Chapter 1. Probability

1.1. Introduction

In our quest to describe the properties of materials with statistical accuracy, we must first learn the language of statistics. Statistics itself is an applied field built upon the theory of probability, a mathematical discipline. Therefore, it is essential that we familiarize ourselves with the most basic premises of probability theory.

This chapter introduces the vocabulary of probability, counting rules, and rules for the probabilities of intersections, unions and conditional relationships.

1.2. Vocabulary

In this section, we introduce a the vocabulary required for talking about probability.

Set

A set is a collection of objects or outcomes. Braces denote a set. For example, the set of letters in the alphabet is

$$S = \{a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z\} \tag{1.1}$$

The set of integers greater than 1 and less than 5 is

$$S = \{2,3,4\} \tag{1.2}$$

Element

An element is one member of a set. For example, ‘a’ is an element of the set defined in equation (1.1).

Rule

The elements of a set can frequently be described by a rule. For example, we can rewrite the set of integers greater than 1 and less than 5 as

$$S = \{n \in I | 1 < n < 5\} \tag{1.3}$$

The symbol, $\in$, means 'is an element of' and the pipe reads as 'such that", so this statement should be read as the set including all *n* that are elements of the set of integer numbers (*I*), such that *n* is greater than 1 and less than 5.

Some sets must be written with a rule because they have a large or infinite number of elements that cannot be listed explicitly. For example, the set including all *x* that are elements of the set of real numbers ($\Re$), such that *x* is greater than 1 and less than 5 is written as

$$S = \{x \in \Re | 1 < x < 5\} \tag{1.4}$$

Another example of a rule is the set of ordered pairs (*x*,*y*) such that the satisfy a specific equation.

$$S = \{(x, y) | x^2 + y^2 \leq 4\} \tag{1.5}$$

Here the number of real solutions to the equation is infinite. Also an element of this example is an ordered pair.

Subset

A subset is part of a larger set. For example, the set of vowels are a subset of the set of letters, given in equation (1.1).

$$V = \{a, e, i, o, u\} \tag{1.6}$$

Null set

The null set is a set with no elements in it. The null set is written,

$$N = \{\emptyset\} \tag{1.7}$$

Complement

The complement of a subset, *A*, of a set, *S*, is defined as A' and includes all elements of *S* that are not in *A*. For example, the complement of the set of vowels, equation (1.6), from the set of letters, equation (1.1), is the set of consonants,

$$V' = \{b, c, d, f, g, h, j, k, l, m, n, p, q, r, s, t, v, w, x, y, z\} \tag{1.8}$$

The complement of the set, *S*, is the null set, *N*. The complement of the null set, *N*, is the entire set, *S*. The complement of a complement of a subset, *A*, is *A*, $(A')' = A$.

Sample Space

When we specifically apply this vocabulary to experiments, we say that the set of all possible outcomes of a statistical experiment is called the sample space, *S*. Several examples follow. When you flip a coin once, the sample space is heads or tails, *S*={H,T}. When you toss a six-sided die, the sample space is a number from 1 to 6, *S*={1,2,3,4,5,6}. When you flip two coins, the sample space is *S*={HH,HT,TH,TT}.

Event

An event is a subset of a sample space. Example, in the experiment of flipping two coins, where the sample space is *S*={HH,HT,TH,TT}, possible events include getting two heads, *B*={HH}, getting the same result twice, *B*={HH,TT}, getting a different result each time, *B*={HT,TH}, or getting any result, *B*=*S*, or getting no result, $B = \{\emptyset\}$. (In a properly run experiment, the probability of getting the null set, should be zero.)

Intersection

The intersection of two events *A* and *B*, denoted by the symbol, $A \cap B$, is the event containing all elements in both *A* and *B*. The key operating word for the intersection is "and". The elements in the intersection are in *A* and *B*. For example,

If *B*={HH,TT} and *A*={HH,HT,TH}, then $A \cap B = \{HH\}$.
If *B*={HH,TT} and *A*={HT,TH} then $A \cap B = \{\emptyset\}$.
If *B*={HH,TT} and *A*={HH,TT}, then $A \cap B = A = B$.

Mutually exclusive, or disjoint

Two events *A* and *B* are mutually exclusive if there intersection is the null set, $A \cap B = \{\emptyset\}$, that is, if *A* and *B* have no common elements. For example: $A \cap A' = \{\emptyset\}$, that is, an event and its complement are by definition mutually exclusive events.

Union

The union of two events *A* and *B*, denoted by the symbol, $A \cup B$, is the event containing all elements in either A or B. The key operating word for the union is "or". The elements in the union are in *A* or *B*. For example, $A \cup A' = S$, that is, the union of an event and its complement is by definition mutually the sample space.

Venn Diagrams

Venn diagrams are a graphical way to express sets and events. The bounding box of the Venn Diagram contains the entire sample space, S. In Figure 1, we provide several examples of the use of Venn Diagrams in graphically interpreting various combinations of intersections, unions, and complements of sets.

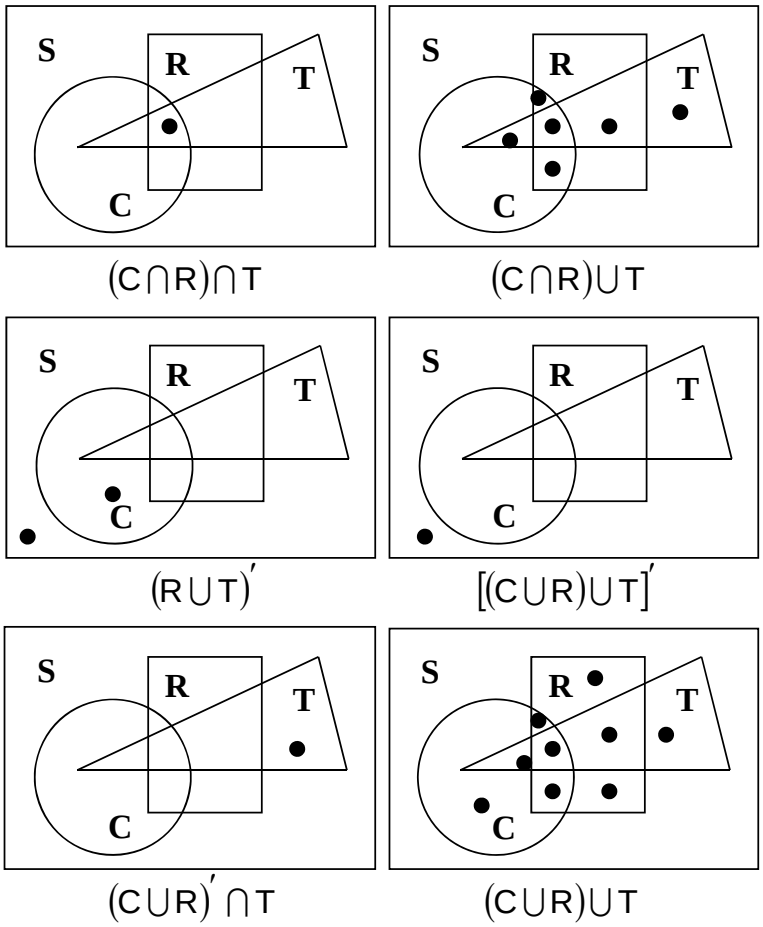

Figure 1. Visualizing probabilities through the use of Venn Diagrams.

1.3. Counting Rules

We need counting rules in probability because the probability of an event *A* is the ratio of the number of elements in A over the number of elements in the sample space *S*.

$$P(A) = \frac{\#\,of \text{ elements in } A}{\#\,of \text{ elements in } S} \tag{1.9}$$

Therefore, we need to know how to count the number of elements in *A* and *S*. We will study three counting rules:

1. generalized multiplication rule
2. permutations of distinct objects rule

3. combinations of distinct objects rule

Independence

Later in this text we will provide a formal definition of independence. However, we need a qualitative understanding of independence now. By independence, we mean that the result of the one experiment (or portion of an experiment) does not depend on the result of another experiment (or portion of an experiment). This is an important distinction and can be illustrated with the following simple example. Consider a bag containing 5 black marbles and 5 white marbles. Consider an experiment where we randomly draw three marbles out of the bag, but we do so one marble at a time and replace the marble bag in the bag after each drawing. Clearly, the probability of getting a white marble on any single draw is always 0.5 via equation (1.9), because there are always 5 white and 5 black marbles in the bag.

$$P(W) = \frac{\#of \text{ elements in } W}{\#of \text{ elements in } S} = \frac{5}{10} = 0.5 \qquad (1.10)$$

Thus sequential drawings of marbles with replacement are independent operations. The probability that we draw a white marble on the second draw is independent of the result of the first draw.

Now consider a second experiment in which we randomly draw three marbles from the same bag containing 5 black marbles and 5 white marbles, however this time we do so without replacing the marble. Therefore, the number of marbles changes each time we draw a marble. The probability of the first marble being white is still 0.5 as was thee case above. However, the probability of drawing a white marble on the second draw depends on the result of the first draw. If the first draw was a white marble, then only 4 of the 9 remaining marbles are white and the probability of getting a white marble on the second draw given that we chose a white marble on the first draw is 4/9. On the other hand, if the first draw was a black marble, then 5 of the 9 remaining marbles are white and the probability of getting a white marble on the second draw given that we chose a black marble on the first draw is 5/9. Thus in this second experiment, each drawing is not independent.

Generalized Multiplication rule

If an operation can be performed in n_1 ways, and if for each of these, a second independent operation can be performed in n_2 ways, and for each of the first two, a third independent operation can be performed in n_3 ways, and so forth, the sequence of *k* operations can be performed in

$$\#\text{ of ways} = n_1 n_2 n_3 ... n_k \quad \text{for } k \text{ operations} \qquad (1.11)$$

For example, if you consider the set of elements composed of one coin toss, followed by one six-sided die roll, followed by drawing from a hat containing *m* names. The number of elements in the set is 2*6**m*. In this case it is clear that the operations are independent of each other. Flipping the coin does not change probability of each outcome on a die.

Permutations

A permutation is an arrangement of all or part of a set of objects. A permutation is a grouping of elements arranged in a particular way. For example, how many ways can you order the letters A, B & C. The six permutations are ABC, ACB, BAC, BCA, CAB and CBA. All sequences contain the same letters but in different orders. The key concept in permutations is that 'order matters'.

The number of permutations of *n* distinct objects is *n*! That is read as "n factorial".

$$n! = n(n-1)(n-2)(n-3)...3 \cdot 2 \cdot 1 \tag{1.12}$$

The factorical only applies to non-negative integers. By definition, the factorial of zero is 1,

$$0! \equiv 1 \tag{1.13}$$

The number of permutations of *n* distinct objects taken *r* at a time, where $r \leq n$, is

$${}_nP_r = \frac{n!}{(n-r)!} \tag{1.14}$$

For the example above, where we ordered three letters, $n = 3$ and r = 3, so the result is

$${}_3P_3 = \frac{3!}{(3-3)!} = \frac{3!}{0!} = \frac{3 \cdot 2 \cdot 1}{1} = 6 \tag{1.15}$$

When is the formula applicable? This formula applies when the order of a result is important and the objects are distinct. If the order doesn't matter, then there is only one way to take three letters from a set of three letters, namely a set that contains A, B and C. We shall discuss shortly how to count the number of ways when order doesn't matter. If the objects are not distinct, the permutation formula again does not apply. For example, if our three letters are A, B & A, even if order matters, we have only 3 arrangements, AAB, ABA and BAA. The number of arrangements are reduced because two of the elements were indistinguishable. We shall discuss shortly how to count the number of ways for indistinct objects.

As another example, we can ask how many ways can a group schedule 3 different meetings on any of five possible dates? The answer is ${}_5P_3$ = 60. How did we know to use the equation for permuations? The key tip-off was the word "different". This means the meetings are distinguishable and order matters.

Quick Calculation of Permutations by hand

When *n* becomes large but *r* is small, it can be difficult to compute the permutation of ${}_nP_r$. Consider the case where *n*=200 and *r*=2. Then

$$ {}_nP_r = \frac{200!}{(200-2)!} = \frac{200!}{198!} \tag{1.16}$$

Our calculators cannot compute the factorial of 200 or 198. The numbers are too large. However, we can still obtain the number of permutations, if we consider that

$$ 200! = 200 \cdot 199 \cdot 198! \tag{1.17}$$

Then we have

$$ {}_nP_r = \frac{200!}{(200-2)!} = \frac{200 \cdot 199 \cdot 198!}{198!} = 200 \cdot 199 = 39800 \tag{1.18}$$

In general, we have

$$ \begin{aligned} {}_nP_r &= \frac{n!}{(n-r)!} = \frac{n \cdot (n-1) \cdot (n-2) \ldots (n-r+1)(n-r)!}{(n-r)!} \\ &= n \cdot (n-1) \cdot (n-2) \ldots (n-r+1) = \prod_{i=n-r+1}^{n} i \end{aligned} \tag{1.19}$$

Codes for both the naïve and better implementations of permutations are provided in the Subroutines section of this chapter.

Combinations

A combination is a grouping of elements without regard to order. The number of combinations of *n* distinct objects taken *r* at a time, where $r \leq n$, is

$$ \binom{n}{r} = \frac{n!}{r!(n-r)!} \tag{1.20}$$

The key for combinations is that order doesn't matter. In our example where we had three letters A, B. C and we chose three of them (*n*=3, *r*=3), we saw there were 6 permuations. From equation (1.20), we see that there is only one combination.

$$ \binom{3}{3} = \frac{3!}{3!(3-3)!} = \frac{3!}{3!0!} = 1 \tag{1.21}$$

In other words, there is only one way to take three objects from a pool of three objects if order doesn't matter, namely you take all of them. Again, this formula assumes that all of the objects are distinct.

Quick Calculation of Combinations by hand

When *n* becomes large but *r* is small, it can be difficult to compute the combination, $\binom{n}{r}$. The same cancellation trick that was used for permutations can again be used for combinations. Consider the case where *n*=200 and *r*=2. Then

$$\binom{n}{r} = \frac{n!}{r!(n-r)!} = \frac{200!}{2!(200-2)!} = \frac{200!}{2!\cdot 198!} = \frac{200\cdot 199\cdot 198!}{2!\cdot 198!} = \frac{200\cdot 199}{2!} = 19900 \qquad (1.22)$$

In general, we have

$$\binom{n}{r} = \frac{n!}{r!(n-r)!} = \frac{n!}{x_{big}!\,x_{lit}!} = \frac{n\cdot(n-1)\cdot(n-2)\ldots(x_{big}+1)x_{big}!}{x_{big}!\,x_{lit}!} = \frac{n\cdot(n-1)\cdot(n-2)\ldots(x_{big}+1)}{x_{lit}!} \qquad (1.23)$$

Here, we need to cancel either *r*! or (*n*-*r*)!, whichever is larger. To make this clear we introduced two new variables where x_{big} is the larger of *r* and *n*-*r* and x_{big} is the smaller. A code for this implementation of the combination is provided in the Subroutines section of this chapter.

In order to emphasize the difference between permutations and combinations, let us look at a few examples. We have already seen that considering the set of letters, A, B & C, for n = 3 and r = 3, that there are six permutations and one combination. If we only select two of the three letters, then we have ${}_3P_2$ =6 permutations. They are {AB, AC, BA, BC, CA, CB}. We have $\binom{3}{2} = 3$ combinations. They are {AB,BC, AC}.

Permutations of indistinct objects

In some cases, you have both distinct and indistinct objects. In this case, you must combine the permutation and combination formulae. The number of distinct permutations of *n* things where there are n_1 of one kind, n_2 of the second kind, up to n_k of the k^{th} kind is:

$$permutations\_of\_indistinct\_objects = \frac{n!}{n_1!n_2!...n_k!} \qquad (1.24)$$

For example, how many ways can you arrange all elements of the following set {A,AA,B,B,C}?

$$\frac{6!}{3!2!1!} = \frac{720}{1} * \frac{1}{12} = 60 \tag{1.25}$$

Consider an example where we have a three letter passcode that can be composed of any of the 26 letters in the alphabet.

Q: How many combinations are there if each letter is used no more than once?

A: This problem requires the permutation formula with $n = 26$ and $r = 3$, because each object is distinct.

$$_{26}P_3 = \frac{26!}{(26-3)!} = \frac{26!}{23!} = \frac{26 \cdot 25 \cdot 24 \cdot 23!}{23!} = 26 \cdot 25 \cdot 24 = 15{,}600 \tag{1.26}$$

Q: How many combinations are there if each letter can be reused?

A: This problem requires the general multiplication rule because the choice of each letter is independent.

$$\#\,of\ ways = 26 \cdot 26 \cdot 26 = 17{,}576 \tag{1.27}$$

1.4. Probability

The probability of an event *A* is the sum of the weights of all sample points in *A*. By weights, we mean the number of times a particular outcome is represented. The weight of *A* in the set {A,A,B} is 2. The weight of *B* is one.

The probability of an event *A* must satisfy the following conditions. First, the probability is always bounded between zero and one.

$$0 \le P(A) \le 1 \tag{1.28}$$

There is no such thing as a negative probability. There is also no probability greater than 100%. Second, the probability of the null set is zero.

$$P(\varnothing) = 0 \tag{1.29}$$

This is another way of saying that the experiment will result in some outcome. Third, the probability of getting something in the sample space is one.

$$P(S) = 1 \tag{1.30}$$

Since the sample space contains all possible outcomes, the outcome has to lie within the sample space.

Consider the following example. On an ordinary six-sided die, any number is equally likely to turn up. The probability of getting any particular number is 1/6, which is between 0 and 1. The probability that you don't get any result (the null set) when you roll the die is zero. The problem that you get something in the set S={1,2,3,4,5,6} is one.

Now consider an unbalanced six-sided die that is weighted to preferentially yield 6. For example, instead of yielding each number between 1 and 6 1/6 of the time, the die yields 6 half the time, and the rest of the numbers 1/10 of the time. All of the probabilities are still bounded between 0 and 1. The sum of the weights 5*(0.1) + 0.5 = 1. Therefore, the probability of getting a 6 is P(6) = 0.5. The probability of any other number, such as 1, is P(1) = 0.1.

Union

The probability of the union of two events A and B is given by

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{1.31}$$

For example, the sample space consists of the letter number pairs, S={A1, A2, B1}. The probability of getting a pair with an A or a pair with a 1 is

$$\begin{aligned} &P(A \cup 1) = P(A) + P(1) - P(A \cap 1) \\ &P(A) = 2/3 \\ &P(1) = 2/3 \\ &P(A \cap 1) = 1/3 \\ &P(A \cup 1) = 2/3 + 2/3 - 1/3 = 1 \end{aligned} \tag{1.32}$$

If A and B are mutually exclusive, then their union is the sum of their probabilities (because their intersection is zero).

$$P(A \cup B) = P(A) + P(B) \tag{1.33}$$

If many events, A_1 A_2….A_n are mutually exclusive, then their union is the sum of their probabilities.

$$P(A_1 \cup A_2 \cup ... \cup A_n) = \sum_{i=1}^{n} P(A_i) \tag{1.34}$$

If many events, A_1 A_2….A_n, are mutually exclusive and include all of the sample space,

$$P(A_1 \cup A_2 \cup ... \cup A_n) = \sum_{i=1}^{n} P(A_i) = P(S) = 1 \tag{1.35}$$

The union of three events is given by

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C) \tag{1.36}$$

For example, consider the set of ten objects S = {cat, dog, wolf, tiger, oak, elm, maple, opal, ruby, pearl}.

Q: What is the probability that you would randomly select a word that is (A) an animal OR that (B) has 4 letters OR that (C) starts with a vowel.)

A: There are two methods of solution. First, you can pick these out by hand. There are 4 animals, 2 trees that start with vowels, and 2 minerals that have 4 letters. Therefore the probability is 0.8 or 80%. The other method of solution is to use the equation given above to find the probability of the union.

$$\begin{aligned} &P(A) = 4/10; \quad P(B) = 3/10; \; P(C) = 3/10; \\ &P(A \cap B) = 1/10; \quad P(A \cap C) = 0; \; P(B \cap C) = 1/10; \; P(A \cap B \cap C) = 0; \\ &P(A \cup B \cup C) = 4/10 + 3/10 + 3/10 - 1/10 - 0 - 1/10 + 0 = 8/10 \end{aligned} \tag{1.37}$$

Conditional Probability

The probability of an event B occurring when it is know that some event A has already occurred is called a conditional probability, $P(B|A)$, and is read, "the probability of B given A", and is defined by:

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} \qquad \text{for P(A)} > 0 \tag{1.38}$$

Using the above example of ten words, what is the probability that we choose a (B) three letter word given that we know that we have chosen an (A) animal.

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{2/10}{4/10} = 1/2 \tag{1.39}$$

The conditional probability of B given A is different than the conditional probability of A given B,

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \qquad \text{for P(B)} > 0 \tag{1.40}$$

Equations (1.38) and (1.40) can be rearranged in terms of the intersection

$$P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A) \qquad (1.41)$$

If it possible to compute the probability of the intersection of A and B, given information about the conditional probability and the probability of A or B. This equation also provides a relationship between the conditional probability of A given B and the conditional probability of B given A.

The rule for the intersection can be extended to more than two events. If in an experiment, $A_1, A_2, A_3 ... A_k$ can occur then:

$$P(A_1 \cap A_2 \cap ... \cap A_k) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2)...P(A_k \mid A_1 \cap A_2 \cap ... \cap A_{k-1}) \qquad (1.42)$$

This formula stems from the repeated application of the conditional probability rule.

Independence/Multiplicative Rules

If two events are independent, then the outcome of the first experiment does not impact the outcome of the second experiment. Thus the probability of B happening should not be affected by the probability that A happened. In other words, two events are independent if and only if

$$P(B \mid A) = P(B) \quad \text{and} \quad P(A \mid B) = P(A) \quad \text{if } A \text{ and } B \text{ are independent} \qquad (1.43)$$

Substituting eqns (1.43) into equation (1.41) we have that two events are independent if and only if

$$P(A \cap B) = P(A) * P(B) \quad \text{if } A \text{ and } B \text{ are independent} \qquad (1.44)$$

This expression can be used as a way to test for independence, if the probability of the intersection is known. For multiple events, we have that the events are independent if and only if

$$P(A_1 \cap A_2 \cap ... \cap A_k) = \prod_{i=1}^{k} P(A_i) \quad \text{if all } A_i \text{ are independent} \qquad (1.45)$$

Example Problems for Probability

There are three examples given below. In each example, let it be clear that we use three and only three rules! We use the rules for the union, the conditional probability, and the intersection. To restate these rules, we have for two elements:

Union: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ (1.31)

Conditional: $P(B \mid A) = \dfrac{P(A \cap B)}{P(A)}$ for $P(A) > 0$ (1.38)

Intersection: $P(A \cap B) = P(A)P(B \mid A) = P(B)P(A \mid B)$ (1.41)

Example 1: You flip a coin twice. What is the probability of getting heads on the second flip (B) given that you got heads on the first flip (A)? Are the events independent?

Solution:

The probability of getting a head on the first flip is 0.5 The probability of getting a head on the second flip is 0.5. The intersection of A and B from the set {HH, HT, TH, TT} is 0.25.

The conditional probability is thus

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{0.25}{0.5} = 0.5 \quad (1.46)$$

We see that $P(B \mid A) = P(B) = 0.5$ so each flip of the coin is independent.

Example 2: You have a bag with 3 lima beans and 2 pinto beans in it. You draw 2 beans from it randomly without replacement. What is the probability that you draw a lima bean (B) given that you already drew a lima bean on the first draw (A)? Are the events independent?

Solution:

The easiest way to solve this problem is to list all the possible outcomes.

The probability for drawing a lima bean the first time (event A) is 3/5 = 0.6

The probability for drawing a pinto bean the first time is 2/5 = 0.4

If we draw a lima bean the first time, then there are 2 lima beans and 2 pinto beans. In that case the probability for drawing a lima bean the second time is 2/4 = 0.5 and the probability for drawing a pinto bean the second time is 2/4 = 0.5.

If we draw a pinto bean the first time, then there are 3 lima beans and 1 pinto beans. In that case the probability for drawing a lima bean the second time is 3/4 = 0.75 and the probability for drawing a pinto bean the second time is 1/4 = 0.25.

So we have four possible outcomes {LL, LP, PL, PP}.

The probability of each outcome is given by the product of the probabilities of the events in that outcome. (This is the multiplicative rule.)

The probability of LL is 0.6*0.5 = 0.3.

The probability of LP is 0.6*0.5 = 0.3.

The probability of PL is 0.4*0.75 = 0.3.

The probability of PP is 0.4*0.25 = 0.1.

The individual probabilities then for {LL, LP, PL, PP} are {0.3, 0.3, 0.3, 0.1}. With these figures, we can write:

Now event A includes any outcome with L in the first draw, LL and LP.

$$P(A) = P(LL) + P(LP) = 0.3 + 0.3 = 0.6 \quad (1.47)$$

Event B includes any outcome with L in the first draw, LL and PL. The sum of those two probabilities is 0.3+0.3 = 0.6 so

$$P(B) = P(LL) + P(PL) = 0.3 + 0.3 = 0.6 \tag{1.48}$$

The intersection of A and B includes the outcome LL

$$P(A \cap B) = P(LL) = 0.3 \tag{1.49}$$

Given this information, we have that the conditional probability of *B* given *A*, (or the probability that we draw a lima bean on the second draw, given that we drew a lima bean on the first draw) is

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{0.3}{0.6} = 0.5 \tag{1.50}$$

Now, to check for independence, we can evaluate equation (1.43)

$$P(B \mid A) = 0.5 \neq P(B) = 0.6 \tag{1.51}$$

Therefore, the two experiments are not independent. We can equivalently check for independence by examining the intersection of *A* and *B* from equation (1.44).

$$P(A \cap B) = 0.3 \neq P(A) * P(B) = 0.6 * 0.6 = 0.36 \tag{1.52}$$

This also says events *A* and *B* are not independent. In general, there is no need to make both checks. They will always return the same result. Use the one for which the information is more readily available.

Example 3.

In sampling a population for the presence of a disease, the population is of two types: Infected (I) and Uninfected (U). The results of the test are of two types: Positive (P) and Negative (N). In rare disease detection, a high probability for detecting a disease can still lead to more false positives than true positives. Consider a case where a disease infects 1 out of every 100,000 individuals. The probability for a positive test result given that the subject is infected is 0.99. (The test can accurately identify an infected individual 99% of the time.) The probability for a negative test result given that the subject is uninfected is 0.999. (The test can accurately identify that an uninfected individual 99.9% of the time.)

We shall answer the following questions.

(1) For testing a single person, define the complete sample space.

(2) What is the probability of a false negative test result (a negative test result given that the subject is infected)?

(3) What is the probability of being uninfected AND having a negative test result?

(4) What is the probability of testing positive?

(5) Determine rigorously whether testing positive and having the disease are independent.

(6) Determine the percentage of people who test positive who are really uninfected.

(7) In a population of 250 million, with the infection rate given, how many people would you expect to be (a) Infected-test Positive, (b) Infected-test Negative, (c) Uninfected-test Positive, (d) Uninfected-test negative.

Solution:

We are given the following information.

$$P(I) = 10^{-5} \tag{1.53}$$

$$P(N|U) = 0.999 \tag{1.54}$$

$$P(P|I) = 0.99 \tag{1.55}$$

(1) For testing a single person, define the complete sample space.

The sample space is $S = \{IP, IN, UP, UN\}$ where I = Infected, U=Uninfected, P=positive test result, N=negative test result. The Venn Diagram looks like this:

Infected $\cap$ Positive	Infected $\cap$ Negative
Uninfected $\cap$ Positive	UnInfected $\cap$ Negative

When you have a simple sample space like this, you can see some additional constraints on the system, in addition to the union, conditional, and intersection rules. You will need some of these additional constraints to solve the problems below.

For example, if a person tests positive, they are either infected or uninfected. Therefore, using the union rule we have:

$$P(P) = P[(P \cap I) \cup (P \cap U)] = P(P \cap I) + P(P \cap U) - P[(P \cap I) \cap (P \cap U)] \tag{1.56}$$

There is no intersection between being infected and uninfected, therefore:

$$P(P) = P(P \cap I) + P(P \cap U) \tag{1.57}$$

We can write three other analogous constraints based on the union rule,

$$P(N) = P(N \cap I) + P(N \cap U) \tag{1.58}$$

$$P(U) = P(U \cap P) + P(U \cap N) \tag{1.59}$$

$$P(I) = P(I \cap P) + P(I \cap N) \tag{1.60}$$

Also consider that the probability of being infected given a person is positive plus the probability of being uninfected given a person is positive is 1. A person is either infected or uninfected, regardless of whether they tested positive or negative. We can write this as.

$$P(I \mid P) + P(U \mid P) = \frac{P(I \cap P)}{P(P)} + \frac{P(U \cap P)}{P(P)} = \frac{P(P)}{P(P)} = 1 \tag{1.61}$$

Here we used the conditional probability rule and the union rule. This leads to the following constraint and three analogous constraints,

$$P(I \mid P) + P(U \mid P) = 1 \tag{1.61}$$
$$P(I \mid N) + P(U \mid N) = 1 \tag{1.62}$$
$$P(P \mid U) + P(N \mid U) = 1 \tag{1.63}$$
$$P(P \mid I) + P(N \mid I) = 1 \tag{1.64}$$

In solving the problems, below, remember we have this group of rules. There are many ways to solve some of the problems. We just go looking for the one that seems easiest.

(2) What is the probability of a false negative test result (a negative test result given that the subject is infected)?

We want: $P(N|I) = \frac{P(N \cap I)}{P(I)}$ so we need the two factors on the right hand side. We have been given the denominator. In order to find the numerator, we must use the other given:

$$P(P|I) = \frac{P(P \cap I)}{P(I)} = 0.99 \tag{1.65}$$

which rearranges for the intersection of P and I

$$P(P \cap I) = P(I) \cdot P(P|I) = \left(10^{-5}\right)(0.99) = 0.99 \cdot 10^{-5} \tag{1.66}$$

We must realize that the probability of I is the union of IP and IN groups, equation (1.60). So using the definition of the Union, we have:

$$P(I) = P(I \cap P) + P(I \cap N) \tag{1.60}$$

Rearranging yields

$$P(I \cap N) = P(I) - P(I \cap P) = \left(10^{-5}\right) - 0.99 \cdot 10^{-5} = 10^{-7} \tag{1.67}$$

Then we can plug into our original equation:

$$P\left(N \middle| I\right) = \frac{P(N \cap I)}{P(I)} = \frac{10^{-7}}{10^{-5}} = 0.01 \tag{1.68}$$

OR, an alternative solution, relies on us recognizing

$$P\left(P \mid I\right) + P\left(N \mid I\right) = 1 \tag{1.64}$$

and rearranging

$$P\left(N \middle| I\right) = 1 - P\left(P \middle| I\right) = 1 - 0.99 = 0.01 \tag{1.69}$$

(3) What is the probability of being uninfected AND having a negative test result?

We want $P(N \cap U)$. We can obtain this from either:

(a) the UNION RULE:

$$P(N) = P\left[(N \cap I) \cup (N \cap U)\right]$$
$$P(N) = P(N \cap I) + P(N \cap U) - P\left[(N \cap I) \cap (N \cap U)\right]$$
$$P\left[(N \cap I) \cap (N \cap U)\right] = 0$$
$$P(N) = P(N \cap I) + P(N \cap U)$$
$$P(N \cap U) = P(N) - P(N \cap I)$$

but we don't know $P(N \cap I)$ and we don't know $P(N)$, so this doesn't seem immediately helpful.

or (b) the conditional probability rule:

$$P(U \middle| N) = \frac{P(N \cap U)}{P(N)}$$
$$P(N \cap U) = P(N) \cdot P(U \middle| N)$$

but we don't know $P(U \middle| N)$ and we don't know $P(N)$, so again this doesn't seem immediately helpful.

or (c) the conditional probability rule:

$$P(N|U) = \frac{P(N \cap U)}{P(U)} = 0.999$$
$$P(N \cap U) = P(U) \cdot P(N|U) = P(U) \cdot 0.999$$

I like choice (c) because we are given $P(N|U) = 0.999$ and we know

$P(U) = 1 - P(I) = 1 - 10^{-5} = 0.99999$ so

$$P(N \cap U) = P(U) \cdot P(N|U) = (0.99999)(0.999) = 0.99899001 \qquad (1.70)$$

(4) What is the probability of testing positive?

We want $P(P)$. We can find $P(P)$ by either:

(a) the fact that the sum of the probabilities must be one

$P(P) + P(N) = 1$

$P(P) = 1 - P(N)$

but we don't know $P(N)$.

or (b) the conditional probability distribution:

$P(I|P) = \frac{P(P \cap I)}{P(P)}$ but we don't know $P(I|P)$.

pr (c) the conditional probability distribution:

$P(U|P) = \frac{P(P \cap U)}{P(P)}$ but we don't know $P(U|P)$.

or (d) the sum of the probabilities must be one and a different conditional probability:

$P(P) = 1 - P(N)$

$$P(U|N) = \frac{P(N \cap U)}{P(N)}$$

$P(P) = 1 - P(N) = 1 - \dfrac{P(N \cap U)}{P(U|N)}$ but we don't know $P(U|N)$.

or (e) the sum of the probabilities must be one and a different conditional probability:

$$P(P) = 1 - P(N)$$
$$P(I|N) = \frac{P(N \cap I)}{P(N)}$$
$P(P) = 1 - P(N) = 1 - \dfrac{P(N \cap I)}{P(I|N)}$ but we don't know $P(I|N)$.

or (f) the union rule:

$$P(P) = P\left[(P \cap I) \cup (P \cap U)\right]$$
$$P(P) = P(P \cap I) + P(P \cap U) - P\left[(P \cap I) \cap (P \cap U)\right]$$
$$P\left[(P \cap I) \cap (P \cap U)\right] = 0$$
$$P(P) = P(P \cap I) + P(P \cap U)$$

Combine this expression with conditional probabilities that we do know:

$$P(P) = P(I) * P(P \mid I) + P(U) * P(P \mid U)$$

I like choice (f).

$$P(P) = 10^{-5} \cdot 0.99 + 0.99999 \cdot P(P \mid U)$$

We can get the last factor by considering (as we did in part (2))

$$P(P \mid U) + P(N \mid U) = 1 \qquad (1.63)$$

$$P(P|U) = 1 - P(N|U) = 1 - 0.999 = 0.001$$

so

$$P(P) = 10^{-5} \cdot 0.99 + 0.99999 \cdot 0.001 = 0.00100989 \qquad (1.71)$$

(5) Determine rigorously whether testing positive and having the disease are independent.

If $P(P)$ and $P(I)$ are independent, then

$P(P \cap I) = P(P) \cdot P(I)$.

$0.99 \cdot 10^{-5} = 0.00100989 \cdot 10^{-5}$

Testing positive and being infected are not independent. Thank goodness. The entire point of the test is to identify infected individuals.

(6) Determine the percentage of people who test positive but who are really uninfected.

We want $\frac{P(P \cap U)}{P(P)}$.

$$\frac{P(P \cap U)}{P(P)} = \frac{0.99999 \cdot 10^{-3}}{0.00100989} = 0.990196952 = 99\%$$

Despite the high accuracy of the test 99% of those people who test positive are actually uninfected.

(7) In a population of 250 million, with the infection rate given, how many people would you expect to be (a) Infected-test Positive, (b) Infected-test Negative, (c) Uninfected-test Positive, (d) Uninfected-test negative.

From part (5) we know:

$P(P \cap I) = 0.99 \cdot 10^{-5}$

$P(P \cap U) = 0.99999 \cdot 0.001 = 0.99999 \cdot 10^{-3}$

From part (2) we know

$P(N \cap I) = P(I) - P(P \cap I) = \left(10^{-5}\right) - 0.99 \cdot 10^{-5} = 10^{-7}$

From part (3) we know

$P(N \cap U) = P(U) \cdot P(N|U) = (0.99999)(0.999) = 0.99899001$

These four probabilities should sum to 1.0 and they do.
Out of 250 million people, the number who are infected and test positive are: 2475.
Out of 250 million people, the number who are infected and test negative are: 25.
Out of 250 million people, the number who are uninfected and test positive are: 249,997.5
Out of 250 million people, the number who are uninfected and test negative are: 249,747,500.

1.5 Subroutines

Code 1.1. Permutations via a Naïve Implementation

This simple code, perm_naive.m, illustrates how one would numerically compute a permutation. It doesn't use the cancellation trick shown above. It computes the factorial of *n*, then computes the factorial of (*n*-*x*), then returns the quotient. This code won't work for large numbers

```
function f = perm_naive(n,x)
fac1 = 1.0;
if (n > 1)
   for i = n:-1:2
         fac1 = fac1*i;
   end
end
fac2 = 1.0;
if (n-x > 1)
   for i = (n-x):-1:2
         fac2 = fac2*i;
   end
end
f = fac1/fac2;
```

Code 1.2. Permutations with cancellations

This code, perm.m, computes permutations, using the cancellation trick.

```
function f = perm(n,x)
f = 1.0;
if (n > 1 & n >=x)
   for i = n:-1:(n-x+1)
      f = f*i;
   end
end
```

Code 1.3. Combinations (with cancellations)

This code, comb.m, illustrates how one would numerically compute a combination. It invokes the cancellation trick, so it must first determine the larger factorial in the denominator.

```
function f = comb(n,x)
a = n-x;
if (a > x)
   xbig = a;
   xlit = x;
else
   xbig = x;
   xlit = a;
end
if (n > 1 & n >=xbig)
   fnum = 1.0;
```

```
      for i = n:-1:(xbig+1)
         fnum = fnum*i;
      end
      fden = 1.0;
      for i = xlit:-1:2
         fden = fden*i;
      end
   end
   f = fnum/fden;
```

1.6. Problems

Homework problems are posted on the course website.