Chapter 3

Probabilistic Biomedical Knowledge

Facts are stubborn, but statistics are more pliable.

– Mark Twain
attributed

When representing knowledge in terms of symbols and formulas of symbols, uncertainty and probability come up in several different ways. One possibility is that our theory formed by a choice of symbols and a symbol system language like first-order logic, a set of axioms, an inference method, and an interpretation might actually be incomplete, that is, it is not able to infer all the important assertions we will observe in our biomedical or health environment to which the theory applies. Sometimes, this can

be resolved by adding more axioms. However, sometimes it cannot. We need methods for representing knowledge that is uncertain and subject to refinement as more information is obtained, and as well, for modeling the significance of different outcomes that are consequences of our decisions. Thus, the focus here is on *events*, or outcomes, and *decisions* that may lead to various outcomes, for events that are uncertain, either because they are of unknown status or they are actually non-deterministic.

In this chapter, we introduce basic ideas of probability and statistics and then show how they can be applied to build probabilistic reasoning systems. This is followed by an introduction to utility theory and decision modeling. We conclude with a brief introduction to information theory.

Brief mention is made of some ideas from statistics, but the emphasis here is on the mathematics of graphical relationships among variables using conditional probabilities. Such graphs can compactly and meaningfully represent causal relationships in very complex systems with many variables. They are particularly suited to modeling pathology and treatment of disease.

3.1 PROBABILITY AND STATISTICS

A statistician is a person who, with his head in the oven and his feet in the freezer, will say that on the average he feels comfortable.

– anonymous
well-known joke

This section introduces the basic ideas and formulas of probability and some elementary concepts from statistics. More thorough developments can be found in standard probability and statistics textbooks [33, 344] and [231, Parts I and II]. In contrast to the concise but very far-ranging treatment in Wasserman [344] in particular, this chapter should probably be called "Some of Statistics."

3.1.1 Probability

Probability is about *events*. An event is something with a definite outcome that can be observed. Put another way, we are describing the state of a system by specifying the values of a set of variables. The possible values (or combinations of values) of the variables represent all the possible states. Often rather than predicting the detailed outcomes, we can only predict the probability of any particular outcome or observed state. The classic example is a coin which we toss to see if it will land head up or tail up.[1]

In any given situation, the set of all possible elementary outcomes or results of measurements (e.g., in a laboratory, clinical setting, or public health surveillance) is known and specified. This set may be discrete or continuous, numeric or symbolic, and finite or infinite. The set of all possible outcomes or observations is called the *sample space*. For a single coin toss, the sample space has two elements, `heads` and `tails`. For the observation of what base appears in some position in a DNA sequence, there are four elements in the sample space, A, G, C, and T, and many subsets. Each subset can be considered an event, for example, the occurrence of either G or C, or the occurrence of anything but T, and so on.

In some cases, the behavior of the biological or other system we are modeling is inherently non-deterministic, that is, the things we observe are not uniquely predictable from other more basic characteristics. Taking cancer as an example, the process of spread of tumor cells from the primary tumor site to form metastatic lesions is inherently variable or non-deterministic. For any particular cell, we cannot predict where it will go, but we can perhaps build a theory that predicts the frequencies with

1. Actually, many coins do have a head, usually of a significant political figure, a president, monarch, or some such. However, I personally have never seen a coin with a tail depicted on it. The US "buffalo nickel" comes close, but the whole bison (not a buffalo) is depicted, not just the tail. By "tail" we generally mean the opposite side of the coin from the head.

which cells migrate
represent such know
is called the "freque

Another importa
but it is completely
behavior of the syst
we may have some
patient's status is de
not know yet which
not the physical real
reality has not chang
"Bayesian" view.

These two views
an effective agent fo
the treatment (with
patient will respond
characteristics, perha
respond to treatment
non-deterministic bel
of response.

3.1.2 Statistics

In many areas of biol
range of values. We
several). This depenc
for example, one vari
complex, like an S-sh
of absorbed radiation
to estimate the param
no existing theoretica
idea of generating the
and medical theories,
imagination, though g
models, compute thei
of probabilistic theori
different.

An important goal
computing their predi
ticular observation ha
are for the various val
involving ensembles
of identical systems,
number that came up
heads should approxi
experimental procedu
probability. One migh
is, "what sampling me

In order for such pr
systems tested or repli
that statistics addresse

which cells migrate to one or another location. So, we need probability-based modeling methods to represent such knowledge. This view that probability represents the reality of non-deterministic events is called the "frequentist" view.

Another important case is the situation where there may be an underlying deterministic behavior, but it is completely beyond our reach for now, and we must model our uncertain knowledge about the behavior of the system, as distinct from its actual behavior. Again referring to cancer as an example, we may have some diagnostic tests (e.g., for prostate cancer), which can be used as evidence. The patient's status is definite but unknown, that is, he either has a prostate cancer or not, but we just do not know yet which is the case. In this situation, we are modeling our *beliefs* about the situation, not the physical reality. As more evidence accumulates, our beliefs will be updated but the physical reality has not changed, only our knowledge about it has changed. This is called the "subjectivist" or "Bayesian" view.

These two views interact when we need to make decisions. We know that interleukin-2 can be an effective agent for treating kidney cancer, but only 10–20% of the patients so treated respond to the treatment (with tumor growth arrest, regression, or complete ablation). We do not know which patient will respond (although one might imagine that at some future time we could discover which characteristics, perhaps genetic markers, would predict whether a patient with kidney cancer would respond to treatment or not). To decide whether to apply the treatment, we use what is known about the non-deterministic behavior of the treatment, and what we believe about a particular patient's likelihood of response.

3.1.2 Statistics

In many areas of biology and medicine, we describe observations in terms of *variables* that take on a range of values. We hypothesize that one variable has a functional dependency on another (or possibly several). This dependency typically reflects some kind of biological law. The relation could be linear, for example, one variable is proportional to another, with possibly a constant offset. It could be more complex, like an S-shaped curve. The percentage of living cells that are killed by radiation as a function of absorbed radiation dose is an example of an S-shaped curve. Many statistical methods are designed to estimate the parameters of such a function or even to estimate the shape itself, where there are few or no existing theoretical ideas to suggest a functional shape. However, despite the attractiveness of the idea of generating theories automatically through data analysis, there is no way to do this. Biological and medical theories, like their physical counterparts, are descriptions created principally by human imagination, though grounded in observation. Just as with theories based in logic formalism, we create models, compute their entailments, and check these predictions against the available data. In the realm of probabilistic theories or models, the process is the same and only the inference methods are somewhat different.

An important goal of statistics is to provide a mathematical structure for formulating such theories, computing their predictions, and deriving their parameters from experimental data. Because any particular observation has a particular value, how can we determine experimentally what the probabilities are for the various values? We cannot measure probabilities directly. Two approaches are possible, one involving ensembles and the other involving time. In the ensemble approach, we have a large number of identical systems, for example, a large number of identical coins. We toss them all and count the number that came up heads and the number that came up tails. The fraction of the total that came up heads should approximate the probability of a "heads" result, and by definition (and guaranteed by the experimental procedure just described), the probability of a "tails" result should be 1.0 minus the heads probability. One might wonder how such experiments can really provide the information. The question is, "what sampling method provides an *estimate* of the probabilities in the probability distribution?"

In order for such procedures to yield an accurate approximation to the true probability, the number of systems tested or replicated event observations must be sufficiently large. Another important question that statistics addresses is how large a data set must be in order to claim that a result is significant.

3.1.3 The Laws of Probability

A *probability function* is a function that assigns a number between 0.0 and 1.0 to each subset of elements in the sample space. The term "event" is used to refer to any particular subset, including the subsets that each have only a single member, a single elementary outcome. Equation 3.1 expresses this requirement that the range of the probability function is between 0.0 and 1.0. This is the first *axiom* of probability. If x represents an event, and $P(x)$ represents the probability that event will occur, then

$$\text{Axiom 1} \qquad 0.0 <= P(x) <= 1.0 \tag{3.1}$$

The probability of an event tells us what is the likelihood that the event will occur, out of all possible events. A probability of 0.0 means that the event will *never* occur, while a probability of 1.0 means that the event will *certainly* occur. For subsets with several outcomes as members, the probability function's value is the probability that *any* one of the outcomes occurs. Thus, events can be overlapping or mutually exclusive (disjoint). For the coin toss, the head and tail outcomes are mutually exclusive, that is, when we toss the coin it *always* comes up one way or the other, not both (we are not considering the possibility that it lands on edge – we will return to this later). With mutually exclusive events, there is no overlap. Because these events cover all possible outcomes, the probabilities must add up to 1.0, that is, one of them certainly happens.

In the case of the coin toss event space, the requirement that the sum of all probabilities must be 1.0 is expressed in Equation 3.2. Here, x is the event where the coin landed head up, and $\neg x$ where it landed head down (i.e., *not* head up).

$$P(x) + P(\neg x) = 1.0 \tag{3.2}$$

The coin toss is an example of a Boolean sample space – it has only two elementary outcomes. Biomedical examples abound: five-year survival in cancer patients, presence or absence of an epidemic outbreak, presence or absence of diseases in organisms, etc. When there are n possible (mutually exclusive) outcomes, x_i, Equation 3.2 generalizes to Equation 3.3. An example would be the selection of a colored ball from a bowl containing many balls, each of which is colored with one of n different colors. In this example, each of the x_i is one of the events, that is, that a particular colored ball was drawn out. A biological example is the particular base that appears at a specified location in the genome of an organism. In that case, $n = 4$, because there are four possible bases that could appear at any point in a DNA sequence.

$$\text{Axiom 2} \qquad \sum_{i=1}^{n} P(x_i) = 1.0 \tag{3.3}$$

Equation 3.3 is the *second* axiom of probability. It reflects the intuition that of all possible outcomes of an event, one of them certainly will occur.

Anything we measure or observe in biology, medicine, and public health can be formulated as a sample space, such as overall health state (Excellent, Good, Fair, Poor), identity of organism found in a blood culture or other bacterial culture (many possible states for this one), presence or absence of a point mutation in a particular place in a genome, immunization status of a child, etc. In addition to these *discrete* sample spaces, one can also have *continuous* observations, that is, the outcome is the measurement of a number (or numbers) that can vary continuously. Examples include medical laboratory test values, such as the creatinine clearance level, blood glucose level, percent oxygen saturation in blood, and rate of occurrence of new cases of a reportable disease (the number of cases is a discrete variable, but the rate may have a fractional component, and time is a continuous variable). Although we will briefly touch on continuous variables, most often the way a continuous variable is handled is to divide up its range into segments, which then become discrete. Typically, diagnostic tests are discretized in this way, so that a variable is assigned a "normal" range, and a test result is then classified as below normal, normal, or above normal.

In the case that we have multiple outcomes that are possible, one could have many subsets of outcomes, with overlapping elements, as well as individual outcomes. In the example of which base occurs at a location in a DNA sequence, there are four possible outcomes, which we will call G, C, A, and T, as usual. Recall from Chapter 1, Section 1.1.2, that the different bases can be grouped by how many hydrogen bonds they form with each other. The A and T bases form two hydrogen bonds together, while G and C bases form three hydrogen bonds. So, we might ask what is the probability of occurrence of either G or C, rather than each alone. In general, to fully specify a probability function, we need to specify what are the probabilities of every possible subset of outcomes. Fortunately, there is a third axiom that constrains this function, namely that the probability of the union of disjoint subsets is the sum of the probabilities of the individual subsets. If we have a series of events (which could be subsets of more than one element, as in the DNA example), which we will call e_1, e_2, etc. and they are each pairwise disjoint (which is equivalent to saying no element is a member of more than one subset), then

$$\text{Axiom 3} \qquad P(\cup_{i=1}^{k} e_i) = \sum_{i=1}^{k} P(e_i) \tag{3.4}$$

Equation 3.4 is the third axiom of probability. It can also be expressed in a slightly different way, that the probability of an event consisting of a subset of *elementary* events, that is, single outcomes, is the sum of the probabilities of the individual outcomes.

When we consider the probability of two sets of outcomes occurring, it may be that the two sets overlap, or they may be disjoint. In the case that they are disjoint, for example, the probability of a base being either A or T and the probability of it being G or C, it is easy to see that the probability of the union of the two sets is the sum of the probabilities. If the two overlap, for example, the set of G and C (the triple bonding pairs) and the purines (those are A and G), the sum rule does not apply. If we add the probabilities of the two subsets, we will be counting the probability of occurrence of the overlapping part twice. So, denoting the two subsets as S_1 and S_2, the relation is instead,

$$P(S_1 \cup S_2) = P(S_1) + P(S_2) - P(S_1 \cap S_2), \tag{3.5}$$

where we have subtracted the overlap. This can be easily proved from the three axioms (see [344, page 6]).

In medical diagnosis, it is common for a patient to have multiple symptoms and possibly multiple diagnoses. These can occur separately or together. The sample space in cases like this is the *product* space of the individual outcomes. This means that each element of the sample space is a combination of outcomes for each of the separate measurements or outcomes. A medical example would be the problem of diagnosing pneumonia. One set of outcomes is the presence or absence of pneumonia. Another is whether the chest X-ray had positive findings or was negative. Yet another is whether the patient was experiencing cough or not. So, the sample space contains eight elementary events, as follows:

$$\begin{aligned}
x_1 &= \neg\text{pneum} \wedge \neg\text{xray} \wedge \neg\text{cough} \\
x_2 &= \neg\text{pneum} \wedge \neg\text{xray} \wedge \text{cough} \\
x_3 &= \neg\text{pneum} \wedge \text{xray} \wedge \neg\text{cough} \\
x_4 &= \neg\text{pneum} \wedge \text{xray} \wedge \text{cough} \\
x_5 &= \text{pneum} \wedge \neg\text{xray} \wedge \neg\text{cough} \\
x_6 &= \text{pneum} \wedge \neg\text{xray} \wedge \text{cough} \\
x_7 &= \text{pneum} \wedge \text{xray} \wedge \neg\text{cough} \\
x_8 &= \text{pneum} \wedge \text{xray} \wedge \text{cough}.
\end{aligned}$$

Note that the event corresponding to the occurrence of one of the three items without considering the others is a subset of the above. So, the event symbolized by xray is the subset consisting of x_3, x_4, x_7, and x_8, and its probability would be the sum of the probabilities of those elementary events.

3.1.4 Conditional Probability

When multiple events occur, we can also express relations between them by measuring or computing the *conditional* probability that one event will occur given that another has already been determined to occur. The conditional probability that event x will occur, given that event y has already occurred, is denoted by $p(x|y)$.

The relations among the conditional probabilities, the individual probabilities, and the probability of both events is expressed in Equations 3.6 and 3.7. These equations provide a *definition* of conditional probability.

$$P(x|y) = \frac{P(x \cap y)}{P(y)} \tag{3.6}$$

$$P(y|x) = \frac{P(y \cap x)}{P(x)}. \tag{3.7}$$

Several facts about Equations 3.6 and 3.7 should be noted. First, the definitions only apply when $P(y)$ and $P(x)$ are not zero. This is reasonable: if event y cannot occur, it does not make sense to talk about the probability of another event, given that y has occurred, and so on. Second, if x and y are disjoint, then they do not occur together and $P(x|y)$ should be 0, which is consistent with the equation. Finally, the two conditional probabilities are *not* equal in general, but since $P(x \cap y) = P(y \cap x)$, one can eliminate it from the two equations, giving a relation between the two conditional probabilities.

$$P(x|y)\,P(y) = P(y|x)\,P(x) \tag{3.8}$$

Then, we can solve the equation for one of the conditional probabilities in terms of the other.

$$P(x|y) = \frac{P(y|x)\,P(x)}{P(y)} \tag{3.9}$$

This formula is known as Bayes' rule or Bayes' theorem, named after Rev. Thomas Bayes (1702–1761) who first derived it. We will use this formula extensively to support *probabilistic reasoning*.

One thing that can be done easily is to generalize this to the case of an arbitrary number of events, x_i. The joint probability for all the events, $P(x_1 \cap \ldots \cap x_{n-1} \cap x_n)$, is related by a kind of chain rule to the conditional probabilities, as follows:

$$P(x_1 \cap \ldots \cap x_{n-1} \cap x_n) = P(x_1|x_2 \cap \ldots \cap x_n)\,P(x_2|x_3 \cap \ldots \cap x_n) \ldots P(x_{n-1}|x_n)\,P(x_n) \tag{3.10}$$

This can be derived from the definition (generalized) of conditional probability. However, $2^n - 1$ numbers are still required (the one constraint remains that they all add up to 1). Later, we will see that the requirements for determining the full joint probability can be drastically reduced in situations where the events and their relationships satisfy additional constraints.

Conditional probabilities are often used to express causal relations between events, that is, when y occurs we expect most of the time that x will occur also, through some causal mechanism. However, we can define conditional probabilities whether or not we believe there is a causal relation between the two variables. In fact, the formulas (3.6 and 3.7) are symmetric with respect to x and y (although the two conditional probabilities are not equal) so there is no way to infer causality from the formulas or from raw data. Instead, we construct a causal theory to test it, and if its predictions match the data, where no other theory does as well, we may accept the causal relation provisionally as true.[2]

When we hypothesize that y *causes* x, the conditional probability $P(x|y)$ is usually closer to 1.0 than to 0.0. The fact that it is not exactly 1.0 accounts for the non-deterministic nature of this causal relationship. For example, the presence of certain kinds of pollen usually causes allergy symptoms, but it is possible for the pollen to be present without the symptoms, even in the same individual.

2. Often, there is only anecdotal data to support the theory, but for many people this is tantalizingly appealing, leading to the often used expression, "Coincidence? I think not."

3.1.5 Independence

Another idea concerning events is the idea of *independence*. If two events are independent, the probability that both occur is simply the product of the individual probabilities. The occurrence of one has no effect on the probability of occurrence of the other, in some cases. This is a *definition*, and it is stated mathematically in Equation 3.11.

$$P(x \cap y) = P(x)P(y) \tag{3.11}$$

In general, one can only determine if two events are independent by doing the calculations (if possible). Alternatively, one can posit the independence of two events as a statement about the biological or other process you are observing and modeling. In the pneumonia example mentioned above, we would expect that the three phenomena, presence of pneumonia, positive X-ray, and cough are not independent. However, there are phenomena we believe really are independent, and we could build a statement to that effect (some form of Equation 3.11) into our biomedical domain theory. Then, we would derive consequences from that assumption.

The one condition under which you can determine if two events are independent or not is the situation where the two events are disjoint. In that case, they cannot be independent, since the occurrence of one means that the other could not occur. Events being disjoint means that $P(x \cap y)$ is 0.0, but each of the individual probabilities are not 0, so their product is not 0, and Equation 3.11 will not be satisfied.

One further fact to note is that independence does *not* mean that either of the conditional probabilities are 0. In fact just the opposite, if either of the conditional probabilities *are* 0, the events are disjoint, because then the probability of both occurring must be 0. Thus, the two events are *not* independent. If the conditional probabilities are *not* 0, the events might or might not be independent. If they are independent, combining Equations 3.11 and 3.6 (for example) gives

$$P(x|y) = P(x) \tag{3.12}$$

$$P(y|x) = P(y) \tag{3.13}$$

that is, the conditional probability of an event is independent of the other event.

It may happen that an event is not independent of another event in general, but given that a third event happens, the first two may become independent. Stated in terms of conditional probabilities, for events x, y, and z, we say that x and y are *conditionally independent* given z if any one of the following conditions holds:

1. $P(x|z) = 0$
2. $P(y|z) = 0$
3. $P(x|y \cap z) = P(x|z)$ *and* $P(x|z) \neq 0$ *and* $P(y|z) \neq 0$

This is a mathematical property of the probabilities, not something that has physical or biological significance. Its importance in the process of probabilistic reasoning is that a problem in which this condition is satisfied will be enormously simplified in terms of the number of data we need to calculate anything. In practice, one usually assumes conditional independence where it seems appropriate to the phenomena being modeled, rather than examining the data to determine it because often there is not sufficient data, and the idea is to use what we know to create a predictive model.

Thus, the biological or medical or health phenomenon that the probabilities represent may provide justification for claiming that events are conditionally independent. In the example of pneumonia, we would expect that there is a causal relation between the presence of pneumonia and a positive finding on a chest X-ray, as well as one between the presence of pneumonia and the presence of a cough. Without knowing whether pneumonia is present or not, there may well be a correlation between cough and positive X-ray. The probability of a positive X-ray given cough likely will not be the same as the probability of a positive X-ray in general. However, once it is known that pneumonia is present, the probability of the positive X-ray likely *will* be the same, whether or not cough is present, that is, positive X-ray and cough would be conditionally independent given the presence of pneumonia.

So, once again, we see that the mathematical and logical framework is the language in which we will express our theories of biology, medicine, and public health processes, but the particular shape these theories take must come from intuition, not just automatically from computations on data.

3.1.6 Random Variables and Estimation

Up to now we have been discussing the properties of probability values in relation to events, subsets of a sample space. When our sample space consists of a composite involving a number of (possibly related) observations, as in the pneumonia example, the sample space as a whole is difficult to describe. In addition, in many real-world applications, we do not have an easy way to define the sample space, but we can identify some property or properties of the sample space, perhaps subsets that can be well defined in terms of the values assigned to some variable. Most often in clinical trials or in biological experiments, we define variables, and we are interested in how one variable might depend on the other. For example, a biologist is interested in the possible effect of a gene function on metabolism of a certain cell type. She prepares cells with the gene present and cells without the gene and observes whether there is any difference in the observable processes in the two cell groups. The presence or absence of the gene is a binary or Boolean variable, and there is some other variable measuring metabolic function of the cells with perhaps many possible values. The events being observed are complex, but the subsets defined by the variable values are what is of interest, not the details behind them.

So, the idea of a *random variable* is that it provides an abstraction layer to some extent, over the sample space. Given a sample space and a probability function, the definition is as follows: a random variable is a function that assigns a real number to each element of the sample space. In mathematical notation, if ω is the sample space, R is the space of real numbers, and X is a random variable, then X is a function or mapping, as follows:

$$X : \omega \rightarrow \mathrm{R} \tag{3.14}$$

There could be many different mappings of the same sample space, corresponding to the fact that in an experiment, there are many ways to group the events or assign a measurement to an event. An example is the staging of tumors. When a patient comes to the cancer clinic for radiation treatment, the radiation oncologist classifies the tumor according to its anatomic site, size, cell type, whether the cells look close to normal or have very deformed nuclei, etc. From this information, they assign a stage, typically a number between 1 and 4. The data collected in a clinical trial are organized according to staging. The elementary events are the conditions of the individual patients, but in this case, stage would be a random variable on which the data analysis would be based. Questions to ask would be: how does the rate of tumor ablation (or even cure) depend on stage, for a given treatment regimen, or how does the recurrence rate depend on stage.

Probability Distribution Functions

Since the random variable assigns values to events in the sample space and we have assumed we have a probability function for the sample space, we can define *distribution functions* for the random variable.

First, we define the *cumulative distribution function* (CDF). It is a function F whose domain is the real numbers and whose range is the interval [0,1]. The value of $F(x)$ will be the probability of the event consisting of all the points in the sample space for which the random variable X is less than or equal to x, that is,

$$F(x) = P(X \leq x), \tag{3.15}$$

where we have written $X \leq x$ to denote the subset of the sample space for which the value of X is less than or equal to x. This is well defined, since a random variable assigns real numbers to points in the sample space, so any real number will define a subset of the sample space. It could be the empty set if the random variable maps no outcomes in the sample space to a particular real number or range of real

numbers. Every such subset has some probability assigned, so the function P has a value, and it will be between 0 and 1, by Axiom 1.

The CDF has some useful properties that are easy to see. It is non-decreasing, since as x increases, more of the sample space is included and so the probability either stays the same or increases. As x tends to $-\infty$ the value of $F(x)$ tends to 0, since less and less of the sample space is included. As x tends to ∞ the value of $F(x)$ tends to 1, since eventually all of the sample space ω is included and $p(\omega)$ is defined to be 1.

We define two other functions for X, depending on whether X is *discrete* or *continuous*. We say X is discrete if the set of possible values for X is discrete, that is, countable.[3] For a discrete random variable we can define the *probability mass function* (sometimes called just the *probability function*), as the probability for the event consisting of all sample space points for which the value of X is one of these discrete values, x_i, labeled by an index, i.

$$f(x_i) = P(X = x_i) \tag{3.16}$$

The CDF is related to the probability mass function in a simple way in this case. The CDF for a value x can be obtained from f by adding up all the values of $f(x_i)$ for x_i up to and including x.

$$F(x) = \sum_{x_i \leq x} f(x_i) \tag{3.17}$$

In the case of a continuous random variable, the possible values will vary in some range, possibly the entire real line from $-\infty$ to ∞. To have a continuous random variable, there needs to exist a function called the *probability density function* (PDF), $f(x)$, satisfying three conditions, analogous to the probability mass function.

1. for all x, $f(x) \geq 0$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$
3. for all real numbers a and b, where $a \leq b$, the probability that x is in the range from a to b is $P(X \in (a, b)) = \int_a^b f(x)dx$

It follows from this definition that the CDF for a continuous random variable is just the probability of being in the range from $-\infty$ to x, or

$$F(x) = \int_{-\infty}^{x} f(t)dt. \tag{3.18}$$

Note that the PDF, $f(x)$, is *not* the probability of the random variable X having the value x. In fact, that probability is 0. Unlike the probability mass function, the PDF can have values much greater than 1 and can even be infinite. What matters is its integral over some finite range.

Most of the time, we write statistical formulas in terms of the random variables, and we forget about the underlying sample space. In general, we are going to be writing formulas in terms of the probability mass function or density function.

Several probability distribution functions have become standard and widely used. Which one is applicable to describe a particular biomedical situation is a modeling choice that may be motivated by what you believe or hypothesize about the variables you are modeling. Here, we just describe a few of the most common.

One of the simplest discrete distributions is the *Uniform* distribution. In the discrete case, the probability mass function describes a variable that takes on one of some fixed number, n, of values, all with equal probability. If x is one of the values, $1, \ldots, n$, then $f(x) = 1/n$, and it is 0 otherwise. The analog for the continuous case is similar, except that instead of a fixed *number* of values, the variable

3. A set is countable if it is finite, or if it is infinite but able to be mapped one to one to the positive integers.

takes on a fixed *range* of values, from a to b inclusive. In this case, if $x \in [a, b]$, then the PDF takes on the constant value, $f(x) = 1/(b - a)$, and it is 0 otherwise.

A popular contraption on display at the Pacific Science Center in Seattle, Washington, is a large box, tall and wide but shallow in depth, with horizontal pegs set out in a vertical triangular array, equally spaced, with apex at the top, so that in each succeeding row, each peg in that row is just under the midpoint between the two upper pegs, or in the case of the end pegs arranged so that the peg above is above the midpoint of the two pegs just below. At the top, a conveyor drops hard wooden balls one at a time on the top peg. Each ball cascades down through the peg array to the bottom where there are (a discrete set of) bins to collect the balls. Figure 3.1 illustrates this schematically.

Roughly, the probability of going to the right or left when a ball hits a peg is 1/2, and the distribution at the bottom, the probability of a ball landing in a particular bin, is a special case of the *binomial distribution*. In general, the probability of one or the other choice at each point can vary from 0 to 1. For n levels and probability p at each choice point (or peg, in the illustration), the probability mass function is

$$f(x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

for $x = 0, \ldots, n$ and 0 otherwise. The symbol $\binom{n}{x}$ is the *binomial coefficient*,

$$\binom{n}{x} = \frac{n!}{x!(n - x)!}$$

In this formula, n and p are parameters of the distribution function. Most distribution functions are specified with parameters. One of the goals of statistics is to provide methods for estimating the parameters of a distribution from measured or observed data, where you have hypothesized that the thing being observed is modeled by the distribution you have chosen.

There is a continuous analog of the binomial distribution, called the *Normal*, or *Gaussian*, distribution. The PDF for the Normal distribtion is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x - \mu)^2}{2\sigma^2}\right\}, \tag{3.19}$$

where μ can have any value, and σ has to be positive. This function has its maximum at $x = \mu$, and it is symmetric with respect to $x = \mu$. It is the well known "bell-shaped curve." When μ is 0 and σ is 1, it is called the *standard Normal distribution*. Applying Equation 3.18 would give the CDF corresponding to the Normal distribution, but this integral cannot be done in closed form. Its shape, however, is well known, a sigmoid curve, starting from 0 at $-\infty$, rising quickly in the vicinity of μ and then asymptotically approaching 1. The CDF has a value of 0.5 at $x = \mu$, and is also symmetric about $x = \mu$. The Normal distribution is important not only because it has some nice mathematical properties,

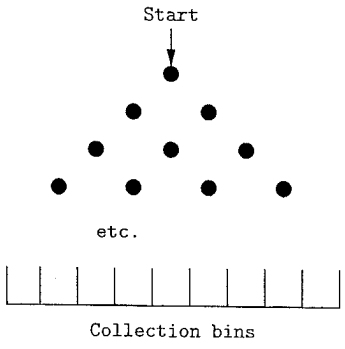


FIGURE 3.1 A contraption that illustrates the binomial distribution, for the special case that $p = 0.5$.

but also because of the Central Limit Theorem, which says essentially that as we make larger and larger numbers of observations, the probability distribution of the mean of all the observations behaves more and more like a Normal distribution.

Expectation and Variance

The expectation of a random variable provides some information about the distribution function for that random variable. If we know or hypothesize what the form of the distribution function is, for example, one of the standard well-known ones, there will be a relation between the parameter or parameters that specify the distribution function and the expectation. So, knowing the expectation we could recover the parameters and thus the full distribution function. Thus estimating the expectation from actual observations is very important.

The expectation of a random variable is analogous to computing an average. However, in general, the different values that the random variable can take are not all equally probable. If we were to make a series of observations and they occurred with frequencies corresponding to the probabilities of each, when we average our observed values, we will expect to get a *weighted* average, not the midpoint of the range of possible values. Thus, we define the *expectation* of a random variable, X, as the weighted average, where we multiply each value by the probability of it occurring. The following formula applies to a discrete random variable, X,

$$E(X) = \sum_{i=1}^{n} x_i f(x_i) \tag{3.20}$$

where $f(x_i)$ is the value of the probability mass function at $X = x_i$.

For a continuous random variable, we would use the PDF, and do an integral instead of a sum.

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx \tag{3.21}$$

For the example distributions described here, the expectations can be computed in terms of the parameters. For the Uniform distribution, in the discrete case, it is $n/2$ and in the continuous case for the interval $[a, b]$, the expectation is simply $(a + b)/2$, the midpoint of the interval. For the Binomial distribution, it is np, and for the Normal distribution, it is μ.

A reasonable estimate of the expectation of a discrete random variable is the *mean* of a collection of observational values. The observational values will (in a large enough sample) occur with repetitions, and a frequency histogram of these values should approximate the probability distribution.

$$\mu = \frac{1}{n} \sum_{i=1}^{n} M_i \tag{3.22}$$

Here, we have used μ to denote the estimated value of the expectation. This (mean value) sometimes is written $<M>$. The Law of Large Numbers is an important theorem in statistics, which says that if we treat each observation as a sample of a random variable, and thus the mean is also a sample of a random variable, as the number of observations gets larger and larger, the probability distribution for the mean as a random variable gets more and more concentrated around the expectation of the distribution. This does not say anything about the shape of the distribution; the shape is addressed by the Central Limit Theorem, previously mentioned.

In addition to the mean, it will be useful to compute the expected value (EV) of functions of a random variable, for example, polynomials or other functions. They are computed in the same way, namely, if $Y = g(X)$, we compute the expectation of Y by

$$E(Y) = \int_{-\infty}^{\infty} g(x) f(x) dx \tag{3.23}$$

One such function, the *variance*, is especially useful. The variance, σ^2, of a distribution is a measure of how wide or narrow it is. It is the expectation of the square of the deviation of the random variable value from the mean. For a discrete random variable, it is

$$\sigma^2 = \sum_{i=1}^{n} (x_i - \mu)^2 f(x_i) \tag{3.24}$$

and for a continuous variable, it is

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx \tag{3.25}$$

We can also get an *estimate* of the variance of a distribution, by averaging the squared deviation of the measurements or observations from the mean value. We compute the squared deviation because the deviation can be in either direction, positive or negative, and the average deviation will be 0.[4] This average squared deviation is also called the variance, and its square root is called the *standard deviation*. We take the square root to get a number whose measurement units are the same as the measurements themselves, and the number then can be used as an indicator of how far from the mean the measurements typically will be found.

$$\sigma^2_{\text{est}} = \frac{1}{n} \sum_{i=1}^{n} (M_i - \mu)^2 \tag{3.26}$$

and of course $\sigma = \sqrt{\sigma^2}$. We can code this pretty simply for later use, for example, in a syndromic surveillance system. For simplicity we represent the measurements as an array, `m`, and remember that in computer programs, arrays are zero-indexed, while in mathematical expressions, sequences are usually one-indexed. In the variance function, `mu` is the computed mean, μ, and the variance function returns both the variance σ^2 and the standard deviation σ as multiple values.

Knowing how to do iteration, one might be inclined to write the expressions for mean and variance as `do` constructs. The mean, for example, would look like this:

```
(defun mean (m)
  (let ((n (1- (array-dimension m 0))))
    (/ (do ((i 0 (1+ i))
            (stop n)
            (result 0))
           ((> i stop) result)
         (incf result (aref m i)))
       n)))
```

Expressions with sums and products occur often in probability and statistics. Rather than having to write out these `do` loops, it would be nice to have `sum` and `product` operators that can take expressions and transform them into iterative code as above. The Common Lisp standard does not include such operators, but we can write macros to do these jobs. Following an example in Graham,[5] we can define these operators, and then use them as if they were built-in. They follow the form of the `for` macro in Graham, but each does the sum or product operation as part of the definition, so that makes their use simpler even than using the `for` macro. Here is an implementation of the `sum` macro, which evaluates the body of code that is input, for each value of the symbol passed in as `var` from the specified `start`

4. It is simple to prove this, and it is left as an exercise for the reader.

5. We introduced the basic idea of how to write a macro with `defmacro` in Chapter 2, on page 114. A brief introduction to macros can also be found in Graham [94, Chapter 10].

value to the stop value, inclusive, summing up the results of evaluating body each time through as a progn.

```
(defmacro sum (var start stop &rest body)
  (let ((gstop (gensym))
        (gresult (gensym)))
    `(do ((,var ,start (1+ ,var))
          (,gstop ,stop)
          (,gresult 0))
         ((> ,var ,gstop) ,gresult)
       (incf ,gresult
             (progn ,@body)))))
```

The let form creates some unique local variables using gensym, and then uses them in the iteration. The symbols gstop and gresult are not literally used. Their *values* (the symbols created by gensym) are the symbols (variables) actually used in the do expression that the macro creates when it is called. Those are initialized to the values passed in with the macro call. This is to avoid the possibility that any particular name we might choose there would be also in the body code that the user of this macro provides. That is called variable capture and it is a source of subtle errors. On the other hand, the value of the first parameter, var, is exactly the name of a variable used for the iteration, and it is intended that it appear in the body.

When this new definition is used, it will be translated by macro expansion into an expression very similar to the first implementation above. In using this macro the symbol i above is the value of var in the macro definition, the number 0 becomes the value of start, and the array-dimension function is used to get the stop value (one less than the size of the array). The result of macro expansion of a sum expression, as it would be used in the mean function, can be seen by using the Common Lisp function macroexpand-1. Its input is the expression as we would code it, and the transformed code is shown following, in all uppercase.

```
> (macroexpand-1
    '(sum i 0 n (aref m i)))

(DO ((I 0 (1+ I))
     (#:G16 N)
     (#:G17 0))
    ((> I #:G16) #:G17)
  (INCF #:G17
    (PROGN
      (AREF M I))))
```

The symbol G16 corresponds to stop in the original code, and G17 corresponds to result in the original code. These symbols are unique symbols that were created automatically by the macro expansion process. The result of the macro expansion is pretty much the same as the original do expression that we wrote. So, having this macro, we can just *use* it instead of writing out complicated do forms. The macro expander translates our code into the more complicated version for us. Although the body of this example is just the aref expression, the body of the macro call can be any sequence of expressions, as if the sum macro were just built into Common Lisp. So, the mean and variance functions can be coded very simply and readably.

```
(defun mean (m)
  (let ((n (1- (array-dimension m 0))))
    (/ (sum i 0 n (aref m i))
       n)))
```

```
(defun variance (m mu)
  (let* ((n (1- (array-dimension m 0)))
         (sigma-squared
           (/ (sum i 0 n (expt (- (aref m i) mu) 2))
              n)))
    (values sigma-squared (sqrt sigma-squared))))
```

Simple algebra can be used to derive an alternate form of the formula for variance. The squared expression is expanded, then you can use Equation 3.22 to simplify the result.

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} M_i^2 - \mu^2 \tag{3.27}$$

The first term is just the average of the squared measurements, $<M^2>$. The second term is the square of the mean, $<M>$, so the formula becomes the difference between the mean of the square and the square of the mean.

$$\sigma^2 = <M^2> - <M>^2 \tag{3.28}$$

When handling large data sets, some attention must be given to efficiency of the arithmetic. In this case, we have two forms for the same computation. For a given n, one may be faster or they may be about the same. This is not a question of whether the sum macro is faster than the handwritten do form because they end up the same code. It is about which algebraic formula involves fewer operations. It is left as an exercise for the reader to determine which is computationally cheaper.

Multiple Random Variables: The Joint Probability Distribution

In general, a domain has multiple random variables, not just one. The pneumonia example illustrates this situation. The sample space is a Cartesian product of the individual sample spaces. Instead of a single random variable we model the situation with three random variables, and the distribution functions are functions of multiple variables.

- An *atomic event* is an assignment of particular values to all the variables, $X_1, \ldots X_n$.
- The table of values is the joint probability distribution $P(X_1, \ldots X_n)$.
- Often, each variable X_i is a Boolean, but this is not required.

An example of a simple joint probability distribution with just two variables, High PSA and Prostate Cancer, is shown in Table 3.1. A very common test that is used for screening for prostate cancer is the PSA test. PSA stands for "Prostate-Specific Antigen," a protein that may be present in blood in very small concentrations, but in the presence of prostate cancer the concentration is higher. A PSA below 4 ng per milliliter of blood serum is considered normal. A PSA between 4 and 10 that is rising over time is a good indicator of the need for more extensive investigation. Above 10, an even higher level of concern is appropriate, but for our purposes we just model PSA as a Boolean event, with the two states, normal and high (above 4). The joint probability distribution table shows the probability for each combination of values of each of the variables. In general, for n Boolean variables, the joint probability distribution will have 2^n entries. For variables that can have more than two values, the table gets correspondingly larger. The size of the table is the product of the number of values each random variable can have.

There are some constraints on the values in the joint probability distribution table. First, since every entry in the table is the probability of one of the possible elementary outcomes, and the table covers all the outcomes, the sum of all the numbers in the table has to add up to 1.0. Table 3.1 satisfies this

TABLE 3.1 A Simple Joint Probability Distribution for Prostate Cancer and High PSA

	High PSA	¬ High PSA
Prostate Cancer	0.25	0.07
¬Prostate Cancer	0.03	0.65

requirement.[6] Second, we expect to see some dependence between events in the table, that is, if the PSA test has diagnostic use, there should be some correlation between the presence of prostate cancer and a positive test, and vice versa, between the absence of prostate cancer and a negative test. This also appears to be the case in Table 3.1.

The PSA test is typical in medical practice in that it is not *pathognomonic*. A *pathognomonic* finding unequivocally signals the presence of a particular cause. One of the most common pathognomonic findings is the presence of two independent heart beat sounds in a woman. For sure, this indicates that the woman is pregnant.[7] Most tests have some level of accuracy that is well known. However, the notion of "accuracy" needs to have some precision as it can easily be misunderstood.

To describe the accuracy of a test, we need two numbers, not just one. First, the *sensitivity* of a test is defined to be the probability that a patient will test positive, given that the disease is present. This is the *true positive rate* relative to the disease positive population. It is the true positives divided by the total patients with the disease (the true positives plus the false negatives). In the prostate cancer example above, it is $0.25/(0.25 + 0.07)$ or 0.78. The sensitivity tells you something about how well the test indicates the presence of the disease, but it does not tell you anything about the behavior of the test in the absence of the disease, that is, the false positives. For that, the *specificity* is defined to be the *true negative rate*, that is, the probability of a negative test when the patient does not have the disease. The specificity is computed by dividing the true negatives by the sum of the true negatives and false positives. For the prostate cancer example, it is 0.96 so it seems that the test is pretty specific.

Typically, we do not know the joint probability distribution function but we do know the sensitivity and specificity of a test. Since there is one numeric constraint on the table values (they have to add up to 1.0), if we had one more number, we could compute the table. That number is the overall rate at which the disease occurs in a particular population. This is called the *prevalence* or prior probability. The sum of the numbers in the row representing presence of prostate cancer should be the probabilility overall of occurrence of prostate cancer in some specific population (of men), the population used to generate the numbers in the table. From Table 3.1, this is about 0.32, meaning that we expect that 32% of the men in that group will experience prostate cancer. In the next section, we will take another look at how to use Bayes' rule in this and in more complex problems.

3.2 APPLICATION AND GENERALIZATION OF BAYES' RULE

Bayes' rule is the basis for modeling a large number of complex phenomena from a simple causal relation between two variables to a large and complex network of relationships among events. The events and the causal relationships among them form a graph with the events as nodes in the graph and

6. The values in this table are *not* exactly the values found in the medical literature. In reality, there is wide variability in the definition and interpretation of PSA test results.

7. This does not mean that the various chemical pregnancy tests are unnecessary. The pregnancy is well along by the time a fetal heart beat can be detected, and it may well be useful to know much sooner.

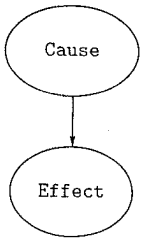


FIGURE 3.2 A simple causal relationship diagram.

the relationships as directed arcs or edges connecting the nodes. The simplest possible graph is one with two nodes linked by a single arc, as shown in Figure 3.2.

The PSA test for prostate cancer is an example of this kind of relationship. It is a practical application rather than a scientific hypothesis. It is known that high PSA levels are associated with prostate tumors, so we can say that there is a causal link with prostate cancer being the cause and the high PSA state being the effect. This is a case where we observe the positive test result and would like to draw some inference about the probability of the presence of prostate cancer.

In a rather different kind of application, we may wish to use the graph as a predictor of the effect, rather than a way to ascertain the cause. An example with many nodes and links is Figure 3.3, which is an example of a Bayes Net for modeling the prognosis of prostate cancer. In this example, it is already known that the patient has prostate cancer and we wish to estimate (predict) the probability of effective treatment, that is, disease control.

The prognosis is dependent on findings about the disease, as well as treatment parameters (radiation dose to the tumor) and the lymph nodes at risk for metastasis. Control of prostate cancer is dependent on controlling the local disease – the tumor cells within the prostate itself – and any cells which have migrated to the nearby lymph nodes. The PSA level, Gleason score, and tumor stage are three clinical tests which describe the extent of the disease, and along with radiation dose, are predictive of cure rate. The nodes "Local Tumor Control" and "Lymph Node Control" each have states "yes" and "no." The node "Disease Control" reflects the fact that an individual whose lymph nodes are not cured has a 90% chance of further spread, regardless of the state of "Local Tumor Control," while if "Lymph Node

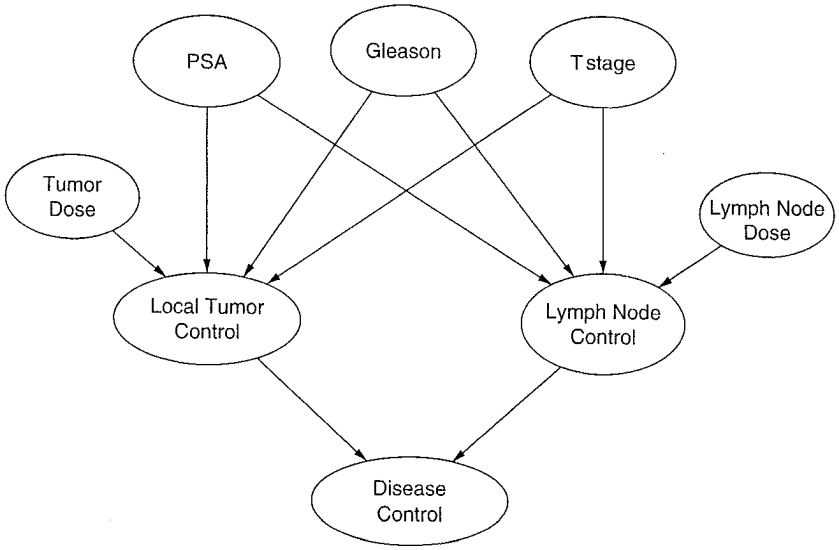


FIGURE 3.3 A Bayes net for computing a prognosis for radiotherapy treatment of prostate cancer (courtesy of Mark Phillips and Wade Smith).

Control" is "yes," only 1/3 of patients for whom "Local Tumor Control" is "no" will develop distant metastasis.

So the first consideration is how to use Bayes' rule to compute one conditional probability when the other is known. The example is for a Boolean sample space, but the basic Bayes formula easily generalizes to the case of a sample space like the DNA bases, where there are many mutually exclusive possible events. Then, we examine the properties of the larger networks or graphs, called Bayes nets. The particular model shown in Figure 3.3 will be described in more detail in Chapter 9 along with other applications of the ideas in Part I to radiation therapy.

3.2.1 Simple Bayesian Inference

In most circumstances, the joint probability distribution table is not available. However, the prostate cancer example illustrates a common situation in medicine. We have some test for the presence or absence of a disease. Usually what is known about the test is its sensitivity and specificity, collectively called the *accuracy* of the test. One might think if the test is 90% accurate, and you test positive, the probability that you have the disease is 90%. This is incorrect. The 90%, or probability of 0.9, is the probability of testing positive given that you have the disease, not the probability that you have the disease, given that you tested positive. The sensitivity and specificity are the conditional probabilities, respectively, of testing positive given that you have the disease, and of testing negative, given that you do not have the disease. These numbers are *not* the same as the probabilities we are really interested in, namely the probability that you have the disease given that you test positive, or, conversely, the probability that you do not have the disease given that you test negative.

Fortunately, Bayes' rule (Equation 3.9) provides a way to compute one conditional probability from the other. In the prostate cancer case, let us suppose that we start with the following information.

- The PSA test for prostate cancer has a sensitivity of 0.78 and a specificity of 0.96, which would seem pretty useful.
- The overall rate of occurrence of prostate cancer in the male population being tested is 0.32 as mentioned earlier.

From this information, Bayes' rule provides a way to compute the probability of having the disease given a positive test. However, we need one more piece of information, the overall rate of occurrence of positive PSA tests, which is the denominator in Equation 3.9. We do *not* need to go back to the joint probability distribution table. This quantity is related to the other quantities we already know. The event of positive PSA is the union of two things, the event of positive PSA *and* having prostate cancer, and the event of positive PSA *and not* having prostate cancer. Since those two are disjoint, the probability of positive PSA is the sum of these two probabilities. Using the definition of conditional probability, we can write them in terms of conditional probabilities, giving

$$P(\text{psa}) = P(\text{psa}|\text{pc})\,P(\text{pc}) + P(\text{psa}|\neg\text{pc})\,P(\neg\text{pc}) \tag{3.29}$$

In Equation 3.29, psa is the event of having a positive PSA test, and pc is the event of having prostate cancer. More generally, the formula is

$$P(y) = P(y|x)\,P(x) + P(y|\neg x)\,P(\neg x), \tag{3.30}$$

where x and y are events (in the prostate cancer case, x is pc and y is psa). However, we *still* do not have all the numbers. In particular, we do not have $P(y|\neg x)$ or $P(\neg x)$. These are related to the numbers we do have because Axiom 1 says that $P(\neg x)$ is just $1 - P(x)$, and similarly, $P(y|\neg x)$ is $1 - P(\neg y|\neg x)$. We know $P(\neg y|\neg x)$, which is the specificity. So, the final version of Bayes' rule becomes

$$P(x|y) = \frac{P(y|x)\,P(x)}{P(y|x)\,P(x) + [1 - P(\neg y|\neg x)][1 - P(x)]} \tag{3.31}$$

This equation now uses only the three numbers we typically would have available to us. So, doing the arithmetic for the prostate problem gives

$$\frac{0.78 \times 0.32}{(0.78 \times 0.32) + (0.04 \times 0.68)} = 0.90, \tag{3.32}$$

and we can conclude that the probability of prostate cancer given a positive test is 0.90, not 0.78. So, it really pays to do the arithmetic, as the answer is significantly different from the naive (and incorrect) answer.

Now, let us suppose the test is much more accurate, but the disease we are testing for is very much more rare, that is, the numbers are 0.99 for both sensitivity and specificity, but the disease prevalence is 0.0001 (1 in 10,0000). Then, the calculation gives, for a positive test result,

$$\frac{0.99 \times 0.0001}{(0.99 \times 0.0001) + (0.01 \times 0.9999)} = 0.0098 \tag{3.33}$$

This is dramatically different. The test is 99% accurate, but the probability of having the disease if the test is positive is only about 1%. This result comes about because the disease is so rare that the occurrence of false positives overwhelmingly obscures the true positives.

Even more important, note that the sensitivity and specificity are properties of the relation between the test and the disease, and they do not change from one population to the other, but the prevalence may vary greatly from one population to another, and this will strongly affect the interpretation of test results.

Where do the numbers come from? The accuracy of diagnostic tests is the subject of medical (clinical) research. It is part of the work of public health practitioners to collect data to estimate the prevalence of various important diseases, so for many diseases also the overall prevalence or occurrence rate is known.

Implementing a formula like Equation 3.31 is straightforward in most computer programming languages. Here is a version in Common Lisp.

```
(defun disease-given-test (sensitivity specificity prevalence)
  (/ (* sensitivity prevalence)
     (+ (* sensitivity prevalence)
        (* (- 1 specificity) (- 1 prevalence)))))
```

This is the simplest possible application. It is easy to generalize the formula to the case where there are many possible (mutually exclusive) cause states rather than just presence or absence of a disease.

3.2.2 Non-Boolean Variables

In some cases, the causal hypothesis is not a Boolean variable but can have any one of many mutually exclusive values, which we denote as c_i. We can generalize Bayes' rule to handle this case. For each causal condition c_i, we have a probability, $P(c_i)$. The conditional probabilities for effect e, given each value of the cause variable c_i, similarly are represented by a conditional probability, $P(e|c_i)$. For each possible value of the causal variable, we would know the conditional probability of a positive effect. We would also need to know the prevalence rate or occurrence rate of each causal value. From this, we can compute the probability of each of the causal values, given a positive effect.

This kind of situation occurs when a causal variable has continuous values but is typically divided up into discrete intervals for purposes of simplifying the calculations. An example of such a variable would be number of years of smoking, with the effect being any of the several diseases known to be linked with smoking (heart disease, *COPD*, emphysema, etc.). Rounding to whole numbers of years is a way to group the data in discrete bins, as is lumping still further into five-year intervals.

In this case, the prior probability of the effect can be computed as a sum over all the cause values, of the products of the prior and conditional probabilities, just as for the Boolean case where we summed over the two cases, true and false.

$$P(e) = \sum_i P(e|c_i)P(c_i) \tag{3.34}$$

Having a way to compute $P(e)$, we can now compute the conditional probability of each causal value given a single effect. This, together with Equation 3.34, is one form of a generalized Bayes' theorem:

$$P(c_i|e) = \frac{P(e|c_i)P(c_i)}{P(e)} \tag{3.35}$$

Now, instead of simple true and false, we have a collection of values and probabilities, indexed by an integer labeling the values. A straightforward way to implement this computation is to represent the probabilities as arrays indexed by the value number, i. Our function to compute $P(c_i|e)$ will have as inputs the index, `i`, the conditional probability array, which we will call `sensitivity` because its elements are the sensitivities of the effect to the occurrence of each causal value, and the array containing the prior probability of each causal value, which we will call `prevalence`. We will also need to implement Formula 3.34 as a function summing over products of elements of the two arrays. Now, we can see it was worth the trouble to define a `sum` operator because we can use it in the implementation of Formula 3.34. To be general, we would obtain the array dimension of the `prevalence` array to use as the iteration upper bound. Using the `sum` macro defined on page 197, the `effect-prob` function, implementing the formula for $P(e)$ looks like this:

```
(defun effect-prob (sensitivity prevalence)
  (sum i 0 (1- (array-dimension prevalence 0))
       (* (aref sensitivity i) (aref prevalence i))))
```

Formula 3.35 can now be coded very simply by using the function just defined for $P(e)$. Of course one could also write the body out in the following, but the layered design approach is likely to pay off as the formulas get further generalized.

```
(defun cond-cause (i sensitivity prevalence)
  (/ (* (aref sensitivity i) (aref prevalence i))
     (effect-prob sensitivity prevalence)))
```

You might wonder what happened to the problem of having to incorporate the specificity and the rule that probabilities add up to 1.0. They are present. The Boolean case is just the case of two causal factors, one the negative of the other. In the more general case, what was specificity, or the true negative rate, is now for each of the possible causal values the true positive rate for that value. So one could think of specificity in the Boolean case as the sensitivity for testing negative.

3.2.3 Bayes Nets

A more complex graph such as Figure 3.3 represents a much larger collection of variables, each with a number of states. The joint probability distribution for such a graph is large, since every combination of values for each variable needs an entry. For the graph in Figure 3.3, the Tumor Dose node has five states, the PSA node has six (unlike our simple binary PSA test), the Gleason score has four states, and T-stage has three. The rest each are Boolean nodes, so, altogether, the joint probability table for this set of variables would have $5 \times 6 \times 4 \times 3 \times 2 \times 2 \times 2 \times 2$ or 2,880 states. To glean that much information from clinical trials data or retrospective data would be simply impractical. The point of the graph is that we can reasonably hypothesize causal dependencies among the variables, and thus the graph represents

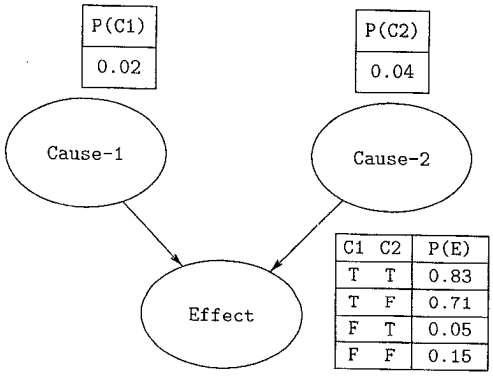


FIGURE 3.4 A graphical representation of an effect with two causal factors, showing the prior probabilities of the causes and the conditional probability for the effect given each combination of causes.

a kind of theory of prostate cancer, relating findings and prognosis. The arcs connecting the nodes represent causal or predictive links. The PSA, Gleason score, and T-stage are all predictive of local tumor control and lymph node control. Tumor dose (from radiation treatment) also influences local tumor control, and whether we irradiate the lymph nodes will have an influence on lymph node control.

In the case of diagnosis, we still needed three numbers to determine everything we would want to compute, though they are more convenient than the joint probability table numbers. In this case, however, provided certain conditions are met, we can compute anything we want starting with far less than the full joint probability table.

First, consider a node with only incoming arcs, as in Figure 3.4. Each of the three variables is Boolean. We would like to compute the probability of the effect given the presence or absence of the causes. This is actually another version of the non-Boolean case just considered.

Although `Cause-1` and `Cause-2` are not mutually exclusive, the set of combinations of values for each does form a set of mutually exclusive events. In this case, the events are the four combinations of presence (`T`) or absence (`F`) for the two causes, that is, the set `[TT,TF,FT,FF]`. So, for this graph, we can just use Formula 3.34 to compute $P(E)$, the probability of `Effect`, and Formula 3.35 if we observe the presence or absence of `Effect` and want to know, for example, $P(c_1)$, the probability of one of the causes. Of course, $P(c_1)$ is the sum of two terms, $P(c_1 = T \cap c_2 = T)$ and $P(c_1 = T \cap c_2 = F)$, but each can be computed from Formula 3.35.

In general, however, when we use a graphical representation to denote a causal model of a biological, clinical, or public health phenomenon, we will have a more complex graph. In particular, we will have causal factors that lead to two or more effects, as well as multiple causes for a single effect, as illustrated in Figure 3.5.

In these cases, the probabilities are much more difficult to compute. However, if each node is conditionally independent of all its non-parent nodes, given its parent nodes, we can write the joint probability in terms of relations of each local node to its parent nodes. For most reasonable graphs, this drastically reduces the number of probabilities that are needed.

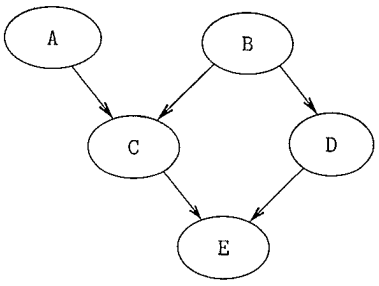


FIGURE 3.5 A graph with nodes that have both parents and children.

To illustrate this idea we start with the simplest possible case, where a single cause has two effects. How do we combine them to estimate the conditional probability of a cause? We will use the example described earlier of testing for prostate cancer as an illustration. In addition to the High PSA test result, we consider the percent free PSA. Many observational variables have been considered as diagnostic markers for prostate cancer, but no single one seems to have both the sensitivity and specificity that is desired. Considering two (somewhat independent) variables should give some improvement, but we need a formula for updating our previous result of how to compute the probability of prostate cancer given a high PSA value.

Most of the PSA observed in serum is bound to a protein inhibitor, alpha-1-antichymotrypsin. The percentage of PSA that is unbound (non-complexed) is referred to as the percentage of "free" PSA, a fraction of the total. In diagnostic testing, two numbers are reported, the total PSA concentration and the percentage of it that is "free." The percentage of free PSA seems to be lower when prostate cancer is present. When the PSA is between 4 and 10, if the fraction of free PSA is less than 0.10, the probability of having prostate cancer is considerably higher than it is if the free PSA fraction is higher.

Here, rather than distinguishing all the levels and combinations, we just use simple categories. In the formulas, to be concise, we abbreviate "Prostate Cancer" as "PC," "High PSA," that is, greater than 4, as just "psa," and "abnormally low fraction of free PSA," that is, less than 0.10, as "fpsa." Figure 3.6 shows diagrammatically the relationships. We have filled in the conditional probability tables for each node in the diagram. The Prostate Cancer node just has one entry, the prior, or overall, probability of occurrence of prostate cancer. The High PSA node conditional probability table is calculated from the joint probability distribution in Table 3.1. We have not provided the corresponding joint distribution for fpsa, but only the conditional probabilities of observing fpsa when prostate cancer is present and when prostate cancer is not present.

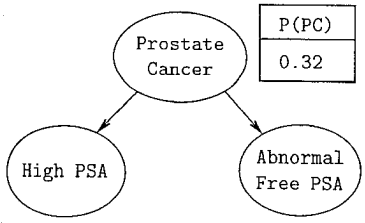
Essentially, the conditional probabilities for each of the tests give the true positive and false positive rates. From these and Bayes' rule, we can compute what we are really interested in; that is, given a positive test, what is the probability of having prostate cancer.

$$P(\text{pc}|\text{PSA}) = \frac{(0.78)(0.32)}{(0.78)(0.32) + (0.044)(0.68)} = 0.89 \tag{3.36}$$

This result is slightly different from the previous one because the false positive rate has an additional decimal place, giving a slightly lower result.

$$P(\text{pc}|\text{fpsa}) = \frac{(0.8)(0.32)}{(0.8)(0.32) + (0.15)(0.68)} = 0.715 \tag{3.37}$$

Each test separately gives some significant indication, but we would expect that both tests together would give a stronger prediction. So, suppose we have high PSA *and* abnormal free PSA? How do we



PC	P(PSA)
T	0.78
F	0.044

PC	P(FPSA)
T	0.8
F	0.15

FIGURE 3.6 A Bayes net for prostate cancer.

combine the information from both tests? We can use Bayes' rule here too,

$$P(\text{pc}|\text{psa} \cap \text{fpsa}) = \frac{P(\text{psa} \cap \text{fpsa}|\text{pc})\,P(\text{pc})}{P(\text{psa} \cap \text{fpsa})} \tag{3.38}$$

but we do not have the conditional probabilities required here, the joint probability of both tests being positive or negative etc. given the presence (or absence) of prostate cancer. In this case, there are only a few combinations, but, in general, there are (as we saw with the prognostic graph) a large number of combinations, and direct observation of the joint causal probabilities simply will not scale to larger graphs.

Instead, we observe that the two tests are actually independent, conditioned on the presence or absence of the disease. If you have the disease, the residual effects on the test outcomes would be expected to be independent of each other. Thus, we could claim that Equation 3.12 applies to this case, so that

$$P(\text{fpsa}|\text{psa} \cap \text{pc}) = P(\text{fpsa}|\text{pc}) \tag{3.39}$$

and

$$P(\text{psa}|\text{pc} \cap \text{fpsa}) = P(\text{psa}|\text{pc}) \tag{3.40}$$

More generally, for multiple effects e_j and a single cause, c

$$P(e_i|c) = P(e_i|e_j, c) \tag{3.41}$$

Then, applying this to Equation 3.10, we get

$$P(e_1 \cap e_2 \cap \ldots \cap e_j|c) = P(e_1|c)\,P(e_2|c) \ldots P(e_j|c) \tag{3.42}$$

Now, we can substitute for $p(\text{psa} \cap \text{fpsa}|\text{pc})$, giving

$$P(\text{pc}|\text{psa} \cap \text{fpsa}) = \frac{P(\text{psa}|\text{pc})\,P(\text{fpsa}|\text{pc})\,P(\text{pc})}{P(\text{psa} \cap \text{fpsa})} \tag{3.43}$$

Now, all that remains is to simplify the denominator. This is a *normalizing* factor, just like the case with multiple causes and a single effect. It is the sum of the numerator in Equation 3.43 and the corresponding expression for when prostate cancer is not present, that is,

$$\begin{aligned} P(\text{psa} \cap \text{fpsa}) = {} & P(\text{psa}|\text{pc})\,P(\text{fpsa}|\text{pc})\,P(\text{pc}) + \\ & P(\text{psa}|\neg\text{pc})\,P(\text{fpsa}|\neg\text{pc})\,P(\neg\text{pc}) \end{aligned}$$

Substitution in Equation 3.43 then gives the final formula for the prediction for the combined tests:

$$P(\text{pc}|\text{psa} \cap \text{fpsa}) = \frac{P(\text{psa}|\text{pc})\,P(\text{fpsa}|\text{pc})\,P(\text{pc})}{P(\text{psa}|\text{pc})\,P(\text{fpsa}|\text{pc})\,P(\text{pc}) + P(\text{psa}|\neg\text{pc})\,P(\text{fpsa}|\neg\text{pc})\,P(\neg\text{pc})} \tag{3.44}$$

In the previous section, we saw how to generalize to a multi-valued hypothesis, or many possible causes, rather than just a Boolean. Now, we see how to generalize to many pieces of evidence, when the items of evidence are conditionally independent, that is, we have a graph node that has no parents but several children.

$$P(c|e_i) = \frac{P(c)\prod_i P(e_i|c)}{P(c)\prod_i P(e_i|c) + P(\neg c)\prod_i P(e_i|\neg c)} \tag{3.45}$$

In this formula, we have a *product* expression instead of a sum. We can support this by using a macro implementing a `product` operator, exactly analogous to the `sum` macro. Coding of this macro and implementation of Formula 3.45 are left as an exercise for the reader.

Now, all that remains is to write an expression for the joint probability distribution for an arbitrary Bayesian graph, like the ones in Figures 3.5 and 3.3. This is not simply a matter of combining the two formulas, but a theorem about such graphs [231, page 39] provides a solution to the problem.

If the graph has the property that each node (or variable) is conditionally independent of all its *non-descendents*, given its *parents*, then the joint probability distribution is the product of all the conditional distributions of each node given values for its parents. This means we can consider each node with its parents, independent of the rest of the graph. So, for the graph in Figure 3.5, we could propose a joint probability distribution of the following form:

$$P(E, C, D, A, B) = P(E|C \cap D)P(C|A \cap B)P(D|B)P(A) \tag{3.46}$$

Such graphs are called Bayes nets. They can be used to model both diagnostic and predictive problems. There are two big challenges in using these ideas. First, the topology of the network itself cannot be derived from the data but generally comes from intuitions about relations among the variables. These are supported by the data, otherwise one would not have the conditional probabilities, but there are direct dependencies and indirect dependencies. For example, in Figure 3.3 disease control is indirectly related to T-stage, but one could also draw the graph with a direct connection there. Second, it is possible that for many nodes the local conditional probabilities are not known but other data are available, and one can induce the missing information with suitable advanced methods [231].

3.3 UTILITY AND DECISION MODELING

De gustibus non disputandum.

– anonymous
Latin proverb

There's no accounting for taste.

– Mr. Spock
Star Trek episode, "The Trouble With Tribbles"

To make reasonable practical decisions in clinical care and in public health, we will need to be able to incorporate into our models such factors as the relative importance of outcomes or effects, in cases where there are choices with multiple outcomes affected simultaneously by the choices. Again, referring to cancer, choices of treatments will often require deciding between a treatment that is more effective but more damaging to normal body functions and a treatment that is less effective but less debilitating. How important is each of these outcomes to an individual or to a population more broadly? We need ways to represent these importance factors, usually called *utilities*. A utility is a measure of how useful or important an outcome is. Once we have a framework for defining utilities, it is a considerable challenge to determine what are their values for any particular individual. *Utility theory* is a framework for creating models incorporating these values. Combining utility theory with probability models constitutes *decision theory*. We can use decision theory to represent trade-off and compute the consequences of one choice as compared to another.

Utility is a single number that expresses the desirability of a state of the world. *Expected utility* combines the probability of any outcome with its utility. A *normative* theory of decision-making asserts that a rational individual should make the choice that maximizes expected utility. This section presents two ways of representing and computing with utilities and probabilities, namely *decision trees* and *influence diagrams*.

A decision tree is a graph with a tree structure whose root or trunk is a decision to be made, where the branches represent the alternatives to be considered. Branching further from each decision option are different kinds of events that may happen when that decision choice is taken. These junctions are called chance nodes. There may be further chance nodes. Eventually, the leaf ends of the tree represent outcomes with which some value can be associated. An example is shown in Figure 3.7. The structure of the tree comes from defining the problem to be solved. The values must be evoked from the person

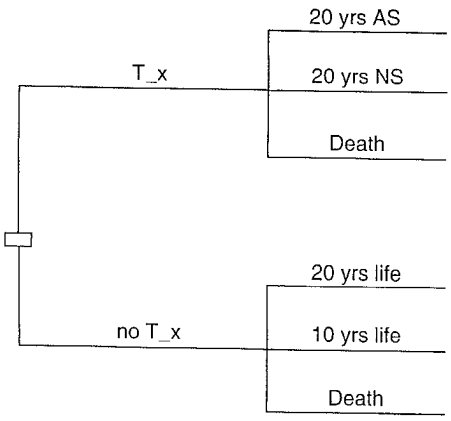


FIGURE 3.7 A decision tree representation of Mr. X's dilemma.

or population for whom the decision needs to be made. At each chance node, there are probabilities for each outcome or branch. The probabilities are estimates based on the conditions represented at that point, for example, the likelihood of success of a surgery and the probability of recurrence of a tumor.

Once the tree is created and values assigned, the *expected value* (EV) at each node is computed. For an outcome, the value or utility is multiplied by the probability of that outcome at that point in the tree (the same outcome could appear in several branches). At a chance node, the EVs on each branch are summed to get the EV for the node. At the chance nodes,

$$\mathrm{EV} = \sum_i p_i \times \mathrm{EV}_i.$$

This process is repeated until each decision option has an EV assigned. Then, we know which option has the most EV and can make a rational choice.

These decision trees are different from another kind of tree structure which is also sometimes called a decision tree, but should really be called a *classification* tree. A classification tree has no probabilities or EV. It is a representation of a logic formula that incorporates many factors in a decision. In this case, the tree represents a series of conditional alternatives. At the base, it might be, "If PSA is higher than 4 and rising, get a biopsy. If not, then do nothing and retest in 6 months." Each of those branches then has a further conditional, for example, the biopsy branch might have, "If biopsy shows active cancer, get treatment." This branch in turn may have several treatments. At each point, an unequivocal decision is made. Such tree structures can provide an efficient summary of what to do when the number of variables is large. There are also learning algorithms for constructing trees from examples, but there is not sufficient space to discuss them here.

An influence diagram extends the idea of a Bayes net. In addition to having probability nodes that represent random variables, it can have utility nodes, whose values can represent the utilities mentioned above and decision nodes as in a decision tree. It is a very compact representation but complicated to work with.

3.3.1 A Decision Analysis Vignette

These ideas can best be illustrated by a somewhat realistic medical example.[8]

> Mr. X has prostate cancer. His treatment options include chemotherapy, radiation, or surgery. He is against surgery and says that radiation therapy is not an option. He is well advanced in

8. The material in this section is contributed by Richard Boyce, from his class notes in a class taught by Jason Doctor, on clinical decision-making.

age, and the cancer is a non-aggressive type so he is willing to consider not treating it at all. Chemotherapy comes with a chance of immediate death and may result in loss of his natural voice. Should he choose treatment or let nature run its course?

This is an artificial example contrived to represent a clinical decision involving a great deal of uncertainty. The course that Mr. X's illness will take is simply unknown as is the outcome of treatment. Uncertainty is a common feature of medical decisions at all levels; from individual care to public health to health policy. The study of decision-making under uncertainty asks: "Which decision, from a set of decisions with uncertain outcomes, will result in the greatest *expected utility*?"

The normative theory of decision-making can be applied to clinical decision-making by decision analysis. We will walk through a simple decision analysis using the example above.

3.3.2 Graph Structure and Probability Assignment

First, we need to structure the problem by deciding what the decision options and possible outcomes are. Possible outcomes can come from an expert, the literature, and/or clinical databases. Here, we assume that the literature provides the following outcomes:

- *Decision options:* Treat or not treat.
- *Possible outcomes:*
  - Treat (T_x) – survive up to 20 years with loss of voice, survive up to 20 years with no loss of voice, or immediate death.
  - Not treat ($\bar{T}_x$) – survive up to 20 years or immediate death.

We will use two *normative* techniques for representing this decision analysis problem. A decision tree representation is shown in Figure 3.7. An influence diagram representation is shown in Figure 3.8. These models were made up for this example. In practice, the choice of model depends on factors such as the type of information available and the nature of the clinical decision. For example, different health states can be dynamic or static over time with implications on the decision model to use (e.g., dynamic states may be better modeled using Markov models).

Now, we need to assign probabilities to the outcomes. Like outcomes, probabilities can come from an expert, the literature, and/or clinical databases. The probabilities we assign need to fit within the constraints of probability theory. For example, the probabilities of all sub-branches of a particular outcome cannot add up to more than one. This may mean that we will need to perform some adjustment to the rough figures we get from our sources because the probabilities garnered from disparate sources may not neatly meet the probability theory's constraints. One significant reason to use probability theory is that probabilistic beliefs that do not satisfy these axioms can lead to non-rational decisions at least some of the time. The interested reader is referred to [237] for more discussion.

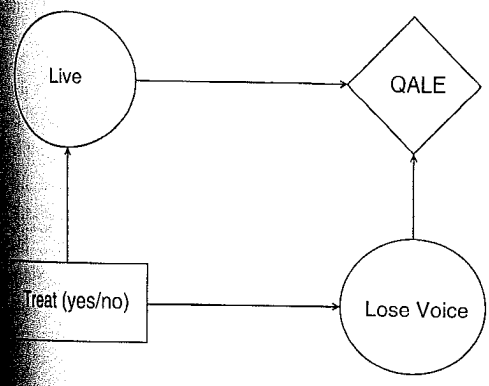


FIGURE 3.8 An influence diagram representation of Mr. X's dilemma.

TABLE 3.2 Probabilities for Possible Outcomes

	Outcome	Probability
$\bar{T}_x$	Live 10 years	.8
	Live 20 years	.1
	Immediate death	.1
T_x	Live 20 years with artificial speech (AS)	.6
	Live 20 years with normal speech (NS)	.2
	Immediate death	.2

In this case, let us assume that we can assign the probabilities shown in Table 3.2 to outcomes.

3.3.3 Determining Utilities

Now, we need to determine and assign the patient's utilities for the various outcomes. Simply put, a utility is an interval scale measure of the value an individual or group of individuals place on health states. It is a number that represents the subjective value of a given health state for one or more people affected by the decision. The derivation of utility is the focus of much discussion and research around *descriptive* and *normative* models for eliciting utility. *Descriptive* models attempt to acquire utilities based on how people really make valuations. They include prospect theory and rank-dependent theory [63]. Descriptive theories are designed to account for well-known heuristics and biases that individuals are subject to during decision-making.

Normative models make the assumption that people follow expected utility in making decisions. That is, their value for each possible outome is the product of their utility for each state and the probability of the state's occurrence. There are at least three different normative ways to elicit utilities. The best one to use depends on the focus population of the decision analysis (individual versus community), what methods make the most sense to the patient, and the time/resources that the decision analyst has to elicit utilities.

Time Trade-off (TTO): The utility a patient has for a given health state is assigned the ratio of amount of time at which a patient is indifferent between X years in perfect health and Y years in that health state: $\frac{X}{Y}$. For example, if a patient is indifferent between 10 years of perfect health and 20 years of blindness then his/her utility for that health state is $\frac{X}{Y} = \frac{10}{20} = 0.5$. The *constant proportional trade-off* assumption says that if a patient would give up x years of life to avoid y years in a given health state, (s)he would be also be willing to give up cx years of life to avoid cy years in that health state. TTO combined with the *constant proportional trade-off* assumption asserts that utility is linear with time. In reality, a person's utility for a health state may not be linear over time, and in such cases, TTO does not hold.

Standard Gamble (SG): Present a patient with a choice $-X$ years in a particular health state or a p chance of Y years in an improved health state versus a $1-p$ chance of immediate death. Adjust the probability until the patient is indifferent between the certainty and the gamble. For example, the result of this process might show that the patient has no preference between 20 years in a poor health state and a 0.5 chance of 20 years full health versus a 0.5 chance of immediate death. The probability at which the patient is indifferent is his/her utility for the given health state. SG assumes independence between the utility of survival duration and the utility of health status. Since this may not always hold, multiple utilities can be derived for different durations of a health state approximating the decision-maker's utility curve.

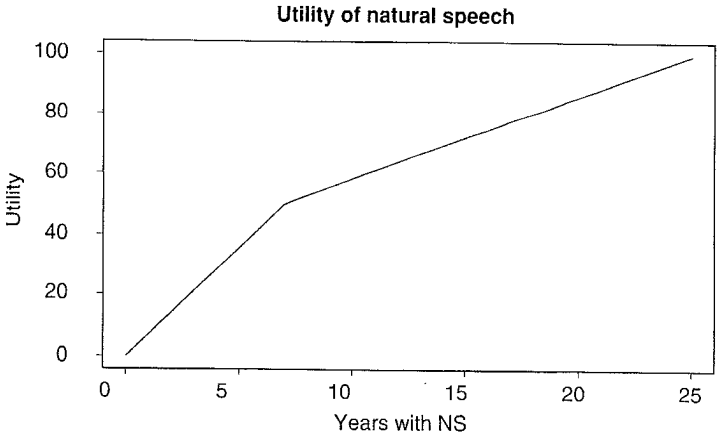


FIGURE 3.9 Patient's utility of natural speech.

Utility-based questionnaire: There are standard, validated survey tools that provide utilities for specific health states over time for certain populations. For example, the "Health Utilities Index" or the "EQ-5D."

Let us suppose that we have two sets of utilities for Mr. X. One set, for years of life with natural speech (NS), were derived using standard gamble technique:

- $U_{ns}(0) = 0$ – Mr. X's utility for death is 0.
- $U_{ns}(7) = 50$ – Mr. X is indifferent to accepting 7 years of normal speech (NS) with 100% certainty and taking a gamble with $p = 0.5$ of obtaining 25 years of normal speech versus 0 year of normal speech.
- $U_{ns}(25) = 100$.

Let us assume that we also have Mr. X's utility for artificial speech (AS) but it was derived using TTO. Specifically, Mr. X's utility for 10 years of AS is 0.7, that is, he is indifferent to 10 years of AS and 7 years of NS ($\frac{X}{Y} = \frac{7}{10} = 0.7$). Also, his utility for 25 years of AS is 1 or he is indifferent between 25 years of AS and 12.5 years of NS ($\frac{X}{Y} = \frac{12.5}{25} = 0.5$).

Using this relationship and the utilities derived from both methods we can establish the rough utility curves for both years of life with NS and years of life with AS. Without going into the details, the utility curves in Figures 3.9 and 3.10 were derived from the following relationships:

$$U_{ns} = \frac{100 - 50}{25 - 7}x + (50 - 2.8(7)) = 2.8x + 30.4 \tag{3.47}$$

$$U_{ns}(7) = U_{as}(10) = 50 \tag{3.48}$$

$$U_{ns}(12.5) = U_{as}(25) = 2.8(12.5) + 30.4 = 65.4 \tag{3.49}$$

$$U_{as} = \frac{65.4 - 50}{25 - 10}x + (50 - 1.03(10)) = 1.03x + 39.7 \tag{3.50}$$

3.3.4 Computation of Expected Values

With these figures for probabilities of outcomes and utilities of states we can go back and add this information into our decision tree and influence diagram and solve for the decision that maximizes expected utility. Figures 3.11 and 3.12 show the two representations, this time with values assigned.

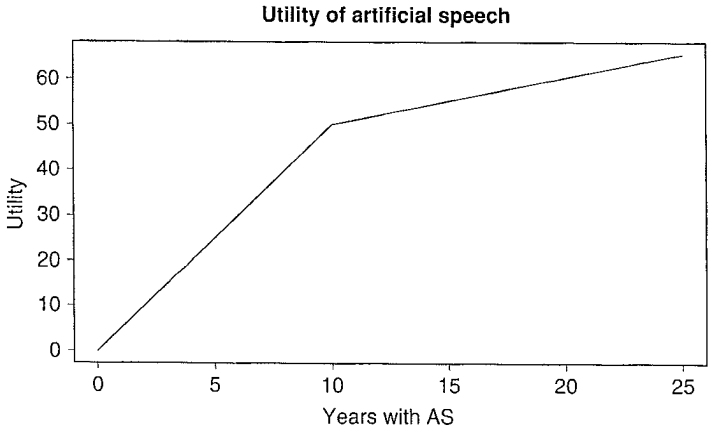


FIGURE 3.10 Patient's utility of artificial speech.

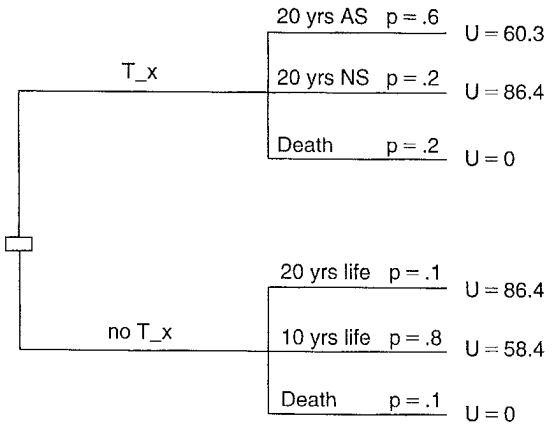


FIGURE 3.11 A decision tree representation of Mr. X's dilemma with utilities and probabilities.

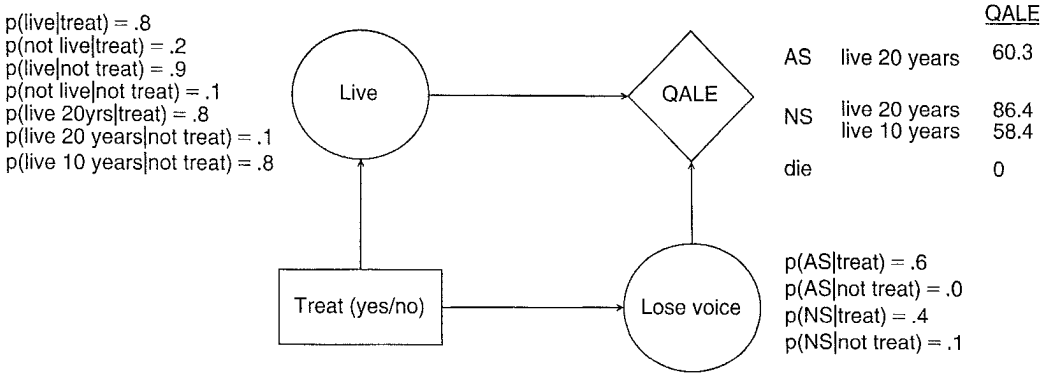


FIGURE 3.12 An influence diagram representation of Mr. X's dilemma with utilities and probabilities.

Both representations of Mr. X's decision problem can now be solved to determine the choice that will maximize expected utility. Solving the influence diagram is a bit involved and the reader is referred to [243] or [237] for details. The decision tree is easy to solve as it simply involves summing the products of the probabilities and utilites of each branch from leaf to root. This process, explained earlier, is called "folding back"the decision tree.

$$T_x = .6(U_{as}(20)) + .2(U_{ns}(20)) + .2(0) = 53.5 \tag{3.51}$$

$$\bar{T}_x = .1(U_{ns}(20)) + .8(U_{ns}(10)) + 0 = 55.4 \tag{3.52}$$

The process is shown in Equations 3.51 and 3.52. The result is that the decision not to treat gives slightly more expected utility than the decision to treat.

3.4 INFORMATION THEORY

Another way to think of evidence is that it is information or a message, which decreases our uncertainty about an outcome or assertion in the same way as information coming through a communication channel. Thanks to the pioneering work of Claude Shannon [297], it is possible to quantify the information content in a message in terms of probabilities. The idea that the information content of a message could be quantified made a huge change in our ability to design communication systems. This relatively new field, *information theory*, has also found some use in representing biomedical knowledge and interpreting data. Its significance to biology and medicine is not yet clear, but information theory provides yet another way to look at representing uncertain or non-deterministic knowledge. In this section, we give a very brief introduction to information theory.

3.4.1 Encoding of Messages

Before radio, people used telegraphy to encode messages by transmitting changes in electric current or voltage over a wire. Later, this same coding scheme was used with continuous wave radio transmissions. In more recent times, the invention of computer network communication protocols significantly increased the interest in transmitting messages at high speed over wires, using digital signaling techniques, such as Ethernet [119, 215]. Of course, it was not long before high-speed digital data transmission was implemented using radio technology, in the form of wireless local area networks [120], and digital cellular telephone technology.

- Morse code represents each letter of the alphabet by a group of dots and dashes, for example, . for E, . . . for S, -.-. for C, --- for O, and so on. The symbols used for encoding are the dot and the dash. In radio transmissions, these were implemented as a short continuous wave transmitted signal and a longer one (ideally the long one, or dash, was three times the length of the short one). Skilled telegraphers used a telegraph key, a simple hand-operated switch, to make the signals and a receiver that generated a tone corresponding to the incoming signal.
- The ASCII code represents each letter by a pattern of seven binary digits, written as 0 or 1, for example, 1000101 for E, 1010011 for S, 0110001 for the digit 1, and so on. In this case, the symbols are just 1 and 0. In electric circuits, the two symbols would be represented by two different voltages sufficiently different enough to ensure that the value is unambiguous.
- Other codes have been used to encode text on computer tapes and disks (SIXBIT, EBCDIC). On magnetic media (which also use ASCII), the magnetization direction and strength are used as a realization of the binary digits, 0 and 1.

No matter what the encoding scheme is, we transmit messages by using *some* scheme. Even natural language can be considered an encoding scheme, and human speech and writing are message sources.

The basic symbols which are used to encode messages can be considered to be the common alphabetic characters. However, one can also consider words to be symbol units, in which case there will be many more individual symbols, but possibly higher efficiency of message transmission.

In order to quantify information content in messages, we need to characterize the source. What kinds of messages are possible? How often do different symbols occur? A message source may have completely unpredictable behavior or it may be very regular. The idea of a "regular" message source can be given precise meaning.

- A *stationary source* is one in which the probability of a symbol occurring in a message (averaged over all possible messages) does not depend on the position in the message.
- An *ergodic* source is a stationary source in which the time-averaged occurrence rate of a symbol is the same as its ensemble average.

Information theory considers only ergodic message sources. In general, we are considering message sources that only send meaningful messages, not random nonsense. So, the symbols do not occur with equal probability, and also they do not occur in random order. However, information theory is not so much concerned with the *grammar* of the message, as its encoding. Choosing an efficient encoding can help increase the rate of message transmission. When noise is present, some of the symbols received will not be the ones that were sent, so some of the message is lost. It is, as we will see, possible to make up for this by including some redundancy in the encoding, so that errors can be detected and corrected.

In English, the letter E has a frequency of about 0.13, while W has only 0.02. Similarly, a table of word frequencies can be built up by counting words in large samples of text. Shannon [297] experimented with random text generated by selecting words according to their frequency in English text. His first-order approximation only used the base frequencies. His second-order approximation used conditional probabilities, that is, given a particular word, what is the probability of a particular word following it? You can see that the second-order process already starts to resemble real text. To really generate stochastic English text, one would want to incorporate a grammar-based generator, not only with word probabilities but also with structure. The problem in the second-order text is that it tends to wander, that is, has no larger structure, as one might expect.

- First-order approximation

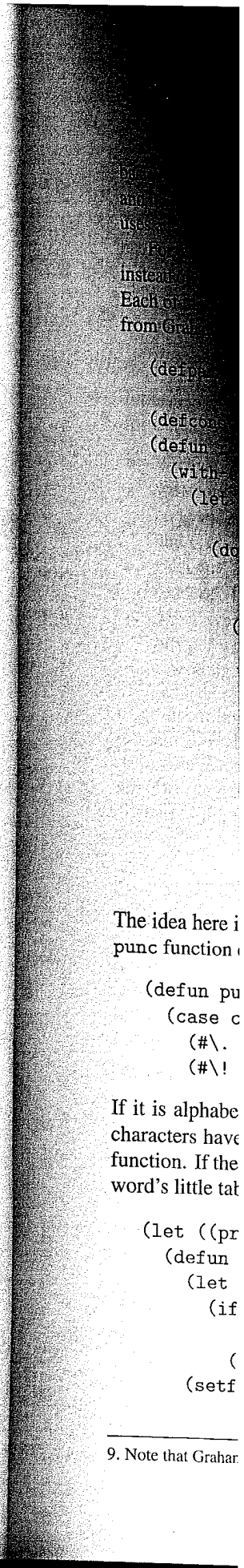
```
REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME
CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE
TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE
MESSAGE HAD BE THESE
```

- Second-order approximation

```
THE HEAD ON AND IN FRONTAL ATTACK ON AN ENGLISH
WRITER THAT THE CHARACTER OF THIS POINT IS
THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE
TIME OF WHO EVER TOLD THE PROBLEM FOR AN
UNEXPECTED.
```

Pierce [255] did a more extended experiment approximating fourth-order correlation by having friends write single words based on the last three. One of the more appealing sentences that resulted was "It happened one frosty look of trees waving gracefully against the wall."

It is relatively easy to create a profile of a body of text by simply counting words. Using a large enough text, assuming it is representative, such a table would give an estimate of the first-order word probabilities. When applied to a specific document rather than determining the frequency table for English text in general, this is called the Zipf distribution. It is somewhat characteristic of the document and useful for indexing. The `item-count` function described in Chapter 1 on page 34 expresses the



basic idea that each word found in sequence in the text is added to the word table if not already there, and its count is incremented if it is already there. In Chapter 4, there is a slightly improved version that uses a hash table instead of a list.

For the second-order experiment, the word table should have an entry for each word in the text, but instead of a single number, the entry should be a list of words that directly follow that word in the text. Each of those would have with it the frequency with which it follows the indexed word. Here is code from Graham [94, page 140] that will process a file of text, filtering for punctuation.[9]

```
(defparameter *words* (make-hash-table :size 10000))

(defconstant maxword 100)
(defun read-text (pathname)
  (with-open-file (s pathname :direction :input)
    (let ((buffer (make-string maxword))
          (pos 0))
      (do ((c (read-char s nil :eof)
              (read-char s nil :eof)))
          ((eql c :eof))
        (if (or (alpha-char-p c) (char= c #\'))
            (progn
              (setf (aref buffer pos) c)
              (incf pos))
          (progn
            (unless (zerop pos)
              (see (intern (string-downcase
                            (subseq buffer 0 pos))))
              (setf pos 0))
            (let ((p (punc c)))
              (if p (see p)))))))))
```

The idea here is to process a character at a time and determine if it is alphabetic, or punctuation. The punc function does this by straight comparision.

```
(defun punc (c)
  (case c
    (#\. '|.|) (#\, '|,|) (#\; '|;|)
    (#\! '|!|) (#\? '|?|) ))
```

If it is alphabetic put it in a buffer, that is, build up the current word. If non-alphabetic, and some characters have accumulated, look up the previous word in the hash table. This is the job of the see function. If the previous word is in the hash table, check if the current word has an entry in the previous word's little table. If so, increment it, and if not, make one.

```
(let ((prev '|.|))
  (defun see (symb)
    (let ((pair (assoc symb (gethash prev *words*))))
      (if (null pair)
          (push (cons symb 1) (gethash prev *words*))
        (incf (cdr pair))))
    (setf prev symb)))
```

[9] Note that Graham uses car and cdr instead of first and rest. Once again, "there's no accounting for taste."

The see function captures the variable prev, which becomes a private but persistent variable, in which the previous word can be stored (and updated each time). The variable prev is only accessible from within see. This is a nice example of information hiding while still preserving global state.

Once the table is created and populated, new random text with second-order word correlation can be generated by randomly picking a word, then randomly picking a successor from the list associated with it. That successor then becomes the current word, it is looked up and the process continues from word to word until the requested number of words are generated. Again from Graham, the code for generating text:

```
(defun generate-text (n &optional (prev '|.|))
  (if (zerop n) (terpri)
      (let ((next (random-next prev)))
        (format t "~A " next)
        (generate-text (1- n) next))))

(defun random-next (prev)
  (let* ((choices (gethash prev *words*))
         (i (random (reduce #'+ choices :key #'cdr))))
    (dolist (pair choices)
      (if (minusp (decf i (cdr pair)))
          (return (car pair))))))
```

Text produced this way has no meaning, but many people have a propensity to find meaning in everything, and so it can be the basis for much entertainment. Nevertheless, the point here is that the table represents all the possible symbols in a message and something about the frequency of appearance. This will be the key to quantifying the information content of a message, for purposes of analyzing the message encoding, transmission, and decoding process, aside from whatever meaning it might have.

3.4.2 Entropy and Information

Borrowing from the field of physics, in particular statistical mechanics, there is a relation between the intrinsic order or organization of a message source and the amount of information it can provide. When there are lots of possible messages all with equal probability, that is, almost total disorder, getting a message is very informative. On the other hand, if there is only one possible message, that is, total order, when we receive it we have not gotten any new information. To quantify the variety of states of a message source, we use the concept of *entropy*. The entropy of a message source is a measure of the information gained by receiving a message. For a message of one symbol, chosen from two possibilities, with probabilities p_1 and p_2 (note that $p_1 + p_2 = 1$), the entropy is defined as

$$H = -(p_1 \log p_1 + p_2 \log p_2)$$

where usually we use base 2 for the logarithm function, giving the entropy in units of *bits*. When $p_1 = p_2 = 0.5$, $H = 1$, and we say the entropy is one bit. When the message is a choice from among n symbols, the entropy in bits per symbol is defined as:

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{3.53}$$

In the case that we have n symbols all of equal probability (recall that this is the discrete uniform distribution), then $p_i = 1/n$. Every term in the sum is $\log_2 n/n$ and there are n terms, so the entropy is $\log_2 n$. So, to represent 32 different symbols, we would need 5 bits per symbol, and to represent 64 different symbols, we need 6 bits per symbol. If the number of symbols is between those two, the

entropy includes a fraction between 5 and 6. It would seem useful to have a function to compute the entropy for a list of probabilities corresponding to symbols from some source.[10]

```
(defun entropy (probs)
  (- (apply #'+
       (mapcar #'(lambda (p)
                   (* p (log p 2)))
               probs))))
```

By aggregating the symbols into larger groups, we may be able to encode messages more efficiently, for example, by assigning codes to blocks of characters or to words. Pierce [255] provides some figures for English text, as follows:

- With about 50 typewriter keys, the entropy per symbol is around 5.62 bits. With 6 bits per character we waste some but not much information. Some early computer systems used a 6 bits per character code and supported only uppercase alphabetic characters along with the 10 digits, special symbols, and control characters.
- Encoding all possible blocks of 3 characters, with each block assigned a code, this is slightly reduced.
- At an average word length of 4.5 characters plus a space, with 5 bits per character, we have 27.5 bits per word.
- Encoding words instead of characters, using a vocabulary of about 16,000 words, allows this to be reduced to 14 bits per word.

Shannon found by experiment that the entropy of English text is about 9 bits per word or around 1.7 bits per character (including the space). This may seem awfully small, until you recall the many humorous examples of text with letters left out, which is still quite readable. For example, hr is a hlarius xmpl of txt wth ltrs lft ot bt wch is stil qt rdabl. The popularity of text messaging with cellular telephones also supports this result.

3.4.3 Efficient Encoding

Having determined the entropy of a message source, many questions are possible:

- Is the entropy a limit on how efficient a coding scheme can be? Possibly with data compression, we can get arbitrarily efficient transmission if the message is large enough.
- What effect will noise or errors in transmission have on the rate of transmission of information?
- Is it possible to detect and correct transmission errors, thus guaranteeing reliable communication?
- If there is a limit on how efficient a coding scheme can be, can it actually be achieved? If so, how? What's the best way to do this?

The answer to the first question is "yes." Shannon's noiseless channel theorem states that the transmission rate, the channel capacity, and the entropy of a source are all related. Channel capacity is the maximum rate at which a channel can transmit information, denoted by C. It is measured in bits per second. Every communication mechanism has a channel capacity. Shannon's theorem then states that it is possible to transmit from a source of entropy H bits per symbol through such a channel, at the average rate of

$$\frac{C}{H} - \varepsilon$$

10. If `probs` were an array, it would be simpler to use the `sum` operator previously defined, but for small sets of numbers, lists, and the usual list operators seem straightforward. If you are by now an ambitious Lisper, try writing a macro, `mapsum`, that is like `sum` but operates efficiently on lists, instead of iterating on an index.

symbols per second. It is *not* possible to transmit at an average rate $> C/H$. Note that the rate at which information can be transmitted from a given source through a given channel depends on *both* the source entropy and the channel capacity.

So, what about data compression? A lossless data compression algorithm is one that is *reversible*, that is, the output of the compression can always be put through a decompression algorithm to recover a faithful copy of the original data. Such compression algorithms are most effective in reducing the size of files, when the files contain highly redundant data, that is, their information content is low. Examples include medical images in which there are large areas of constant pixel values, for example, large regions that are black because they are images of air. Files containing data that have higher entropy will not compress well.

The effect of noise is that some received symbols will not be the same as the corresponding transmitted symbols. When this happens, it reduces the rate at which information can be transmitted. Shannon derived a corresponding theorem for the noisy channel. In a noisy channel, the source has entropy,

$$H(x) = -\sum p(x) \log p(x)$$

but the output may have a different one,

$$H(y) = -\sum p(y) \log p(y)$$

$H(y)$ depends on both the input and the noise (transmission errors). If we knew both the transmitted and received messages, we could compute the entropy of the combination of transmitting x but receiving y.

$$H(x, y) = \sum_x \sum_y p(x, y) \log p(x, y) \tag{3.54}$$

We can write this in a clearer way by defining the notion of *conditional entropy*, somewhat analogous to conditional probability. Conditional entropy is defined from the probability of receiving any y given that x was transmitted.

$$H_x(y) = -\sum_x \sum_y p(x) p_x(y) \log p_x(y)$$

that is, given x what entropy is associated with y? Similarly, we define

$$H_y(x) = -\sum_x \sum_y p(y) p_y(x) \log p_y(x)$$

that is, given y what entropy is associated with x? Then, we can write Equation 3.54 as,

$$H(x, y) = H(x) + H_x(y) = H(y) + H_y(x).$$

The quantity $H_y(x)$ is called the "equivocation." The rate at which information is transmitted over the channel is then

$$R = H(x) - H_y(x).$$

Shannon proved that with a discrete channel of capacity C and a discrete source of entropy per second H, if $H < C$, there exists a coding system to transmit information with arbitrarily small error rate. If $H > C$, it is possible to achieve equivocation that is less than $H - C + \varepsilon$, but there is no method of coding that will give an equivocation less than $H - C$. This is Shannon's *Noisy Channel Theorem.*

Now, we turn to the problem of choosing an encoding for messages. It turns out that there is an optimal method of generating an encoding of arbitrary sets of symbols using binary digits, when the source has a discrete, finite set of symbols. It is called Huffman encoding, and we will illustrate how it works by following a very small artificial example.

TABLE 3.3
Symbols a
a Public H
Problem

Disease
Influenza
Measles
Pneumonia
Lung-ca
Prostate-ca
Breast-ca
Anthrax
Kidney-ca

Influenza (.55
Measles (.13
Pneumonia (.12
Lung–ca (.10
Prostate–ca (.04
Breast–ca (.03
Anthrax (.02
Kidney–ca (.01

Consider a hypothetical source of messages formed from a very small vocabulary. The vocabulary consists of the names of eight reportable diseases, and the messages consist of sequences of the disease names. Each day various clinics send messages to a central databank. The messages contain the names of each of the diseases that turned up in their clinic, in order, once for each patient, so there can be repeats. The public health agency will of course use these messages to create prevalence values, perhaps even with geographical maps. Table 3.3 lists a small set of disease names (the symbols or vocabulary of this message source).

The entropy per word can be computed from Equation 3.53. It turns out to be 2.073. Because we have eight choices, we could assign the eight possible combinations of three binary digits, one combination per word. This encoding scheme uses 3 bits per symbol. However, it can be done much more efficiently. The diagram in Figure 3.13 shows the derivation of a Huffman encoding, a method to generate a maximally efficient encoding for a discrete, finite source.

The procedure involves connecting the two lowest probabilities, assigning the upper branch a 1 and the lower a 0. Repeat this until all are connected in a full tree. In the example in Figure 3.13, the branches go up in an orderly fashion, but with different probabilities, the connections may well cross each other. The procedure is to do a pair at a time. Once the graph is generated, the code for each symbol is the sequence of binary digits on its path going to the left from the apex of the graph. For the top symbol, `influenza`, its code is the single digit, 1. The code for `prostate-ca` is 0001. Table 3.4 shows the result of computing the Huffman encoding for all the symbols in this message source.

Adding the numbers in the last column of Table 3.4 gives the average number of digits per word, 2.09, which is much closer to the entropy of the source than 3. This might seem counterintuitive because

TABLE 3.3 An Example Set of Symbols and Probabilities for a Public Health Communication Problem

Disease	Probability
Influenza	0.55
Measles	0.13
Pneumonia	0.12
Lung-ca	0.10
Prostate-ca	0.04
Breast-ca	0.03
Anthrax	0.02
Kidney-ca	0.01

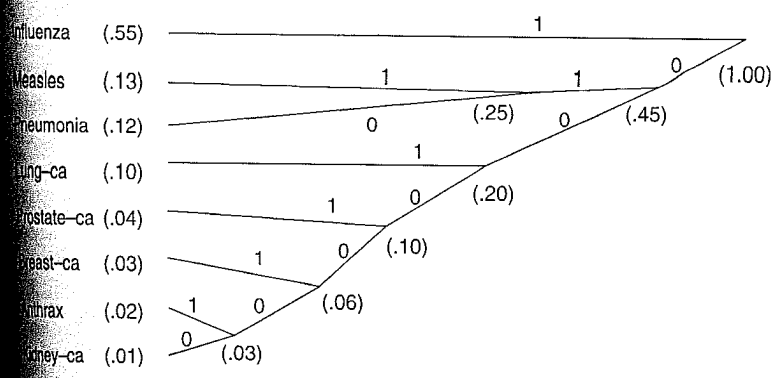


FIGURE 3.13 Algorithm for Huffman encoding.

TABLE 3.4 A Huffman Encoding for the Disease Transmission Problem

Word	Probability p	Code	Digits (N)	Np
Influenza	0.55	1	1	0.55
Measles	0.13	011	3	0.39
Pneumonia	0.12	010	3	0.36
Lung-ca	0.10	001	3	0.30
Prostate-ca	0.04	0001	4	0.16
Breast-ca	0.03	00001	5	0.15
Anthrax	0.02	000001	6	0.12
Kidney-ca	0.01	000000	6	0.06

you would think that with only about 2 bits per symbol, you could encode only four symbols. This is true if all the symbols have equal probability. In the case of the message source we are considering here, some symbols occur only rarely, so it is efficient to use as many as 6 bits per symbol. The savings of being able to encode the most frequently occurring symbol with only one bit makes up for it, on the average. So, the efficiency of a variable length code is precisely related to the unequal probabilities of occurrence of different symbols. It is left as an exercise to implement the Huffman encoding procedure as a function on a list of symbol probabilities.

3.4.4 Error Detection and Correction

It appears that it is possible to transmit a reliable stream of information over a noisy channel. Indeed the functioning of all modern digital technology depends on this. The Noisy Channel Theorem suggests that the amount of redundancy required is related to the amount of noise, or, alternately, the expected error rate. Here are brief descriptions of some methods for providing redundancy in binary-encoded data.

- A "parity" bit can allow detection of single bit errors but not correction. On tapes and disks (and memory) this can trigger a "reread."
- Repeating bits (sending three identical bits) can allow correction of single bit errors (1 in 3), but this drops the transmission rate to a third.
- "Check digits" allow correction of single bit errors at lower cost, depending on the expected error rate. This technique is used in memory modules and in network transmission.

3.4.5 Information Theory in Biology and Medicine

The implementation of the genetic code in cells provides some conceptually simple mechanisms for error correction. The fact that the DNA in a cell is double stranded enables repair when a single strand is broken and some part of its sequence is lost. The complementary strand can assimilate nucleotides because of the matching process and can restore the missing section. Even double strand breaks can sometimes be repaired.

In addition, the genetic code itself has some redundancy, as we saw in Chapter 1, Table 1.5. Many amino acids are represented by several codons. There is some correlation between the frequency of appearance of amino acids in proteins and the number of codons that code for them. When a single point mutation occurs in the third base in a codon, it often does not affect which amino acid results in that position in the generated protein sequence. The biological significance of this is not well understood, but there is active research in this area.

Classification trees, mentioned earlier, can be induced, or automatically generated from examples. A well-known algorithm for generating an optimal classification tree is the ID3 algorithm [260, 261]. In deciding at any point in the tree, which variable should be considered next, it chooses the one that has maximum entropy with respect to the data set, that is, which variable will by itself classify the most examples or reduce the uncertainty about the remaining examples in the training set. Examples of application of classification tree learning to the solution of medical problems include neonatal intensive care [333] and recurrence of prostate cancer [358].

Entropy plays a role in medical image processing as well, where it is used as a measure of correlation between image data sets. The problem is that different imaging studies (PET, CT, MRI) on the same patient need to be aligned with each other to correlate the physical locations of findings on the different kinds of images. This process is called *image registration*. Many algorithms for finding a suitable mapping or alignment have been developed [203, 204, 340]. The most popular and successful ones use the entropy of the data sets as a cost function to search for a transformation between them that maximizes the correlation of pixel values. These methods have been applied to the problem of automatically generating a target volume for planning radiation therapy treatment of cancer [325–327]. Chapter 9 describes this area in more detail.

Another application for error-correcting codes is to ensure that a transmitted message has not been modified in transit. This is particularly important for software systems used for electronic medical records, and other systems that transmit, receive, and store sensitive or confidential data. Providing checksums, that is, some kind of arithmetic function of all the bits in the message, can help detect alterations.

Several interesting articles provide for further exploration of the application of information theory in biology [2, 3, 287].

3.5 SUMMARY

Modeling uncertainty and non-deterministic processes adds considerably to our ability to make biomedical knowledge computable. Although the mathematics of probabilistic reasoning is quite different from that of logic, the themes are similar, that is, one can build predictive models, compute their entailments, and verify or reject hypotheses by comparing with experimental or observed data. One can build network representations that can explain findings and perform classification tasks.

The MYCIN experiments [38, 299] were noteworthy for the fact that MYCIN combined a probability-like representation of medical knowledge with a rule-based system. In MYCIN, as the rules were applied, the degree of certainty of a list of working hypotheses would increase or decrease. This is an example of a rule-based *production system*, in which the action of a rule can be any operation, not just the addition of an assertion to a set of propositions.

Another example of work combining probabilistic and logic-based systems is Daphne Koller's work combining frames and Bayes nets [178]. There is a lot to be done, but it points to the possible eventual unification of these seemingly different frameworks for biomedical theories.