Chapter II Probability

2.1 Basic Concepts

The discipline of statistics deals with the ***collection*** and ***analysis of data***. When measurements are taken, even seemingly under the same conditions, the results usually vary. Variability is a fact of life, and proper statistical methods can help us understand data collected under inherent variability.

The term ***experiment*** refers to the process of obtaining an observed result of some phenomenon, and a performance of an experiment is called a ***trial*** of the experiment. An observation result, on a trial of the experiment, is called an ***outcome***. An experiment, the outcome of which cannot be predicted with certainty, but the experiment is of such a nature that the collection of every possible outcome can described prior to its performance, is called a ***random experiment***. The collection of all possible outcomes is called the outcome space or the ***sample space***, denoted by S.

2.2 Algebra of Sets

If each element of a set A_1 is also an element of set A_2, the set A_1 is called a ***subset*** of the set A_2, indicated by writing $A_1 \subset A_2$. If $A_1 \subset A_2$ and $A_2 \subset A_1$, the two sets have the same elements, indicated by writing $A_1 = A_2$. If a set A has no elements, A is called the ***null set***, indicated by writing $A = \varnothing$. Note that a null set is a subset of all sets.

The set of all elements that belong to at least one of the sets A_1 and A_2 is called the ***union*** of A_1 and A_2, indicated by writing $A_1 \cup A_2$. The union of several sets, $A_1, A_2, A_3, \ldots$ is the set of all elements that belong to at least one of the several sets, denoted by $A_1 \cup A_2 \cup A_3 \cup \cdots$ or by $A_1 \cup A_2 \cup A_3 \cup \cdots \cup A_k$ if a finite number k of sets is involved.

The set of all elements that belong to each of the sets A_1 and A_2 is called the ***intersection*** of A_1 and A_2, indicated by writing $A_1 \cap A_2$. The intersection of several sets, $A_1, A_2, A_3, \ldots$ is the set of all elements that belong to each of the several sets, denoted by $A_1 \cap A_2 \cap A_3 \cap \cdots$ or by $A_1 \cap A_2 \cap A_3 \cap \cdots \cap A_k$ if a finite number k of sets is involved.

The set that consists of all elements of S that are not elements of A is called the ***complement*** of A, denoted by A^c. In particular, $S^c = \varnothing$. Given $A \subset S$, $A \cap A^c = \varnothing$, $A \cup A^c = S$, $A \cap S = A$, $A \cup S = S$, and $(A^c)^c = A$.

Commutative Law: $A \cup B = B \cup A$; $A \cap B = B \cap A$.

Associate Law: $(A \cup B) \cup C = A \cup (B \cup C)$; $(A \cap B) \cap C = A \cap (B \cap C)$.

Distributive Law: $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$;
$(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$.

De Morgan's Laws: $\left(\bigcup_{i=1}^{n} A_i\right)^c = \bigcap_{i=1}^{n} A_i^c$; $\left(\bigcap_{i=1}^{n} A_i\right)^c = \bigcup_{i=1}^{n} A_i^c$.

An ***event*** is a subset of the sample space *S*. If *A* is an event, we say that *A* has ***occurred*** if it contains the outcome that occurred. Two events A_1 and A_2 are called ***mutually exclusive*** if $A_1 \cap A_2 = \varnothing$.

Example 2.2-1: If the experiment consists of flipping two coins, then the sample space consists of the following four points: $S = \{(H,H),(H,T),(T,H),(T,T)\}$. If *A* is the event that a head appears on the first coin, then $A = \{(H,H),(H,T)\}$ and $A^c = \{(T,H),(T,T)\}$. ◘

Example 2.2-2: If the experiment consists of tossing one die, then the sample space consists of the 6 points: $S = \{1,2,3,4,5,6\}$. If *A* is the event that the die appears an even number, then $A = \{2,4,6\}$ and $A^c = \{1,3,5\}$. ◘

2.3 Probability

One possible way of defining the probability of an event is in terms of its ***relative frequency***. Suppose that an experiment, whose sample space is *S*, is ***repeatedly*** performed under exactly the ***same conditions***. For each event *A* of the sample space *S*, we define #(*A*) to be the number of times in the first *n* repetitions of the experiment that the event *A* occurs. Then $P(A)$, the probability of the event *A*, is defined by

$$P(A) = \lim_{n\to\infty} \frac{\#(A)}{n}.$$

How do we know that $\#(A)/n$ will converge to some constant limiting value that will be the same for each possible sequence of repetitions of the experiment? One way is that the convergence of $\#(A)/n$ to a constant limiting value is an **assumption**, or an **axiom**, of the system. However, it seems to be a very complex assumption and does not at all seem to be a prior evident that it need be the case. In fact it is more reasonable to assume a set of simpler and more self-evidence axioms about probability and then attempt to prove that such a constant limiting frequency does in some sense exist. This latter approach is the modern axiomatic approach to probability theory that we shall adopt.

For a given experiment, *S* denotes the sample space and $A_1, A_2, A_3, \ldots$ represent possible events. We assume that a number $P(A)$, called the ***probability*** of *A*, satisfies the following three axioms:

(**Axiom 1**) $P(A) \ge 0$ for every *A*;

(**Axiom 2**) $P(S) = 1$;

(**Axiom 3**) $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$ if $A_i \cap A_j = \varnothing$ for $i \neq j$.

Example 2.3-1: Consider an experiment to toss a fair coin. If we assume that a head is equally to appear as a tail, then we would have $P(\{H\}) = P(\{T\}) = 1/2$. ◘

Example 2.3-2: If a die is rolled and we suppose that all six sides are equally likely to appear, then we would have $P(\{1\}) = P(\{2\}) = P(\{3\}) = P(\{4\}) = P(\{5\}) = P(\{6\}) = 1/6$. From Axiom 3 it would thus follow that the probability of rolling an even number would equal

$$P(\{2,4,6\}) = P(\{2\}) + P(\{4\}) + P(\{6\}) = 1/2.$$ ◘

In fact, using these axioms we shall be able to prove that if an experiment is repeated over and over again then, with probability 1, the proportion of times during which any specified event A occurs will equal $P(A)$. This result is known as the strong law of large numbers.

<u>Some Properties of Probability:</u>

(1) For each $A \subset S$, $P(A) = 1 - P(A^c)$.

Proof: Since $A \cup A^c = S$ and $A \cap A^c = \varnothing$, $1 = P(S) = P(A \cup A^c) = P(A) + P(A^c)$. ◘

(2) The probability of the null set is zero; that is $P(\varnothing) = 0$.

Proof: Since $\varnothing^c = S$, $P(\varnothing) = 1 - P(S) = 0$. ◘

(3) If A and B are subsets of S such that $A \subset B$, than $P(A) \le P(B)$.

Proof: Since $A \subset B$, $B = A \cup (A^c \cap B)$ and $A \cap (A^c \cap B) = \varnothing$. We have

$P(B) = P(A) + P(A^c \cap B)$. Because $P(A^c \cap B) \ge 0$, $P(B) \ge P(A)$. ◘

(4) For each $A \subset S$, $0 \le P(A) \le 1$.

Proof: Since $\varnothing \subset A \subset S$, $0 = P(\varnothing) \le P(A) \le P(S) = 1$. ◘

(5) **(The Additive Law of Probability)** If A and B are subsets of S, then
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Proof: Note that $A \cup B = A \cup (A^c \cap B)$ and $B = (A \cap B) \cup (A^c \cap B)$.
Since $A \cap (A^c \cap B) = \varnothing$ and $(A \cap B) \cap (A^c \cap B) = \varnothing$,
$P(A \cup B) = P(A) + (A^c \cap B)$ and $P(B) = P(A \cap B) + P(A^c \cap B)$.
Replace $P(A^c \cap B)$ by $P(B) - P(A \cap B)$. ◘

****

(6) If *A*, *B*, and *C* are subsets of *S*, then

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

Proof: Omitted.

(7) If $A_1, A_2, A_3, \ldots$ are events, then $P\left(\bigcup_{i=1}^{\infty} A_i\right) \le \sum_{i=1}^{\infty} P(A_i)$. (**Boole's Inequality**)

Proof: Let $B_1 = A_1$, $B_2 = A_2 \cap A_1^c$, and in general $B_i = A_i \cap \left(\bigcup_{j=1}^{i-1} A_j\right)^c$. It follows that

$\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} B_i$ and $B_1, B_2, B_3, \ldots$ are mutually exclusive. Since $B_i \subset A_i$, $P(B_i) \le P(A_i)$.

Hence, $P\left(\bigcup_{i=1}^{\infty} A_i\right) = P\left(\bigcup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} P(B_i) \le \sum_{i=1}^{\infty} P(A_i)$. A similar result holds for finite unions, i.e., $P(A_1 \cup A_2 \cup \cdots \cup A_k) \le P(A_1) + P(A_2) + \cdots + P(A_k)$. ◘

(8) If $A_1, A_2, \ldots, A_k$ are events, then $P\left(\bigcap_{i=1}^{k} A_i\right) \ge 1 - \sum_{i=1}^{k} P(A_i^c)$. (**Bonferroni's Inequality**)

Proof: Applied $\bigcap_{i=1}^{k} A_i = \left(\bigcup_{i=1}^{k} A_i^c\right)^c$, together with $P(A_1 \cup \cdots \cup A_k) \le P(A_1) + \cdots + P(A_k)$. ◘

Thus far we have interpreted the probability of an event of a given experiment as being a measure of how frequency the event will occur when the experiment is continually repeated. However, there are also other uses of the term probability. Probability can be interpreted as a measure of the individual's ***belief***. Furthermore, it seems logic to suppose that a "measure of belief" should satisfy all of the axioms of probability.

2.4 Conditional Probability

A major objective of probability modeling is to determine how likely it is that an event *A* will occur when a certain experiment is performed. However, there are numerous cases in which the probability assigned to *A* will be affected by knowledge of the occurrence or nonoccurrence of another event *B*. In such an example, we will use the terminology "***conditional probability of A given B***." The notation $P(A \mid B)$ will be used to distinguish between this new concept and ordinary probability $P(A)$.

We consider only those outcomes of the random experiment that are elements of *B*; in essence, we take *B* to be a sample space, called the ***reduced sample space***. Since *B* is now the sample space, the only elements of *A* that concern us are those, if any, that are also elements of *B*, that is, the elements of $A \cap B$. It seems desirable, then, to define the symbol $P(A \mid B)$ in such a way that $P(B \mid B) = 1$ and $P(A \mid B) = P(A \cap B \mid B)$.

Moreover, from a relative frequency point of view, it would seem logically inconsistent if we did not require that the ratio of the probabilities of the events $A \cap B$ and B, relative to the space B, be the same as the ratio of the probabilities of these events of these events relative to the space *S*; that is, we should have

$$\frac{P(A \cap B \mid B)}{P(B \mid B)} = \frac{P(A \cap B)}{P(B)}.$$

Hence, we have the following suitable definition of conditional probability of A given B $P(A \mid B)$.

Definition 2.4-1: The conditional probability of an event *A*, given the event *B*, is defined by

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \qquad \text{provided } P(B) \neq 0. \quad ◘$$

Moreover, we have

(1) $P(A \mid B) \geq 0$.

(2) $P(A_1 \cup A_2 \cup \cdots \mid B) = P(A_1 \mid B) + P(A_2 \mid B) + \cdots$, provided $A_1, A_2, \cdots$ are mutually exclusive sets.

(3) $P(B \mid B) = 1$.

Note that relative the reduced sample space B, conditional probabilities defined by above satisfy the original definition of probability, and thus conditional probabilities enjoy all the usual properties of probability on the reduced sample space.

Example 2.4-1: A hand of 5 cards is to be dealt at random and without replacement from an ordinary deck of 52 playing cards. The conditional probability of an all-spade hand (*A*), relative to the hypothesis that there are at least 4 spades in the hand (*B*), is, since $A \cap B = A$,

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(B)} = \frac{\binom{13}{5} \Big/ \binom{52}{5}}{\left[\binom{13}{4}\binom{39}{1} + \binom{13}{5}\right] \Big/ \binom{52}{5}}. \quad ◘$$

Theorem 2.4-1 (The Multiplicative Law of Probability): For any events *A* and *B*,

$$P(A \cap B) = P(A)P(B \mid A) = P(B)P(A \mid B). \quad ◘$$

Example 2.4-2: A bowl contains eight chips. Three of the chips are red and the remaining five are blue. Two chips are to be drawn successively, at random and without replacement. We want to compute the probability that the first draw results in a red chip (*B*) and the second draw results in a blue chip (*A*). It is reasonable to assign the following probabilities:

$$P(B) = 3/8 \qquad \text{and} \qquad P(A \mid B) = 5/7.$$

Thus, under these assignments, we have $P(A \cap B) = P(B)P(A \mid B) = (3/8)(5/7) = 15/56$. ◘

Definition 2.4-2: For some positive integer *k*, let the sets $A_1, A_2, \ldots, A_k$ be such that

(1) **(Mutually Exhaustive)** $S = A_1 \cup A_2 \cup \cdots \cup A_k$.

(2) **(Mutually Exclusive)** $A_i \cap A_j = \varnothing$ for $i \neq j$.

Then the collection of sets $\{A_1, A_2, \ldots, A_k\}$ is said to be a ***partition*** of *S*. ◘

Note that if *B* is any subset of *S* and $\{A_1, A_2, \ldots, A_k\}$ is a partition of *S*, *B* can be decomposed as follows: $B = (B \cap A_1) \cup (B \cap A_2) \cup \cdots \cup (B \cap A_k)$.

Theorem 2.4-2 (Total Probability): Suppose that $A_1, A_2, \ldots, A_k$ is a **partition** of *S* such that $P(A_i) > 0$, for $i = 1, 2, \ldots, k$. Then for any event *B*,

$$P(B) = \sum_{i=1}^{k} P(A_i)P(B \mid A_i).$$

Proof: Since the events $A_1 \cap B,\ A_2 \cap B, \ldots, A_k \cap B$ are mutually exclusive, it follows that

$$P(B) = \sum_{i=1}^{k} P(B \cap A_i)$$

and the theorem results from applying the Multiplicative Law of Probability to each term in this summation. ◘

Theorem 2.4-3 (Bayes' Rule): Suppose that $A_1, A_2, \ldots, A_k$ is a **partition** of *S* such that $P(A_i) > 0$, for for $i = 1, 2, \ldots, k$. Then for any event *B*

$$P(A_j \mid B) = \frac{P(A_j)P(B \mid A_j)}{\sum_{i=1}^{k} P(A_i)P(B \mid A_i)}.$$

Proof: from the definition of the conditional probability, and the Multiplicative Theorem, we have

$$P(A_j \mid B) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(A_j)P(B \mid A_j)}{P(B)}.$$

The theorem follows by replacing the denominator with the total probability $\sum_{i=1}^{k} P(A_i)P(B \mid A_i)$. ◘

Example 2.4-3: In a certain factory, machine I, II, and III are all producing springs of the same length. Of their production, machine I, II, and III produce 2, 1, and 3% defective springs, respectively. Of the total production of springs in the factory, machine I produces 35%, machine II produces 25%, and machine III produces 40%. If one spring is selected at random from the total springs produced in a day, the probability that it is defective, in an obvious notation, equals

$$P(D) = P(\mathrm{I})P(D \mid \mathrm{I}) + P(\mathrm{II})P(D \mid \mathrm{II}) + P(\mathrm{III})P(D \mid \mathrm{III})$$

$$= 0.35 \times 0.02 + 0.25 \times 0.01 + 0.4 \times 0.03 = 0.0215 .$$

If the selected spring is defective, the conditional probability that it was produced by machine III is, by Bayes' rule,

$$P(\mathrm{III} \mid D) = \frac{P(\mathrm{III})P(D \mid \mathrm{III})}{P(D)} = \frac{0.4 \times 0.03}{0.0215} = \frac{120}{215}$$

Note that I, II, and III are mutually ***exclusive*** and ***exhaustive*** events. ◘

Example 2.4-4: In answering a question on a multiple-choice test, a student either knows the answer or guess. Let p be the probability that the student knows the answer and $1-p$ the probability that the student guesses. Assume that a student who guesses at the answer will be correct with probability $1/m$, where m is the number of multiple-choice alternatives. What is the conditional probability that a student knew the answer to a question, given that he or she answered it correctly?

Solution: Let C and K denote, respectively, the events that the student answers the question correctly and the events that he or she actually knows the answer. Now

$$P(K \mid C) = \frac{P(K \cap C)}{P(C)} = \frac{P(K)P(C \mid K)}{P(K)P(C \mid K) + P(K^c)P(C \mid K^c)} = \frac{p}{p + (1/m)(1-p)} = \frac{mp}{1 + (m-1)\mathrm{p}}$$

Thus, for example, if $m = 5$, $p = 0.5$, then the probability that a student knew the answer to a question he or she correctly answered is $5/6$. ◘

2.5 Independence

Definition 2.5-1: Two events A and B are called ***<u>independent events</u>*** if

$$P(A \cap B) = P(A)P(B) .$$

Otherwise, A and B are called ***<u>dependent events</u>***. ◘

Example 2.5-1: A card is selected at random from an ordinary deck of 52 playing cards. If E is the event that the selected card is an ace and F is the event that it is spade, then E and F are independent. This follows because $P(E \cap F) = 1/52$, whereas $P(E) = 4/52$ and $P(F) = 13/52$. ◘

Theorem 2.5-1: If A and B are events such that $P(A) > 0$ and $P(B) > 0$, then A and B are independent if and only if either of the following holds:

$$P(A \mid B) = P(A) \qquad\qquad P(B \mid A) = P(B) .$$ ◘

Note that some textbooks use the **Theorem 2.5-1** as the definition of independent events.

There is often confusion between the concepts of ***independent events*** and ***mutually exclusive events***. Actually, they are quite different notions, and perhaps this is seen best by comparisons involving conditional probabilities. Specifically, if A and B are mutually exclusive, then $P(A \mid B) = P(B \mid A) = 0$, whereas for independent non-null events the conditional probabilities are nonzero as noted by **Theorem 2.5-1**. In other words, the property of being mutually exclusive involves a very strong form of ***dependence***, since, for non-null events, the occurrence of one event precludes the occurrence of the other event.

Theorem 2.5-2: Two events A and B are independent if and only if the following pairs of events are also independent:

(1) A and B^c .

(2) A^c and B .

(3) A^c and B^c .

Proof: Left as an exercise. ◘

Definition 2.5-2: The k events $A_1, A_2, \ldots, A_k$ are said to be independent or mutually independent if for every j = 2, 3, …, k and every subset of distinct indices $i_1, i_2, \ldots, i_r$,

$$P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_r}) = P(A_{i_1})P(A_{i_2})\cdots P(A_{i_r}) . \quad ◘$$

Suppose A, B, and C are three mutually independent events, according to the definition of mutually independent events, it is not sufficient simply to verify pair-wise independence. It would be necessary to verify $P(A \cap B) = P(A)P(B)$, $P(A \cap C) = P(A)P(C)$, $P(B \cap C) = P(B)P(C)$, and also $P(A \cap B \cap C) = P(A)P(B)P(C)$.

Example 2.5-2: Three events that are pair-wise independent but not independent

Consider the tossing of a coin twice, and define the three events, A, B, and C, as follows:

A is “head on the first toss.”

B is “heads on the second toss.”

C is “exactly one head and one tail (in either order) in the two tosses.”

Then we have

$$P(A) = P(B) = P(C) = 0.5$$

and $P(A \cap B) = P(B \cap C) = P(C \cap A) = 0.25 = 0.5 \times 0.5$.

It follows that any two of the events are pair-wise independent. However, since the three events cannot all occur simultaneously, we have

$$P(A \cap B \cap C) = 0 \neq 1/8 = P(A)P(B)P(C) .$$

Therefore the three events, *A*, *B*, and *C*, are not independent. ◘

Example 2.5-3: Two independent events for which the conditional probability of *A* given *B* is not equal to the probability of *A*.

This surprising but trivial example is possible due to the fact that the conditional probability $P(A \mid B)$ is undefined when $P(B) = 0$ and so cannot be equal to $P(A)$. A concrete example is provided by the choice $A = S$, the entire sample space, and $B = \varnothing$. They are independent because

$$P(A \cap B) = P(\varnothing) = 0 = P(A)P(B).$$

However, $P(A) \neq P(A \mid B)$ because this conditional probability is not defined. ◘

Example 2.5-4: Independent trials, consisting of rolling a pair of fair dice, are performed. What is the probability that an outcome of 5 appears before an outcome of 7 when the outcome of a roll is the sum of the dice?

Solution: If we let E_n denote the event that no 5 or 7 appears on the first $n-1$ trials and a 5 appears on the *n*th trial, then the desired probability is

$$P\left(\bigcup_{n=1}^{\infty} E_n\right) = \sum_{n=1}^{\infty} P(E_n).$$

Now, since *P*{5 on any trial} = 4/36 and *P*{7 on any trial} = 6/36, we obtain, by the independence of trials

$$P(E_n) = (1 - 10/36)^{n-1}(4/36)$$

and thus

$$P\left(\bigcup_{n=1}^{\infty} E_n\right) = \left(\frac{1}{9}\right)\sum_{n=1}^{\infty}\left(\frac{13}{18}\right)^{n-1} = \left(\frac{1}{9}\right)\left(\frac{1}{1-(13/18)}\right) = \frac{2}{5}.$$ ◘

This result may also have been obtained by using conditional probabilities. If we let *E* be the event that a 5 occurs before a 7, then we can obtain the desired probability, *P*(*E*), by conditioning on the first trial, as follows: Let *F* be the event that the first trial results in a 5; let *G* be the event that it results in a 7; and let *H* be the event that the first trial results in neither a 5 nor a 7. Since $E = (E \cap F) \cup (E \cap G) \cup (E \cap H)$, we have

$$\begin{aligned} P(E) &= P(E \cap F) + P(E \cap G) + P(E \cap H) \\ &= P(F)P(E \mid F) + P(G)P(E \mid G) + P(H)P(E \mid H) \end{aligned}$$

However,

$$P(E \mid F) = 1$$

$$P(E \mid G) = 0$$

$$P(E \mid H) = P(E).$$

The first two equalities are obvious. The third follows because, if the first outcome results in neither a 5 nor a 7, then at that point the situation is exactly as when the problem first started; namely, the experimenter will continually roll a pair of fair dice until either a 5 or 7 appears. Furthermore, the trials are ***independent***; therefore, the outcome of the first trial will have no effect on subsequent rolls of the dice. Since

$$P(F) = \frac{4}{36}$$

$$P(G) = \frac{6}{36}$$

$$P(H) = \frac{26}{36},$$

we see that

$$P(E) = \frac{1}{9} + P(E)\left(\frac{13}{18}\right) = \frac{2}{5}.$$

Note that the answer is quite intuitive. That is, since a 5 occur on any roll with probability 4/36 and a 7 with probability 6/36, it seems intuitive that the odds that a 5 appears before a 7 should be 6 to 4 against. The probability should be 4/10, as indeed it is.

The same argument shows that if E and F are mutually exclusive events of an experiment, then, when independent trials of this experiment are performed, the E will occur before the event F with probability

$$\frac{P(E)}{P(E) + P(F)}.$$