Background Information

Introduction to Probability

Environmental health risks are often expressed in terms of probability. Thus, an understanding of the basic concepts of probability is critical when making sound choices regarding such issues. The aspects of probability most relevant to risk are discussed here.

Probability is the likelihood that a random event will occur. Probability is expressed as a fraction between 0 and 1, with 0 indicating an event will not occur and 1 indicating it will definitely occur, or as a percentage greater than or equal to 0% and less than or equal to 100%. Thus the probability of event A occurring, p(A), is:

$$0 \le p(A) \le 1$$
$$0\% \le p(A) \le 100\%$$

Probability Space

The roll of a die is a good example of a random process. It is impossible to accurately predict which side will land face up. However, you can expect that each of the six sides will land face up approximately one-sixth of the time.

In a case such as the roll of the die where each of the possible outcomes is equally likely, the probability that any one of the possibilities will occur is easy to calculate. First we determine the probability space, which is the set of all possible outcomes. For the roll of the die, the probability space is: {1, 2, 3, 4, 5, 6}.

For equally likely outcomes, the probability that any one of the possible outcomes will occur is:

$$p(A) = \frac{N(A)}{N}$$

where N(A) is the number of times outcome A appears in the probability space and N is the number of outcomes that can occur.

Now, suppose we want to know the probability that a roll of the die will result in a 4, which we can refer to as p(4):

$$p(4) = \frac{N(4)}{N} = \frac{1}{6} = 16.7\%$$

This technique becomes more complex when the probabilities of several outcomes (collectively called an event) are considered. For example, if you roll a die twice, each roll has six possible outcomes so the total number of outcomes from the two rolls is 6 x 6, or 36.

(This is known as the multiplication principle.) The probability space for two rolls is shown in Table 1. Each cell in the table represents an event. For example, the first cell in the upper-left corner shows an event in which each roll resulted in a 1.

Table 1: Probability Space for Two Dice

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

In the darker shaded cell, we can see the probability of rolling first a 3 and then a 4 is:

$$p(3,4) = \frac{N(3,4)}{N} = \frac{1}{36} = 2.8\%$$

In the lighter shaded cells, we see the probability of rolling a 5 and a 1 in either order is:

$$p(5,1;1,5) = \frac{N(5,1;1,5)}{N} = \frac{2}{36} = \frac{1}{18} = 5.6\%$$

Inferential Statistics

So far we have only considered situations in which determining N(A) and N is relatively easy. These situations are distinguished by the following features: they are repeatable; they involve an event that, in different trials, sometimes happens and sometimes does not; and, while each trial has one of a number of equally likely outcomes, the event is a particular subset of these outcomes. Thus the event of rolling a die twice and getting a 5 and a 1 in either order is repeatable; it sometimes happens and sometimes does not; and it is a subset of two equally likely outcomes out of the total of 36.

But how can we come up with the probability that, for example, a member of a population will test positive for a virus? If the population is very small, everyone could be tested, but then the probability is no longer of interest because the results are known. (Probability is useful only for situations about which we have incomplete knowledge of the outcomes.)

Inferential statistics are used to draw conclusions about a larger population of people or things using data from a smaller sample of the population. These conclusions are useful when studying the entire population is impossible. Instead, a representative sample is picked randomly from the population. The test results from this sample are used to determine probabilities for the population.

Because the sample is only a portion of the whole population, the best available estimates of important parameters must be selected and the reliability of these estimates must be established. For example, the best estimate of the population mean is the mean of a sample of the population. The reliability of this sample mean as a representative of the population mean depends on how closely the sample represents the population, which is not easy to determine. The best way to promote reliability of the estimates is to use a sample that is truly random.

Conditional Probability

In our discussion of probability space, we determined that the probability that two rolls of a die would produce first a 3 and then a 4 was 1/36, or 2.8%. What happens to the probability if the result of the first throw is known?

Table 2: Probability Space for Two Dice When First Throw Is a 3

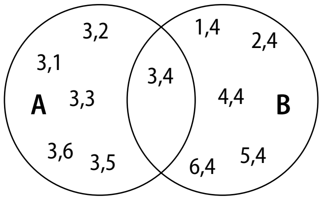
(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

If the result of the first throw is something other than a 3, then the probability of rolling 3 and then 4 becomes 0. However, if the result of the first throw is a 3, the probability space consists of those six outcomes which begin with a 3: {3,1; 3,2; 3,3; 3,4; 3,5; 3,6}. The probability of tossing the combination of {3,4} is now:

$$p(3,4) = \frac{N(3,4)}{N} = \frac{1}{6} = 16.7\%$$

This situation is an example of conditional probability.

Before we consider conditional probability further, we need to understand the concept of set intersection. The intersection of sets A and B, which is symbolized by "A∩B," contains only the events in both A and B. The intersection in our example of two tosses of the die would be A∩B = {3,4}. One way to demonstrate this is using a Venn diagram.



The circle labeled A represents all the possibilities where the first throw is a 3, and the circle labeled B represents all the possibilities where the second throw is a 4. Each of these contains a total of six possibilities, including {3,4}, the intersection set of A and B.

Recalling that N is the total number of possible outcomes, the probability of the intersection set of A and B occurring is:

$$p(A \cap B) = \frac{N(A \cap B)}{N} = \frac{1}{36} = 2.8\%$$

We can also calculate the conditional probability of B happening when A has already happened, p(B|A), as follows:

$$p(B|A) = \frac{(A \cap B)}{(A)} = \frac{N(A \cap B)}{N(A)}$$

The vertical line in p(B|A) means "given that." The probability of event B occurring given that event A has already occurred is the probability of both A and B happening (that is, the probability of the intersection) divided by the probability of A happening. In our example, the probability of the second toss yielding a 4 after the first has produced a 3 is:

$$p(B|A) = \frac{N(A \cap B)}{N(A)} = \frac{1/36}{1/6} = \frac{6}{36} = \frac{1}{6} = 16.7\%$$

Notice that in this case, p(B|A) = p(B). This is true because the two events, A and B, are independent. The probability of B happening is the same whether or not A happens. This will be true only when $p(A \cap B) = p(A) \times p(B)$.

For example, suppose that studies have shown that the probability of a resident of a city having a particular virus in his or her system is 2.00%. A clinical test for this virus yields 3.00% false negatives (test negative although the person has the virus) and 5.00% false positives (test positive although the person does not have the virus).

What is the conditional probability of a person having the virus given that he or she tested positive? The formula tells us that it is equal to the probability of a person having the virus and testing positive, $p(A \cap B)$ (the probability of the intersection of the two events), divided by the probability of having a positive test, p(A).

Assume a population of 10,000 people. The number of people in this population who have the virus and test positive is:

$$N(A\cap B) = (10{,}000)(0.02)\times(1-0.03) = (10{,}000)(0.02)\times(0.97) = 194$$

The number of people who don't have the virus but test positive is:

$$(10{,}000)(1-0.02)\times(0.05) = (9{,}800)\times(0.05) = 490$$

The number of individuals testing positive is the sum of these two numbers:

$$N(A) = 194+490 = 684$$

So the conditional probability of having the virus given a positive test is:

$$p(B|A) = \frac{N(A\cap B)}{N(A)} = \frac{194}{684} = 0.284 \text{ or } 28.4\%$$

Testing positive for the virus thus increases the probability that an individual has the virus from 2.00% to 28.4%. In other words, if a person tested positive, they still have a greater than 70% probability of not having the virus. Why would anyone use such a test? Tests that yield 5% false positives may be called for in the initial screening of a population—cases in which a more definitive test is very expensive or more painful and time consuming. In practice, a positive on this first test would call for a follow-up test on the possibility of infection.

References

Dieffenbach, R. and Fisher, A. Department of Mathematics and Statistics, Miami University Middletown, OH. Personal communication, 2001.

Downing, D. and Clark, J. *Statistics the Easy Way,* 2nd ed.; Barron's Educational Series: Hauppage, NY, 1989.

Hector C. Parr Website. Essays on Maths and Physics: Probability. www.c-parr.freeserve.co.uk/hcp/prob.htm (accessed February 20, 2001).

University of Wisconsin-Milwaukee Website. Directories. Personal Pages. Eric S. Key's Home Page. Lecture 5: Conditional Probability. www.uwm.edu/People/ericskey/361F98/L05/index.html (accessed February 20, 2001).

Walsh, J. *True Odds: How Risk Affects Your Everyday Life;* Merritt: Santa Monica, CA, 1996.