a Markov chain can be used to estimate $E[f(X)]$, where the expectation is taken over its stationary distribution $\phi(\cdot)$. This would seem to provide the solution to the initial problem; however, first the construction of a Markov chain such that its stationary distribution $\phi(\cdot)$ is precisely the distribution of interest $\pi(\cdot)$ needs to be established. There are two widely established ideas which are presented in the next two subsections – dimension reduction by conditioning (Gibbs sampling approach) and acceptance and rejection (Metropolis-Hastings algorithm).

The practical appeal of simulation methods, including MCMC, is that, given a set of random draws from the posterior distribution, one can estimate all summary statistics from the posterior distribution directly from the simulations. MCMC methods have been widely successful because they allow one to draw simulations from a wide range of distributions, including many that arise in statistical work, for which simulation methods were previously much more difficult to implement. To apply the MCMC method in a particular problem a sampler has to be constructed. There are two central ideas for building samplers and are guided by the two most common methods – the Gibbs sampler and the Metropolis-Hastings algorithm which are described in sections **§ 4.2** and **§ 4.3**. These algorithms are not in competition with each other and in fact complement each other. The Metropolis within Gibbs algorithm is identified as a conceptual blend in which one parameter is updated at a time; this was the initial proposed algorithm which can be automated (Gilks et al. 1996).

Gibbs Sampling

The Gibbs sampling algorithm holds the central idea of "dimension reduction by conditioning" (Tierney and Mira, 1999) and is best described mathematically by its transition kernel (Gilks et al. 1996). Suppose $\pi(\cdot)$ is the joint distribution of $\{X_t, t = 1,...,n\}$ where $t \geq 1$, a Markov chain. Update each component based on the value of each other component

$$\pi(X_t \mid X_{t-1}) = \pi(X_t \mid X_1,...,X_{t-1},X_{t+1},...,X_n)$$

These distributions are called full conditionals. The transition kernel describes the density of going from one point *X* to another point *Y* (Gilks et al. 1996, Brooks et al. 1998):

$$K(X,Y) = \prod_{i=1}^{d} \pi(Y_{\cdot i} \mid \{Y_{\cdot j}, j < i\},\{X_{\cdot j}, j < i\})$$

This is a product of the conditional densities of the individual steps required to produce an iteration of a *d*-dimensional Gibbs sampler. A thorough theoretical discussion is omitted for brevity sake.

The Gibbs sampling algorithm is quite simpler in concept and thus, a simple example is featured below. Let $\boldsymbol{\theta}$ be 3-dimensional and the parameters of a Bayesian model. The algorithm follows:

1. Start with a set of initial values: $\boldsymbol{\theta}^{(0)}$.

2. Given the *m-th* update, get the $(m+1)th$ sample by:

$$\theta_1^{(m+1)} \sim f(\theta_1 \mid \theta_2 = \theta_2^{(m)}, \theta_3 = \theta_3^{(m)}, \text{data})$$
$$\theta_2^{(m+1)} \sim f(\theta_2 \mid \theta_1 = \theta_1^{(m)}, \theta_3 = \theta_3^{(m)}, \text{data})$$
$$\theta_3^{(m+1)} \sim f(\theta_3 \mid \theta_1 = \theta_1^{(m)}, \theta_2 = \theta_2^{(m)}, \text{data})$$

3. The joint posterior is invariant, therefore:

$$\bar{f} = \frac{1}{n-m} \sum_{t=m+1}^{n} f(\boldsymbol{\theta}^{(m)}) \rightarrow E_{\infty}[f(\boldsymbol{\theta})].$$

It should be noted that when this sampling was developed by Geman and Geman (1984) its purpose was to sample from a Gibbs distribution, which is a distribution over a lattice; however since its inception it has expanded over many distributions and thus, being called a Gibbs sampling algorithm is a bit of a misnomer. The Gibbs sampler is a special case of the Metropolis-Hastings algorithm which is explained next.

Metropolis-Hastings Algorithm

The notion that if the chain is irreducible and the expectations exist, then the sample path averages converge to the expectations is the basic idea proposed by Metropolis et al. (1953) and was extended by Hastings (1970). The central idea to the Metropolis-Hastings algorithm is acceptance (or proposals) and rejection (Tierney and Mira, 1999). This algorithm proposes a new point on the Markov chain which is either accepted or rejected. If the point is accepted, the Markov chain moves to a new point. If it is rejected, the Markov chain remains in the same state. By choosing the acceptance probability correctly, a Markov chain which has $\pi(\cdot)$ as a stationary distribution is created.

Let S be a state space with probability distribution $\pi(\cdot)$ on S. Then choose a proposal distribution $\{q_{ij} : i, j, \in S\}$ with $q_{ij} \geq 0$ and $\sum_{j \in S} q_{ij} = 1$ for each $i \in S$. Given that $X_n = i, X_{n+1}$ is computed as the M-H algorithm follows:

1. Choose $Y_{n+1} = j$ according to the Markov chain $\{q_{ij} : i, j, \in S\}$
2. Set $\alpha_{ij} = \min\left\{\frac{\pi(j)q_{ij}}{\pi(i)q_{ij}}\right\}$

3. With probability $\begin{cases} \alpha_{ij}, \text{ let } X_{n+1} = Y_{n+1} = j \text{ (accept proposal)} \\ \text{Otherwise, let } X_{n+1} = X_n = i \text{ (reject proposal)} \end{cases}$

As desired, this will create a chain with stationary distribution $\pi(\cdot)$. The Metropolis-Hastings algorithm can also be used for continuous random variable by using densities and continuous proposal distributions, but is not explored further here. It is also possible to extend the concept to a multidimensional case. The "Gibbs sampler", is a special case of the multivariate Metropolis-Hastings algorithm where the proposals *q* are the full conditionals and the acceptance probability is always 1 as explained above.

Convergence Issues

A properly derived and implemented MCMC method will produce draws from the joint posterior density once it has converged to stationarity $\phi(\cdot)$. As a result, a primary concern in applying such methodology is determining that it has in fact essentially converged; that is, that after an initial burn-in period (designed to remove dependence of the simulated chain on its starting location), all further samples may be safely though of as coming from the stationary distribution. This is complicated by two factors. First, since what is produced at convergence is not a single value but a random sample of values, somehow the natural variability in the converged chain and the typically greater variability in the pre-convergence samples must be distinguishable. Second, since the sampled results come from a Markov chain, they will typically be serially correlated.

Although MCMC methods have effectively revolutionized the field of Bayesian statistics over the past decade, any inference based upon MCMC output relies critically upon the assumption that the Markov chain being simulated has achieved a steady state or "converged". Many techniques have been developed for trying to determine whether or not a particular Markov chain has converged, and the paper by Brooks and Roberts (1998) reviews some of these methods with an emphasis on the mathematics underpinning these techniques in an "attempt to summarize the current 'state-of-play' for convergence assessment". These are not elaborated here for brevity purposes.

A round table discussion held at the Joint Statistical meetings in August of 1996 with a panel of experts in the field published parts of the discussion in which they discuss more practical applications and procedures they follow. As various diagnostics are available, these experts mentioned that use of autocorrelation plots and trace plots are amongst the most useful and practical. The former is simply a plot demonstrating the path of the parameter; while the latter is simply the internal correlations. These diagnostics are often used to estimate the degree of mixing in a simulation which is the extent to which a simulated chain traverses the entire parameter space (Kass et al. 1998).

Section 12.5: Integrating Integration & Optimization Methods

To summarize, some of the aspects addressed in the introductory section on Bayesian framework (**§ 12.2**) such as awkward posterior distributions and further distributional

complexity introduced by model parameters would (otherwise) require subtle and sophisticated or analytic approximation; however dimensionality problems typically put out of reach the implementation of other sophisticated approximation techniques as those introduced in **§ 12.3**. Markov chain Monte Carlo methods put to ease much of these constraints which have otherwise limited Bayesian data analysis with the included aspects of providing general functions of models parameters and alleviating much awkward predictive inference. Thus, the problem of computing expectations or making probabilistic inferences – in this instance is remedied.

In problems with optimization requirements such as general loss functions, such computational effort is not sufficient. That is, combining problems of optimization and integration are required for decision theoretic computations. For general loss functions, the objective is to determine the alternative *a** which minimizes the expected loss. There are two operations: (1) compute the expected loss; and (2) minimize the expected loss.

French and Insua (2000) review several optimization methods which are useful within statistical decision theory. It is also important to note that some of the integration or sampling methods previously described also have an optimization step to compute maximum likelihood estimators, for instance. Thus, it only seems plausible to have Monte Carlo optimization methods to solve such problems. Adaptive methods which try to accomplish such a task have been developed; however, they are not without their limitations.

An alternative approach was introduced by Shao (1989) in the statistical decision theory literature; and is referred to as sample path optimization. The strategy follows:

1. Select a sample $\theta_1,...,\theta_m \sim p_\theta(\theta \mid x)$.

2. Solve the optimization problem

$$\min_{a \in A} \frac{1}{m} \sum_{i=1}^{m} l(a, \theta_i)$$

yielding $a_m(\theta)$.

By the law of large numbers, it holds that $a_m(\theta)$ converges to a minimum expected loss alternative *a**. This approach can be extended to any form of the loss function and thus presents some alternatives to the computational problem.

This section has introduced some of the main computational tools used in statistical decision theory. These problems require both powerful integration and optimization methods as well as algorithms that achieve both simultaneously.

Section 13: Analytic Approaches to Solving Decision Problems

Part I, Part II and Part IV introduced the framework for decision models and the construction of graphical models, respectively. The previous section emphasized some of the computational issues which arise in decision problems. However, this was removed from within the sequential framework described in Part III and which is eminent in decision tree structures. This section develops upon the concepts introduced in earlier sections and relates them to more analytic approaches to solving the problems. The emphasis is placed on dynamic programming and progresses into its important as a computational tool for decision problems. Sensitivity analysis is addressed last yet is important as means of checking the sensitivity of the output, and the implications of the findings and possible inconsistencies.

Section 13.1: Dynamic Programming

The computational methods presented in section 12 neglected decision problems which involved a sequential approach or which had been transformed such that they could be assessed sequentially. Sequential problems lend themselves to another extension of optimization methods. Dynamic programming is a mathematical technique which allows for a sequence of interrelated decisions to be made. And further, it provides a systematic procedure for determining optimal combinations of decisions.

Hillier and Lierberman (2001) present the basic features that characterize dynamic programming problems. These include:

1. The problem can be divided into stages, with a decision required at each stage.
2. Each stage has a number of states associated with the beginning of that stage.
3. The consequence of the decision at each stage is to transform the current state to a state associated with the beginning of the next stage.
4. The solution procedure is designed to find an optimal choice for the overall problem.
5. Bellman's principle (Bather, 2000): Given the current state, an optimal policy for the remaining stages is independent of the decisions adopted in previous stages.
6. The solution begins by finding an optimal choice for the last stage.

Dynamic programming is defined by a recursive relationship. The general notation used to describe a dynamic programming problem is summarized below:

N = number of stages
n = index for current stage
θ_n = current state for stage n
x_n = decision variable for stage n
x_n^* = optimal value of $x_n \mid \theta_n$
$f_n(\theta_n, x_n)$ = contribution of stages $n, n+1, ..., N$ to objective function

The recursive relationship will lead to two possible forms, depending on the problem:

$$f_n^*(\theta_n) = \max_{x_n}\{f_n(\theta_n, x_n)\}$$

or

$$f_n^*(\theta_n) = \min_{x_n}\{f_n(\theta_n, x_n)\}$$

This recursive procedure induces a backward form of mathematical induction in order to achieve the required task. This approach (as dynamic programming) was first outlined by Bellman, and was coined such to describe the techniques which he had brought together to study a class of optimization problems involving sequential decisions (Bather, 2000).

Sequential problems and subsequently dynamic programming problems can be classified into two groups: deterministic and stochastic. The former approach suggests that the state at the next stage is completely determined by the state and decision at the current stage. The latter approach (which is probabilistic) implies that a probability distribution will dictate what the next state will be. This probabilistic approach is more conducive to our needs and is associated with solving decision tree diagrams.

Consider the diagram above illustrating the general structure of a stochastic dynamic programming problem. Let *S* denotes the number of possible states at stage *n* + 1. The system goes to state *i* with probability p_i for $i = 1,2,...,S$, given state θ_n and decision x_n at stage *n*. At state *i*, the cumulative information of stage *n* to the objective function is represented by C_i. This figure can be extended in such a manner for all possible stages; and diagrammatically has the form of a decision tree.

With the inclusion of a probability distribution dictating the consequence of the next state, means that the precise form of the objective function will differ slightly from the one given above. It can now be written as:

$$f_n(\theta_n, x_n) = \sum_{i=1}^{S} p_i[C_i + f_{n+1}^*(i)]$$

with

$$f_{n+1}^*(i) = \min_{x_{n+1}} f_{n+1}(i, x_{n+1}),$$

where the minimization is taken over all feasible values of x_{n+1}.

In statistical decision theory[1], it might be said that the transition from one stage to another is controlled by a sequence of actions, given the state θ_n at stage *n*, the choice of action a_n determines the probability distribution of the next state. Essentially because, we are

[1] Bather provides a concise introduction to both deterministic and probabilistic dynamic programming and its application to utility theory.

dealing with a Markov system, only the current state is important. The state and action variables may be discrete or continuous and the range of possible choices of an action may vary state to state. The essential characteristic of such modeling (in summary) is that given θ_n at stage *n*, the choice of action or decision x_n determines both the probability distribution of the next state θ_{n+1} and the expected cost[2] of the transition $\theta_n \to \theta_{n+1}$.

Section 13.2: Bellman's Principle and the Bayesian Perspective

Recall that one of the fundamental elements of decision problems is the loss function and in dynamic programming it is this objective function we tend to want to minimize. Now consider an additive function which by definition is monotonic and separable. (See Part II on Utility Theory.) That is if the first few terms are fixed then minimizing its whole sum is equivalent to minimizing its individual terms - a property which holds when taking expectations. This expounds the aforementioned Bellman Principle of Optimality:

> The optimal sequential decision policy for the problem which begins with $P_\theta(\cdot)$ as the decision maker's prior for θ and has *R* stages to run must have the property that is, at any stage $n < N$, the observations $X_1 = x_1, ..., X_n = x_n$ have been made, then the continuation of the optimal policy must be the optimal sequential policy for the problem beginning with $P_\theta(\cdot \mid x_1, ..., x_n)$ as the prior and having $N - n$ stages to run. (French and Insua, 2000)

French and Insua make it a point to emphasize the importance of the decision maker's prior in Bellman's principle. Since the prior describes the state of knowledge, it also describes the state of the decision making process.

For simplicity of notation, let $\pi = P_\theta(\cdot)$ and $r_n(\pi)$ be the Bayes risk of the optimal policy with at most *n* stages left to run with knowledge π. Then the situation in which the decision maker has no option to make an observation and must choose an action can be expressed as:

$$r_0(\pi) = \min_{a \in A} E_\theta[l(a, \theta)]$$

where the expectation over θ is taken with respect to π. Next consider that the decision maker is given the chance to make an observation. The decision maker has two options:

1. Take an action $a \in A$ without making an observation at an expected loss of $r_0(\pi)$.
2. Make a single observation *X* at cost γ and choose an action $a \in A$ in light of her current knowledge $\pi(X)$ at an expected loss of $E_X[r_0(\pi(X))]$.

Bellman's Principle assumes that after the observation(s), the optimal Bayes action is taken (French and Insua, 2000). Accordingly, this can be extended to *n* observations

[2] This refers to the expected loss function.

which then give her option (1) stated above or an extension of option (2) which would result in observations made for the remaining n -1 stages with an expected loss of

$$E_X[r_{n-1}(\pi(X))].$$

Thus, for $n = 1,2,...,N$

$$r_n(\pi) = \min\{r_0(\pi), \gamma + E_X[r_{n-1}(\pi(X))]\}.$$

Note the form of this recursion formula. This defines the dynamic programming or backward induction form introduced in the previous subsection. Thus, dynamic programming allows for both the calculation of $r_n(\pi)$ and the characterization of the optimal policy. This following section now extends the basic principles demonstrated in these earlier subsections to illustrate the analytic approach to solving decision problems as applied to utility theory – or more precisely in the maximization of utilities.

Section 13.3: Analysis in Extensive Form

There are two basic modes of analysis which can be utilized to determine which course of action will maximize the decision maker's utility: the extensive form of analysis and the normal form. The two forms are mathematically equivalent and lead to identical results; each has something to contribute to the insight into the decision problem and have separate technical advantages in certain situations. This section focuses on the extensive form which equates the use of decision trees more readily.

Backwards Induction

Consider a simple decision tree whereby an initial decision results in a chance and a subsequent action results in a final state. The extensive form of analysis proceeds by working backwards from the end of the decision tree (the right side of the tree) to the initial starting point: instead of starting by asking which experiment *e* the decision maker should choose, the procedure starts by asking which terminal act he should choose if he had already performed a particular experiment *e* and observed a particular outcome *x*. Even at this point, with a known history (*e*, *x*), the utilities of the various possible terminal acts are uncertain because θ which will be chosen by chance at the end is still unknown; but this difficulty is easily resolved by treating the utility of any *a* for given (*e*, *x*) as a random variable $u(e, x, a, \tilde{\theta})$[3] and applying the operator $E''_{\theta|x}$ which takes the expected value of $u(e, x, a, \tilde{\theta})$ with respect to the conditional measure $P''_{\theta|x}$. Symbolically, we can compute for any given history (e, x) and any terminal act *a*

$$u^*(e,x,a) \equiv E''_{\theta|x} u(e,x,a,\tilde{\theta});$$

[3] Recall notation introduced in Section 1 for clarification. Notation for unknown quantity is described as a random variable in this case.

this is the utility of being at the juncture (e, x, a) looking forward, before chance has made a choice of θ (Raiffa and Schlaifer, 2000).

Since the decision maker's objective is to maximize his expected utility, he will, if faced with a given history (e, x), wish to choose the a (or one of the as if more than one exist) for which $u^*(e, x, a)$ is greatest; and the utility of being at the terminal with history (e, x) and the choice of a still to make is

$$u^*(e, x) \equiv \max_a u^*(e, x, a).$$

After $u^*(e, x)$ has been computed in this way for all possible histories (e, x), the problem of the initial choice of an experiment can be addressed. At this point, the initial move, the utilities of the various possible experiments are uncertain only because x which will be chosen by chance is still unknown, and this difficulty is resolved in exactly the same way that the difficulty in choosing a given (e, x) was resolved: by putting a probability measure over the chance's moves and taking expected values. In other words, $u^*(e, \tilde{x})$ is a random variable at the initial decision point because $\tilde{x}$ is a random variable, and therefore it can be defined for any e

$$u^*(e) \equiv E_{x|e} u^*(e, \tilde{x})$$

where $E_{x|e}$ expects with respect to the marginal measure $P_{x|e}$.

Again, the decision maker will wish to choose the e for which $u^*(e)$ is greatest; and therefore may suggest that the utility of being at the initial decision node with the choice of e still to make is

$$u^* \equiv \max_e u^*(e) = \max_e E_{x|e} \max_a E''_{\theta|x} u(e, \tilde{x}, a, \tilde{\theta}).$$

This procedure of working back from the outermost branches of the decision tree to the base of the trunk is often called "backward induction". More descriptively it could be called a process of "averaging out and folding back".

Section 13.4: Analysis in Normal Form

The final product of the extensive form of analysis presented in the previous section can be thought of as a description of the optimal strategy consisting of two parts:

1. A prescription of the experiment e which should be performed,
2. A decision rule prescribing the optimal terminal act a for every possible outcome x of the chosen e.

The whole decision rule for the optimal *e* can be simply determined from the part of the analysis which determined the optimal *a* for every *x* in *X*; and incidentally these same results also enable the optimal decision rule to accompany any other *e* in *E*, even though the *e* in question is not itself optimal.

The normal form of analysis, also has as its end product the description of an optimal strategy, and arrives at the same optimal strategy as the extensive form but via a different route. Instead of first determining the optimal action *a* for every possible outcome *x*, and thus implicitly defining the optimal decision rule for any *e*, the normal form of analysis starts by explicitly considering every possible decision rule for a given *e* and then choosing the optimal rule for that *e*. This can be done for $\forall e \in E$ such that an optimal *e* can be found as in the extensive form of analysis.

Recall the decision rules as presented in **Sections 2** and **3**. Mathematically, a decision rule δ for a given experiment *e* is a mapping which carries *x* in *X* into $\delta(x)$ in *A*. Given a particular strategy (e,δ) and a particular pair of values (x,θ), the decision maker's act as prescribed by the rule will be $a = \delta(x)$ and his utility will be $u(e,x,\delta(x),\theta)$; but before the experiment has been conducted and its outcome observed, $u(e,\tilde{x},\delta(x),\tilde{\theta})$ is a random variable because $\tilde{x}$ and $\tilde{\theta}$ are random variables.

The decision maker's objective is therefore to choose the strategy (e,δ) which maximizes his expected utility

$$u^*(e,\delta) \equiv E_{\theta,x|e}\, u(e,\tilde{x},\delta(\tilde{x}),\tilde{\theta}).$$

This double expectation will actually be accomplished by iterated expectation and the iterated expectation can be carried out in either order: we can first expect over $\tilde{\theta}$ holding $\tilde{x}$ fixed and then over $\tilde{x}$, using the same measures $P_{x|e,\theta}$ and P'_θ.

If *e* and δ are given and $\tilde{\theta}$ is held fixed, then by taking the expectation of $u[e,\tilde{x},d(\tilde{x}),\theta]$ with respect to $P_{x|e,\theta}$ we obtain

$$u^*(e,\delta,\theta) \equiv E_{x|e,\theta}\, u[e,\tilde{x},d(\tilde{x}),\theta],$$

which will be called the conditional utility of (e,δ) for a given state θ. Next taking the expected value over $\tilde{\theta}$ with respect to the unconditional measure P'_θ, we obtain

$$u^*(e,\delta) = E'_\theta u^*(e,\delta,\tilde{\theta}),$$

which can be referred to as the unconditional utility of (e,δ).

For any particular experiment *e*, the decision maker is free to choose the decision rule δ whose expected utility is greatest; and therefore it may be said that the utility of any experiment is

$$u^*(e) \equiv \max_d u^*(e,\delta).$$

After computing the utility of every $e \in E$, the decision maker is free to choose the experiment wit the greatest utility such that

$$u^* \equiv \max_e u^*(e) = \max_e \max_d E'_\theta E_{x|e,\theta} u[e,\tilde{x},\delta(\tilde{x}),\tilde{\theta}].$$

Section 13.5: Equivalence of the Extensive and Normal Form[4]

The extensive and normal forms of analysis will equivalent if and only if they assign the same utility to every potential *e* in *E*, i.e. if the formula

$$u_n{}^*(e) = \max_d E'_\theta E_{x|e,\theta} u[e,\tilde{x},\delta(\tilde{x}),\tilde{\theta}]$$

derived as $u^*(e) \equiv \max_d u^*(e,\delta)$ by the normal form of analysis agrees for all *e* with the formula

$$u_e{}^*(e) = E_{x|e} \max_a E''_{\theta|x} u(e,\tilde{x},a,\tilde{\theta})$$

derived as $u^*(e) \equiv E_{x|e} u^*(e,\tilde{x})$ by the extensive for. The operation $E'_\theta E_{z|e,\theta}$ is equivalent to the expectation over the entire possibility space $\Theta \times X$ and is therefore equivalent to $E_{z|e} E''_{\theta|x}$. It follows that the formal result as presented above can be written as

$$u^*(e) = \max_d E'_{x|e} E''_{\theta|x} u[e,\tilde{x},\delta(\tilde{x}),\tilde{\theta}],$$

and it is then obvious that the best δ will be the one which for every *x* maximizes

$$E''_{\theta|x} u[e,x,\delta(x),\tilde{\theta}].$$

This, however, is exactly the same thing as selecting for every *x* an a_x which satisfies

$$E''_{\theta|x} u(e,x,a_x,\tilde{\theta}) = \max_a E''_{\theta|x} u(e,x,a,\tilde{\theta})$$

[4] See Raiffa and Schlaiffer (2000) for further details on extensive and normal form of analysis. This has been adapted from the author's book on analysis of statistical decision problems.

as stated by the extensive form of analysis. Letting $\delta^*(x)$ denote the optimal decision rule selected by the normal form, it has just been shown that $\delta^*(x) = a_x$ and that the formulas $u_n^*(e)$ and for $u_e^*(e)$ are equivalent.

Thus, in order to choose the best *e* and must therefore evaluate $u^*(e)$ for all *e* in *E*, the extensive and normal forms of analysis require exactly the same inputs of information and yield exactly the same results even though the intermediate steps in the analysis are different. If, however, *e*, is fixed and one wishes merely to choose an optimal terminal action *a*, the extensive forma has the merit that one has only to choose an appropriate action for the particular *x* which actually materializes; there is no need to find the decision rule which selects the best action for every *x* which might have occurred but in fact did not occur.

Section 13.6: Combination of Formal and Informal Analysis[5]

The general model stated in the overview of modeling a decision problem is often criticized on the grounds that utilities cannot be rationally assigned to the various possible (e, x, a, θ) combinations because the costs, profits, or in general the consequences of these combinations would not be certain even if θ were known. In principle, such criticisms represent nothing but an incomplete definition of the state space Θ which can be made rich enough to include all possible pairs of value of "these" unknowns. The decision maker's uncertainties about these values can then be evaluated together with his other uncertainties in the probability measure which is assigned to Θ, and the analysis of the decision problem can the proceed as before.

Consider the state θ expressed as a doublet $(\theta^{(1)}, \theta^{(2)})$ so that the state space is of the form $\Theta = \Theta^{(1)} \times \Theta^{(2)}$. Thus, $\theta^{(1)}$ might be the parameter of a Bernoulli process while $\theta^{(2)}$ might be the cost of the product (for example). In terms of the original decision tree in figure 4, a play was a 4-tuple (e, x, a, θ) and utilities were assigned directly to each (e, x, a, θ). If θ is split into $(\theta^{(1)}, \theta^{(2)})$, a play is 5-tuple $(e, x, a, \theta^{(1)}, \theta^{(2)})$; utilities are assigned to each $(e, x, a, \theta^{(1)}, \theta^{(2)})$ and the utility of any $(e, x, a, \theta^{(1)})$ is the expected value of the random variable $(e, x, a, \theta^{(1)}, \tilde{\theta}^{(2)})$, the expectation being taken with respect to the conditional measure on $\Theta^{(2)}$ given the history $(e, x, a, \theta^{(1)})$. This asserts the idea that it is possible to further consider the split of a tree by only conducting a partial analysis (and partial construction of a tree) based on $\theta^{(1)}$ while holding $\theta^{(2)}$ constant.

Besides cutting the decision tree before it is logically complete, the decision maker may rationally decide not to make a complete formal analysis of even a truncated tree which he has constructed. Thus if *E* consists of two experiments e_1 and e_2, $u^*(e_1)$ may be formally worked out by evaluating

[5] Raiffa and Schlaiffer (2000), French and Insua (2000), and Hillier and Lieberman (2001).

$$u^*(e_1) = E_{x|e_1} \max_a E''_{\theta|x} u(e_1, \tilde{x}, a, \tilde{\theta});$$

but after the formal analysis of e_1 is completed, it may be concluded without any formal analysis at all that e_2 is not as good as e_1 and adopt e_1.

That behavior is perfectly consistent with the principle of choice described in **Sections 2** and **3** and readdressed in **Section 6**. Before making a formal analysis of e_2, the decision maker can think of the unknown quantity $u^*(e_2)$ as a random variable $\tilde{v}$. If the number *v*, resulting from a formal analysis was known, was greater than the known number $u^*(e_1)$, e_2 would be adopted rather e_1 and the value of the information that $\tilde{v} = v$ could be measured by the difference $v - u^*(e_1)$ in the decision maker's utility which results from this change of choice. If on the contrary *v* were less than $u^*(e_1)$, the decision maker would adhere to the original choice of e_1 and the information would have been worthless.

In other words, the random variable can be defined as

$$\max\{0, \tilde{v} - u^*(e_1)\} = \textit{value of information} \text{ regarding } e_2,$$

and before expounding such information at the cost of making a formal analysis of e_2 the decision maker may prefer to compute its expected value by assigning a probability measure to $\tilde{v}$ and then expecting with respect to this measure. If the expected value is less that the cost, the decision maker will quite rationally decide to use e_1 without formally evaluating $v = u^*(e_2)$. Operationally one usually does not formally compute either the value or the cost of information on e_2: these are subjectively assessed. The computations could be formalized, of course, but ultimately direct subjective assessments must be used if the decision maker is to avoid an infinite regress.

Before closing this section and the subject of incomplete analysis, it should be said that completely formal analysis and completely intuitive analysis are not the only possible methods of determining a utility such as $u^*(e_2)$. In many instances it is possible to make a partial analysis in order to gain some insight but at a less prohibitive cost than a full analysis entails.

Section 13.7: Sensitivity Analysis[6]

Sensitivity analysis is an essential element of decision analysis. The principle of sensitivity analysis is also directly applied to areas such as meta-analysis and cost-effectiveness analysis most readily but is not confined to such approaches. This section

[6] See Petitti (2000), Hillier and Lieberman (2001), and Clemen (1996).

simply describes the overall purpose of sensitivity analysis and describes one-way and its expansion to n-way analysis as applied to decision analysis on a very general level.

Sensitivity analysis evaluates the stability of the conclusions of an analysis to assumptions made in the analysis. When a conclusion is shown to be invariant to the assumptions, confidence in the validity of the conclusions of the analysis is enhanced. Such analysis also helps identify the most critical assumptions of the analysis (Petitti, 2000).

An implicit assumption of decision analysis is that the values of the probabilities and of the utility measure are the correct values for these variables. In one-way sensitivity analysis, the assumed values of each variable in the analysis are varied, one at a time, while the values of the other variables in the analysis remain fixed. When the assumed value of a variable affects the conclusion of the analysis, the analysis is said to be "sensitive" to that variable. When the conclusion does not change, when the sensitivity analysis includes the values of the variables that are within a reasonable range, the analysis is said to be "insensitive" to that variable.

If an analysis is sensitive to the assumed value of a variable, the likelihood that the extreme value is the true value can be assessed qualitatively; perhaps, weighting the benefit of one strategy over the other under the extreme assumption.

An extension of the one-way sensitivity analysis is the threshold analysis. In such a case, the value of one variable is varied until the alternative decision strategies are to have equal outcomes, and there is no benefit of one alternative over the other in terms of estimated outcome. The threshold point is also called the break-even point at the decision is a "too-up". That is, neither of the alternative decision options being compared is clearly favored over the other. Threshold analysis is especially useful when the intervention is being considered for use in groups that can be defined a priori based on the values of the variable that is the subject of the threshold analysis.

In two-way sensitivity analysis, the expected outcome is determined for every combination of estimates of two variables, while the values of all other variables in the analysis are held constant at baseline. It is usual to identify the pairs of values that equalize the expected or expected utility of the alternatives and to present the results of the analysis graphically. It is simpler to interpret the results of a two-way sensitivity analysis with the aid of graphs.

In n-way sensitivity analysis, the expected outcome is determined for every possible combination of every reasonable value of every variable. N-way sensitivity analysis is analogous to n-way regression and is seemingly difficult to interpret and is not discussed any further in this report.

It is usual to do one-way sensitivity analysis for each variable in the analysis. The highest and the lowest values within reasonable range of values are first substituted for the baseline estimate in the decision tree. If substitution of the highest or the lowest

value changes the conclusions, more values within the range are substituted to determine the range of values.

In an analysis with many probabilities, there are numerous combinations of two and three variables, and the computational burden of doing all possible two-way and three-way sensitivity analysis is large. For this reason, it is not usually feasible to conduct two-way sensitivity analysis for all possible combinations let alone n-way. The choice of variables for multiple way analysis requires considerable judgment, and therefore is not laid down by any fast and hard rules. However, the graphical approach applied in these instances also lends itself to the discussion of dominance considerations. In section 3, the notion of dominant alternatives was introduced and can be considered a type of sensitivity analysis. A graphical approach, in this instance suggests when one alternative may supercede another and suggests possible implications.

Another area of emphasis for sensitivity analysis is the sensitivity with respect to the prior, when applying Bayesian methodology. These are typically performed when there is little imprecision un the loss because a standard choice of inference loss such as the quadratic was adopted, for instance. It is usually suggested to start with this case as (a) it is simplest and most thoroughly studied and (b) provide many insights for more general problems. It is also notably one of the most difficult elements to assess.

The overview of statistical decision theory provided in this report places has tried to maintain a balance between the classical and the Bayesian approach. In simple terms, the solution of a statistical decision problem proceeds by modeling a decision maker's judgments by means of the loss function, probability model of the observation process and a prior, and then uses these to identify a 'good' decision rule. Thus, this presents a variety of reasons as to why a sensitivity analysis should be conducted – that is, checking the sensitivity of the decision rule (output) with respect to the model and decision maker's judgments (inputs). In other words, the objective is to check the impact of the loss function, the prior and the model on the Bayes decision rule or Bayes alternative, and their posterior expected loss.

Section 14: Prototype Example

The conceptual ideas presented thus far have laid the foundation of statistical decision theory. The purpose of the following section is to present a prototype example which exemplifies the key ideas and methods discussed in this report. This example has been illustrated in many statistical decision theory books in various forms. It is illustrated in this report and develops in much of the same way the methodological sections of this report have.

Section 14.1: Summary of the Problem

An oil company owns a tract of land that may contain oil. A consulting geologist has reported to management that it is believed that there is a 1 in 4 chance of oil. Because of this prospect, another oil company has offered to purchase the land for $90,000. However, the land-owning oil company is considering holding the land in order to drill for oil itself. The cost of drilling is $100,000. If oil is found, the resulting expected revenue will be $800,000, so the company's expected profit (after deducing the cost of drilling) will be $700,000. A loss of $100,000 will be incurred if the land is dry (no oil).

Table 14.1: Prospective profits of oil company

Status of Land	**Payoff**	
Alternative	**Oil**	**Dry**
Drill for oil	$700,000	-$100,000
Sell the land	$90,000	$90,000
Chance of status		1 in 4

Table 14.1 summarizes the prospective profits for this land-owning company. This oil company is operating without much capital so a loss of $100,000 would be quite serious. Deciding whether to drill or sell also hinges on the option to conduct a detailed seismic survey of the land to obtain a better estimate of the probability of finding oil. Thus, first we present this prototype problem of decision without experimentation and then with experimentation and conclude with ways of refining the evaluation of the consequences of the various possible outcomes.

Section 14.2: Summary of the Decision Modeling Framework

The proceeding sections have explained the decision modeling framework. Here we re-cap the framework before proceeding with the prototype example. Simply, statistical decision problems refer to those problems containing data or observations on the state of nature. We first consider those cases without that information; that is problems of making decisions in the absence of data (or without experimentation). This lays a framework for decision analysis, for the problem at hand, which is later extended to incorporate data or decision problems with experimentation and various other decision methodologies.

Framework for Decision Analysis:

1. *Decision maker* needs to choose one of the *alternatives:*
$a \in A$
2. *Nature* would choose one of the possible *states of nature*:
$\theta \in \Theta$
3. Each combination of an action and state of nature would result in a *payoff/decision*, which is given as one of the entries in a *payoff/decision table*.
$c(a,\theta) \in C$
4. This payoff/decision table should be used to find an optimal action for the decision maker according to an appropriate criterion.

An additional element needs to be incorporated into this framework – prior probabilities. Many decision makers would generally want to incorporate additional information to account for the relative likelihood of the possible states of nature. This information is translated into a probability distribution where the state of nature is considered to be a random variable - making up the prior probability distribution.

In general terms, the decision maker must choose an action from a set of possible actions. The set contains all the feasible alternatives under consideration for how to proceed with the problem of concern. This choice of an action must be made in the face of uncertainty, because the outcome will be affected by random factors that are outside the control of the decision maker. These random factors determine what situation is referred to as a possible state of nature. For each combination of an action and a state of nature, the decision maker knows what the resulting payoff (or consequence) would be. The payoff is a quantitative measure of the value to the decision maker of the consequences of the outcome. Although in many instances the payoff is a monetary gain, there are other measures that can be used (as explained in utility theory). If the consequences of the outcome do no become completely certain even when the state of nature is given, then the payoff becomes an expected value (in the statistical sense) of the measure of the consequences. The payoff table can be used to determine an optimal action according to an appropriate criterion which suits the beliefs of the decision maker.

The decision problem faced by the oil company can now be placed within the decision framework just presented. This can be summarized in the following decision[7] table:

Table 14.2: Formulation of the problem within the framework of decision analysis

Status of Land	**State of Nature Θ**	
Alternative *A*	θ_1	θ_2
a_1	c_{11}=700	c_{12}= -100
a_2	c_{21}=90	c_{22}=90
Prior Probability	0.25	0.75

Section 14.3: Decision Making without Experimentation

[7] Payoff table and decision table can be used interchangeably.

This section illustrates the decision making process without experimentation. A formulation of the problem being assessed has two possible actions under consideration: drill for oil or sell the land. The possible states of nature are that the land contains oil and that it is does not. Since the consulting geologist has estimates that there is a 1 in 4 chance of oil, the prior probabilities of the two states of nature are 0.25 (for oil) and 0.75 (for no oil). ***Table 14.1*** has been redesigned in ***Table 14.2*** with the payoff units in thousands of dollars of profit. This payoff table will be used to determine the optimal action under the three main criterions discussed here.

1. Minimax Criterion:

The minimax criterion was described in section 2. Its rationale is that it provides the best guarantee of the payoff that will be obtained. That is, one determines, for each action *a*, the maximum loss over the various possible states of nature:

$$M(a) = \max_{\Theta} l(\theta, a),$$

and this provides an ordering among the possible actions. In words, for each possible action, find the minimum payoff over all possible states of nature. Next find the maximum of these minimum payoffs. Choose the maximum of the minimum payoff gives the maximum.

The application of this criterion to the prototype example suggests that selling the land is the optimal action to take. Regardless of what the true state of nature turns out to be for the problem, the payoff from selling the land cannot be less than 90, which provides the best available guarantee. Thus, this criterion provides the pessimistic viewpoint that regardless of which action is selected, the worst state of nature for that action is likely to occur, so one should choose the action which provides the best payoff with its worst state of nature.

2. Maximum Likelihood Criterion:

The maximum likelihood function focuses on the most likely state of nature. Recall a simple general definition of the likelihood function: For the observed data, *x*, the function $l(\theta) = f(x \mid \theta)$, considered as a function of θ, is called the likelihood function. The intuitive reason for the name "maximum likelihood function" is that a θ for which $f(x \mid \theta)$ is small, in that *x* would be more plausible occurrence if $f(x \mid \theta)$ were large. In simplicity, the steps in this approach begin by identifying the most likely state of nature (largest prior probability). For this state of nature, find the action with the maximum. Choose this decision.

The application of this criterion to the prototype example indicates that the *dry* state has the largest prior probability. In Table 10.2, in the dry cell column, the sell alternative has the maximum payoff, so the choice is to sell the land.

The appeal of this criterion is that the most important state of nature is most likely one, so the action chosen is the best one for this particularly important state of nature. Basing the decision on the assumption that this state of nature will occur tends to give a more, the criterion does not rely on questionable subjective estimates of the probabilities of the respective states of nature other than identifying the most likely state. However, the major drawback of this criterion is that it completely ignores some relevant information. No state of nature besides the most likely one is considered. Therefore, a problem with many possible states of nature, the probability of the most likely one may be quite small and or show little difference between "quite likely" states of nature.

3. Bayes' Decision Rule (Expected Monetary Value Criterion):

The Bayes decision rule was introduced in **Section 2**. In this instance, the probability weight assigned to each state of nature θ, the loss incurred for a given action incurs the expected value:

$$B(a) = \sum_i g(\theta_i) l(\theta_i, a) .$$

This approach uses the best available estimates of the probabilities of the respective states of nature (current prior probabilities) and calculates the expected value of the payoff for each of the possible actions. Choose the action with the maximum expected payoff.

In this application of the criterion, it can be easily determined that the optimal action is to drill. The expected payoffs can be calculated from ***Table 14.2*** directly as follows:

$$\begin{aligned} E[c_1] &= g(\theta_1) l(\theta_1, a_1) + g(\theta_2) l(\theta_2, a_1) \\ &= 0.25(700) + 0.75(-100) \\ &= 100 \\ E[c_2] &= g(\theta_1) l(\theta_1, a_2) + g(\theta_2) l(\theta_2, a_2) \\ &= 0.25(90) + 0.75(90) \\ &= 90 \end{aligned}$$

Since 100 > 90, the alternative action to drill is selected. Note that this choice differs from the other two preceding criteria.

The advantage to the Bayes' decision rule is that it incorporates all the available information, including payoffs and the best available estimates of the probabilities of the respective states of nature. It is sometimes argued that these estimates of the probabilities are largely subjective and so are too shaky to be trusted. Nevertheless, under many circumstances, past experience and current evidence enables one to develop reasonable estimates of the probabilities. The methodology of including such information was described in **Section 3**. Before applying this approach to the problem at hand, we will consider the use of sensitivity analysis to assess the effect of possible inaccuracies in the prior probabilities.

Section 13 briefly discussed the role and importance of sensitivity analysis in various applications to study the effect if some of the elements included in the mathematical model were not correct. The decision table in ***Table 14.2*** representing the payoffs is the mathematical model of concern. The prior probabilities in this model are most questionable and will be the focus of the sensitivity analysis to be conducted; however, a similar approach could be applied to the payoffs given in the table.

Basic probability theory suggests that the sum of the two prior probabilities must equal 1, so increasing one of these probabilities automatically decreases the other one by the same amount, and vice versa. The oil company's management team feels that the true "chances" of having oil on the tract of land are more likely to lie between the range from 0.15 to 0.35. Thus, the corresponding prior probability of the land being dry would range from 0.85 to 0.65, respectively.

Conducting a sensitivity analysis in this situation requires the application of the Bayes' decision rule twice – once when the prior probability of the oil is at the lower end (0.15) of this range and next when it is at the upper end (0.35). When the prior probability is conjectured to be 0.15, we find

$$
\begin{aligned}
E[c_1] &= g(\theta_1)l(\theta_1, a_1) + g(\theta_2)l(\theta_2, a_1) \\
&= 0.15(700) + 0.85(-100) \\
&= 20 \\
E[c_2] &= g(\theta_1)l(\theta_1, a_2) + g(\theta_2)l(\theta_2, a_2) \\
&= 0.15(90) + 0.85(90) \\
&= 90
\end{aligned}
$$

and when the prior probability is thought to be 0.35, the Bayes' decision rule finds that

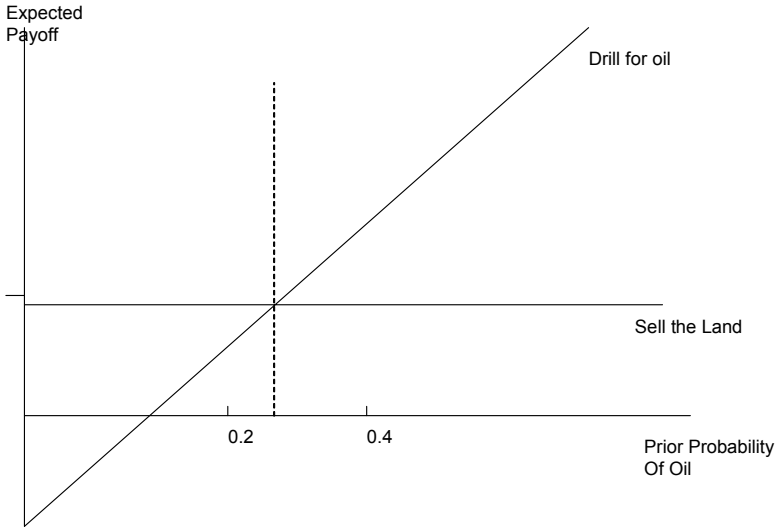
$$
\begin{aligned}
E[c_1] &= g(\theta_1)l(\theta_1, a_1) + g(\theta_2)l(\theta_2, a_1) \\
&= 0.35(700) + 0.65(-100) \\
&= 180 \\
E[c_2] &= g(\theta_1)l(\theta_1, a_2) + g(\theta_2)l(\theta_2, a_2) \\
&= 0.35(90) + 0.65(90) \\
&= 90.
\end{aligned}
$$

Thus, the decision is very sensitive to the prior probability of oil as the expected payoff to drill shifts from 20 to100 to180 when the prior probabilities are 0.15, 0.25, and 0.35, respectively. Thus, if the prior probability of oil is closer to 0.15, the optimal action would be to sell the land rather than to drill for oil as suggested by the other prior probabilities. This suggests that it would be more plausible to determine the true value of the probability of oil.

Let p = prior probability of oil. Then the expected payoff from drilling for any p is

$$E[c] = 700p - 100(1-p)$$
$$= 800p - 100.$$

Figure 14-1:



A graphical display of how the expected payoff for each alternative changes when the prior probability of oil changes for the oil company's problem of whether to drill or sell is illustrated in ***figure 14-1***. The point in this figure where the two lines intersect is the threshold point where the decision shifts from one alternative (selling the land) to the other (drill for oil) as the prior probability increases. Algebraically, this is simple to determine as we set

$$E[c_1] = E[c_2]$$
$$800p - 100 = 90$$
$$p = \frac{190}{800} = 0.2375,$$

where c_1 is the expected payoff of the action to drill and subsequently c_2 is the expected payoff of the action to sell the land. Thus, the conclusion should be to sell the land if $p <$ 0.2375 and should drill for oil if $p > 0.2375$. Because, the decision for the oil company decides heavily on the true probability of oil, serious consideration should be given to conducting a seismic survey to estimate the probability more accurately. This is considered in the next subsection.

Section 14.4: Decision Making with Experimentation

Here another element is added to the decision analysis framework – posterior probabilities. This additional 'element' allows for testing or experimentation to be

conducted; thus, improving the preliminary estimates of the probabilities of the respective states of natures, i.e. the prior probabilities.

For this problem, an available option before making a decision is to conduct a detailed seismic survey of the land to obtain a better estimate of the probability of oil. The cost is $30,000. A seismic survey obtains seismic soundings that indicate whether the geological structure if favorable to the presence of oil. Dividing the possible findings of the survey into the following two categories:

USS: unfavorable seismic soundings; oil is unlikely, and
FSS: favorable seismic soundings; oil is likely to be found.

Based on past experience, if there is oil, then the probability of unfavorable seismic soundings (USS) is

$$P(\text{USS} \mid \text{State} = \text{Oil}) = 0.4, \text{so} \quad P(\text{FSS} \mid \text{State} = \text{Oil}) = 1 - 0.4 = 0.6 .$$

Similarly, if there is no oil (state of nature = Dry), then the probability of unfavorable seismic soundings is estimated to be

$$P(\text{USS} \mid \text{State} = \text{Dry}) = 0.8, \text{so} \quad P(\text{FSS} \mid \text{State} = \text{Dry}) = 1 - 0.8 = 0.2.$$

This data is used to find the posterior probabilities of the respective states of nature given the seismic readings.

The main 'criterion' implemented here is Bayes Theorem. Now suppose before choosing an action, an outcome from an experiment is observed X = x. Therefore for each i=1,2,..,n, the corresponding posterior probability is

$$P(\theta = \theta_i \mid X = x) = \frac{P(X = x \mid \theta = \theta_i) P(\theta = \theta_i)}{\sum_{k=1}^{n} P(X = x \mid \theta = \theta_k) P(\theta = \theta_k)}$$

Proceeding in general terms, let

n = number of possible states of nature;
$P(\Theta = \theta_i)$ = prior probability that true state of nature is state i, for i =1, 2, …, n;
X = finding from experimentation (a random variable);
x_j = one possible value of the finding;
$P(\Theta = \theta_i \mid X = x_j)$ = posterior probability that true state of nature is state i, given $X = x_j$, for j = 1, 2, …, n.

Thus, the question being addressed is: Given $P(\Theta = \theta_i)$ and $P(X = x_j \mid \Theta = \theta_i)$, for i =1, 2, …, n, what is $P(\Theta = \theta_i \mid X = x_j)$? The answer is solved by following standard formulas of probability theory which states that the conditional probability,

$$P(\Theta = \theta_i \mid X = x_j) = \frac{P(\Theta = \theta_i, X = x_j)}{P(X = x_j)},$$

where

$$P(X = x_j) = \sum_{k=1}^{n} P(\Theta = \theta_k, X = x_j),$$

gives

$$P(\Theta = \theta_i \mid X = x_j) = P(X = x_j \mid \Theta = \theta_i) P(\Theta = \theta_i).$$

Therefore, for i =1, 2, …, n, the desired formula for the corresponding posterior probability is the Bayes theorem as stated above.

Returning to the prototype example and applying this formula, one finds if the finding of the seismic survey is unfavorable, then the posterior probabilities are

$$P(\text{State} = \text{Oil} \mid \text{Finding} = \text{USS}) = P(\Theta = \theta_1 \mid X = x_1) = \frac{0.4(0.25)}{0.4(0.25) + 0.8(0.75)} = \frac{1}{7},$$

$$P(\text{State} = \text{Dry} \mid \text{Finding} = \text{USS}) = P(\Theta = \theta_2 \mid X = x_1) = 1 - \frac{1}{7} = \frac{6}{7}.$$

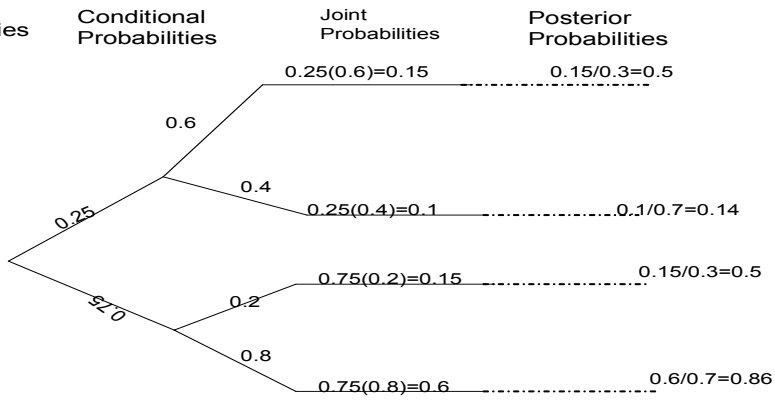
Note that USS = x_1 and FSS = x_2. Similarly, if the seismic survey is favorable, then

$$P(\text{State} = \text{Oil} \mid \text{Finding} = \text{FSS}) = P(\Theta = \theta_1 \mid X = x_2) = \frac{0.6(0.25)}{0.6(0.25) + 0.2(0.75)} = \frac{1}{2},$$

$$P(\text{State} = \text{Dry} \mid \text{Finding} = \text{FSS}) = P(\Theta = \theta_2 \mid X = x_2) = 1 - \frac{1}{2} = \frac{1}{2}.$$

Probability Tree diagrams can be a useful and visually appealing way of organizing the outcomes and the corresponding probabilities. Such a diagram is presented in ***Figure 14-2***. Reading the probability diagram from left to right one can see all the probabilities – prior probabilities to conditional probabilities to joint probabilities – leading to the posterior probabilities. After these computations have been completed, Bayes' decision rule can be applied again just as before, with the posterior probabilities replacing the prior probabilities. Again using the "payoffs" (in units of thousands of dollars) from ***Table 14.2*** and subtracting the cost of experimentation, we obtain the results shown below.

Figure 14-2:



The expected payoffs if finding is unfavorable seismic soundings:

$$E[c_1(a_1 \mid X = x_1)] = \frac{1}{7}(700) + \frac{6}{7}(-100) - 30$$

$$= -15.7$$

$$E[c_2(a_2 \mid X = x_1)] = \frac{1}{7}(90) + \frac{6}{7}(90) - 30$$

$$= 60$$

And the expected payoffs if the finding is favorable seismic surroundings:

$$E[c_1(a_1 \mid X = x_2)] = \frac{1}{2}(700) + \frac{1}{2}(-100) - 30$$

$$= 270$$

$$E[c_2(a_2 \mid X = x_2)] = \frac{1}{2}(90) + \frac{1}{2}(90) - 30$$

$$= 60$$

Table 14-3: Optimal Policies

Finding from Seismic Survey	Optimal Action	Expected Payoff Excluding Cost of Survey	Expected Payoff Including Cost of Survey
USS	Sell	90	60
FSS	Drill	300	270

Since the objective is to maximize the expected payoff, these results yield the optimal policy shown in ***Table 14-3***. However, this analysis has not yet addressed the issue of whether the expense of conducting a seismic survey is truly valuable or whether one should choose the optimal solution without experimentation. This issue is addressed next.

The **Value of Experimentation** can be determined by using two complementary methods. The first assumes (unrealistically) that the experiment will remove all uncertainty about what the true state of nature is, and then calculates the resulting "improvement" in the expected payoff (ignoring the cost of the experiment). This quantity, the expected value of perfect information, provides an upper bound on the potential value of the experiment. Therefore, if the upper bound is less than the cost of experiment, than the experiment should be forgone. However, if this upper bound exceeds the cost of experiment, then second methods should be implemented. This method calculates the actual "improvement" in the expected payoff (ignoring the cost of experiment) that would result from performing the experiment. Comparing the improvement with the cost indicates whether the experiment should be performed.

1. Expected Value of Perfect Information

Suppose that the experiment could definitely identify the true state of nature thereby providing "perfect" information. Whichever state of nature is identified, you naturally

choose the action with the maximum payoff for the state. Since we do not know in advance the state of nature which will be identified, a calculation of the expected payoff or consequence with perfect information (ignoring the cost of the experiment), requiring the weighting of the maximum payoff for each state of nature by the prior probability of that state of nature is needed.

Expected payoff with perfect information = 0.25(700) + 0.75(90) = 242.5.

Thus, if the oil company could learn more before choosing its action whether the land contains oil, the expected payoff as of now (before acquiring this information) would be $242,500 (excluding the cost of the experiment generation the information.)

To evaluate whether the experiment should be conducted, we calculate the expected value of perfect information (EPVI)[8]:

EPVI = *E*[payoff with perfect information] – *E*[payoff without experimentation]

Thus, since experimentation usually cannot provide perfect information, EVPI can provide an upper bound on the expected value of experimentation. For this prototype example, EVPI = 242.5 – 100 = 142.5, where the expected payoff without experimentation was determined earlier to be 100 under Bayes' decision rule. Since 142.5 far exceeds 30, the cost of experimentation (a seismic survey) is shown to be worthwhile. Thus, we carry on and implement the second method of evaluating the potential benefit of experimentation.

2. Expected Value of Experimentation

Now we want to determine expected increase directly which is referred to as the expected value of experimentation. Calculating this quantity requires first computing the expected payoff with experimentation (excluding the cost of experimentation). Obtaining this latter quantity requires the previously determined posterior probabilities, the optimal policy with experimentation, and the corresponding expected payoff (excluding the cost of experimentation) for each possible finding from the experiment. Then each of these expected payoff needs to be weighted by the probability of the corresponding finding, that is,

$$\text{Expected payoff with experimentation} = \sum_j P(X = x_j)E[c \mid X = x_j],$$

where the summation is taken over all possible *j* and *c* denotes the payoff or consequence. For the prototype example much of the work has been done for the right side of the equation. The values of $P(X = x_j)$ for each of the two possible findings from the seismic survey – unfavorable or favorable – were calculated at the bottom of the probability tree diagram in Figure 14.2 to be $P(x_1) = 0.7$ and $P(x_2) = 0.3$. For the optimal policy with

[8] The value of perfect information is a random variable equal to the payoff with perfect information minus the payoff without experimentation. Thus the expected value of perfect information is the expected value of this random variable.

experimentation, the corresponding expected payoff for each finding can be obtained from Table 14.3 as

$E(c \mid X = x_1) = 90,$
$E(c \mid X = x_2) = 270.$

Then the expected payoff with experimentation is determined to be 0.7(90) + 0.3(300) = 153. Similar to before the expected value of experimentation (EVE) can be calculated as

$$\text{EVE} = 153 = 100 = 53$$

and identifies the potential value of the experimentation. Since this value exceeds 30, the cost of conducting a detailed seismic survey shows valuable.

Section 14.5: Decision Trees

Decision trees are useful in providing a visual display of the sequential decision processes and subsequent consequences. It can help in organizing the computational work that has been described thus far.

This prototype problem involves a sequence of two decisions which each result in possible consequences and in turn resulting in further decisions and consequences. It involves a sequence of two decisions:

1. Should a seismic survey be conducted before an action is chosen?
2. Which action (drill for oil or sell the land) should be chosen?

The corresponding decision tree (with partially added numbers and performing computations) is displayed in ***Figure 14.3.*** Note that this is the final decision tree and was done to save space and time. Recall from **Section 9**, the interpretation of the specific types of nodes. The nodes of the decision tree are referred to as forks, and the arcs are called branches. A decision fork, represented by a square, indicates that decision needs to be made at that point in the process. A chance fork, represented by a circle, indicates that a random event occurs at that point.

In this example, the first decision is represented by a decision fork *a*. Fork *b* is a chance fork representing the random event of the outcome of the seismic survey. The two branches emanating from fork *b* represent the two possible outcomes of the survey. The decisions *c* and *d* follow the possible outcomes and result in the consequences from the random events of *f* and *g*. Similarly, decision fork *e* leads to a chance fork *h*, where again the two branches correspond to the two possible states of nature.

Since the decision tree is filled in with the available information, thus far – some interpretation can be deduced before the actual analysis proceeds. The numbers under or over the branches that are not in parentheses are the cash flows that occur at those branches. The resulting in end payoff is recorded in boldface at the end of the terminal

branch. At each chance fork, the probabilities of random events are recorded in parentheses. From chance fork *h*, the probabilities are the prior probabilities as no seismic survey was conducted. The probabilities emanating from chance fork *b* are the probabilities of these findings while the probabilities emanating from chance fork *f* and *g* are the posterior probabilities of the states of nature, given the finding from the seismic survey.

Performing the Analysis

The backward induction procedure was introduced in **Section 10** and is implemented here for the prototype example. The steps involved can be summarized as follows:

1. Start at the right side of the decision tree and move left one column at a time. For each column, perform either step 2 or 3 depending on whether the forks in that column are chance forks or decision forks.
2. For each chance fork, calculate the expected payoff by multiplying the expected payoff of each branch (shown in boldface to the right of the branch) by the probability of that branch and then summing these products. Record this expected payoff for each decision fork in boldface next to the fork, and designate this quantity as also being the expected payoff for the branch leading to the fork.
3. For each decision fork, compare the expected payoffs of its branches and choose the alternative whose branch has the largest expected payoff. In each case, record the choice on the decision tree by inserting a double dash as a barrier through each rejected branch.

To begin the procedure, consider the rightmost column of forks, namely, chance forks *f*, *g*, and *h*. Applying step 2, their expected payoffs are calculated as

$$E[c_f] = \frac{1}{7}(670) + \frac{6}{7}(-130) = -15.7,$$

$$E[c_g] = \frac{1}{2}(670) + \frac{1}{2}(-130) = 270,$$

$$E[c_h] = \frac{1}{4}(700) + \frac{3}{4}(-100) = 100.$$

Note that $E[c_k]$ denotes the expected payoff or consequence of the k^{th} chance fork. These expected payoffs are placed above these forks as displayed in ***Figure 14.3***. Next, moving one column to the left, are the decision forks *c*, *d*, and *e*. The expected payoff for a branch that leads to a chance fork now is recorded in boldface over that chance fork. Therefore, step 3 can be applied as follows.

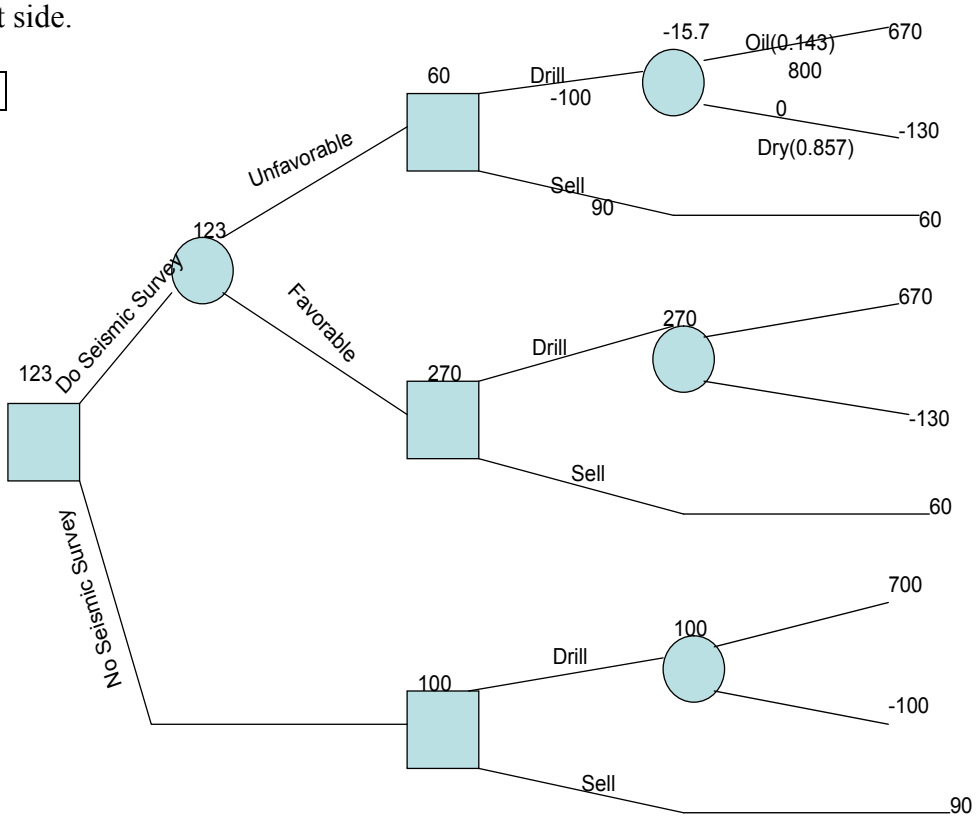
$E[c_c(a_1)] = -15.7,$
$E[c_c(a_2)] = 60.$
$60 > -15.7,$ so choose $a_2 =$ sell.
$E[c_d(a_1)] = 270,$
$E[c_d(a_2)] = 60.$
$270 > 60,$ so choose $a_1 =$ drill.
$E[c_e(a_1)] = 100,$
$E[c_e(a_2)] = 90.$
$100 > 90,$ so choose $a_1 =$ drill.

Next, move one column to the left again and apply step to the chance fork *b*. Here we find that $E[c_b] = 0.7(60) + 0.3(270) = 123$. In the final step, we consider the last decision fork *a* and apply step 3 as before:

$E[c_a(a_{01})] = 123,$
$E[c_a(a_{02})] = 100.$
$123 > 100,$ so choose $a_{01} =$ do seismic survey.

Note a_{01} denotes the action to do a seismic survey where a_{02} denotes the action to not conduct a seismic survey. Now the decision tree is complete where the payoffs are given on the far right side.

Figure 14-3:



The double dashes have blocked off the undesirable paths. According to Bayes' decision rule, follow the open paths from left to right. Thus, the optimal policy is:

- Do the seismic survey.
- If the result is unfavorable, sell the land.
- If the result is favorable, drill for oil.
- The expected payoff (including the cost of the seismic survey) is $123,000.

This was the same unique optimal solution obtained in the earlier section without the decision tree which was shown in ***Table 14.3***. The optimal policy via the backward induction process leads to the same policy or policies via the more formal process.

Section 14.6: Utility Theory

This section will focus on various models of individual rationality which can be categorized into an individual's traits. Thus far, when applying the Bayes' decision rule, we have assumed that the expected payoff or consequence in monetary terms is the appropriate measure of the consequences of taking an action. However, in many situations this assumption is inappropriate. Fortunately, there is a way of transforming monetary values to an appropriate scale that reflects the decision maker's preferences. This scale is referred to as the utility function for money.

A utility function for money illustrates an individual's behavior pertaining to money. A person who is risk-averse will exhibit a decreasing marginal utility function. However, individuals who are risk seekers demonstrate an increasing marginal utility for money. The intermediate case is that of a risk-neutral individual, who prizes money at its face value. Such an individual's utility for money is simply proportional to the amount of money involved. It is possible to exhibit a mixture of these kinds of behavior. For example, an individual might be essentially risk-neutral with small amounts of money, then become a risk seeker with moderate amounts, and then turn risk-averse with large amounts. Moreover, one's attitude toward risk can change over time and circumstances.

The fact that different people have different utility functions for money has an important implication for decision making in the face of uncertainty. When a utility function for money is incorporated into a decision analysis approach to a problem, this utility function must be constructed to fit the preferences and values of the decision maker involved. The key to constructing the utility function for money to fit the decision maker is based on the fundamental properties explained in **Section 6**. The following fundamental property summarizes in words the central notion that under the assumptions of utility theory, the decision maker's utility for money has the property that the decision maker is in different between two alternative courses of action if the two alternatives have the same expected utility.

The scale of the utility function is irrelevant. It is only the relative values of the utilities that matter. All the utilities can be multiplied by any positive constant without affecting which alternative course of action will have the largest expected utility. To summarize

the basic role of utility function in decision analysis is to simply state that when the decision maker's utility function for money is used to measure the relative worth of the various possible monetary outcomes, Bayes' decision rules replaces monetary payoffs by the corresponding utilities. Therefore, the optimal action (or sequence of actions) is the one which maximizes the expected utility. Although only utility functions for money are discussed in this section, they can be constructed for consequences of the alternatives of action that are not monetary.

Returning to the prototype example, recall in the summary of the problem it was stated that the oil company was operating without much capital, so a loss of $100,000 would be steep. The worst case scenario is to conduct a seismic survey and then drill for $100,000 and drill when there is no oil. On the other hand, striking oil is an exciting prospect of earning $700,000. To apply the decision maker's utility function for money to the problem as described earlier, it is necessary to identify the utilities for all possible monetary payoffs. These payoffs and their corresponding utilities are presented in ***Table 14-4*** and can be obtained by the following methodology.

Table14-4: Utilities for oil company

Monetary Payoff	**Utility**
-130	-150
-100	-105
60	60
90	90
670	580
700	600

As a starting point in constructing the utility function, it is natural to let the utility of zero money be zero, so $u(0) = 0$. An appropriate next step is to consider the worst scenario and best scenario and then to address the question of what value of p would make the decision maker indifferent between two alternatives. Suppose the decision maker's choice is $p = \frac{1}{5}$. If $u(M)$ denotes the utility of a monetary payoff of M, this choice of p implies that

$$\frac{4}{5}u(-130) + \frac{1}{5}u(700) = 0 \quad \text{(utility of alternative 1)}.$$

The value of either $u(-130)$ or $u(700)$ can be set arbitrarily to establish the scale of the utility function. By choosing $u(-130) = -150$, this equation yields $u(700) = 600$. To identify u(-100), a choice of p is made that makes the decision maker indifferent between a payoff of -130 with probability p or definitely incurring a payoff of -100. The choice is $p = 0.7$, so

$$u(-100) = p\,u(-130) = 0.7(-150) = -105.$$

To obtain $u(90)$, a value of p is selected that makes the decision maker in different between a payoff of 700 with probability p or definitely obtaining a payoff of 90. The value chosen is $p = 0.15$, so

$$u(90) = p\,u(700) = 0.15(600) = 90.$$

To obtain the decision maker's utility function for money, a smooth curve was drawn through $u(-130), u(-100), u(90),$ and $u(700)$ as shown in ***Figure 10-4***. The values on this curve at $M = 60$ and $M = 670$ provide the corresponding utilities, $u(60) = 60$ and $u(670) = 580$, which completes the list of utilities displayed in ***Table 10.4***. For contrast, the dashed line drawn at a 45° shows the monetary value of the payoffs used exclusively in the preceding sections. Note how $u(M)$ essentially equals M for small values of M, and then how $u(M)$ gradually falls off M for larger values of M. This is a typical of a moderately risk-averse individual.

There are other approaches of estimating $u(M)$; however, they are not necessarily appropriate in this instance. Decision trees can also be used in an identical fashion to the analysis in the preceding section except for substitution utilities for monetary payoffs. For this analysis of a drilling decision we add an extra element, for variety purposes. We also give a more general analysis perspective to end this section.

14.7: Combination of Formal and Informal Analysis

Continuing with the same problem, the desirability of drilling depends on the amount of oil which will be found – oil or no oil. Before making this decision, the oil company wishes to obtain more geological and geophysical evidence by means of a seismographic recording which is quite expensive. It is also assumed that these recordings, if made, will give completely reliable information that one of the tree conditions prevails; (1) there is no subsurface structure, (2) there is an open subsurface structure, or (3) there is a closed subsurface structure. The descriptions of the four spaces, A, Θ, E, X, are summarized in ***Table 14-5*** and the possible sequence of choices are displayed in the decision tree in ***Figure 14-5***.

Table 14.5: Possible Choices

Space	Elements	Interpretations
A	a_1	Drill, do not sell location
	a_2	Do not drill, sell location
Θ	θ_1	Oil
	θ_2	No oil
E	e_1	Do not take seismic readings
	e_2	Take seismic readings
X	x_0	Dummy outcome of e_0
	x_1	e_1 reveals no structure
	x_2	e_2 reveals open structure
	x_3	e_3 reveals closed structure

Assignment of Utilities

The psychological stimulus associated in this problem with a (e, x, a, θ) 4-tuple is highly complicated. Different wells entertain different drilling costs, and different strike produce different quantities and qualities of oil which can be recovered over different periods of time and sold at different prices. Furthermore, each potential consequence of the present drilling venture may interact with future potential drilling ventures such as geological information gained.

There are uncertainties surrounding any particular (e, x, a, θ) complex which must be compared formally or informally. Assume, nevertheless, that the decision maker can assign to utility numbers to reflect his preferences. This presents the idea of cutting the decision tree presented in **Section 4 and 9**.

Assignment of Probabilities

The assignment of a probability measure to the probability space $\Theta \times X$ bears more importance on the conditional (or "posterior") measure $P''_{\theta|x}$ and the marginal measure $P_{x|e}$ than on the complementary measures $P_{x|e,\theta}$ and P'_θ. Previous experience with the amounts of oil found in the three possible types of geological structure (x_1 = no structure, x_2 = open structure, x_3 = closed structure) may make it possible to assign a nearly "objective" measure $P''_{\theta|x}$ to the amount of oil which will be found given any particular experiment result, whereas it would be much less clear what measure P'_θ should be assigned to the amount of oil in the absence of knowledge of the structure. At the same time it will in general be much more meaningful to a geologist to assign a marginal measure $P_{x|e}$ to the various structures and thus to the sample space X than it would be to assign conditional measures $P_{x|e,\theta}$ to X depending on the amount of oil which will be found. The hypothetical measures $P''_{\theta|x}$ and $P_{x|e}$ are shown on those branches of the decision tree in Figure 6.11 which emanate from e_1; the prior probabilities $P'_\theta\{\theta_1\}$=0.2 and $P'_\theta\{\theta_2\}$=0.8 shown on the branches emanating from e_0 were computed from them by use if the formula

$$P'_\theta(\theta_i) = P''_\theta(\theta_i \mid x_1)P_x(x_1 \mid e_1) + P''_\theta(\theta_i \mid x_2)P_x(x_2 \mid e_1) + P''_\theta(\theta_i \mid x_3)P_x(x_3 \mid e_1).$$

Analysis

Since all the data required for analysis appears in the decision tree in ***Figure 14-5***, it can be easily verified that the optimal decision is to pay for seismographic recordings (e_1) and then drill (a_1) if and only if the recordings reveal open (x_2) or closed (x_3) structures. It is worthwhile to pay for the recordings because the expected utility of this decision is 15.25 whereas the expected utility of the optimal act without seismic information is 0.

References

Bather, J. (2000). *Decision Theory*. Chichester: John Wiley & Sons, Inc.

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis. (2nd ed.)*. New York: Springer-Verlag New York Inc.

Brooks, S.P. and Roberts, G.O. (1998). Convergence Assessment Techniques for Markov Chain Monte Carlo. *Statistics and Computing*, 8:319-335.

Buchanan, J.T. (1982). *Discrete and Dynamic Decision Analysis*. Chichester: John Wiley & Sons, Inc.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings Algorithm. *The American Statistician*. 49(4).

Clemen, R.T. (1996). *Making Hard Decisions: An Introduction to Decision Analysis. (2nd ed.)*. Pacific Grove, CA: Brooks/Cole Publishing Company.

Draper D. and Madigan, D. (1997). *The Scientific Value of Bayesian Statistical Methods and Outlook*.

Everitt, B.S. (2002). *The Cambridge Dictionary of Statistics* (2nd ed.). Cambridge, Cambridge University Press.

French, S., & Insua, D.R. (2000). *Statistical Decision Theory*. London: Arnold.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (1995). *Bayesian Data Analysis*. London, Chapman and Hall.

Gilboa, I. & Schmeidler, D. (2001). *Theory of Case-based Decisions*. UK: Cambridge University Press.

Gilks, W.R., Richardson, S. and Spielgelhalter, D.J. (1996). *Markov Chain Monte Carlo in Practice*. London, Chapman and Hall.

Goldstein, H. (2002). *Multilevel Statistical Models* (3rd ed.). London, Arnold Publishers.

Grimmett, G.R. Stirzaker, D.R. (1982). *Probability and Random Processes*. New York: Oxford University Press.

Hastings, N.A.J. & Mello, J.M.C. (1978) *Decision Networks*. UK: John Wiley & Sons, Inc.

Hillier, F.S., & Lieberman, G.J. (2001). *Introduction to Operations Theory. (7th ed.)*. New York: McGraw-Hill.

Jensen, F.V. (2001). *Bayesian Networks and Decision Graphs*. New York: Springer-Verlag New York, Inc.

Jordan, M.I. (editor) (1999). *Learning in Graphical Models*. The Netherlends: Kluwer Academic Publishers.

Kass, R.E., Carlin, B.P., Gelman, A., Neal, R.M. (1998). Statistical Practice: Markov Chain Monte Carlo in Practice: A Roundtable Discussion. *The American Statistician*. 52(2).

Lindgren, B.W. (1971). *Elements of Decision Theory*. New York: The Macmillan Company.

Luce, R. D. (2000). *Utility of Gains and Losses*. New Jersey: Lawrence Erlbaum Associates, Publishers.

MacDonald, I.L. & Zuchchini, W. (1997). *Hidden Markov and Other Models for Discrete-value Time Series*. Boca Raton: Chapman & Hall, Inc.

Marshall, K.T. & Oliver, R.M. (1995). *Decision Making and Forecasting*. New York: McGraw-Hill, Inc.

Muphy, K. (1998). http://www.ai.mit.edu/~murphyk/Bayes/bayes.html

Nau, R. (2002). *Ph.D. Seminar on Choice Theory*. Duke University: The FUQUA School of Buisness.

Neal, R. (1996). *Bayesian Learning for Neural Networks*. New York: Springer-Verlag New York, Inc.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. California: Morgan Kaufmann.

Pearl, J. (1999). *Causality*. http://bayes.cs.ucla.edu/jp_home.html

Petitti, D.B. (2000). *Meta-Analysis, Decision Analysis and Cost-Effectiveness Analysis: Methods for Quantitative Synthesis in Medicine. (2nd ed.)*. New York: Oxford University Press, Inc.

Raiffa, H., Schlaifer, R. (2000). *Applied Statistical Decision Theory*. New York: John Wiley & Sons, Inc.

Ripley, B. (1987). *Stochastic Simulation*. Chichester: John Wiley & Sons, Inc.

Ripley, B. (1997). *Pattern Recognition and Neural Networks*. Chichester: John Wiley & Sons, Inc.

Rivett, P. (1980) *Model Building for Decision Analysis*. Chichester: John Wiley & Sons, Inc.

Tierney, L. and Mira, A. (1999). Some Adapative Monte Carlo Methods for Bayesian Inference. *Statistics in Medicine*, 18:2507-2515.

Weirich, P. (2001). *Decision Space*. Cambridge: Cambridge University Press.

West, D.B. (2001). *Introduction to Graph Theory*. New Jersey: Prentice Hall.

White, D.L. (1976). *Fundamentals of Decision Theory*. New York: American Elsevier Publishing Company, Inc.

Courses and Professors:

STA4276H (2003): MCMC algorithms by Professor Jeffrey Rosenthal.
CHL5223H (2002/3): Applied Bayesian Methods by Professor Michael Escobar (meetings).