

## **Fundamentals of Biostatistical Inference I**

Course TA: Rajarshi Guha Niyogi, guhan003@umn.edu.

Rajarshi's office hours: Mon. 1:30-2:30 & Wed. 1:30-2:30 in Mayo A446.

Tim Hanson office hours: Tues. 11:00-1:00 & Thur. 11:00-12:00 in Mayo A444.

Homework will be due beginning of class (i.e. 1:25) on the days it is due. Please place in a pile on a table at the front of class.

## Section 1.5: Conditional probability

The probability of any event  $A$  can change given that we know some other event  $B$  has occurred.

If we know  $B$  has occurred, then only those outcomes  $s \in B$  are possible. The sample space  $S$  of possible outcomes has been reduced to those elements in  $B$ , and we must accordingly “adjust” probabilities attached to events  $A$  happening.

The probability that  $A$  occurs given that we know  $B$  has occurred is written  $P(A|B)$ .

Let's consider a simple example. Say a 6-sided die is rolled but it rolls too far away for us to see it. A person down the hall can see the die and informs us that it is odd number; we know the event  $B = \{1, 3, 5\}$  has occurred.

- What is the probability of the outcome  $\{1\}$ ? Before the roll all possible outcomes  $\{1, 2, 3, 4, 5, 6\}$  were equally likely; the *new* possible outcomes in  $B = \{1, 3, 5\}$  should also be equally likely now that we've learned  $B$  has happened. There's three elements in  $B$ , so the probability is  $1/3$ . We write

$$P(\{1\}|\{1, 3, 5\}) = 1/3.$$

- What is the probability of the event  $\{1, 2\}$ ? We already know  $\{2\}$  did not occur. So the answer remains the same,

$$P(\{1, 2\}|\{1, 3, 5\}) = 1/3.$$

**def'n:** The **conditional probability** of  $A$  given  $B$ , written  $P(A|B)$ , is

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

provided  $B$  can happen, i.e.  $P(B) > 0$ .

$P(A|B)$  is a probability function as well and has the usual properties.

- $P(A^C|B) = 1 - P(A|B)$ .
- $A_1 \subset A_2$  implies  $P(A_1|B) \leq P(A_2|B)$ .
- $P(A_1 \cup A_2|B) = P(A_1|B) + P(A_2|B) - P(A_1 \cap A_2|B)$ .

Example (Windle, 2003): Let  $D$  be the event that a randomly selected 12th-grader has been drunk in the last 30 days, let  $M$  be the event that a 12th-grader is male. There are  $m = 2$  levels of the experiment “drunk” and  $n = 2$  levels of the experiment “male.” So there are  $mn = 4$  possible outcomes to the experiment (drunk,male), namely  $S = \{dm, df, nm, nf\}$ .

Outcome	$s_i \in S$	$p_i = P(s_i)$
not drunk, male	$nm$	0.302
not drunk, female	$nf$	0.395
drunk, male	$dm$	0.157
drunk, female	$df$	0.146

Let  $D = \{dm, df\}$  be the event that a randomly selected student was drunk in the last 30 days and  $M = \{dm, nm\}$  be the event that a randomly selected student is male. Then

$$P(D) = P(\{dm, df\}) = 0.157 + 0.146 = 0.303.$$

$$P(M|D) = \frac{P(M \cap D)}{P(D)} = \frac{P(\{dm\})}{P(D)} = \frac{0.157}{0.303} = 0.518.$$

$$P(M^C|D) = 1 - P(M|D) = 1 - 0.518 = 0.482.$$

$$P(D|M) = \frac{P(M \cap D)}{P(M)} = \frac{0.157}{0.157 + 0.302} = 0.342.$$

$$P(D|M^C) = \frac{P(M^C \cap D)}{P(M^C)} = \frac{0.146}{0.146 + 0.395} = 0.270.$$

A student is  $0.342/0.270 = 1.27$  times more likely to have gotten drunk in the last 30 days if male rather than female. This number is called a relative risk. See additional book example pp. 16-17.

**Proposition:**  $P(A \cap B) = P(A|B)P(B)$  as long as  $P(B) > 0$ .

Example: an urn contains 3 red balls and one blue ball. Two balls are drawn without replacement. What is the probability they are both red?

Let  $R_1$  and  $R_2$  denote the events of obtaining a red ball on the first and second draws. Both balls are red if the event  $R_1 \cap R_2$  happens.

$$P(R_1 \cap R_2) = P(R_2|R_1)P(R_1) = \frac{2}{3} \times \frac{3}{4} = 0.5.$$

**def'n:** Let  $B_1, B_2, \dots, B_n$  be mutually disjoint ( $B_i \cap B_j = \emptyset$  for  $i \neq j$ ) and such that  $B_1 \cup B_2 \cup \dots \cup B_n = S$ . The set of sets  $\{B_i\}_{i=1}^n$  is called a *partition* of  $S$ .

e.g. Rolling a 6-sided die.  $B_1 = \{1, 5\}$ ,  $B_2 = \{2\}$ ,  $B_3 = \{3, 4, 6\}$  is a partition of  $S = \{1, 2, 3, 4, 5, 6\}$ .

e.g. Experimental outcome is survival time in years of cow infected with Johnne's disease. Let  $S = (0, \infty)$ .  $B_1 = (0, 1]$ ,  $B_2 = (1, 2]$ ,  $B_3 = (2, 3]$ ,  $B_4 = (3, \infty)$  is one (of many) partition of  $S$ .

**Law of total probability:** Let  $B_1, B_2, \dots, B_n$  be a partition of  $S$ .

Then

$$P(A) = \sum_{i=1}^n P(A|B_i)P(B_i).$$

See proof in book. A Venn diagram illustrates this nicely by noting  $P(A|B_i)P(B_i) = P(A \cap B_i)$ .

**Example: Screening for hepatitis C at an STD clinic**

(Weisbord, Trepka, Zhang, Smith, and Brewer, 2003). At an STD clinic in Miami, Florida, patients were screened for hepatitis C using CDC screening criteria in the form of a questionnaire.

Let  $T+$  denote the event that the CDC screening criteria indicates a randomly selected individual has hepatitis C and  $T-$  denote the complimentary event. Let  $D+$  denote the event that a randomly selected individual actually is infected with hepatitis C.

$P(T+ | D+)$  is called the *sensitivity* of the screening test and  $P(T- | D-)$  is called the *specificity*.  $P(D+)$  is the prevalence of hepatitis C in this population.

This study concluded  $P(T + |D+) = 0.61$ ,  $P(T - |D-) = 0.91$  and  $P(D+) = 0.047$ . What is the probability that a screening test comes up positive for a randomly selected individual? Note that an individual can either have the disease or not, so the events  $D+$  and  $D-$  partition the sample space of experimental outcomes.

$$\begin{aligned}P(T+) &= P(T + |D+)P(D+) + P(T + |D-)P(D-) \\&= P(T + |D+)P(D+) + [1 - P(T - |D-)][1 - P(D+)] \\&= 0.61 \times 0.047 + (1 - 0.91)(1 - 0.047) = 0.114.\end{aligned}$$

Note that we are *implicitly* partitioning the sample space into two pieces  $D+$  and  $D-$  without actually defining the sample space.

What *is* the sample space for the experiment (CDC screening test, infection status)?

This is sort of like rolling a 6-sided die and defining the events “even” and “odd.” We know a roll has to be one of the two events without ever actually listing the experimental outcomes in the sample space.

A method for computing conditional probabilities:

**Bayes' rule:** Let  $P(B) > 0$ . Then

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

A more elaborate version using a more general partition  $\{A_1, A_2, \dots, A_n\}$  instead of the simple partition  $\{A, A^C\}$  is in your text on page 20. Proving Bayes rule uses the definition of conditional probability and the law of total probability.

Screening for hepatitis C revisited. Let's say the CDC criteria tells me I'm at risk for hepatitis C, i.e. that  $T+$ . What is the probability that I really have it?

$$P(D+ | T+) = \frac{P(T+ | D+)P(D+)}{P(T+)} = \frac{0.61 \times 0.047}{0.114} = 0.25.$$

There's still only a 1 in 4 chance I've got hepatitis C. But this is much larger than 0.047, the probability before knowing  $T+$ .

Better get a blood test.

More interesting examples pp. 20-22.

## Section 1.6: Independence

Two events  $A$  and  $B$  are independent if knowing that  $B$  occurred does not change the probability that  $A$  has occurred. That is,  $A$  is independent of  $B$  if

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

i.e. if

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

**def'n:**  $A$  is independent of  $B$  if  $P(A \cap B) = P(A)P(B)$ .

Note that  $P(A \cap B) = P(A)P(B)$  if and only if  $P(A|B) = P(A)$  if and only if  $P(B|A) = P(B)$ .

**Property:** If  $A$  independent of  $B$ , then  $A$  independent of  $B^C$ ,  $A^C$  independent of  $B$ , and  $A^C$  independent of  $B^C$ .

e.g. In the hepatitis C example, is  $T+$  independent of  $D+$ ?

No because

$$P(D+ | T+) = 0.25 \neq 0.047 = P(D+).$$

Knowing  $T+$  tells

us something about  $D+$  and knowing  $D+$  tells us something about  $T+$ .

e.g. If  $A$  and  $B$  are disjoint are they independent?

No, they are highly dependent. In fact knowing  $A$  happens tells us that  $B$  *did not happen*. By definition  $P(A|B) = 0$  and  $P(B|A) = 0$ .

Formally:

$$P(A \cap B) = P(\emptyset) = 0 \neq P(A)P(B).$$

**def'n:** A collection of events  $\{A_1, A_2, \dots, A_n\}$  are *mutually independent* if for all subcollections of sizes  $k = 1, 2, \dots, n$   $\{A_{i_1}, A_{i_2}, \dots, A_{i_k}\}$  (i.e.  $\{i_1, i_2, \dots, i_k\}$  is a combination of size  $k$  from  $\{1, 2, 3, \dots, n\}$ ),

$$P(A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_k}) = P(A_{i_1})P(A_{i_2}) \dots P(A_{i_k}).$$

For example,  $\{A_1, A_2, A_3\}$  are mutually independent if all four of the following are true

$$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3), \quad P(A_1 \cap A_2) = P(A_1)P(A_2),$$

$$P(A_1 \cap A_3) = P(A_1)P(A_3), \quad P(A_2 \cap A_3) = P(A_2)P(A_3).$$

**Balls in urns:** Let an urn have 10 red balls and 5 green balls. We will draw three balls from the urn, let  $R_1$ ,  $R_2$ , and  $R_3$  denote the events that a red ball is obtained on the first, second, or third draws respectively.

The sample space is  $S = \{rrr, rrg, rgr, rgg, grr, grg, ggr, ggg\}$ .

$R_2 = \{rrr, rrg, grr, grg\}$ , etc.

We are interested in whether  $\{R_1, R_2, R_3\}$  are mutually independent, which may depend on whether balls are put back after being drawn or not.

Are  $\{R_1, R_2, R_3\}$  mutually independent when balls are drawn and replaced immediately?

Yes. The ball is put back after drawing it the first time so  $P(R_2|R_1) = \frac{2}{3} = P(R_2)$ .

Also  $P(R_3|R_1) = P(R_3|R_2) = \frac{2}{3}$ . These all imply that  $P(R_1 \cap R_2) = P(R_1)P(R_2) = \frac{4}{9}$ . One can further check that  $P(R_1 \cap R_2 \cap R_3) = P(R_3|R_1 \cap R_2)P(R_1 \cap R_2) = \frac{8}{27} = P(R_1)P(R_2)P(R_3)$ .

All 4 conditions for mutual independence among  $\{R_1, R_2, R_3\}$  check out.

Are  $\{R_1, R_2, R_3\}$  mutually independent when balls are drawn and not replaced?

By the law of total probability

$$P(R_2) = P(R_2|R_1)P(R_1) + P(R_2|R_1^C)P(R_1^C) = \frac{9}{14} \times \frac{2}{3} + \frac{10}{14} \times \frac{1}{3} = \frac{2}{3}.$$

Similarly,  $P(R_3) = \frac{2}{3}$ .

When sampled without replacement  $P(R_2|R_1) = \frac{9}{14}$  and  $P(R_1) = \frac{1}{3}$ .

All 4 conditions must check out. We'll start by checking if  $P(R_1 \cap R_2) = P(R_1)P(R_2)$  is true.

$$\begin{aligned} P(R_1 \cap R_2) &= P(R_2|R_1)P(R_1) \\ &= \frac{9}{14} \times \frac{2}{3} = \frac{3}{7} \\ &\neq \frac{4}{9} = P(R_1)P(R_2). \end{aligned}$$

We can stop here;  $\{R_1, R_2, R_3\}$  are not mutually independent.

Homework 1 (part A): 1, 2\*, 3, 4, 5, 9, 11, 12, 13, 14\*, 15, 16\*, 19, 20\*, 21, 22, 25, 27, 30\*, 31, 32, 33, 38\*, 41

Homework 1 (part B): 42\*, 45, 46\*, 47, 49, 50\*, 52, 53, 56\*, 60\*, 63, 67, 70\*, 78 & 79 (if interested).

\*hand in **Thursday**, Sept. 11.

We will attempt some of these in class.