BST 675 – Fall 2010 – Dr. Charnigo

Unit I: Probability

a. Motivating Case Study #1: For which diseases are diagnostic tests useful?

Consider the following six scenarios.

- About 10% of people have disease 1. (Sometimes this is called "prevalence".) Among people who have disease 1, about 80% will test positive for it. (Sometimes this is called "sensitivity".) Among people who do not have disease 1, about 80% will test negative for it. (Sometimes this is called "specificity".)

- About 10% of people have disease 2. Among people who have disease 2, about 90% will test positive for it. Among people who do not have disease 2, about 70% will test negative for it.

- About 10% of people have disease 3. Among people who have disease 3, about 70% will test positive for it. Among people who do not have disease 3, about 90% will test negative for it.

- About 1% of people have disease 4. Among people who have disease 4, about 80% will test positive for it. Among people who do not have disease 4, about 80% will test negative for it.

- About 1% of people have disease 5. Among people who have disease 5, about 90% will test positive for it. Among people who do not have disease 5, about 70% will test negative for it.

- About 1% of people have disease 6. Among people who have disease 6, about 70% will test positive for it. Among people who do not have disease 6, about 90% will test negative for it.

Assuming that the diseases are equally serious, for which of the above scenarios would a positive diagnostic test most alarm you? For which of the above scenarios would a negative diagnostic test most relieve your anxiety?

If your intuition does not provide the answers, you can obtain them from Bayes' Theorem. However, since we have yet to introduce Bayes' Theorem in this course, let us instead perform a thought experiment. Suppose that we gather 1000 people, randomly chosen.

- About 100 of them will have disease 1. Among the 100 who have disease 1, about 80 will test positive and 20 will test negative. Among the 900 who do not have disease 1, about 180 will test positive and 720 will test negative. Therefore, about $80/(80+180) = 31\%$ of the people who test positive will really have disease 1. (Sometimes this is called "positive predictive value".) On the other hand, about $720/(20 + 720) = 97\%$ of the people who test negative will really not have disease 1. (Sometimes this is called "negative predictive value".)

- About 100 of them will have disease 2. Among the 100 who have disease 2, about 90 will test positive and 10 will test negative. Among the 900 who do not have disease 2, about 270 will test positive and 630 will test negative. Therefore, about $90/(90 + 270) = 25\%$ of the people who test positive will really have disease 2, while about $630/(10+630) = 98\%$ of the people who test negative will really not have disease 2.

- About 100 of them will have disease 3. Among the 100 who have disease 3, about 70 will test positive and 30 will test negative. Among the 900 who do not have disease 3, about 90 will test positive and 810 will test negative. Therefore, about $70/(70+90) = 44\%$ of the people who test positive will really have disease 3, while about $810/(30+810) = 96\%$ of the people who test negative will really not have disease 3.

A few remarks can be made at this juncture.

First, in each of the three scenarios discussed thus far, a positive test signals true disease status less than half of the time. This is because the people who do not have the disease far outnumber the people who do have the disease. So, even though people who do have the disease are more likely to test positive for it, in absolute numbers there are more positive tests from people who do not have the disease.

Second, a positive test is more likely to signal true disease status in the third scenario than in the first two scenarios. This may be surprising because only 70% of people with disease 3 test positive, while 80% of people with disease 1 and 90% of people with disease 2 test positive. As noted in the first remark, though, the key is how many people are testing positive who do not have the disease. The third scenario has fewer such people than either of the first two scenarios.

Third, a negative test is extremely likely to signal true freedom from disease in all three scenarios. However, a negative test is marginally more likely to signal true freedom from disease in the second scenario. This may be surprising because only 70% of people who do not have disease 2 test negative. Can you explain this?

We could also work out the details for the last three scenarios, but by now your intuition is sufficiently informed. Compared to the first three scenarios, will a positive test in the last three scenarios alarm you more or less? Will a negative test in the last three scenarios relieve your anxiety more or less?

Are there any circumstances under which a negative test would not relieve your anxiety?

b. Motivating Case Study #2: Infant mortality and Simpson's paradox

Among many facts noted by perinatal epidemiologists are that: (i) infant mortality is higher among infants with low birthweights; (ii) infant mortality is higher among infants whose mothers smoke; and, (iii) low birthweights are more common among infants whose mothers smoke.

Yet, surprisingly, infant mortality is less common among low birthweight infants whose mothers smoke than among low birthweight infants whose mothers do not smoke.

This is an example of Simpson's paradox, which in its general form says that the relationship between two variables within a stratum determined by a third variable may be qualitatively different from the relationship between the same two variables globally. (Epidemiologists often refer to this phenomenon as "confounding". Note that this is not the same as "interaction", which says that the relationship between two variables within one stratum determined by a third variable is qualitatively different from the relationship between the same two variables within a different stratum determined by the third variable.) In our example, the relationship between smoking and infant mortality is negative among infants of low birthweight, even though the relationship between smoking and infant mortality is positive when we do not stratify by birthweight.

To provide a numerical demonstration, suppose that there are 10000 infants, of which 2000 are born to smoking mothers and 8000 are born to non-smoking mothers. Among the 2000 infants born to smoking mothers, suppose that 400 are of low birthweight, 20 experience mortality, and 10 of the 20 experiencing mortality are of low birthweight. Among the 8000 infants born to non-smoking mothers, suppose that 800 are of low birthweight, 40 experience mortality, and 30 of the 40 experiencing mortality are of low birthweight.

Then fact (i) is verified because the mortality rate among low birthweight infants (40/1200) is much higher than the mortality rate among normal birthweight infants (20/8800), fact (ii) is verified because the mortality rate is higher among infants whose mothers smoke (20/2000) than among infants whose mothers do not smoke (40/8000), and fact (iii) is verified because low birthweights are more common among infants whose mothers smoke (400/2000) than among infants whose mothers do not smoke (800/8000). Even so, the mortality rate among low birthweight infants whose mothers smoke (10/400) is lower than the mortality rate among low birthweight infants whose mothers do not smoke (30/800).

c. Definition and axioms of probability (Cf. pp. 21-42 of Larsen and Marx)

Preface. Before defining probability and stating its axioms, we need to introduce some concepts from set theory.

Sample space. Suppose that we conduct an experiment. The set of all possible outcomes from that experiment is referred to as the sample space. When convenient, we use S as a symbol for the sample space.

Example (sample space). Suppose that I flip a coin four times. Critique the following statement: "The sample space is $\{0, 1, 2, 3, 4\}$, where the number refers to how many times the coin lands on heads."

Event. Suppose that every element of a set A is an element of the sample space S, which we write as $A \subset S$. Then we refer to A as an event.

Union, intersection, and complementation operators. Suppose that $A, B \subset S$. We define the union $A \cup B$ as the set containing all elements of S present in A or B or both. We define the intersection $A \cap B$ as the set containing all elements of S present in both A and B. We define the complementation A^c as the set containing all elements of S not present in A.

Example (union, intersection, and complementation operators). Suppose that I roll a six-sided die. The sample space is $\{1,2,3,4,5,6\}$, where the number refers to the result of the roll. Let A be the event that I roll an odd number and B be the event that I roll a multiple of 3. What are $A \cup B$, $A \cap B$, A^c, $A \cup A^c$, $A \cap A^c$, and $(A \cup A^c)^c$?

Properties of operators. Suppose that $A, B, C \subset S$. We have the following.

1. Commutativity: $A \cup B = B \cup A$, $A \cap B = B \cap A$.
2. Associativity: $A \cup (B \cup C) = (A \cup B) \cup C$, $A \cap (B \cap C) = (A \cap B) \cap C$.
3. Distributivity: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$, $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.
4. DeMorgan laws: $(A \cup B)^c = A^c \cap B^c$, $(A \cap B)^c = A^c \cup B^c$.

Example (properties of operators). Critique the following attempt at proof of the first DeMorgan law: "Suppose that $B = A^c$. On the one hand, $A \cup B = S$ and $(A \cup B)^c = S^c = \emptyset$. On the other hand, $B^c = A$ and $A^c \cap B^c = A^c \cap A = \emptyset$. This completes the proof."

Proving that two sets are the same. There is a standard way to prove that two sets are the same: show that each is a subset of the other. For such proofs the following notation is helpful: $\in$ means "is an element of", while $\notin$ means "is not an element of".

Example (proving that two sets are the same). Let us prove the first DeMorgan law. Suppose that $x \in (A \cup B)^c$. Then $x \notin A \cup B$. Hence $x \notin A$ and $x \notin B$. So $x \in A^c$ and $x \in B^c$. Hence $x \in A^c \cap B^c$. This shows that $(A \cup B)^c \subset A^c \cap B^c$. Starting from $x \in A^c \cap B^c$ and working our way back to $x \in (A \cup B)^c$ shows that $A^c \cap B^c \subset (A \cup B)^c$.

General unions and intersections. Suppose that $A_\gamma \subset S$ for all $\gamma \in \Gamma$, where Γ is some non-empty index set. Then we define $\cup_{\gamma \in \Gamma} A_\gamma$ to contain all elements of S present in at least one of the A_γ and $\cap_{\gamma \in \Gamma} A_\gamma$ to contain all elements of S present in all of the A_γ. Common choices for Γ include $\{1,2\}$, which reproduces our earlier definitions, and $\{1,2,3,\ldots\}$, which gives rise to so-called countable unions and intersections.

Example (general unions and intersections). Suppose that $A_i = [0, i]$ for $i \in \{1, 2, 3, \ldots\}$. What are $\cup_{i=1}^{\infty} A_i$ and $\cap_{i=1}^{\infty} A_i$?

Mutually exclusive and collectively exhaustive. Suppose that $A_\gamma \subset S$ for all $\gamma \in \Gamma$, where Γ is some non-empty index set. If $A_i \cap A_j = \emptyset$ for all unequal $i, j \in \Gamma$, then we say that the A_γ are mutually exclusive. If $\cup_{\gamma\in\Gamma} A_\gamma = S$, then we say that the A_γ are collectively exhaustive. If the A_γ are both mutually exclusive and collectively exhaustive, then we say that they constitute a partition of S.

Probability definition and axioms. Consider the simple experiment of flipping a coin. The sample space S is the finite set $\{Heads, Tails\}$. If we were to repeat this experiment an indefinitely large number of times, then we might find that about half of the flips resulted in Heads while about half of the flips resulted in Tails. This might motivate us to define the probability of getting a Heads on a single flip as 1/2 and likewise to define the probability of getting a Tails on a single flip as 1/2. In other words, we might define the probability of an event in an experiment based on the event's relative frequency over an indefinitely large number of repetitions of that experiment.

The relative frequency perspective helps us to describe to the lay person what probability is. However, most statisticians prefer to define probability axiomatically. Therefore, we proceed axiomatically for now. A connection to the relative frequency perspective will be apparent in BST 676, when you encounter the Laws of Large Numbers.

Let $P(A)$ denote the probability of an event $A \subset S$. We want the following conditions or "axioms" to be satisfied for any events $A, A_1, A_2, \ldots \subset S$.

1. $P(A) \geq 0$.
2. $P(S) = 1$.
3. $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ whenever $A_i \cap A_j = \emptyset$ for $i \neq j$.

Unfortunately, there are some complications when S is uncountably infinite. These complications are circumvented by requiring that probabilities be defined and axioms hold not for all events but only for events belonging to a special collection of subsets of S called a "sigma algebra" or a "sigma field". Pursuing the mathematical intricacies will take us too far afield, so let it be tacit in BST 675 that a sigma field is somehow appropriately defined and that probabilities are only provided or requested for events in that sigma field.

Also, we will show a little later that $P(\emptyset) = 0$. Taking this for granted right now, we can conclude from the third axiom that, whenever $A_i \cap A_j = \emptyset$ for $i \neq j$ with $1 \leq i, j \leq M$, we have $P(\cup_{i=1}^M A_i) = \sum_{i=1}^M P(A_i)$ for any positive integer M. (To see this, put $A_i := \emptyset$ for $i > M$.)

Example (Probability definition and axioms). Suppose that $S = \{1, 2, 3, 4, 5, 6\}$, as if I were rolling a six-sided die. How can we define probabilities for events so that the axioms are satisfied?

Calculus of probabilities. Several useful results follow from the axioms. We assume that $A, B, C_1, C_2, \ldots \subset S$.

4. $P(\emptyset) = 0$.
5. $P(A) \leq 1$.
6. $P(A^c) = 1 - P(A)$.
7. $P(B \cap A^c) = P(B) - P(B \cap A)$.
8. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
9. If $A \subset B$, then $P(A) \leq P(B)$.
10. $P(\cup_{i=1}^\infty C_i) \leq \sum_{i=1}^\infty P(C_i)$.
11. If $C_1, C_2, \ldots$ is a partition, then $P(A) = \sum_{i=1}^\infty P(A \cap C_i)$.

Example (Calculus of probabilities). Let us prove result 4. Put $A_i := \emptyset$ for $i \in \{1, 2, \ldots\}$. Then $A_i \cap A_j = \emptyset$ for $i \neq j$ and $\cup_{i=1}^\infty A_i = \emptyset$. By axiom 3 we have $P(\emptyset) = \sum_{i=1}^\infty P(\emptyset)$, which is only possible if $P(\emptyset) = 0$.

Critique the following attempt at proof of result 10: "The left hand side must be less than or equal to 1 by result 5, while the right hand side can be infinite. So the left hand side is less than or equal to the right hand side."

Let us prove result 10. Put $D_1 := C_1$, $D_2 := C_2 \cap C_1^c$, $D_3 := C_3 \cap C_1^c \cap C_2^c$, $D_4 := C_4 \cap C_1^c \cap C_2^c \cap C_3^c$, and so forth.

Then $\cup_{i=1}^\infty D_i = \cup_{i=1}^\infty C_i$. (Suppose that $x \in \cup_{i=1}^\infty C_i$. Let j be the smallest positive integer for which $x \in C_j$. Then $x \in D_j$, so that $x \in \cup_{i=1}^\infty D_i$. This shows that $\cup_{i=1}^\infty C_i \subset \cup_{i=1}^\infty D_i$. How do we know that $\cup_{i=1}^\infty D_i \subset \cup_{i=1}^\infty C_i$?)

Moreover, $D_i \cap D_j = \emptyset$ for $i \neq j$. (Suppose that $x \in D_i \cap D_j$. Then $x \in C_i$ and $x \in C_j$. If $i > j$, then $x \in C_j$ implies $x \notin C_j^c$ and $x \notin C_i \cap C_1^c \cap \cdots \cap C_{i-1}^c = D_i$. This is a contradiction, so there is no such x. Likewise, we obtain a contradiction if $i < j$.)

By axiom 3 we have $P(\cup_{i=1}^\infty D_i) = \sum_{i=1}^\infty P(D_i)$. By result 9 we have $P(D_i) \leq P(C_i)$ for $i \in \{1, 2, \ldots\}$. Since $\cup_{i=1}^\infty D_i = \cup_{i=1}^\infty C_i$, we conclude that $P(\cup_{i=1}^\infty C_i) = P(\cup_{i=1}^\infty D_i) = \sum_{i=1}^\infty P(D_i) \leq \sum_{i=1}^\infty P(C_i)$.

d. Conditional probability and Bayes' Theorem (Cf. pp. 42-69 of Larsen and Marx)

Definition of conditional probability. Let A and B be events. If $P(B) > 0$, then we define the conditional probability of A given B as

$$P(A|B) := P(A \cap B)/P(B).$$

In this context, we sometimes refer to $P(A)$ as an unconditional probability. Intuitively, $P(A|B)$ is an updated version of $P(A)$ given the knowledge that event B has occurred.

Example (definition of conditional probability). Let A denote the event that a randomly selected person smokes, and let B denote the event that a randomly selected person develops lung cancer. Suppose that $P(B) = 0.02$, $P(A) = 0.25$, and $P(B|A) = 0.05$. Then $P(A \cap B) =$, so that $P(A|B) =$. Moreover, $P(A^c \cap B) =$, so that $P(A^c|B) =$. Is the last answer intuitively obvious?

Useful results for conditional probabilities. Assuming that the axioms for unconditional probabilities are satisfied and that $P(D) > 0$, we have the following useful results for conditional probabilities.

1. $P(A|D) \geq 0$.
2. $P(S|D) = 1$.
3. $P(\cup_{i=1}^{\infty} A_i|D) = \sum_{i=1}^{\infty} P(A_i|D)$ whenever $A_i \cap A_j = \emptyset$ for $i \neq j$.
4. $P(\emptyset|D) = 0$.
5. $P(A|D) \leq 1$.
6. $P(A^c|D) = 1 - P(A|D)$.
7. $P(B \cap A^c|D) = P(B|D) - P(B \cap A|D)$.
8. $P(A \cup B|D) = P(A|D) + P(B|D) - P(A \cap B|D)$.
9. If $A \subset B$, then $P(A|D) \leq P(B|D)$.
10. $P(\cup_{i=1}^{\infty} C_i|D) \leq \sum_{i=1}^{\infty} P(C_i|D)$.
11. If $C_1, C_2, \ldots$ is a partition, then $P(A|D) = \sum_{i=1}^{\infty} P(A \cap C_i|D)$.

Example (useful results for conditional probabilities). To verify result 3, we write

$$P(\cup_{i=1}^{\infty} A_i|D) = P(\cup_{i=1}^{\infty}(A_i \cap D))/P(D) = \qquad\qquad = \sum_{i=1}^{\infty} P(A_i|D).$$

The justification for the second equality is that, if $A_i \cap A_j = \emptyset$, then

$$(A_i \cap D) \cap (A_j \cap D) = (A_i \cap A_j) \cap D = \emptyset \cap D = \emptyset.$$

To verify result 9, we note that $A \cap D \subset B \cap D$. (How do we know this?) Then $P(A \cap D) \leq P(B \cap D)$, whence

$$P(A|D) = P(A \cap D)/P(D) \leq P(B \cap D)/P(D) = P(B|D).$$

Iterating conditional probabilities. Suppose that $P(B \cap C) > 0$. Then

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

Moreover, since the left side is also equal to $P(A \cap B|C)P(C)$, we see that

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

(How do we know that $P(C) > 0$?) The above formulas may be useful in situations when the conditional probabilities $P(A|B \cap C)$ and $P(B|C)$ are intuitively obvious while $P(A \cap B|C)$ and $P(A \cap B \cap C)$ are less readily apparent.

Example (iterating conditional probabilities). I am dealt three cards (without replacement) from a well-shuffled standard 52-card deck. Let C be the event that the suit of the first card is diamonds, B be the event that the suit of the second card is diamonds, and A be the event that the suit of the third card is diamonds. Then $P(C) = 13/52$, $P(B|C) =$, and $P(A|B \cap C) =$. So $P(A \cap B|C) =$ and $P(A \cap B \cap C) =$.

Bayes' Theorem. Let $A_1, A_2, \ldots$ be a partition of S such that $P(A_i) > 0$ for $i \in \{1, 2, \ldots\}$. Then for any event B and any $i \in \{1, 2, \ldots\}$ we have

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{\infty} P(B|A_j)P(A_j)}.$$

A similar result also holds for a finite partition $A_1, A_2, \ldots, A_k$,

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

In particular, with $k = 2$ we have (upon a minor change in notation)

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}.$$

This last version of Bayes' Theorem provides a recipe for computing $P(A|B)$ if we have available $P(B|A)$ (and two other probabilities — why two rather than three?).

Example (Bayes' Theorem). Bayes' Theorem is obvious from the definition of conditional probability if we can verify that the denominator is $P(B)$. (Equality of the denominator to $P(B)$ is sometimes called the Law of Total Probability.) Suggestions?

Again, let A denote the event that a randomly selected person smokes, and let B denote the event that a randomly selected person develops lung cancer. As before, suppose that $P(A) = 0.25$ and $P(B|A) = 0.05$. But now suppose that we are not given $P(B)$. Rather, suppose that we are told that 99% of non-smokers do not develop lung cancer. How can we find $P(A|B)$ with just the information provided here?

e. Independence (Cf. pp. 69-85 of Larsen and Marx)

Independence of two events. Suppose that A and B are events such that $0 < P(A) < 1$, $0 < P(B) < 1$, and $P(A|B) = P(A)$. Thus, knowing that B has occurred does not lead us to revise the probability that A will occur. In this case, multiplying both sides of $P(A|B) = P(A)$ by $P(B)$ yields

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

Moreover, since $P(A \cap B)$ can also be expressed as $P(B|A)P(A)$, we see that $P(B|A) = P(B)$. If any one of these three equivalent conditions holds —

$$P(A|B) = P(A) = P(A|B^c),$$
$$P(A \cap B) = P(A)P(B), \quad \text{or}$$
$$P(B|A) = P(B) = P(B|A^c)$$

— we say that A and B are independent. (How are we able to go from $P(A|B) = P(A)$ to $P(A|B) = P(A) = P(A|B^c)$ and from $P(B|A) = P(B)$ to $P(B|A) = P(B) = P(B|A^c)$?)

If no restrictions on $P(A)$ or $P(B)$ are made, so we are not sure whether $P(A|B)$ and $P(B|A)$ are defined, then we characterize independence by the equation $P(A \cap B) = P(A)P(B)$. Also, if A and B are independent, then so are A^c and B^c, A^c and B, and A and B^c.

Example (independence of two events). Let A and B be two events. If $P(A) = 0$, then A and B are independent. (How do we know this?) If $P(B) = 1$, then A and B are independent. (How do we know this?)

To verify that independence of A and B implies independence of A^c and B, we write

$$P(A^c \cap B) = P(B) - P(A \cap B) = P(B) - P(A)P(B) = P(B)P(A^c).$$

We can similarly prove that independence of A and B implies independence of A^c and B^c and of A and B^c.

Suppose once more that I am dealt three cards (without replacement) from a well-shuffled standard 52-card deck. Let C be the event that the suit of the first card is diamonds, B be the event that the suit of the second card is diamonds, and A be the event that the suit of the third card is diamonds. Are C and B independent?

Independence of three or more events. Let $A_1, A_2, \ldots, A_n$ be events such that for any subcollection $A_{i_1}, A_{i_2}, \ldots, A_{i_k}$ we have

$$P(\cap_{j=1}^k A_{i_j}) = \prod_{j=1}^k P(A_{i_j}).$$

Then we say that $A_1, A_2, \ldots, A_n$ are independent. Independence of n events thus entails not only

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)P(A_2)\cdots P(A_n)$$

but also

$$P(A_1 \cap A_2) = P(A_1)P(A_2),$$
$$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3),$$
$$P(A_4 \cap A_{18}) = P(A_4)P(A_{18}),$$

and many other equalities.

f. Counting (Cf. pp. 85-123 of Larsen and Marx)

Preface. We consider the problem of counting how many ways there are to select r objects from a collection of n objects, where r and n $(> r)$ are positive integers, along with the implication for calculating probabilities.

Objects replaceable, selection order relevant. First consider the situation in which objects can be replaced and the order in which objects are selected matters. An example of this is choosing a four-digit PIN number. Here we have $r = 4$ and $n = 10$. The same digit can appear more than once (3934 is a valid four-digit PIN number), and the order of the digits matters (3934 is not the same as 9334). In this example we see readily that there are $10000 = 10^4$ possibilities for a four-digit PIN number.

The general rule is that there are n^r possibilities when objects can be replaced and the order in which objects are selected matters.

Objects irreplaceable, selection order relevant. Now consider the situation in which objects cannot be replaced and the order in which objects are selected matters. An example of this is being asked to choose your favorite and second favorite colors from among red, orange, yellow, green, blue, and purple. Here we have $r = 2$ and $n = 6$. The same color cannot appear more than once (red cannot be both your favorite and your second favorite color), and the order of the colors matters (red as a favorite and yellow as a second favorite is not the same as yellow as a favorite and red as a second favorite). In this example we see readily that there are 30 possibilities.

The general rule is that there are $n \times (n-1) \times \cdots \times (n-r+1) = n!/(n-r)!$ possibilities when objects cannot be replaced and the order in which objects are selected matters. Here we use $n!$ (read "n factorial") to represent $n \times (n-1) \times (n-2) \times \cdots \times 1$ for any positive integer n. By convention $0! = 1$.

Objects irreplaceable, selection order irrelevant. Next consider the situation in which objects cannot be replaced and the order in which objects are selected does not matter. An example of this is a lottery ticket in which you win by matching the numbers on 6 balls drawn without replacement from a vat containing 44 balls. Here we have $r = 6$ and $n = 44$. The same number cannot appear twice (once ball 1 is removed from the vat, ball 1 cannot be drawn from the vat again), and the order of the numbers does not matter (if your lottery ticket shows $7, 11, 12, 18, 35, 42$, you still win if the balls are drawn in the order $42, 35, 18, 12, 11, 7$). In this example there are $(44 \times 43 \times 42 \times 41 \times 40 \times 39)/(6 \times 5 \times 4 \times 3 \times 2 \times 1) = 44!/\{6!\ 38!\}$ possibilities. The division by 6! avoids overcounting (without it we would have 6! copies of $7, 11, 12, 18, 35, 42$ since $7, 11, 12, 18, 35, 42$ is the same as $42, 35, 18, 12, 11, 7$, which is the same as $7, 12, 35, 11, 18, 42$, and so forth).

The general rule is that there are $n!/\{r!(n-r)!\} =: \binom{n}{r}$ (read "n choose r") possibilities when objects cannot be replaced and the order in which objects are selected does not matter.

Objects replaceable, selection order irrelevant. Finally consider the situation in which objects can be replaced and the order in which objects are selected does not matter. An example of this is if we list from lowest to highest your scores on the five written assignments this semester. Here we have $r = 5$ and $n = 101$ because the possible scores are $0, 1, 2, \ldots, 100$. You can get the same score more than once (all of you are hoping to get more than one 100), but the order in which you acquire the scores does not matter (100, 95, 90, 85, 80 and 90, 85, 80, 95, 100 can both be listed as 80, 85, 90, 95, 100). Although not obvious, in this example there are $105!/\{100!\ 5!\}$ possibilities.

The general rule is that there are $\binom{n+r-1}{r}$ possibilities when objects can be replaced and the order in which objects are selected does not matter.

Implication for probability calculations. Counting techniques are useful in problems where the sample space S is finite, if all elements of S are equally likely — and that is a big if!

Indeed, suppose that $S = \{s_1, s_2, \ldots, s_n\}$ for some positive integer n and that $P(\{s_i\}) = 1/n$ for $i \in \{1, 2, \ldots, n\}$. We apply probability axiom 3 to find that $P(A) = \sum_{s_i \in A} P(\{s_i\}) = \sum_{s_i \in A} 1/n = card(A)/card(S)$, where *card* denotes the cardinality of (i.e., number of elements in) a set.

Example (Implication for probability calculations). Consider being dealt a five-card poker hand from a standard deck of 52 playing cards (there are 13 denominations — ace, king, queen, etc. — in each of 4 suits — hearts, diamonds, spades, clubs). Assuming that the deck is well shuffled, what is the probability of being dealt a "full house" (a hand with a pair — two cards of matching denomination — and a triple — three cards of matching denomination)?

To answer this question, we note that the order in which the cards are dealt does not matter and that they are drawn from the deck without replacement. Letting S denote the set of all possible five-card poker hands, we find that $card(S) = \binom{52}{5} = 2598960$.

Letting A denote the set of all five-card poker hands containing a pair and a triple, we can find $card(A)$ by noting that there are   ways to specify the denomination for the pair,   ways to specify the denomination for the triple,   ways to specify suits for the pair, and   ways to specify suits for the triple. Multiplying these numbers together gives us $card(A) = 3744$.

Since there is no reason to believe that any one poker hand is more likely than another, we can apply the formula $card(A)/card(S)$ to conclude that the probability of being dealt a full house is $3744/2598960 \approx 0.144\%$.

g. Resolution of motivating case studies

Our first motivating case study provided numerical examples suggesting that the positive predictive value of a diagnostic test depends on both the specificity of the diagnostic test and the disease prevalence. Having encountered Bayes' Theorem, we are now in a position to formally describe the relationships among positive predictive value, specificity, and disease prevalence.

Let A denote the event that a randomly selected person has the disease and B the event that a randomly selected person tests positive for the disease. Symbolically, the disease prevalence is $P(A)$. (If this assertion seems entirely obvious, then please pause and think about it for a moment. The disease prevalence is the proportion of people in the population who have the disease, whereas $P(A)$ is the probability that a randomly selected person has the disease. Thus, in equating a proportion to a probability, I am implicitly appealing to the relative frequency interpretation of probability.)

Symbolically, what is the sensitivity? the specificity? the positive predictive value?

Let us use the abbreviations PR, SE, SP, and PPV for prevalence, sensitivity, specificity, and positive predictive value. Assuming that $0 < SP, SE, PR < 1$, we apply Bayes' Theorem to obtain

$$PPV = \frac{SE \times PR}{SE \times PR + (1 - SP) \times (1 - PR)}. \tag{1}$$

Four remarks can be made about formula (1).

First, for fixed values of SE and SP, PPV is an increasing function of PR. As PR approaches 0, so does PPV. (Is that intuitively plausible?) As PR approaches 1, so does PPV.

Second, for fixed values of SE and PR, PPV is an increasing function of SP. As SP approaches 0, PPV approaches a number that is no larger than PR. (Is that intuitively plausible?) As SP approaches 1, so does PPV.

Third, for fixed values of SP and PR, PPV is an increasing function of SE. As SE approaches 0, so does PPV. (Is that intuitively plausible?) As SE approaches 1, PPV approaches a number that is no smaller than PR.

Fourth, if PR is small (as is often the case), then $(1 - PR)$ is large and thus SP has a stronger impact on PPV than SE. This relates to our earlier observation that positive predictive value is deflated by false positives among healthy people, which occur often when the diagnostic test has poor specificity.

Our second motivating case study described Simpson's paradox and provided a numerical example in the context of infant mortality. Having defined conditional probabilities and enumerated computational techniques for them, we are now in a position to take a more formal look at Simpson's paradox.

As an aid to interpretation in what follows, imagine that A represents infant mortality, B represents a smoking mother, and C represents low birthweight. Thus, A^c represents infant survival, B^c represents a non-smoking mother, and C^c represents normal birthweight.

Simpson's paradox says that the overall association between endpoint A and risk factor B may be qualitatively different from the association between endpoint A and risk factor B in a stratum defined by confounder C. Supposing, for example, that the former association is positive while the latter association is negative, we have

$$P(A|B) > P(A|B^c) \tag{2}$$

and

$$P(A|B \cap C) < P(A|B^c \cap C). \tag{3}$$

We already know from our numerical example in the context of infant mortality that (2) and (3) can hold simultaneously, and we may be able to speculate about the underlying biological reasons in that context, but let us look for some mathematical insights.

We have

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{4}$$

and

$$P(A|B \cap C) = \frac{P(A \cap B \cap C)}{P(B \cap C)} = \frac{P(C|A \cap B)P(A \cap B)}{P(C|B)P(B)}. \tag{5}$$

Combining (4) and (5), we find that

$$P(A|B \cap C) = P(A|B) \times \frac{P(C|A \cap B)}{P(C|B)}. \tag{6}$$

Likewise, we find that

$$P(A|B^c \cap C) = P(A|B^c) \times \frac{P(C|A \cap B^c)}{P(C|B^c)}. \tag{7}$$

Taking note of (6) and (7), a necessary condition for (2) and (3) to hold simultaneously is that

$$\frac{P(C|A \cap B)}{P(C|B)} < \frac{P(C|A \cap B^c)}{P(C|B^c)}, \tag{8}$$

which can be rearranged as

$$\frac{P(C|A \cap B)}{P(C|A \cap B^c)} < \frac{P(C|B)}{P(C|B^c)}. \tag{9}$$

The right side of (9) can be interpreted as a relative risk, but this is not the relative risk in the way we usually think of it. Rather, the right side of (9) is the relative risk of the confounder C being present in relation to risk factor B.

The left side of (9) is the relative risk of the confounder C being present in relation to risk factor B, given that the endpoint A occurs.

Thus, the version of Simpson's paradox in (2) and (3) manifests because the relative risk of the confounder C being present in relation to risk factor B is attenuated by the occurrence of the endpoint A.

Relation (8) also provides some insight. While neither side of (8) may be interpreted as a relative risk, the left side may be interpreted as the factor by which the probability of the confounder C being present given the presence of risk factor B is amplified when we are also told that the endpoint A occurs, whereas the right side may be interpreted as the factor by which the probability of the confounder C being present given the absence of risk factor B is amplified when we are also told that the endpoint A occurs.

Revisiting our numerical example, we have

$$P(A|B) = 20/2000 = 0.01, \quad P(A|B^c) = 40/8000 = 0.005,$$

$$P(A|B \cap C) = 10/400 = 0.025, \quad \text{and} \quad P(A|B^c \cap C) = 30/800 = 0.0375.$$

We also see readily that

$P(C|B) =$ $\qquad$ $P(C|B^c) =$

$P(C|A \cap B) =$ $\qquad$ and $\quad P(C|A \cap B^c) =$

In our numerical example, knowledge that an infant has experienced mortality increases the probability of low birthweight by a multiplicative factor of 2.5 when the mother smokes but by a multiplicative factor of 7.5 when the mother does not smoke. When an infant born to a non-smoking mother experiences mortality, the causes of mortality must be other than smoking. However, when an infant born to a smoking mother experiences mortality, smoking may be one of the causes. Causes of mortality other than smoking are more strongly associated with low birthweight than smoking, accounting for the greater frequency of low birthweight accompanying mortality of infants born to non-smoking mothers than mortality of infants born to smoking mothers.