Lecture Notes on Probability and Statistics

Larry Wasserman
Carnegie Mellon University
Fall, 2001

CHAPTER 1. Probability

1.1. Introduction

This course is about probability and statistics. Probability is the mathematical language for quantifying uncertainty. Statistical inference is the process of inferring unobservables from observables. Prediction, classification, clustering and estimation are all special cases of statistical inference. Data analysis, machine learning and data mining are various names given to the practice of statistical inference, depending on the context.

1.2. Sample Spaces and Events

Our discussion of probability begins with the notion of a *sample space* S, the set of possible outcomes in an experiment. Points s in S are called *sample outcomes* or *realizations. Events* are subsets of S.

EXAMPLE 1.2.1. If we toss a coin twice then $S = \{HH, HT, TH, TT\}$. The event that the first toss is heads is $A = \{HH, HT\}$.

EXAMPLE 1.2.2. Let s be the outcome of a measurement of some physical quantity (such as temperature). Then $S = \mathcal{R} = (-\infty, \infty)$. The event that the measurement is larger than 0 but less than or equal to 1 is $A = (0, 1]$.

If we repeat an experiment with sample space S n times, then the sample space for the repeated experiment is

$$S^n = \{s = (s_1, \ldots, s_n) : \; s_i \in S\}.$$

If we repeat an experiment with sample S infinitely many times then the sample space for the repeated experiment is

$$S^\infty = \{s = (s_1, s_2, \ldots,) : \; s_i \in S\}.$$

EXAMPLE 1.2.3. If we toss a coin forever then $S^\infty = \{s : \; s = (s_1, s_2, s_3, \ldots,), \; s_i \in \{H, T\}\}$.

Given an event A, we let $A^c = \{s \in S;\ s \notin A\}$ denote the complement of A. Informally, A^c can be read as "not A." The union of events A and B is defined $A \cup B = \{s \in S;\ s \in A \text{ or } s \in B\}$ which can be thought of as "A or B." The intersection of A and B is $A \cap B = \{s \in S;\ s \in A \text{ and } s \in B\}$ read "A and B." Sometimes we write $A \cap B$ as AB. Define the *set difference* $A - B = A \cap B^c$ and *symmetric difference* $A \Delta B = (A - B) \cup (B - A)$. We say that $A_1, A_2, \ldots$ are *disjoint* or are *mutually exclusive* if $A_i \cap A_j = \emptyset$ whenever $i \neq j$. For example, $A_1 = [0,1), A_2 = [1,2), A_3 = [2,3), \ldots$ are disjoint. The sets $A_1, A_2, \ldots$ form a *partition* of S if they are disjoint and $\cup_{i=1}^{\infty} A_i = S$.

Given an event A, define the *indicator function of* A by

$$I_A(s) = \begin{cases} 1 & \text{if } s \in A \\ 0 & \text{if } s \notin A. \end{cases}$$

If $A_1, A_2, \ldots$ is a (possibly infinite) sequence of events, define $\cup_{i=1}^{\infty} A_i = \{s \in S;\ s \in A_i \text{ for some } i\}$ and $\cap_{i=1}^{\infty} A_i = \{s \in S;\ s \in A_i \text{ for all } i\}$.

EXAMPLE 1.2.4. Let $S = \mathcal{R}$ and let $A_i = [0, 1/i)$ for $i = 1, 2, \ldots$. Then $\cup_{i=1}^{\infty} A_i = [0,1)$ and $\cap_{i=1}^{\infty} A_i = \{0\}$. If instead we define $A_i = (0, 1/i)$ then $\cup_{i=1}^{\infty} A_i = (0,1)$ and $\cap_{i=1}^{\infty} A_i = \emptyset$.

Given $A_1, A_2, \ldots$, define

$$\text{limsup}_n A_n = \cap_{k=1}^{\infty} \cup_{n=k}^{\infty} A_n$$

which is the set of s such that $s \in A_n$ for infinitely many n. We also write $\text{limsup}_n A_n \equiv \{A_n,\ \text{i.o.}\}$ which is read "A_n occurs infinitely often." Define

$$\text{liminf}_n A_n = \cup_{k=1}^{\infty} \cap_{n=k}^{\infty} A_n$$

which is the set of s such that $s \in A_n$ for all but finitely many n. We also write $\text{liminf}_n A_n \equiv \{A_n,\ \text{ult.}\}$ which is read "A_n occurs ultimately." From the definition, it can be shown that $\text{liminf}_n A_n \subset \text{limsup}_n A_n$.

EXAMPLE 1.2.5. Let B and C be events. Let $A_n = B$ if n is odd and $A_n = C$ if n is even. We claim that $\text{limsup}_n A_n = B \cup C$. Let $W_k = \cup_{n=k}^{\infty} A_n$. Then $W_k = B \cup C$ for all k. Hence, $\text{limsup}_n A_n = \cap_{k=1}^{\infty} W_k = B \cup C$. Now consider $\text{liminf}_n A_n$. Let $V_k = \cap_{n=k}^{\infty} A_n$. Then $V_k = B \cap C$ for all k. Thus, $\text{liminf}_n A_n = \cup_{k=1}^{\infty} V_k = B \cap C$.

If $\text{limsup}_n A_n = \text{liminf}_n A_n = A$ then we say that A_n converges to A and we write $A_n \to A$ or $\lim_n A_n = A$. A sequence of sets is *monotone increasing* if $A_1 \subset A_2 \subset \cdots$ in which case $\lim_n A_n = \cup_{n=1}^\infty A_n$. A sequence of sets is *monotone decreasing* if $A_1 \supset A_2 \supset \cdots$ in which case $\lim_n A_n = \cap_{n=1}^\infty A_n$.

THEOREM 1.2.6. *If A_n is monotone increasing then* $\lim_n A_n = \cup_{n=1}^\infty A_n$. *If A_n is monotone decreasing then* $\lim_n A_n = \cap_{n=1}^\infty A_n$.

PROOF. Suppose that A_n is monotone increasing i.e. $A_1 \subset A_2 \subset \cdots$. Let $A = \cup_{n=1}^\infty A_n$. For each k, $\cup_{n=1}^\infty A_n$ is also equal to A. Therefore, $\text{limsup}_n A_n = \cap_{k=1}^\infty \cup_{n=k}^\infty A_n = \cap_{k=1}^\infty A = A$. Now, for each k, $\cap_{n=k}^\infty A_n = A_k$. So, $\text{liminf}_n A_n = \cup_{k=1}^\infty \cap_{n=k}^\infty A_n = \cup_{k=1}^\infty A_k = A$. The proof for monotone decreasing sets is left as an exercise.

1.3. Probability

We assign a real number $P(A)$ to every event A, called the *probability* of A.[1] We also call P a *probability distribution* or a *probability measure.* Think of a probability as a function P that assigns a real number $P(A)$ to each event A. To qualify as a probability, P has to satisfy three axioms:

Axiom 1: $P(A) \geq 0$ for every A
Axiom 2: $P(S) = 1$
Axiom 3: If $A_1, A_2, \ldots$ are disjoint then $P(\cup_{i=1}^\infty A_i) = \sum_{i=1}^\infty P(A_i)$.

There are many interpretations of $P(A)$. The two leading interpretations are frequencies and degrees of beliefs. For the frequency interpretation, $P(A)$ is the long run proportion of times that A is true in repetitions. For example, if we say that the probability of heads is $1/2$, when mean that if we flip the coin many times then the proportion of times we get heads tends to $1/2$ as the number of tosses increases. An infinitely long, unpredictable sequence of tosses whose limiting proportion tends to a constant is an idealization, much like the idea of a straight line in geometry. The degree-of-belief interpretation is that $P(A)$ measures an observer's strength of belief that A is true.

[1] For technical reasons it is not always possible to assign a probability to every event A if the sample space is large, such as the whole real line. Instead, we assign probabilities to a limited class of set called a σ-algebra which is closed under complements and countable unions. Events in the σ-algebra are called measurable events.

In either interpretation, we require that Axioms 1 to 3 hold. The difference in interpretation will not matter much until we deal with statistical inference. There, the differing interpretations lead to two schools of inference: the frequentist and the Bayesian schools. We defer discussion until later.

If A is a finite set, let $|A|$ denote the number of elements in A. The *uniform probability* is defined by $P(A) = |A|/|S|$.

EXAMPLE 1.2.4. Tossing a die. Let $S = \{1, \ldots, 6\}$. If the die is well balanced then it seems reasonable to assign $P(A) = |A|/6$. Hence, if $A = \{1, 2\}$ then $P(A) = 2/6$.

One can derive many properties of P from the axioms. Here are a few:

$$\begin{array}{rcl} P(\emptyset) & = & 0 \\ A \subset B & \Longrightarrow & P(A) \leq P(B) \\ 0 \leq & P(A) & \leq 1 \\ P(A^c) & = & 1 - P(A) \\ A \cap B = \emptyset & \Longrightarrow & P(A \cup B) = P(A) + P(B). \end{array}$$

A less obvious property is given in the following Lemma.

LEMMA 1.2.5. *For any events A and B, $P(A \cup B) = P(A) + P(B) - P(AB)$.*

PROOF. Write $A \cup B = (AB^c) \cup (AB) \cup (A^cB)$ and note that these events are disjoint. Hence, making repeated use of the fact that P is additive for disjoint events, we see that

$$\begin{aligned} P(A \cup B) &= P((AB^c) \cup (AB) \cup (A^cB)) \\ &= P(AB^c) + P(AB) + P(A^cB) \\ &= P(AB^c) + P(AB) + P(A^cB) + P(AB) - P(AB) \\ &= P((AB^c) \cup (AB)) + P((A^cB) \cup (AB)) - P(AB) \\ &= P(A) + P(B) - P(AB) \diamond \end{aligned}$$

EXAMPLE 1.2.6. Two coin tosses. Let H_1 be the event that heads occurs on toss 1 and let H_2 be the event that heads occurs on toss 2. Assuming all outcomes are equally likely, $P(H_1 \cup H_2) = P(H_1) + P(H_2) - P(H_1 H_2) = \frac{1}{2} + \frac{1}{2} - \frac{1}{4} = 3/4$.

THEOREM 1.2.6. (Continuity of Probabilities.) If $A_n \to A$ then $P(A_n) \to P(A)$.

1.4. Probability on Finite Sample Spaces

Suppose that the sample space is finite: $S = \{s_1, \ldots, s_n\}$. Define $p_i = P(\{s_i\})$. Then, $p_i \geq 0$, $\sum_{i=1}^n p_i = 1$ and $P(A) = \sum_{i; s_i \in A} p_i$. For example, if we toss a die twice, then S has 36 elements: $S = \{(i,j);\ i,j \in \{1, \ldots 6\}\}$. If each outcome is equally likely, then $P(A) = |A|/36$. The probability that the sum of the dice is 11 is $2/36$ since there are two outcomes that correspond to this event.

In general, if we have a finite sample space and if each outcome is equally likely, then

$$P(A) = \frac{|A|}{|S|},$$

which we earlier called the *uniform probability distribution.* To compute probabilities, we need to be able to count the number of points in an event A. Methods for counting points, called combinatorial methods, exist and some are quite tricky. We needn't delve into these in any great detail. We will however, we need a few facts from counting theory that will be useful later. Given n objects, the number of ways of ordering these objects is $n! = n(n-1)(n-2) \cdots 3 \cdot 2 \cdot 1$. For convenience, we define $0! = 1$. We also define

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

read "n choose k", which is the number of distinct ways of choosing k objects from n. For example, if we have a class of 20 people and we want to select a committee of 3 students, then there are

$$\binom{20}{3} = \frac{20!}{3!17!} = \frac{20 \times 19 \times 18}{3 \times 2 \times 1} = 1140$$

possible committees. We note the following properties:

$$\binom{n}{0} = \binom{n}{n} = 1 \quad \text{and} \quad \binom{n}{k} = \binom{n}{n-k}.$$

1.5. Probabilities on the Real Line

We shall often use probabilities defined on the real line. For example, if a draw a person at random and measure their blood pressure then $S = \mathcal{R}$. Also, if we imagine that a computer produces numbers with infinite precision, then random number generators involve probabilities on $\mathcal{R}$. Of course, this is an idealization but it is helpful to work with real valued outcomes.

Given a probability P on $\mathcal{R}$, define a function F called the *cumulative distribution function* or *distribution function* or *cdf* given by

$$F(t) = P((-\infty, t]).$$

THEOREM 1.5.1. *Let P have cdf F and let Q have cdf G. If $F(t) = G(t)$ for all t then $P(A) = Q(A)$ for all A.*[2]

The above result says that a probability on the real line is completely determined by the probability of events of the form $(-\infty, t]$. That is one reason why cdf's are important.

THEOREM. 1.5.2. A function F mapping the real line to $[0, 1]$ is a cdf for some probability if and only if it satisfies the following three conditions:
(i) F is non-decreasing i.e. $s_1 < s_2$ implies that $F(s_1) \leq F(s_2)$.
(ii) F is normalized: $\lim_{s\to-\infty} F(s) = 0$ and $\lim_{s\to\infty} F(s) = 1$.
(iii) F is right-continuous, i.e. $F(s) = F(s^+)$ for all s, where

$$F(s^+) = \lim_{\substack{y\to s \\ y>s}} F(y).$$

PROOF. Suppose that F is a cdf. Let us show that (iii) holds. Showing (i) and (ii) is similar. Let s be a real number and let $y_1, y_2, \ldots$ be a sequence of real numbers such that $y_1 > y_2 > \cdots$ and $\lim_i y_i = s$. Let $A_i = (-\infty, y_i]$ and let $A = (-\infty, s]$. Note that $A = \cap_{i=1}^{\infty} A_i$ and also note that $A_1 \supset A_2 \supset \cdots$. Because the events are monotone, $\lim_i P(A_i) = P(\cap_i A_i)$. Thus,

$$F(s) = P(A) = P(\cap_i A_i) = \lim_i P(A_i) = \lim_i F(y_i) = F(s^+).$$

Proving the other direction namely, that if F satisfies (i), (ii) and (iii) then it is a cdf, uses some deep tools in analysis.

[2]Technically, $P(A) = Q(A)$ for every measurable event A.

The *point mass* distribution is the distribution that puts all its probability at one point. Formally, we say that P is a point mass at s if its distribution F is

$$F(t) = \begin{cases} 0 & t < s \\ 1 & t \geq s. \end{cases}$$

The *uniform distribution on [a,b]*, denoted by Unif(a, b) puts equal probability on all points between a and b. The distribution is

$$F(t) = \begin{cases} 0 & t < a \\ \frac{t-a}{b-a} & a \leq t \leq b \\ 1 & t > b. \end{cases}$$

A function f is called a *probability density function* or pdf, if $f(s) \geq 0$ for all s and $\int_{-\infty}^{\infty} f(s)ds = 1$. A pdf defines a probability distribution P by way of $P(A) = \int_A f(s)ds$. The corresponding cdf is $F(s) = \int_{-\infty}^{s} f(t)dt$. Conversely, given a cdf, its density (if it exists) is given by $f(s) = F'(s)$.

We shall have much more to say about cdf's and pdf's later.

1.6. Independent Events

Two events A and B are said to be independent if $P(AB) = P(A)P(B)$. Independence can arise in two distinct ways. Sometimes, we explicitly assume that two events are independent. For example, in tossing a coin twice, we usually assume the tosses are independent which reflects the fact that the coin has no memory of the first toss. In other instances, we derive independence by verifying that $P(AB) = P(A)P(B)$ holds. For example, in tossing a fair die, let $A = \{2, 4, 6\}$ and let $B = \{1, 2, 3, 4\}$. Hence, $A \cap B = \{2, 4\}$, $P(AB) = 2/6 = P(A)P(B) = (1/2) \times (2/3)$ and so A and B are independent. In this case, we didn't assume that A and B are independent it just turned out that they are. In fact, the independence is less than obvious here. Intuitively, A and B are independent because knowing A does not change the probability of B. We shall make this point more precise when we discuss conditional probability.

Suppose that A and B are disjoint events, each with positive probability. Can they be independent? No. This follows since $P(A)P(B) > 0$ yet $P(AB) = P(\emptyset) = 0$. Except in this special case, there is no way to judge independence by looking a the sets in a Venn diagram.

The events $A_1, \ldots, A_k$ are defined to be independent if, for any $I \subset \{1, \ldots, k\}$ we have that

$$P(\cap_{i \in I} A_i) = \prod_{i \in I} P(A_i).$$

For example, three events A_1, A_2 and A_3 are independent if they satisfy

$$\begin{aligned} P(A_1 A_2) &= P(A_1)P(A_2) \\ P(A_1 A_3) &= P(A_1)P(A_3) \\ P(A_2 A_3) &= P(A_2)P(A_3) \\ P(A_1 A_2 A_3) &= P(A_1)P(A_2)P(A_3). \end{aligned}$$

EXAMPLE. 1.6.1. Toss a fair coin 10 times. Let $A =$"at least one Head." Let T_j be the event that tails occurs on the j^{th} toss. Then

$$\begin{aligned} P(A) &= 1 - P(A^c) \\ &= 1 - P(\text{all tails}) \\ &= 1 - P(T_1 T_2 \cdots T_{10}) \\ &= 1 - P(T_1)P(T_2) \cdots p(T_{10}) \quad \text{independence} \\ &= 1 - \left(\frac{1}{2}\right)^{10} \\ &\approx .999. \end{aligned}$$

EXAMPLE. 1.6.2. A series if queries is sent to two search engines. Suppose that search engine 1 successfully finds what we are looking for with probability 1/3 while search engine 2 successfully finds what we are looking for with probability 1/4 (and assume independence). We begin by sending a query to engine 1, then 2, then 1, then 2, etc. We stop as soon as one of the search engines succeeds. What is the probability that engine 1 succeeds before engine 2. Let E denote the event of interest. Let A_j be the event that the first success is by search engine 1 and that it occurs on trial number j. Note that $A_1, A_2, \ldots$ are disjoint and that $E = \cup_{j=1}^{\infty} A_j$. Hence,

$$P(E) = \sum_{j=1}^{\infty} P(A_j).$$

Now, $P(A_1) = 1/3$. Next we find A_2. Now, A_2 occurs if we have the sequence 1 misses, 2 misses, 1 succeeds. This has probability $P(A_1) = (2/3)(3/4)(1/3) = (1/2)(1/3)$. Following this logic we see that $P(A_j) = (1/2)^{j-1}(1/3)$. Hence,

$$P(E) = \sum_{j=1}^{\infty} \frac{1}{3}\left(\frac{1}{2}\right)^{j-1} = \frac{1}{3}\sum_{j=1}^{\infty}\left(\frac{1}{2}\right)^{j-1} = \frac{2}{3}.$$

Here we used that fact that, if $0 < r < 1$ then $\sum_{j=k}^{\infty} r^j = r^k/(1-r)$.

1.7. Conditional Probability

Assuming that $P(B) > 0$, we define the conditional probability of A given that B has occurred by

$$P(A|B) = \frac{P(AB)}{P(B)}.$$

One interpretation is that this is a rule for updating our probability of A once we know that B has occurred. Another interpretation is that, upon repetitions, $P(A|B)$ is the fraction of times A occurs among those in which B occurs.

In our die example, consider the event $A = \{1\}$ so that $P(A) = 1/6$. Suppose we are now told that the outcome is odd. Then, we would no longer use $P(A) = 1/6$ but rather $P(A|B) = P(AB)/P(B) = (1/6)/(3/6) = 1/3$ where $B = \{1, 3, 5\}$.

A few important facts. First, for any fixed B such that $P(B) > 0$, $P(\cdot|B)$ is a probability i.e. it satisfies the three axioms. For example, $P(A|B) \geq 0$, $P(S|B) = 1$ and if $A_1, A_2, \ldots$ are disjoint then $P(\cup_{i=1}^{\infty} A_i|B) = \sum_{i=1}^{\infty} P(A_i|B)$. But it is in general **not** true that $P(A|B \cup C) = P(A|B) + P(A|C)$. The rules of probability apply to the left of the bar.

Second, in general it is **not** the case that $P(A|B) = P(B|A)$. People get this confused all the time. For example, the probability of spots given you have measles is 1 but the probability that you have measles given that you have spots is not 1. In this case, the difference between $P(A|B)$ and $P(B|A)$ is obvious. But there are cases where it is less obvious.

EXAMPLE. 1.7.1. Consider a medical test for a disease D. Suppose that $P(\{D, +\}) = .0081$, $P(\{D, -\}) = .009$, $P(\{D^c, +\}) = .0900$ and $P(\{D^c, -\}) = .9010$. Apply the definition of conditional probability to conclude that $P(+|D) =$

$P(+,D)/P(D) = .0081/(.0081{+}.0009) = .9$ and $P(-|D^c) = P(-,D^c)/P(D^c) = .9010/(.9010 + .0900) \approx .9$. Apparently, the test is fairly accurate. Sick people yield a positive 90 percent of the time and healthy people yield a negative about 90 percent of the time. Suppose you go for a test and get a positive. What is the probability you have the disease. Most people answer .90. The correct answer is $P(D|+) = P(+,D)/P(+) = .0081/(.0081 + .0900) = .08$. The lesson here is that probability is not intuitive. You need to compute the answer numerically; don't trust your intuition.

Suppose that A and B are independent events. Then,

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A).$$

Thus, independence implies that $P(A|B) = P(A)$ and that $P(B|A) = P(B)$. So another interpretation of independence is that knowing B doesn't change the probability of A.

From the definition of conditional probability we can write $P(AB) = P(A|B)P(B)$ and also $P(AB) = P(B|A)P(A)$. Often, these formulae give us a convenient way to compute $P(AB)$ when A and B are not independent.

EXAMPLE. 1.7.2. Draw two cards from a deck, without replacement. Let A be the event that the first draw is Ace of Clubs and let B be the event that the second draw is Queen of Diamonds. Then $P(A,B) = P(A)P(B|A) = (1/52) \times (1/51)$.

EXAMPLE. 1.7.3. There are three cards. The first is green on both sides, the second is red on both sides and the third is green on one side and red on the other. We choose a card at random and we see on side (also chosen at random). If the side we see is green, what is the probability that the other side is also green? Many people intuitively answer 1/2 because there are only two possible cards left after we see one side is green i.e. the red-red card is eliminated. The correct answer is 2/3. To see this, let A be the event that we have the green-green card and let B be the event that we observe green. Note that $A \subset B$. Also note that $P(B) = 1/2$. Thus, $P(A|B) = P(AB)/P(B) = P(A)/P(B) = (1/3)/(1/2) = 2/3$.

INTERLUDE: A PUZZLE. You have probably heard it before. Now you can solve it rigorously. It is called the "Monty Hall Problem."

A prize is placed at random between one of three doors. You pick a door. Let's say you pick door number 1. Now Monty Hall opens another door (either 2 or 3) and shows you that it is empty. He then gives you the opportunity to keep door 1 or switch to the other unopened door. Should you stay or switch?

Intuition suggests it doesn't matter. There are now two doors left. One has the prize. So there is an equal chance for winning whether you switch or not. But this reasoning is wrong.

Before proceeding, I should say that this example has created a lot of confusion. People argue about the answer because they are answering different questions. If you play the game repeatedly and never switch then you will win 1/3 of the time. Now note that a switcher wins if and inly if a non-switcher loses. Conversely, a non-switcher wins if and only if the switcher loses. Since non-switcher wins with probability 1/3, therefore, switcher wins with probability 2/3. If you don't believe it, try it.

So the frequency probability of winning if you switch is 2/3. But another question is this: in a given game, what is the probability you will win if you switch. To be precise, let suppose door 2 was revealed to be empty. Let A be the event that door 3 has the prize. Now find $P(A|B)$. This number is not well defined unless we have further information. In particular, if the prize happens to be behind door 1, Monty can show you either door 2 or door 3. $P(A|B)$ depends on how he chooses which door to show you. If you chooses each with equal probability then one can show again that $P(A|B) = 2/3$. I leave the details to the interested reader. Note: carefully specify the sample space.

1.8. Bayes' Theorem

Bayes' theorem is a simple theorem that allows us to update probabilities in light of new information. First, we need a preliminary result.

THEOREM 1.8.1. (The Law of Total Probability.) *Let $A_1, \ldots, A_k$ be a partition of S. Then, for any event B, $P(B) = \sum_{i=1}^k P(B|A_i)P(A_i)$.*

PROOF. Define $C_j = BA_j$ and note that $C_1, \ldots, C_k$ are disjoint and that $B = \cup_{j=1}^k C_j$. Hence,

$$P(B) = \sum_j P(C_j) = \sum_j P(BA_j) = \sum_j P(B|A_j)P(A_j)$$

since $P(BA_j) = P(B|A_j)P(A_j)$ from the definition of conditional probability.

Now we can state and prove Bayes' theorem, which is sometimes a useful tool for computing probabilities. Expert systems and "Bayes' nets" make extensive use of Bayes' theorem.

THEOREM 1.8.2. *Bayes' Theorem.* Let $A_1, \ldots, A_k$ be a partition of S such that $P(A_i) > 0$ for each i. Let B be an event such that $P(B) > 0$. Then, for each $i = 1, \ldots, k$,

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_j P(B|A_j)P(A_j)}.$$

PROOF. We apply the definition of conditional probability twice, followed by the law of total probability:

$$P(A_i|B) = \frac{P(A_iB)}{P(B)} = \frac{P(B|A_i)P(A_i)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_j P(B|A_j)P(A_j)}.\diamond$$

Sometimes $P(A_i)$ is called the "prior probability of A_i" and $P(A_i|B)$ is called the "posterior probability of A_i." To some extent, Bayes' theorem is an expression of how probabilities are updated as information is acquired. Later, we shall see that Bayes' theorem plays a major role in one school of statistical inference called "Bayesian inference." The other major school of inference is called "frequentist inference." We'll discuss the advantages and disadvantages of each later. For now we need only know that Bayes' theorem is a useful device in probability theory.

EXAMPLE. 1.8.2. Suppose a patient has a cough and we want to diagnose there disease status. For simplicity, suppose there are only three mutually exclusive disease states: $A_1 =$ "no disease, $A_2 =$ "cold" and $A_3 =$ "allergy." In real examples, the list of diseases is much longer and we have to be careful to specify them in such a way that they form a partition. Suppose we are given the following information: $P(A_1) = .3$, $P(A_2) = .5$ and $P(A_3) = .2$. of course, $.3 + .5 + .2 = 1$. We are also told that $P(B|A_1) = .3$, $P(B|A_2) = .9$, $P(B|A_1) = .8$. (Note: $.3+.8+.8 \neq 1$!) What is the probability that the person has a cold? Bayes' theorem yields,

$$P(A_2|B) = \frac{.9 \times .5}{(.3 \times .3) + (.9 \times .5) + (.8 \times .2)} = .64.$$