

Mathematical Modeling of Infectious Diseases, UCSF

Travis C. Porco, instructor

Lecture 1

Version 1.1—modified 8 Jan 2015, time 1226.

1 Independence

Suppose we consider the association of a test E with a disease D . The concept of *sensitivity* should be familiar: it is the probability of testing positive given the disease. We write

$$s = P(E|D).$$

In this case, we let D stand for the event that the person under consideration has the disease.

To go further, we need to talk a bit about probability.

Formal probability theory is built on top of a formalism of set theory, and events are defined to be members of a certain class of subsets of the *sample space*. We won't go into this into much detail. However, it is often useful to know the definition of a few set theoretic concepts, such as union, intersection, complement, and disjoint. We will also not go into formal detail into what exactly is meant by a random variable, but will be content with the intuitive sense of it.

The properties of probability are meant to reflect properties of relative frequency. Suppose an experiment can succeed or fail, and we perform it a number N of times, and count the number N_S of successes. The relative frequency is $f_S = N_S/N$. If the experiment succeeds every time, the relative frequency is 1; if it fails every time, 0. Probability can be interpreted as a long run limit of relative frequencies, and so probability must always be in the range 0 thru 1.

Suppose you roll a die. The sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$. In this simple problem, every possible subset of Ω can have a probability assigned to it. So we can ask what the probability of $A = \{1, 3\}$ is, or what the probability of $B = \{2, 5\}$ is.

Exercise 1. What is $P(\Omega) = P(\{1, 2, 3, 4, 5, 6\})$? ■

A crucial axiom of probability is called *additivity*: if two sets don't have any elements in common, then the probability of the union is the sum of the probabilities. These are called disjoint unions.

For a fair die, all elementary outcomes (elements of the sample space) have the same probability, and we can see that that probability has to be $1/6$, since $P(\{1\}) = P(\{2\}) = \dots = P(\{6\})$ and $P(\{1\}) + \dots + P(\{6\}) = 1$.

Exercise 2. Let D be the event that a person has a disease, and $\bar{D}$ be the event that the person does not have the disease. Show $P(D) + P(\bar{D}) = 1$. ■

We could consider D to be a random variable taking the value 1 when the person has the disease and 0 when they do not; then D would be an *indicator variable*. Similarly, we could let E be an indicator of whether or not the person tested positive. Then we could write $s = P(E = 1|D = 1)$.

Similarly, the definition of *specificity* is

$$f = P(\bar{E}|\bar{D})$$

the probability you test negative given you don't have the disease. Here we let $\bar{E}$ be the event the person does not have the disease. If we were using indicator variables, we would write $f = P(E = 0|D = 0)$.

We will be stern in our usage of the vertical bar, though in the literature people can be somewhat lax about it. When we use it, it will always refer to conditioning on a random variable or event or something; it should not be used to mean simply a functional dependence. In elementary probability, we have this definition:

$$P(E|D) = \frac{P(E \cap D)}{P(D)}$$

(provided $P(D) > 0$).

Optional exercise. Imagine that we consider a long run of people coming into the clinic, and for each person, we know their true disease status and their test result. If each person's results are considered an "experiment" in the probabilistic sense, what is the sample space? Let $f_{E|D}$ be the relative frequency of positive tests among those who really have the disease, let $f_{E \cap D}$ be the relative frequency of people who have both positive test results and true disease (out of all experiments), and let f_D be the fraction of people with disease. Show $f_{E|D} = f_{E \cap D} / f_D$. Argue in favor of the definition of conditional probability given above, based on long run limiting relative frequency. ■

Consider the event of a positive test, E . If you have a positive test, you either have the disease or you do not; we can write formally $E = E \cap D \cup E \cap \bar{D}$. In other words, we broke up the event of a positive test into two *mutually exclusive and exhaustive* events, the first being $E \cap D$, and the second being $E \cap \bar{D}$. By additivity,

$$P(E) = P(E \cap D) + P(E \cap \bar{D}).$$

Since

$$P(E \cap D) = P(D)P(E|D)$$

and similarly

$$P(E \cap \bar{D}) = P(\bar{D})P(E|\bar{D}),$$

$$P(E) = P(D)P(E|D) + P(\bar{D})P(E|\bar{D}).$$

This is an example of the so-called *law of total probability*, and we'll see it again and again.

Exercise 3. Let $P(D)$ be the prevalence π . Let $\theta = P(E)$, the fraction testing positive. Show

$$\theta = \pi s + (1 - \pi)(1 - f),$$

where s is the sensitivity and f is the specificity. Suppose the fraction testing positive is known from a survey to be 0.001. What is an implied bound on the specificity in this population? ■

Note that we can write

$$P(E \cup D) = P(E|D)P(D)$$

and

$$P(E \cup D) = P(D \cup E) = P(D|E)P(E).$$

Thus

$$P(D|E)P(E) = P(E|D)P(D),$$

so

$$P(D|E) = \frac{P(E|D)P(D)}{P(E)},$$

a result known as Bayes' Theorem.

Optional exercise. Let the prevalence $P(D) = \pi = 0.01$, the sensitivity be $s = P(E|D) = 0.99$, and the specificity be $f = P(\bar{E}|\bar{D}) = 0.98$. What is the predictive value of a positive, $P(D|E)$? ■

For two events D and E , we say they are independent if $P(D|E) = P(D)$. If this is true, then

$$P(D \cap E) = P(D|E)P(E) = P(D)P(E).$$

For two independent events, the probability of both occurring is the product of the probabilities. If $P(D|E) = P(D)$, then by Bayes Theorem, we have $P(E|D) = P(D|E)P(E)/P(D) = P(D)P(E)/P(D) = P(E)$, so it goes both ways.

Exercise 4. If D is independent of E , show $\bar{D}$ is independent of E . ■

For more than two events, we have to be more careful. For three events A , B , and C , these events are independent if the chance of every possible intersection is the product of the probabilities: $P(A \cap B) = P(A)P(B)$, $P(A \cap C) = P(A)P(C)$, $P(B \cap C) = P(B)P(C)$, and $P(A \cap B \cap C) = P(A)P(B)P(C)$.

Exercise 5. Suppose A , B , and C are binary events, that either happen or they don't. The sample space can be classified in a two by two by two table. Find a way to assign the probabilities of in this

table so that A and B are independent, A and C are independent, and B and C are independent, but A, B, and C are not independent. (For instance, make $P(A \cap B) = P(A)P(B)$, $P(A \cap C) = P(A)P(C)$, and $P(B \cap C) = P(B)P(C)$, but make sure $P(A \cap B \cap C) \neq P(A)P(B)P(C)$, etc.)

Exercise 6. Given a two by two table with cell probabilities $P(AB) = P(A \cap B)$, $P(A\bar{B})$, $P(\bar{A}B)$, and $P(\bar{A}\bar{B})$, the *odds ratio* is

$$o = \frac{P(AB)P(\bar{A}\bar{B})}{P(A\bar{B})P(\bar{A}B)}.$$

Show that the odds ratio equals 1 when and only when A and B are independent. ■

Exercise 7. Given a two by two by two table representing events A, B, and C, show

$$P(A|BC) = \frac{P(AB|C)P(C)}{P(BC)}.$$

Exercise 8. Two events B and C are conditionally independent given A if $P(BC|A) = P(B|A)P(C|A)$. Assume that B is conditionally independent of C given A. First show given this assumption, $P(B|AC) = P(B|A)$. Then show that

$$P(A|BC) = \frac{P(B|A)P(A|C)}{P(B|C)}.$$

Exercise 9. Suppose a person has a sexual partner and that the probability that the partner is infected with HIV is π . Suppose that the probability of infection given that the partner is infected is β . What is the probability of infection in this partnership? ■

Exercise 10. Show that $P(A \cup B \cup C) \leq P(A) + P(B) + P(C)$. Suppose a person has a needlestick injury with blood contaminated with Hepatitis B, Hepatitis C, and HIV; suppose the probability of HBV transmission is 0.3 from any needlestick, HCV transmission is 0.03 from any needlestick, and HIV transmission is 0.003 from any needlestick. Show the total probability of any infection is less than or equal to 0.333, no matter what the independence or nonindependence between the infections is. ■

Worked exercise. Suppose a needle is being—improperly!—reused between patients, contrary to acceptable policy. Suppose the prevalence of infection in the patient pool is p and the needle is used on a random patient. Suppose that a needle is first rinsed, decontaminating it with probability b . Then, if it is used on an infected patient, the needle becomes contaminated. Find the chance the needle is contaminated after one round, given that it was or was not infected to begin with.

Solution. The needle is rinsed, then used. If the needle was not contaminated, rinsing is irrelevant; the needle becomes contaminated with probability p . If the needle was contaminated, then it is either decontaminated or it was not. It is decontaminated with probability b and not decontaminated with probability $1 - b$. Given decontamination, the probability of recontamination from the patient is p ; given no decontamination, the patient's status is assumed irrelevant. We use the law of total probability to get $bp + 1 - b$.

2 The binomial distribution

A random variable that takes on two values, 0 and 1, is called a Bernoulli random variable. Suppose a Bernoulli random variable has probability p of being 1; 1 will conventionally be called a success.

The count of successes in N independent and identical Bernoulli trials is called the Binomial distribution. The Binomial probability formula is

$$P(X = x) = \binom{N}{x} p^x (1 - p)^{N-x}.$$

Exercise 11. Suppose two parents are carriers of the sickle gene and that the chance of passing the gene to the next generation is $1/2$. What is the probability distribution of the number of children out of 4 who would be born homozygous (received the gene from both parents)? Assume independent and identical risk. ■

Exercise 12. Suppose a person has n sexual partners chosen at random from a pool of partners in which the prevalence is π and the risk of transmission given a positive partner is β . Show the probability of infection is $1 - (1 - \beta p)^n$. ■

3 Expected values

The expected value of a random variable is defined as

$$EX = \sum_x xP(X = x).$$

This is a kind of large number generalization of the sample mean.

Exercise 13. Show the expectation of a Bernoulli random variable is p . ■

The expectation of any sum of random variables is always the sum of the expectations. Let's show this for two. Let X and Y be jointly distributed. Say these are discrete variables, and that X takes values $x_1, x_2, \dots$ and Y takes values $y_1, y_2, \dots$. Then

$$\begin{aligned} E[X + Y] &= \sum_x \sum_y (x + y)P(X = x, Y = y) = \sum_x \sum_y xP(X = x, Y = y) + \sum_x \sum_y yP(X = x, Y = y) \\ E[X + Y] &= \sum_x x \sum_y P(X = x, Y = y) + \sum_y y \sum_x P(X = x, Y = y) \\ E[X + Y] &= \sum_x xP(X = x) + \sum_y yP(Y = y) = E[X] + E[Y]. \end{aligned}$$

In advanced books they do these things carefully to make sure you can sum the series and change the order of summation and whatnot.

Exercise 14. It has been estimated that the fraction of wild-type tubercle bacilli which harbor a resistance mutation to isoniazid, a first-line drug, is 10^{-6} , and the fraction that harbor mutations to rifampicin is 10^{-8} . Assuming these are independent, what is the probability of harboring mutations conferring resistance to both (and by definition being multidrug resistant)? Assume that a cavitary lesion contains 10^9 organisms. What is the expected number with isoniazid mutations? Rifampicin mutations? MDR (both isoniazid and rifampicin mutations)? ■

Here's another important fact:

$$E[kX] = kE[X]$$

where k is some constant.

Here is another—suppose X and Y are independent. Then $E[XY] = E[X]E[Y]$. Why:

$$E[XY] = \sum_x \sum_y xyP(X=x, Y=y) = \sum_x \sum_y xP(X=x)yP(Y=y) = \sum_x xP(X=x) \sum_y yP(Y=y) = E[X]E[Y].$$

Exercise 15. Let X and Y be jointly distributed with $P(X=0, Y=0) = 0$, $P(X=0, Y=1) = 1/2$, $P(X=1, Y=0) = 1/2$, $P(X=1, Y=1) = 0$. Here, $P(X=1) = 1/2$ and $P(Y=0) = 1/2$, but $P(X=1, Y=0) \neq 1/2 \times 1/2$, so X and Y are NOT independent. Compute $E[XY]$ from the definition. What is $E[X]E[Y]$? ■

Exercise 16. Since the number of successes in N independent identical Bernoulli trials is the sum of their values, every 1 being a success, show that if X is a Binomial with N trials and p successes per trial, $EX = Np$. ■

Exercise 17. Suppose 51 people receive a needlestick injury with hepatitis C contaminated blood. Assume independent and identical risk, and that 3 individuals were infected as a result. Compute the probability of this event for $p = 0$, $p = 0.001$, $p = 0.01$, $p = 3/51$, $p = 0.1$, $p = 0.5$. ■

The probability of the data as a function of unknown parameters is called the likelihood. It can be shown for the Binomial model that the value of the unknown success probability that maximizes the likelihood is the relative frequency $\hat{p} = X/N$.

We can also take the expected value of a function of a random variable. In general,

$$E[g(X)] = \sum_x g(x)P(X=x).$$

An important example is the second moment $E[X^2]$. Note that the second moment is always bigger than the square of the mean:

$$E[X^2] \geq (E[X])^2.$$

It's easier to see this backwards:

$$E[(X - E[X])^2] = E[X^2 - 2XE[X] + (E[X])^2].$$

$$E[(X - E[X])^2] = E[X^2] + E[-2XE[X]] + E[(E[X])^2]$$

$$E[(X - E[X])^2] = E[X^2] - 2(E[X])E[X] + (E[X])^2$$

$$E[(X - E[X])^2] = E[X^2] - (E[X])^2$$

But the left can't ever be negative; it's an expectation of things that are never negative. So

$$E[X^2] - (E[X])^2 \geq 0$$

As you probably know, $E[(X - E[X])^2]$ is called the *variance*, so we've just shown it has to be nonnegative.

Suppose X and Y are independent. Then the variance of the sum is the sum of the variances. But in general variances don't add.

$$\text{var}[X + Y] = E[(X + Y)^2] - (E[X + Y])^2 = E[X^2 + 2XY + Y^2] - (E[X] + E[Y])^2$$

$$\text{var}[X + Y] = E[X^2] + 2E[XY] + E[Y^2] - (E[X])^2 - 2E[X]E[Y] - (E[Y])^2$$

We assumed independence, so $EXY = EXEY$, which means two terms cancel, leaving us with

$$\text{var}[X + Y] = E[X^2] - (E[X])^2 + E[Y^2] - (E[Y])^2 = \text{var}[X] + \text{var}[Y].$$

Exercise 18. Show the variance of a Bernoulli random variable is $p(1 - p)$. Use this and the rule for summing variances of independent variables to show the variance of a Binomial with N trials and p per trial is $\text{var}X = Np(1 - p)$. ■

We need another fact. What is $\text{var}(kX)$ where k is some constant? We know

$$\text{var}(kX) = E[(kX)^2] - (E[kX])^2,$$

so

$$\text{var}(kX) = E[k^2X^2] - (kE[X])^2 = k^2E[X^2] - k^2(E[X])^2 = k^2(E[X^2] - (E[X])^2) = k^2\text{var}(X).$$

So the variance of a multiple is the square of that multiple times the original variance.

Let's consider the idea of estimating a Binomial probability by using the observed relative frequency. In other words, say we observed a certain number of successes out of N Bernoulli trials, and you want to estimate the unknown p , the success probability.

We can think of taking whatever we observe and dividing by N , which is just the observed relative frequency. If we imagine repeating the experiment over and over, this relative frequency is a random variable: $\hat{p} = X/N$.

What is its expected value? Here, $E[\hat{p}] = E[X/N] = (1/N)E[X]$, since N is a constant. But X is a Binomial random variable. So, we have $E[\hat{p}] = (1/N)E[X] = (1/N)Np = p$. The expected value of the estimator $\hat{p}$ is the thing we are estimating, p . We don't know what p is in any real application of this, but whatever it is, the expected value of $\hat{p}$ for some study of size N is p . This is a statistical property called *unbiasedness*; the relative frequency is an unbiased estimator of the unknown success probability.

What is the variance of $\hat{p}$? Here, $\text{var}(\hat{p}) = \text{var}(X/N) = (1/N)^2 \text{var}(X)$, since $1/N$ is a constant. Then, from the variance of a Binomial, we have $\text{var}(\hat{p}) = (1/N)^2 Np(1-p) = p(1-p)/N$. The standard deviation is just the square root of that.

Exercise 19. Suppose you believe that a certain drug will work around 30 percent of the time, and you want to estimate this to within a standard deviation of plus or minus five percent in absolute terms. About how large does N have to be? ■

Exercise 20. How long did this assignment take you? We want future assignments to take about six hours. ■