Introduction to Probability

A Paper by Chuck Easttom

Introduction

After descriptive statistics, where one gets a clear picture of the data and any trends indicated by the data, probability is the next branch of statistics. Probability is concerned with predicting how likely a certain event is to occur. This area of mathematics is used in all facets of modern life from political predictions, to marketing, to quantum mechanics. This predictive aspect of statistics is one of the primary reasons people study statistics. Before we can delve deeply into probability a few concepts must be understood.

What is Probability

The first task is to define what probability is. In simplest terms it is a ratio between the number of outcomes of interest divided by the number of possible outcomes. For example if I have a deck of cards consisting of 52 cards, made up of 4 suits of 13 cards each, the probability of pulling a card of a given suit is 13/52 or ¼ = .25 (Aczel 1999). Probabilities are always between zero and one. Zero indicates absolutely no chance of an event occurring. For example if I have removed all thirteen clubs from a deck, the odds of then pulling a club are zero. A probability of 1.0 indicates the event is certain. For example if I remove all the cards except for hearts, then the probability of drawing a heart is 1.0.

Basic Set Theory

Probability often uses set theory, therefore a basic understanding of the essentials of set theory is necessary before we can continue. Let us begin by stating that a set is simply a collection of elements. An empty set is one containing no elements, and the universal set is the set containing all elements in a given context denoted by S (Gullberg 1997). The compliment of a set is the set

containing all the members of the universe set that are not in set A. This is denoted by a capital A with a bar over it or by A^c

Now let us briefly look at relationships between sets. First we have the union of two sets, denoted by A U B. The union is the set of all elements that exist in either of the two sets. For example if set A contains {1,2, 5} and set B contains {8,4,2} then the union of the two sets is {1,2,4,5,8}. The intersection of two sets is the set of elements contained in both sets. It is much like the binary operation OR. In our preceding example the intersection of set A and B would simply be {2}. The compliment of a set are all elements of the universal set that are not members of the set in question(Stoll 1979).

This is just a very elementary introduction to set theory, however it will come into play very briefly when we discuss dependent probability.

Basic Probability

In the introduction to this paper we discussed essentially what probability is. A technical definition would be " Probability is a measure of uncertainty. The probability of event A is a numerical measure of the likelihood of the event's occurring"(Aczel 1999). We also pointed out that the probability of an event must lie between zero and one. It should also be noted that there exist other basic probability rules we must consider

Basic probability Rules

- The probability of any event will be between zero and one, 0 <= P <= 1.0.
- Probability of the complement of an event (remember that set theory plays a role in probability) is equal to 1- probability of the event. Or put another way: P($\bar{A}$) = 1 - P(A). What this means is that if the probability of a given event A is .45, then its compliment is 1 - .45 or .55

- Rule of unions: The probability of a union of events s the probability of event A plus the probability of event B minus the probability of their intersection (or joint probability).
- Joint probability of independent events: This is simply the probability of event A multiplied by the probability of event B.

  P(A and B) = P(A) * P(B)

  So if two events are independent and event A has a probability of .45 and event B has A probability of .85 then the probability of both events occurring is .45 * .85 =.3825
- For two mutually exclusive events the probability of their union is simply the probability of event A + the probability of event B. P(A U B) = P(A) + P(B)

These basic rules are important to probability and should be committed to memory by any student who wishes to successfully study probability.

Conditional Probability

Conditional probability refers to the likelihood of an event occurring given some other event occurring. The likelihood of event A occurring, given event B has occurred is equal to the probability of the intersection of event A and B divided by the probability of event B, or:

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$$

This rule obviously is not referring to situations where event B must follow A, but where event A can lead to event B. For example if it is cold there is a certain probability that I will wear a jacket, but it does not absolutely follow that I will wear a jacket. Consider the following table (Jones 2004)

The question, "Do you smoke?" was asked of 100 people. Results are shown in the table.

.	Yes	No	Total

Male	19	41	60
Female	12	28	40
Total	31	69	100

- What is the probability of a randomly selected individual being a male who smokes? This is just a joint probability. The number of "Male and Smoke" divided by the total = 19/100 = 0.19
- What is the probability of a randomly selected individual being a male? This is the total for male divided by the total = 60/100 = 0.60. Since no mention is made of smoking or not smoking, it includes all the cases.
- What is the probability of a randomly selected individual smoking? Again, since no mention is made of gender, this is a marginal probability, the total who smoke divided by the total = 31/100 = 0.31.
- What is the probability of a randomly selected male smoking? This time, you're told that you have a male - think of stratified sampling. What is the probability that the male smokes? Well, 19 males smoke out of 60 males, so 19/60 = 0.31666...
- What is the probability that a randomly selected smoker is male? This time, you're told that you have a smoker and asked to find the probability that the smoker is also male. There are 19 male smokers out of 31 total smokers, so 19/31 = 0.6129 (approx)

Independent events

Independent events are events whose probability has no relationship at all. Put another way, two events are independent if the following are true (and conversely the following statements are true if the two events are independent):

- P (A | B) = P(A)
- P (B | A) = P(B)

The intersection of two or more independent events is just the product of their separate probabilities.

Bayes Theorem

Thomas Bayes was a clergymen in the 18th century whose work has been very influential in statistics and probability. Bayes's Theorem is a mathematical formula used for calculating conditional probabilities. It is the basis *Bayesian* approaches to epistemology, statistics, and inductive logic. The Theorem's central insight is simply that a hypothesis is confirmed by any body data that its truth renders. The probability of a hypothesis *H* conditional on a given body of data *E* is the ratio of the unconditional probability of the conjunction of the hypothesis with the data to the unconditional probability of the data alone.

Definition.

The probability of *H* conditional on *E* is defined as $\mathbf{P}_E(H) = \mathbf{P}(H \,\&\, E)/\mathbf{P}(E)$, provided that both terms of this ratio exist and $\mathbf{P}(E) > 0$.

That definition may seem a bit convoluted to the novice, so lets illustrate it with an hypothetical example. Assume a randomly chosen American who was alive on January 1, 2000. According to the United States Center for Disease Control, roughly 2.4 million of the 275 million Americans alive on that date died during the 2000 calendar year. Among the approximately 16.6 million senior citizens (age 75 or greater) about 1.36 million died. Now consider our hypothesis that our subject died during the 2000 caldendar year. Essentially the unconditional probability of the hypothesis that our subject died during 2000, *H*, is just the population-wide mortality rate $\mathbf{P}(H) = 2.4\text{M}/275\text{M} = 0.00873$. To find the probability of J. Doe's death conditional on the information, *E*, that he was a senior citizen we must divide the sample space into two different areas (though more than two can be used). We divide the probability that

he or she was a senior who died, $\mathbf{P}(H \,\&\, E) = 1.36\text{M}/275\text{M} = 0.00495$, by the probability that he or she was a senior citizen, $\mathbf{P}(E) = 16.6\text{M}/275\text{M} = 0.06036$. Thus, the probability of our subjects death given that he was a senior citizen is $\mathbf{P}_E(H) = \mathbf{P}(H \,\&\, E)/\mathbf{P}(E) = 0.00495/0.06036 = 0.082$. Notice how the size of the *total* population factors out of this equation, so that $\mathbf{P}_E(H)$ is just the proportion of seniors who died. Bayes theory allows us to work with conditional probabilities more efficiently. Expressed as a formula, Bayes theory is

$$\mathbf{P}_E(H) = [\mathbf{P}(H)/\mathbf{P}(E)]\, \mathbf{P}_H(E)$$

A baysian examination of conditional probability allows one to evaluate the predictive value of certain factors. Statisticians refer to the inverse probability $\mathbf{P}_H(E)$ as the "likelihood" of H on E. It expresses the degree to which the hypothesis *predicts* the data given the background information codified in the probability **P** (Joyce 2003). In the example discussed above, the condition that our subject died during 2000 is a fairly strong predictor of senior citizenship. Indeed, the equation $\mathbf{P}_H(E) = 0.57$ tells us that 57% of the total deaths occurred among seniors that year. Bayes's theorem lets us use this information to compute the probability of our subject dying given that he was a senior citizen. We do this by multiplying the "prediction term" $\mathbf{P}_H(E)$ by the ratio of the total number of deaths in the population to the number of senior citizens in the population, $\mathbf{P}(H)/\mathbf{P}(E) = 2.4\text{M}/16.6\text{M} = 0.144$. The result is $\mathbf{P}_E(H) = 0.57 \times 0.144 = 0.082$, just as expected.

Bayes's Theorem is of value in calculating conditional probabilities because inverse probabilities are typically both easier to ascertain and less subjective than direct probabilities. People with different views about the unconditional probabilities of E and H often disagree about E's value as an indicator of H. Even so, they can agree about the degree to which the hypothesis predicts the data if they know any of the following intersubjectively available facts: (a) E's

objective probability given H, (b) the frequency with which events like E will occur if H is true, or (c) the fact that H logically entails E. Scientists often design experiments so that likelihoods can be known in one of these "objective" ways. Bayes's Theorem then ensures that any dispute about the significance of the experimental results can be traced to "subjective" disagreements about the unconditional probabilities of H and E.

When both $\mathbf{P}_H(E)$ and $\mathbf{P}_{\sim H}(E)$ are known an experimenter need not even know E's probability to determine a value for $\mathbf{P}_E(H)$ using Bayes's Theorem.

Bayes's Theorem (2nd form):

$$\mathbf{P}_E(H) = \mathbf{P}(H)\mathbf{P}_H(E) \,/\, [\mathbf{P}(H)\mathbf{P}_H(E) + \mathbf{P}(\sim H)\mathbf{P}_{\sim H}(E)]$$

In this form Bayes's theorem is particularly useful for inferring causes from their effects since it is often fairly easy to discern the probability of an effect given the presence or absence of a putative cause.

Special Forms of Bayes's Theorem

Bayes's Theorem can be expressed in a several differant forms. Each is useful for different purposes. One version employs what is often called the *relevance quotient* or *probability ratio*. This is the factor $\boldsymbol{PR}(H, E) = \mathbf{P}_E(H)/\mathbf{P}(H)$ by which H's unconditional probability must be multiplied to get its probability conditional on E. Bayes's Theorem is equivalent to a simple symmetry principle for probability ratios.

Probability Ratio Rule: $\boldsymbol{PR}(H, E) = \boldsymbol{PR}(E, H)$

The term on the right provides one measure of the degree to which *H predicts E*. If we think of $\mathbf{P}(E)$ as expressing the "baseline" predictability of E given the background information codified in $\mathbf{P}$, and of $\mathbf{P}_H(E)$ as E's predictability when H is added to this background, then $\boldsymbol{PR}(E, H)$ captures the degree to which knowing H makes E more or less predictable relative to the

baseline: ***PR***(*E*, *H*) = 0 means that *H* categorically predicts ~*E*; ***PR***(*E*, *H*) = 1 means that adding *H* does not alter the baseline prediction at all; ***PR***(*E*, *H*) = 1/**P**(*E*) means that *H* categorically predicts *E*.

Another commonly encountered form of Bayes's Theorem is referred to as *Odds Rule*. In popular terminology, the "odds" of a hypothesis is its probability divided by the probability of its negation: **O**(*H*) = **P**(*H*)/**P**(~*H*). So, for example, a football team whose odds of winning a particular race are 8-to-2 has a 8/10 chance of winning and a 2/10 chance of losing. Contrary to popular thought, probability and odds are not necessarily the same.

Odds Ratio Rule. ***OR***(*H*, *E*) = $\mathbf{P}_H(E)/\mathbf{P}_{\sim H}(E)$

The Central Limit Theorem

Some texts will discuss the central limit theorem along with descriptive statistics, but this is a mistake. As you will see it clearly is most applicable to probability. The central limit theorem states that given a distribution with a mean m and variance s2, the sampling distribution of the mean approaches a normal distribution with a mean and variance/N as N, the sample size, increases. The central limit theorem explains why many distributions tend to be close to the normal distribution.

Lets consider a hypothetical example. Consider a set of independent random variables X1, X2, …XN. Let each X have an arbitrary probability distribution $P(x_1, \ldots, x_N)$ with mean μ_i and a variance σ_i^2. Then the normal form of the variable X

$$X_{\text{norm}} \equiv \frac{\sum_{i=1}^{N} x_i - \sum_{i=1}^{N} \mu_i}{\sqrt{\sum_{i=1}^{N} \sigma_i^2}}$$

has a limiting cumulative distribution function which approaches a normal distribution.

Poisson Process

The Poisson process is named after Simeon Poisson, and is one of the most important random processes in probability theory (Seigrest 2004). It is widely used to model random points in time and space, such as the times of radioactive emissions. Several important probability distributions arise naturally from the Poisson process including the Poisson distribution, the exponential distribution, and the gamma distribution. The process is used as a foundation for building a number of other, more complicated random processes. A Poisson process is a process satisfying the following properties:

1. The numbers of changes in nonoverlapping intervals are independent for all intervals.

2. The probability of exactly one change in a sufficiently small interval $h \equiv 1/n$ is $P = \nu h \equiv \nu/n$, where ν is the probability of one change and n is the number of trials.

3. The probability of two or more changes in a sufficiently small interval h is essentially 0.

In the limit of the number of trials becoming large, the resulting distribution is called a Poisson distribution

Consider a process in which certain points occur randomly in time. The phrase points in time is generic and could represent any activity such as The times when a piece of radioactive material emits particles. It turns out that under some basic assumptions that deal with independence and uniformity in time, a *single*, one-parameter probability model governs all such random processes. Because of this fact the Poisson process is one of the most important in probability theory.

In the limit of the number of trials becoming large, the resulting distribution is called a Poisson distribution. Given a Poisson process, the probability of obtaining exactly n successes in N trials is given by the limit of a binomial distribution

$$P_p(n|N) = \frac{N!}{n!(N-n)!}p^n(1-p)^{N-n}.$$

Viewing the distribution as a function of the expected number of successes

$$\nu \equiv Np$$

instead of the sample size N for fixed p, equation (2) then becomes

$$P_{\nu/N}(n|N) = \frac{N!}{n!(N-n)!}\left(\frac{\nu}{N}\right)^n\left(1-\frac{\nu}{N}\right)^{N-n},$$

Letting the sample size N become large, the distribution then approaches

$$\begin{aligned} P_\nu(n) &= \lim_{N\to\infty} P_B(n) \\ &= \lim_{N\to\infty} \frac{N(N-1)\cdots(N-n+1)}{n!}\frac{\nu^n}{N^n}\left(1-\frac{\nu}{N}\right)^N\left(1-\frac{\nu}{N}\right)^{-n} \\ &= \lim_{N\to\infty} \frac{N(N-1)\cdots(N-n+1)}{N^n}\frac{\nu^n}{n!}\left(1-\frac{\nu}{N}\right)^N\left(1-\frac{\nu}{N}\right)^{-n} \\ &= 1\cdot\frac{\nu^n}{n!}\cdot e^{-\nu}\cdot 1 \\ &= \frac{\nu^n e^{-\nu}}{n!}, \end{aligned}$$

which is known as the Poisson distribution (Weisstein 2004). Note that the sample size N has completely dropped out of the probability function, which has the same functional form for all values of ν.

As expected, the Poisson distribution is normalized so that the sum of probabilities equals 1, since

$$\sum_{n=0}^{\infty} P_\nu(n) = e^{-\nu} \sum_{n=0}^{\infty} \frac{\nu^n}{n!} = e^{-\nu} e^{\nu} = 1.$$

The ratio of probabilities is given by

$$\frac{P_\nu(n=i+1)}{P(n=i)} = \frac{\frac{\nu^{i+1} e^{-\nu}}{(i+1)!}}{e^{-\nu}\nu^i i!} = \frac{\nu}{i+1}.$$

The Poisson distribution reaches a maximum when

$$\frac{dP_\nu(n)}{dn} = \frac{e^{-\nu} n(\gamma - H_n + \ln \nu)}{n!} = 0,$$

where γ is the Euler-Mascheroni constant (note this constant is of particular interest in number theory, you can learn more about it at http://en.wikipedia.org/wiki/Euler-Mascheroni_constant) and H_n is a harmonic number, leading to the transcendental equation

$$\gamma - H_n + \ln \nu = 0,$$

Markov Chains/Processes

A **Markov chain** is a discrete-time stochastic process (note In the mathematics of probability, a **stochastic process** is a random function. For more details see http://en.wikipedia.org/wiki/Stochastic_process) with the Markov property. The Markov property is essentially that the distant past is irrelevant given knowledge of the recent past. These chains are named after A.A. Markov, who produced the first results (1906) for these processes

A Markov chain is a sequence $X_1, X_2, X_3, \ldots$ of random variables. The range of these variables, is called the *state space*, the value of X_n being the state of the process at time n. If the conditional

$$P(X_{n+1} = x|X_0, X_1, X_2, \ldots, X_n) = P(X_{n+1} = x|X_n).$$

Where x is some state of the process. The identity above identifies the **Markov property**.

Perhaps a simpler way to put this is to say that **a Markov process is** A random process whose future probabilities are determined by its most recent values. A stochastic process $x(t)$ is called Markov if for every n and

$$t_1 < t_2 \ldots < t_n$$

we have

$$P(x(t_n) \leq x_n|x(t_{n-1}), \ldots, x(t_1)) = P(x(t_n) \leq x_n|x(t_{n-1})).$$

This is equivalent to

$$P(x(t_n) \leq x_n|x(t) \text{ for all } t \leq t_{n-1}) = P(x(t_n) \leq x_n|x(t_{n-1}))$$

Foot Notes

Aczel, Amir (1999) *Complete Business Statistics*, Irwin McGraw Hill.

Gullberg ,Jan (1997) *Mathematics From the Birth of Numbers*, W. W. Norton & Company.

Jones, James (2004) http://www.richland.cc.il.us/james/

Joyce, James (2003) http://plato.stanford.edu/entries/bayes-theorem/

Seigrist Kyel (2004) http://www.math.uah.edu/stat/poisson/

Stoll, Robert (1979) *Set Theory and Logic*, Dover Publications

Weisstein. Eric W.(2004) "Poisson Process." From *MathWorld*--A Wolfram Web Resource.

http://mathworld.wolfram.com/PoissonProcess.html