11 Probability Theory II

> *Krusty*: You people are pigs. [cries] I personally am gonna spit in every fiftieth burger.
> *Homer*: I like those odds.
>
> From: *The Simpsons*

11.1 P(A or B)

In the previous chapter we discussed two ways of assigning probabilities to events – one based on counting and the other deriving from empirical observations. Regardless of the method selected, it is often possible to use these computed probabilities to find probabilities of other related events. These calculations are accomplished using additional rules that we address in this chapter. Consider the following example.

Example 11.1: There are four main human blood types denoted by *O, A, B* and *AB.* What is the probability that a random donor will be a suitable source for a type *B* recipient?

Solution:

The blood types do not occur with equal frequency. Suppose empirical studies have shown that these appear in the population with relative frequencies of 0.45, 0.40, 0.10, and 0.05 for types *O*, *A*, *B* and *AB*, respectively. In addition, it is known that a person who has type *B* blood can receive blood only from a donor who has type *O* or type *B.*

The information regarding the frequency of occurrence for the four types can be interpreted as a probability statement according to the Relative Frequency Method described in Chapter 10. Therefore $P(O) = .45$, $P(A) = .40$, etc. Thus, amongst 1000 donors we would expect to find about 450 of type *O* (using the Frequency Rule) and 100 of type *B*. Since a person of type *O* cannot have type *B* and vice-versa, there are altogether about 550 persons whose type would be either (*O* or *B*). Therefore, the event in question occurs for approximately 550 out of 1000 random donors, or with probability 0.55. Observe that this answer is simply the sum $P(B) + P(O)$.■

In Example 11.1 the events *O* and *B* could not occur simultaneously for any donor, and this was crucial to analyzing the probability. Such events are called *mutually exclusive* and they play an important role in probability theory.

Definition 11.1: Events $E_1, E_2, \ldots E_n$ are called (pairwise) *mutually exclusive* if no two of them can occur at the same time. In other words, for any pair of events $P(E_i \text{ and } E_j) = 0$.■

Following the idea of Example 11.1 we obtain an important rule for mutually exclusive events.

Rule 11.1 (Law of Exclusives): If $E_1, E_2, \ldots E_n$ are pairwise mutually exclusive then the probability that E_1 or E_2 … or E_n will occur is $P(E_1)+P(E_2)+\cdots+P(E_n)$. This is also the probability that at least one of the events $E_1, E_2, \ldots E_n$ will occur.■

Example 11.2: What is the probability that a game of “craps” will terminate on the initial toss?

Solution:

In exercise 20 of Chapter 10 we discussed the game of “craps”. A person will win at this game if her first toss of two dice gives a sum of 7 or 11. A sum of 2, 3 or 12 produces an immediate loss. Other outcomes result in the game continuing.

There are five outcomes that lead to the game’s termination on the initial toss. These are sums of 7, 11, 2, 3 or 12 respectively. We denote these events by S_7, S_{11}, S_2, S_3 and S_{12}. Obviously these outcomes are pairwise mutually exclusive. Therefore, the probability that at least one will occur may be computed using the Law of Exclusives. Using Table 10.1 in Chapter 10, the sum of the five probabilities is $\frac{6}{36}+\frac{2}{36}+\frac{1}{36}+\frac{2}{36}+\frac{1}{36}=\frac{12}{36}=\frac{1}{3}$.■

It is important to verify the hypothesis that the events $E_1, E_2, \ldots E_n$ are mutually exclusive. Failure to observe it will lead to erroneous results, though that may not be as apparent as in the next example.

Example 11.3: What is the probability that tossing a single cubical die will produce either an even number or a number greater than 2?

Solution:

If we let E_1 represent the event that an even number is produced and E_2 that we toss a number greater than 2, then we are interested in $P(E_1 \text{ or } E_2)$. If we incorrectly assume that the events are mutually exclusive, the Law of Exclusives would imply that $P(E_1 \text{ or } E_2)=P(E_1)+P(E_2)$. However, $P(E_1)=1/2$ and $P(E_2)=2/3$, so the sum would yield a probability of 7/6. This exceeds one and therefore cannot represent the probability of an event (See the Zero-One Rule in Chapter 10). Let us figure out why our formula failed in this case and how we might correct it.

The event E_1 consists of the outcomes 2, 4, 6, while E_2 consists of 3, 4, 5, 6. The event (E_1 or E_2) consists of any of these outcomes, namely 2, 3, 4, 5, 6 and therefore should have probability 5/6. The outcomes in which we toss a 4 or toss a 6 satisfy both conditions, i.e. the event (E_1 and E_2), but they should only be listed once in the event (E_1 or E_2), as we have done. However, when we add the two probabilities $P(E_1)+P(E_2)$ these outcomes contribute twice to the probability. To correct the overcount we must subtract $P(E_1 \text{ and } E_2)$ from $P(E_1)+P(E_2)$. This gives $\frac{1}{2}+\frac{2}{3}-\frac{1}{3}=\frac{5}{6}$, which as we have seen is correct.■

We summarize the result of Example 11.3 in a general principle.

Rule 11.2 (Law of Disjunction): If A and B are any pair of events then $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$ **.■**

Remarks:

a) The word "disjunction" is used in logic to refer to a sentence constructed from simpler pieces using the "or" connective. Sentences built using the "and" connective are referred to as "conjunctions".

b) When A and B are mutually exclusive, the event ($A \text{ and } B$) cannot occur. In that case $P(A \text{ and } B) = 0$ and the Law of Disjunction (Rule 11.2) reduces to the Law of Exclusives (Rule 11.1).

c) There are more complicated rules showing how to compute $P(A \text{ or } B \text{ or } C)$, $P(A \text{ or } B \text{ or } C \text{ or } D)$, etc. even when the events are not mutually exclusive.

Example 11.4: In a certain class, 60% of the students passed the first of two exams, while 75% passed the second. Half the class passed both. Find the probability that a randomly selected student passed at least one exam.

Solution:

Let *A* denote the event that a student passed exam 1, *B* that the student passed exam 2. We need to find $P(A \text{ or } B)$. The events A and B are not mutually exclusive so we must use the Law of Disjunction (Rule 11.2). First, we need to determine appropriate values for $P(A)$, $P(B)$, and $P(A \text{ and } B)$. If any student in the class is equally likely to be selected by our random selection procedure, then the Equal Likelihood Principle (Rule 10.2) of Chapter 10 implies that the events *A*, *B*, (A and B) will occur with probabilities $P(A) = 0.60$, $P(B) = 0.75$ and $P(A \text{ and } B) = 0.5$. Thus from the Law of Disjunction (Rule 11.2) $P(A \text{ or } B) = 0.60 + 0.75 - 0.5 = 0.85$ ■

In Example 11.4 we were provided with additional information regarding $P(A \text{ and } B)$. For mutually exclusive events we know this probability is zero. For arbitrary events there is no way to obtain $P(A \text{ and } B)$ using only $P(A)$ and $P(B)$. However, there is an important exception to this, which forms the subject of the next section.

11.2 Independent Events

Example 11.5: Besides blood types as described in Example 11.1, there is another blood protein known as the *Rh* factor. In the general population 85% have the *Rh*+ form and the remaining 15% have the *Rh*- form of this protein. It is known that these same percentages apply to subgroups consisting of persons of a specific blood type. What is the probability that a random blood donor will be of type *A* and *Rh*+?

Solution:

Imagine 1000 randomly selected people coming to donate blood. Since, according to Example 11.1 there is a probability of .40 for a person to have type *A*, approximately 40% of the 1000, or 400, will have type *A*. Of these, according to the information given, 85% will also be *Rh*+. Hence, altogether $400\times.85 = 340$ will have type *A* and be *Rh*+. Since this is the total out of 1000, we estimate the probability of the joint occurrence as 340/1000 = .34. Notice that the final answer is simply the product of the two probabilities $.40\times.85 = .34$.■

The relationship described between blood type and *Rh* factor is an example of the concept of independent events. A more formal definition is

Definition 11.2: *A* is independent of *B* if knowing that event *B* occurs does not affect the frequency (probability) with which *A* occurs.■

To express this somewhat differently, consider two possible outcomes of an experiment, *A* and *B*. *A* is independent of *B* if, when the experiment is repeated many times, the frequency with which *A* occurs among all outcomes equals the frequency with which *A* occurs in only those outcomes satisfying *B* as well. When *A* is independent of *B* we may employ the reasoning described in Example 11.5 to deduce a fundamental formula.

Rule 11.3 (Product Rule of Independence) If event *A* is independent of event *B* then $P(A \text{ and } B) = P(A)\times P(B)$.■

Roughly speaking, independence asserts that knowledge of the occurrence of event *B* provides no additional information regarding the likelihood of occurrence of event *A*. ***Notice that for mutually exclusive events, this is decidedly <u>not</u> the case***. In that situation, when one of the events has occurred we know that the other event is <u>excluded</u> from occurring. In the probabilistic sense, mutually exclusive events are quite the opposite of independent events, in spite of some linguistic similarity in the terminology.

As stated, Definition 11.2 is asymmetric. On the face of it, *A* might be independent of *B* but not the other way around. However, we can easily see this is not the case. Suppose *A* is independent of *B*. The fraction of outcomes satisfying event *A* is approximately $P(A)$. Of these, the portion $P(A \text{ and } B)$ also satisfy *B*. The ratio of these two is

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

using Rule 11.3. Thus *B* appears among those outcomes satisfying *A* in the same proportion as it does among all outcomes, and so *B* is also independent of *A*. The upshot is that we can simply speak of a pair of events as independent, without having to fuss with the somewhat awkward wording of Definition 11.2.

How do we know whether a pair of events is independent? Many times we assume this holds based on our theoretical or intuitive understanding of the processes that generate the events in question. At other times the question of independence must be investigated from a statistical viewpoint. We consider some examples.

Example 11.6: Two unrelated donors appear at a blood donation center. Determine the following probabilities:

a) The probability that both are of type A.

b) The probability that neither is of type A.

c) The probability that at least one is of type A.

Solution:

a) Since the donors are unrelated we may assume that their blood types are independent. Thus, denoting by (A and A) the event that both donors are of type *A*, we have $P(\text{A and A})=P(\text{A})\times P(\text{A})=.40^2=.16$.

b) The event that a donor is not of type *A* is the complement of event A, which we have denoted by A^{c}. Thus, we need to find the probability that each donor satisfies the event A^{c}, which we denote by $(\text{A}^{\text{c}}\text{ and A}^{\text{c}})$. Since knowledge of one person's blood type has no bearing on the probabilities for the other person's type, these complementary events are again independent and so $P(\text{A}^{\text{c}}\text{ and A}^{\text{c}})=P(\text{A}^{\text{c}})\times P(\text{A}^{\text{c}})=.6^2=.36$.

c) The event stated can be described using the “or” construction as “A or A”, where the first “A” stands for the first person having type A, with a similar interpretation for the second “A”. We can then apply the Law of Disjunction (Rule 11.2). $P(\text{A or A}) = P(\text{A}) + P(\text{A}) - P(\text{A and A})$. Using a) gives $P(\text{A or A}) = .64$.■

Example 11.7: A hot water heater is fitted with two pressure relief valves that are supposed to open to vent excessive pressure in the tank, thereby preventing an explosion. Suppose that it is known that the valves have a failure rate of 0.01, meaning that 1 in 100 times when the pressure reaches a specified level the valve fails to open. It is believed that valve failures are independent events. What is the probability that both valves will fail to open at the prescribed pressure?

Solution:

Denoting by F the event that a valve fails, we are interested in the occurrence of the event (F and F). Assuming the hypothesis of independence we have $P(\text{F and F}) = P(\text{F}) \times P(\text{F}) = (0.01)^2$ $= 0.0001$. Thus, there is only a chance of 1 in 10,000 that both valves will fail. Since only one valve need operate correctly to obtain the desired safety feature, the redundant system appears very reliable. Building reliability through redundancy is quite common. However, it is essential to ensure the independent functioning of the redundant systems. In this instance, for example, if the valves failed with age due to internal corrosion, then a simultaneous failure for two old valves may be much more likely than the independence assumption would predict. Questions of this sort raise serious design problems for engineers.■

You certainly believe that when you toss two coins the outcomes of the two tosses are independent events (often phrased as, “A coin has no memory”). However, you decide to do an experiment to see whether empirically this is correct. Using several friends you assemble data from 250 experiments in which two coins are tossed. The data obtained is summarized in the following table, known as a 2×2 contingency table.

		2nd Toss		Row Totals
		H	T	
1st Toss	H	**71**	**52**	123
	T	**63**	**64**	127
Column Totals		134	126	250

Table 11.1

The entries in the table show the number of trials that resulted in the first and second toss outcomes stated in the respective row and column headings. For example, the number 71 indicates that 71 of the 250 double tosses produced a head on both the first and second tosses.

Example 11.8: How well does the data in the contingency table (Table 11.1) support the contention that the outcomes on each toss are independent?

Solution:

If the outcomes on each toss are independent then the probabilities of each of the events tabulated by the boldface entries should be $0.5 \times 0.5 = 0.25$. The Frequency Rule (Rule 10.4) in Chapter 10 would then imply that in the 250 trials we should get approximately $0.25 \times 250 = 62.5$ occurrences of each type. The pattern in the “T” row certainly fits this expectation, but the first row seems a rather poor fit. Unfortunately there is no way to judge such ambiguous cases by our intuition. Statisticians have developed a way of measuring whether the discrepancies from the expected pattern should be regarded as extreme, or as normal variation (often called “sampling error”). This procedure is known as the chi-square test. According to this test, discrepancies as large as those observed could take place using perfect coins in approximately 40% of experiments of the type conducted. Thus there is nothing atypical about your results and the hypothesis of independence is not refuted.■

The last example illustrates the difficulties of establishing statistically that two events are independent. The nature of statistical hypothesis testing is somewhat like the verdict in a trial. "Not guilty" does not mean "innocent". Similarly, the failure of a statistical test to find fault with a probabilistic hypothesis does not mean the hypothesis is correct, rather that the evidence displayed is not strong enough to refute it. These matters will be discussed at greater length in Chapter 16.

So far we have considered independence for a pair of events. However, we often want to apply the concept to more than two events. Although the technical definition is somewhat difficult, the concept is intuitively similar to what was stated earlier. We say that events $E_1, E_2, \ldots E_n$ are independent if the occurrence of any of them has no effect on the probability of occurrence of any other. In this situation we can assert an extension of the multiplication Rule 11.3.

Rule 11.4 (General Product Rule): If events $E_1, E_2, \ldots E_n$ are independent then

$$P(E_1 \text{ and } E_2 \text{ and } \ldots E_n) = P(E_1) \times P(E_2) \times \cdots \times P(E_n).$$■

As an application of this principle consider

Example 11.9: It is very unlikely that a single antiaircraft gun will bring down an attacking plane. Let's suppose the probability of this happening is 0.01. If 40 guns are used to defend a site and the success of each gun is independent of the success or failure of the other guns, what is the probability that an attacking plane will be shot down?

Solution:

We shall assume that the firing of each gun constitute 40 independent events, where the failure probability for each gun is 0.99 and the success probability is 0.01. We would like to find the probability of at least one success from the 40 gunners. As discussed in Chapter 10, the complement of this event is easier to analyze. The complement is the event that all gunners fail. Denoting a failure by the letter F, we want to find the probability of $(\mathrm{F}_1 \text{ and } \mathrm{F}_2 \text{ and } \ldots \mathrm{F}_{40})$, where the subscripts indicate which gun has failed. Using the independence hypothesis we have that $P(\mathrm{F}_1 \text{ and } \mathrm{F}_2 \text{ and } \ldots \mathrm{F}_{40}) = .99^{40} \approx 0.67$. Thus the probability of at least one success is approximately $1 - 0.67 = 0.33$, rather higher than most people would expect but borne out by the experiences of low altitude bombing attacks in both WW II and the Vietnam War.■

11.3 Conditional Probability & Bayes' Theorem

As we go about our daily routines we are constantly reassessing the probabilities of events based on facts that have already occurred. For example, arriving at a city subway station during the morning rush you would expect, with high probability that the next train will arrive within 5 minutes. Arriving at the same station at 10 PM you would give a much lower probability to the latter event. In each case, the assessment of the probability for the event in question is conditioned by your knowledge of some other event, in this case the time of day. This notion is captured mathematically in the concept of conditional probability. To arrive at the appropriate definition we consider a simple example.

Example 11.10: A single die is tossed and you are told that the outcome is ≥ 4. What is the probability that the toss produced an odd number?

Solution:

Let us denote by *O* the event that the toss produces an odd number and by *F* that it results in a value greater than or equal to 4. Knowing that *F* has occurred we know that the only possible outcome was one of the three equally likely tosses of 4, 5, or 6. Only one of these three (a toss of 5) satisfies the criterion for the event *O*. Thus the probability of *O*, conditioned on the occurrence of *F* and denoted by $P(O \mid F)$, is 1/3.■

We can extract the essential idea behind this computation to arrive at a general definition. To obtain $P(O \mid F)$ we found the number of elementary outcomes (three) that satisfied *F*, which we denote by #(*F*). We then found how many of these also satisfied *O*, (one). This is also the number of outcomes satisfying (*O* and *F*), which we denote by #(*O* and *F*). The conditional probability was computed to be the ratio

$$P(O \mid F) = \frac{\#(O \text{ and } F)}{\#(F)}.$$

The latter formula is somewhat asymmetric, expressing a probability on the left and frequency counts on the right. We can convert the frequencies on the right into probabilities by dividing each term in the numerator and denominator by $n = 6$, the number of “elementary” outcomes when we toss a single die (See the Equal Likelihood Principle (Rule 10.2) in Chapter 10). The resulting formula is the basis of our formal definition.

Definition 11.3 (Conditional Probability): For any two events *A* and *B* (with $P(B) \neq 0$) the conditional probability of *A*, assuming *B* has occurred, is the quantity $P(A \mid B)$ defined by

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}. ■$$

Example 11.11: Use Definition 11.3 to compute $P(O \mid F)$ for the experiment and events described in Example 11.10.

Solution:

We have $P(O \mid F) = P(O \text{ and } F)/P(F)$. The event ($O$ and F) is satisfied by the single outcome {5} and therefore $P(O \text{ and } F) = 1/6$. Similarly $P(F) = 3/6 = 1/2$. Thus

$$P(O \mid F) = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{2}{6} = \frac{1}{3}.$$

Note that $P(O \text{ and } F)$ ***cannot be computed using the Product Rule of Independence (***Rule 11.3***), since there is no reason to believe that these two events are independent***. In fact, as we discuss next, the computation here shows that these events are not independent.■

In Definition 11.2 we characterized two events as being independent if knowledge that event B occurs does not affect the likelihood that A occurs. The conditional probability construct can be used to make the wording of this definition more precise.

Rule 11.5: Two events A and B are independent if and only if $P(A \mid B) = P(A)$ or equivalently $P(B \mid A) = P(B)$.■

Rule 11.5 is often taken as a definition of independence since it captures the idea that knowledge of B's occurrence does not affect one's assessment of the probability that A occurs. Although perhaps more intuitive than the Product Rule of Independence (Rule 11.3), Rule 11.5 is in fact logically equivalent to the latter rule (See exercise 21).

Example 11.12: Show that the events O and F described in Example 11.10 are not independent.

Solution:

There are two ways to approach this problem. We can compute $P(O \text{ and } F)$ and see if it equals $P(O)P(F)$ or we can compare $P(O \mid F)$ with $P(O)$. All the probabilities we need have already been computed in Example 11.11. We have $P(O \text{ and } F) = 1/6$, while $P(O)P(F) = (\frac{1}{2})(\frac{1}{2}) = \frac{1}{4}$. Since $P(O \text{ and } F) \neq P(O)P(F)$ the two events are not independent. Alternatively, $P(O \mid F) = \frac{1}{3}$ while $P(O) = \frac{1}{2}$. This shows that knowledge of the occurrence of F affects the likelihood that O occurs. Therefore, we may again conclude that the events are not independent.■

Conditional probabilities are important because they enable us to gage how much the occurrence of one event affects the likelihood of occurrence of another. This type of computation is particularly common in analyzing risk factors for disease or other environmental hazards. We consider two such examples.

Suppose you want to document the extent to which smoking increases the risk of dying from lung cancer. The following data on lung cancer deaths (L) and smoking (S) for the population of the U.S. in 1998 come from the American Cancer Society (www.cancer.org) and the National Center for Health Statistics (www.cdc.gov/nchswww/). As in Table 11.1, the bold entries refer to the number of persons satisfying the conditions in a given row and column.

		L
Smoker Class	S	**140,000**
	S^c	**20,000**
	Total Lung Cancer Deaths	160,000

Table 11.2

Example 11.13: Find the following probabilities and discuss the meaning of each:

a) $P(L)$ b) $P(L \mid S)$ c) $P(L \mid S^c)$

Solution:

a) We interpret $P(L)$ as the probability that a randomly selected person will die from lung cancer in a single year. Since the estimated U.S. population in 1998 was 270 million, the Relative Frequency Method (Definition 10.1) of Chapter 10 yields $P(L) = \dfrac{1.6\times 10^5}{2.7\times 10^8} = .6\times 10^{-3}$. This number is often multiplied by a factor of 100,000 to yield the expected number (60) of lung cancer deaths each year amongst 100,000 persons.

b) From the definition of conditional probability $P(L \mid S) = \dfrac{P(L \text{ and } S)}{P(S)}$. According to the American Cancer Society, the number of adult smokers in the U.S. in 1998 was 47 million. Therefore $P(L \text{ and } S) = \dfrac{1.4\times 10^5}{2.7\times 10^8}$ and $P(S) = \dfrac{4.7\times 10^7}{2.7\times 10^8}$. Canceling the common denominator we obtain $P(L \mid S) = \dfrac{1.4\times 10^5}{4.7\times 10^7} = .003$.

Since $P(L) \neq P(L \mid S)$ the events are not independent. The risk of dying from lung cancer is greater for smokers than it is for a randomly selected person. It would be more informative, however, to compare the risk of dying from lung cancer for smokers with the risk for non-smokers. This will be done in answering c).

The fact that even for smokers the risk of dying from lung cancer in any year is not particularly high might lead you to conclude that smoking is not so risky after all. However, when judging the risk for an individual the relevant statistic is the risk of dying from lung cancer over a lifetime. This cannot be computed exactly based on the data given, though we can make a crude estimate based on ideas in this chapter (see exercise 15). (The Center for Disease Control estimates that a person who begins to smoke as a teenager has a lifetime risk of 1/3 of dying, usually prematurely, from a smoking related disease.) The value of $P(L\,|\,S)$ is of more concern from the point of view of society. Since there are so many smokers, a risk as small as .003 leads to an enormous amount of mortality each year.

c) We can compute $P(L\,|\,S^c)$ as we computed $P(L\,|\,S)$. Since there are 223 million non-smokers ($270\text{ million} - 47\text{ million}$), we obtain the result $P(L\,|\,S^c) = \dfrac{2\times10^4}{2.23\times10^8} = .9\times10^{-4}$. The ratio $RR = \dfrac{P(L\,|\,S)}{P(L\,|\,S^c)}$ is called the *relative risk ratio*. It measures how much more likely a smoker is to die from lung cancer than a non-smoker. We find here that $RR = 33\frac{1}{3}$. Most epidemiologists consider a relative risk ratio in excess of five an indicator of a significant public health hazard.■

When two events are not independent, knowledge of the occurrence of one may increase the likelihood of the other's occurrence. This is exactly the scenario involved in using a medical diagnostic test. Suppose there is a certain disease *D* that can be accurately diagnosed using some complicated and perhaps invasive procedure. A simple and inexpensive diagnostic procedure has been developed for detecting *D*, but this procedure sometimes produces an incorrect diagnosis. Specifically, after testing on a number of patients whose disease status was not in doubt, it was found that

$$P(+\,|\,D^c) = .03 \text{ and } P(-\,|\,D) = .01\,, \tag{11.1}$$

where the symbol $+$ denotes a "positive" response to the test and $-$ denotes a "negative" response. These probabilities are known as the *false positive rate* and *false negative rate* respectively. They can be used to compute the probability of errors in the test. However, the physician and patient are more interested in the conditional probabilities $P(D\,|\,+)$ and $P(D^c\,|\,-)$. The first of these asks for the probability that a person giving a positive test outcome actually has the disease, while the second tells how likely it is that a negative test result truly indicates the absence of the diseases. Clearly it would be desirable for each of these conditional probabilities to be fairly large (close to one). The next example shows how the computation can be done, provided we know the prevalence of the disease in the targeted population.

Example 11.14 (Bayes' Method): Suppose a diagnostic test for a disease satisfies (11.1) and that the disease is known to be prevalent in 2% of the population. Find $P(D\,|\,+)$ and $P(D^c\,|\,-)$.

Solution:

In addition to the false positive and negative rates we are given that $P(D) = .02$. Bayesian analysis (named after the 18th century English mathematician Rev. Thomas Bayes) uses the outcome of the test, its operating characteristics (11.1), and the known prevalence of the disease in the population, $P(D)$, to assess the likelihood of the disease's presence given a positive test outcome.

From the definition of conditional probability we have $P(D\,|\,+)=\dfrac{P(D \text{ and } +)}{P(+)}$. The key to the analysis is the computation of the denominator. The value of the numerator will be obtained as a by-product. The denominator $P(+)$ is the probability that a randomly chosen person will give a positive reaction to the test. We will compute this probability by imagining that we give the test to a large number of people, say 10,000. Using the information given, we compute how many of these folks would be expected to show a positive outcome for the test. The computation is organized in the tree diagram below.

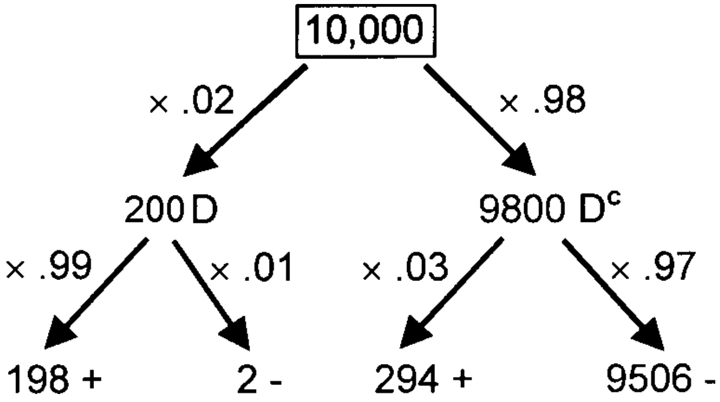

Figure 11.1

The 10,000 individuals are first classified by their disease status. Since the disease has a 2% prevalence the expected number of cases with the disease would be $10,000\times 0.02=200$. This is indicated in the top left branch. A similar computation yields the right branch. Since the group has been partitioned according to its disease status, we can now apply the probabilities determined from the false positive and negative rates to yield the last row.

Specifically, consider the 200 individuals who have the disease. We know that the false negative rate for these persons is 1% ($P(-\,|\,D)=.01$) so this group will yield $.01\times 200=2$ negative test outcomes. The remaining 198 persons must give positive outcomes. (This can also be obtained using that $P(+\,|\,D)=.99$, as shown in the figure.) Similarly 3% of the 9800 persons without the disease will test positive, accounting for the 294 in the last row. The remaining 9506 can be expected to give a negative response.

We can now determine the two probabilities $P(+)$ and $P(D \text{ and } +)$. Starting with 10,000 persons we have computed that we should find $198+294=492$ positive test results. Thus $P(+)=492/10,000=.0492$. This is the probability that a randomly selected person will give a positive test result. $P(D \text{ and } +)$ is found from the number of people who have the disease and give a positive reaction to the test. This is simply the number 198 in the last row, so $P(D \text{ and } +)=.0198$. Thus we obtain that

$$P(D\,|\,+) = \frac{P(D \text{ and } +)}{P(+)} = \frac{.0198}{.0492} \approx .40\,.$$

The reader should show using the same data in Figure 11.1 that

$$P(D^c\,|\,-) = \frac{P(D^c \text{ and } -)}{P(-)} = .99^+ \;\;,$$

meaning that the probability is greater than .99.

Thus in this situation a negative test outcome almost certainly indicates the absence of disease. However, a positive test means only about a 40% chance of disease. Most people are surprised at this result, since the test seems reliable based on the false positive and negative rates. However, a glance at Figure 11.1 reveals the source of the paradox. Since the disease is not very prevalent, most of the positive responses are false positives coming from the large number of persons who do not have the disease. In a situation of this sort a patient with a positive response would require counseling from his or her physician with an explanation as to the meaning of the result. The test should then be redone. As we examine in exercise 19 b) and c), a positive response on a second test provides a much stronger confirmation of the presence of disease.■

To carry out the analysis In Example 11.14 we used the artifice of a fictitious initial population. This is often helpful in understanding an abstract probability computation. The results, however, can usually be formulated in terms of a general probability computation. This is indeed the case here. The reader might want to verify that the following probability formula, usually referred to as Bayes' Theorem, encapsulates the computations in the example. The derivation of the formula can be found in the exercise 20. Generalizations of the formula appear in many books on probability theory.

Rule 11.6 (Bayes' Theorem): Suppose D denotes the presence of a disease and + denotes a positive response to a diagnostic test. If the false positive and false negative rates for the test are known and the overall disease prevalence $P(D)$ is also known, then the posterior probability of disease given a positive test, $P(D\,|\,+)$, is given by

$$P(D\,|\,+) = \frac{P(D)P(+\,|\,D)}{P(D)P(+\,|\,D) + P(D^c)P(+\,|\,D^c)}\,.\blacksquare \qquad (11.2)$$

11.4 Summary

Given the probabilities of known events we may be able to find the probabilities of related, more complicated events. In particular,

- if events $E_1, E_2, \ldots E_n$ are mutually exclusive then

$$P(E_1 \text{ or } E_2 \ldots \text{or } E_n) = P(E_1) + P(E_2) + \cdots + P(E_n)\,.$$

- for arbitrary events E_1 and E_2, not necessarily mutually exclusive, we have

$$P(E_1 \text{ or } E_2) = P(E_1) + P(E_2) - P(E_1 \text{ and } E_2).$$

- if events $E_1, E_2, \ldots E_n$ are <u>independent</u> then

$$P(E_1 \text{ and } E_2 \ldots \text{ and } E_n) = P(E_1) \times P(E_2) \times \cdots \times P(E_n).$$

- for arbitrary events E_1 and E_2 the conditional probability $P(E_1 \mid E_2) = \dfrac{P(E_1 \text{ and } E_2)}{P(E_2)}$. The events are independent if and only if $P(E_1 \mid E_2) = P(E_1)$.

We will see in the next chapter that these rules, together with the counting principles formulated in Chapter 10, have many applications to quantitative genetics, at both the individual and population levels.

11.5 Exercises

1. a) If you know that $P(A) = 0.3$, $P(B^c) = .75$ and *A* and *B* are independent, find $P(A \text{ or } B)$.

   b) If you don't know whether the events *A* and *B* in part a) are independent, explain why you can still assert that $P(A \text{ or } B)$ is at most 0.55.

2. A coin is tossed four times. Show how to use the notion of independence to compute the probability that all four tosses give tails.

3. Referring to the A, B, O , Rh blood type data in Example 11.1 and Example 11.5, find the probability that a randomly selected person has

   a) type A and Rh+

   b) type O, and Rh-

   c) Blood is needed for a person of type A and Rh+. A donor must have type A or O with the same Rh factor. What is the probability that a suitable donor will be found among five randomly selected persons? (Hint: Consider the event that none of the five donors is suitable.)

4. A coin is tossed three times. Determine which of the following outcomes describe mutually exclusive events.

   a) A: all tosses are heads, B: all tosses are tails.

   b) A: at least one toss is a tail, B: at least one toss is a head.

   c) For both a) and b) find $P(A \text{ or } B)$ and $P(A \text{ and } B)$.

d) Which pairs of events A and B, if any, as described in parts a) and b) are independent events?

5. A person is picked at random from a large group and we record the day of the week of his or her 1999 birthday. Assuming any day has equal chance of occurring (not quite true because one day of the week occurs 53 times in the year and the others only 52) find the following probabilities.

   a) The probability that the birthday of a randomly selected person is on a Monday, Wednesday, or Friday.

   b) The probability that two randomly selected persons both have birthdays on Monday.

   c) The probability that at least one of two randomly selected persons will have a birthday on Monday.

   d) The probability that two randomly selected people have birthdays on the same day of the week.

6. Suppose that you have a 40% chance of winning at a certain gambling game and that the outcomes of different games are independent of each other.

   a) What is your chance of winning at least once when you play the game twice?

   b) What is your chance of winning at least once when you play the game ten times?

   c) What are your chances of winning twice when you play three times?

7. In a certain locale the annual rainfall can be classified as follows:

   A = rainfall is < 20 in. per year

   B = rainfall is $\geq$ 20 but < 30 in. per year

   C = rainfall is $\geq$ 30 but < 40 in. per year

   D = rainfall is $\geq$ 40 in. per year

   Suppose the probabilities of these events are:

   $P(A)=.15$ $\quad$ $P(B)=.20$ $\quad$ $P(C)=.30$ $\quad$ $P(D)=.35$

   a) If an adequate rainfall is an annual amount $\geq$ 30 inches, what is the probability of an adequate rainfall?

   b) Assuming the amount of rainfall in one year is independent of the amount in any other year, what is the probability of two consecutive years with less than 20 inches per year?

   c) What is the probability that in at least one of three consecutive years the amount of rainfall will be < 20 inches?

8. A bowl contains 3 red balls and 5 black balls. You pick a ball from the bowl, record its color, and then replace the ball in the bowl (called *sampling with replacement*).

a) If this experiment is performed three times, what is the probability that a black ball will be selected each time?

b) If this experiment is repeated three times what is the probability that the three selections will include both a red ball and a black ball?

9. Suppose the selection process in exercise 8 had been done without replacement (in other words, the balls are discarded after selection). Find the answer to parts a) and b) in that exercise using this sampling scheme? (Hint: You can use either conditional probabilities or counting techniques.)

10. An urn contains 3 red balls, 4 black balls and 2 white balls. You pick two balls from the urn without replacement.

a) Find the probability that both balls are red.

b) Find the probability that both balls are of the same color.

c) Suppose you are told that the balls that were picked were of the same color, but you are not told which color. Your friend argues that the probability that the balls are red must be 1/3, because red is one of the three possible colors. Give a precise formulation of the problem in terms of probabilities and determine whether or not your friend is correct.

11. Suppose there are three medications A, B and C available for the treatment of a certain psychiatric disorder. Clinically it has been shown that A is effective for 40% of patients, B also for 40% and C for 30%. The drugs are known to affect different components of the nervous system and therefore their clinical effectiveness might be independent of each other.

a) Assuming the effects are independent of each other, what would be the probability that at least one of the medications was effective in treating the illness of a randomly chosen person?

b) If the drugs' actions were not independent, would the answer to part a) become larger or smaller or can you not tell? Give some explanation for your answer.

12. A pair of dice is thrown.

a) What is the probability that the sum of the faces equals five? (Call this event A.)

b) What is the probability that the difference between the faces (larger minus smaller) equals three? (Call this event B.)

c) Find $P(\mathrm{A}\,|\,\mathrm{B})$. Are A and B independent? Explain.

13. A college requires all freshmen to take courses X and Y. Records show that 15% receive an A in course X, while only 10% receive an A in course Y. Altogether, 23.5% of the students get an A in one or the other course.

a) What fraction of the students receives an A in both courses?

b) What is the probability that a student who receives an A in course X will also receive an A in course Y?

c) Is receiving an A grade in course X independent of receiving that grade in course Y? Justify your answer.

14. 200 asthmatics were chosen to study the effects of a new anti-asthmatic drug. In the 2×3 contingency table below each patient is classified according to

(i) his or her prior allergic history with anti-asthmatics (row categories) and

(ii) the effect of the new medication on the patient (column categories).

The highlighted boxes give the number of patients falling into the two categories specified in the same row and column.

	Improvement (I)	**No Change (N)**	**Allergic Reaction (AR)**	**Row Totals**
Allergic History (A)		**15**	**15**	60
No Allergic History (A^c)				
Column Totals	140	20		200

a) Fill in the missing entries in the table.

b) Based on the data, what is the probability that a randomly chosen asthmatic will experience an improvement with the new drug?

c) What is the probability that an asthmatic with a history of allergies will have an improvement? Express the answer as a conditional probability.

d) Based on the data would you say that the events *A* and *I* are independent? Justify your answer and explain its medical significance.

15. a) In Example 11.13 we determined that the probability of a smoker dying from lung cancer in a given year is 0.003. What is the probability a smoker does not die from lung cancer during a given year?

b) Based on the answer to a) compute the probability that a smoker does not die from lung cancer over a period of 40 years. What probabilistic assumptions are you making in doing the calculation?

c) Based on b) what is the probability of that a person who smokes for 40 years will die of lung cancer during that time?

16. Occupational exposure to benzene is suspected to be a risk factor in the development of cancers of the lymphatic and circulatory systems (including leukemia). A long-term (1940s - 1970s) study of workers at a Pliofilm Corporation rubber manufacturing plant gave the

following data: (Source: *Should We Risk It?*, D. Kammen & D. Hassenzahl, 1999, Princeton Univ. Press)

# person-years	deaths from **l**eukemia (L)	deaths from other lymphatic & **c**irculatory cancers (C)	**T**otal deaths from lymphatic & circulatory cancers (T)
52,584	14	7	21

(Note: Person-years equals the sum of the number of years each person worked in the plant.)

a) According to this data, for each person-year of working at the plant, what is the probability of dying from leukemia, i.e. find $P(\mathrm{L}|\,\mathrm{W})$, where W represents the condition that a person works at the plant for one year? Compute the analogous probabilities for the other two categories, C and T. How would you make these numbers more understandable to someone without a technical background?

b) The workers in the plant were carefully matched with controls (a so-called case control study) who were similar in age, sex (primarily male), income, prior health history, but had a different occupational background. In this group the probabilities of death from one of the above causes for each person-year of observation were: $P(\mathrm{L}|\,\mathrm{W^c}) = 9\times10^{-5}$, $P(\mathrm{C|W^c}) = 1.4\times10^{-4}$, $P(\mathrm{T|W^c}) = 2.3\times10^{-4}$. Based on these figures determine the relative risk of death for the plant workers in each of the cancer categories. What preliminary conclusions regarding the occupational risk might you draw from the data?

17. In 1854 John Snow, a founding member of the London Epidemiological Society, investigated a recent cholera epidemic in London. He examined the number of deaths from cholera over a seven-week period in houses whose water was supplied by the Southwark & Vauxhall Co. and in those receiving water from the Lambeth Co. One company drew its water from the Thames River upstream from the city, while the other drew it downstream from the city. The data is summarized below.

Water Supply	# of Houses	Cholera Deaths
Southwark & Vauxhall Co.	40,046	1,263
Lambeth Co.	26,107	98
Rest of London	256,423	1422

Source: *Foundations of Epidemiology*, A. M. & D. Lilienfeld, 2nd ed., 1980, Oxford Univ. Press

a) For each water supplier (including the “Rest of London”), compute the number of deaths from cholera per household. Can the latter number be interpreted as a probability?

b) For each water supplier, can we compute from the data the probability that an individual using that supplier will die from cholera? If not, what additional data is needed?

c) Using the answer to a) how would you define a relative risk ratio from this data? On the basis of your relative risk ratio what would you conclude about the safety of each water supply?

d) Based on the answer to c) what might be the source of the cholera infection? What other statistical investigation should be done to rule out other possibilities? (The bacterial origins of cholera were discovered by Robert Koch in 1883.)

18. A disease D is present in 2% of the population. Preliminary trials of a diagnostic test for the disease show a false positive rate of 2% and a false negative rate of 1%. A physician friend of yours notices that in her practice only about half of the patients who test positive turn out to actually have the disease. She was expecting the number to be much higher. How would you explain to her the reason for her observation? Be as precise as possible but do give an explanation, not just a bunch of calculations.

19. a) Suppose in Example 11.14 that the incidence of the disease in the targeted population was 10%. How will this affect the value of the test as a diagnostic tool? Justify your answer with a suitable computation.

b) With the scenario described in Example 11.14 suppose a patient has a positive test result and the doctor decides to administer the test again. Based on the first positive outcome the patient now belongs to the target population of persons who have tested positive. Using the data in Figure 11.1 show that the incidence of disease in this group is 40%.

c) Referring to part b), if the second test is positive, what is the probability that the patient has the disease? If the second test is negative, what is the probability that the patient is free of the disease?

20. a) To prove formula (11.2) we need to show that the denominator is $P(+)$ and the numerator is $P(D \text{ and } +)$. Explain why the event $(+)$ can be decomposed as the disjunction of the two mutually exclusive events $(D \text{ and } +)$ and $(D^c \text{ and } +)$, i.e. show that

$$(+) = (D \text{ and } +) \text{ or } (D^c \text{ and } +) .$$

b) Using the result in a) complete the derivation of formula (11.2).

21. Prove the assertion made in the text that two events A and B satisfy the Product Rule for Independence (Rule 11.3) if and only if $P(A \mid B) = P(A)$.