Probability

Patrick Breheny

September 1

Probability

- Statistical inference – the idea of generalizing about a population from a sample – inherently involves some degree of uncertainty, and "the language of uncertainty is probability" (James Berger)
- People talk loosely about probability all the time, but for scientific purposes, we need to be more specific in terms of defining and using probabilities

Events

- A *random process* is a phenomenon whose outcome cannot be predicted with certainty
- The *sample space* is the collection of all possible outcomes of a random process
- An *event* is an outcome (or collection of outcomes) of a random process:
- Examples:

Random process	Event
Flipping a coin	Obtaining heads
Rolling a die	Obtaining an odd number
Child receives a vaccine	Child contracts polio
Patient takes Clofibrate	Patient survives

Probability: Mathematical definition

Definition: Given a sample space S and collection of events $\mathcal{F}$, a *probability function* is a function P with domain $\mathcal{F}$ that satisfies

1: $P(A) \geq 0$ for all $A \in \mathcal{F}$

2: $P(S) = 1$

3: If $A, B \in \mathcal{F}$ satisfy $A \cap B = \emptyset$, then

$$P(A \cup B) = P(A) + P(B)$$

(Note that $\mathcal{F}$ has to satisfy certain properties for this definition to work – e.g., if A and B are in $\mathcal{F}$, then $A \cup B$ must be in $\mathcal{F}$ – and that this can get tricky if $\mathcal{F}$ is an infinite collection; mathematical statistics courses such as STAT 4100/STAT 5100 discuss these issues in greater detail)

Probability and long-run frequency

- So a probability function takes an event and assigns it a number between 0 and 1; however, the mathematical definition doesn't tell us anything about *how* to assign probabilities to events
- Everyone agrees that the probability of rolling a 1 on a fair die is 1/6 . . . but why?
- In many cases, probability has an uncontroversial interpretation as a *long-run frequency*: the probability of an event occurring is the fraction of time that it would happen if the random process occurs over and over again under the same conditions
- Thus, the probability of rolling a 1 is 1/6 because it happens 1/6 of the time when we roll a die

Long-run frequency

- Suppose that the probability of developing polio for a child who receives a vaccine is 0.00031
- By the long-run frequency interpretation, if we vaccinate 100,000 children, we would expect therefore that 31 of those children will develop polio
- This works the other way too: in our polio study, 28 per 100,000 children who got the vaccine developed polio
- Thus, the probability that a child in our sample who got the vaccine developed polio is 28/100,000=.00028
- Of course, what we really want to know is the probability of a child in the population developing polio (not the sample) – we'll get there

Probability for non-repeatable processes

- Not everyone agrees, however, on what probability means for non-repeatable random processes
- For example, what is the probability that Donald Trump wins the presidential election in 2016?
- Some would argue that it is valid to make probability statements about this event as a reflection of one's subjective belief that the event will happen; others would say that probabilities are not meaningful here because they cannot be objectively established
- We'll return to this point later

The complement rule

- We are often interested in events that are derived from other events, such as complements, unions, and intersections of events; we will now cover the basic rules that allow us to calculate such probabilities, starting with complements
- **Theorem (Complement rule):** For any event A,

$$P(A^C) = 1 - P(A)$$

- This simple but useful rule is called the *complement rule*
- **Example:** Suppose the probability of developing polio is 0.0006. What is the probability of not developing polio?

Introduction

- Next, we turn our attention to unions
- From the definition of probability, we can see that for two events satisfying $A \cap B = \emptyset$, we have $P(A \cup B) = P(A) + P(B)$
- So, at least in some cases, we can find the probability of a union by just adding the probabilities of the two events
- However, this is not true in general

A counterexample

- Let A denote rolling a number 3 or less and B denote rolling an odd number
- $P(A) + P(B) = 0.5 + 0.5 = 1$
- Clearly, however, we could roll a 4 or a 6, which is neither A nor B

The addition rule

- The issue, of course, is that we are "double-counting" events in $A \cap B$
- Simply subtracting $P(A \cap B)$ from our answer corrects this problem
- **Theorem (Addition rule):** For any two events A and B,

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

Mutually exclusive events

- Note that $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ holds for *any* two events, while $P(A \cup B) = P(A) + P(B)$ only holds when $A \cap B = \emptyset$
- A special term is given to the situation when A and B cannot both occur at the same time (i.e., when $P(A \cap B) = 0$): such events are called *mutually exclusive*

Example

- An article in the *American Journal of Public Health* reported that in a certain population, the probability that a child's gestational age is less than 37 weeks is 0.142
- The probability that his or her birth weight is less than 2500 grams is 0.051
- The probability of both is 0.031
- **Exercise:** What is the probability that a child will weight less than 2500 grams or be born at less than 37 weeks?

Failing to use the addition rule

- In the 17th century, French gamblers used to bet on the event that in 4 rolls of the die, at least one "ace" would come up (an ace is rolling a one)
- In another game, they rolled a pair of dice 24 times and bet on the event that at least one double-ace would turn up
- The Chevalier de Méré, a French nobleman, thought that the two events were equally likely

Failing to use the addition rule (cont'd)

His reasoning was as follows: letting A_i denote the event of rolling an ace on roll i and AA_i denote the event of rolling a double-ace on roll i

$$\begin{aligned} P(A_1 \cup A_2 \cup A_3 \cup A_4) &= P(A_1) + P(A_2) + P(A_3) + P(A_4) \\ &= \frac{4}{6} = \frac{2}{3} \\ P(AA_1 \cup AA_2 \cup \cdots) &= P(AA_1) + P(AA_2) + \cdots \\ &= \frac{24}{36} = \frac{2}{3} \end{aligned}$$

This reasoning, of course, fails to recognize that A_1, A_2, ... are not mutually exclusive

Balls in urns

- We now turn our attention to the probabilities of intersections
- Imagine a random process in which balls are placed into an urn and picked out at random, so that each ball has an equal chance of being drawn
- For example, imagine an urn that contains 1 red ball and 2 black balls
- Let R_i denote that the i^{th} ball was red
- Clearly, $P(R_1) = 1/3$, but what about $P(R_1 \cap R_2)$?

Sampling with replacement

- First, suppose that after we draw a ball, we put it back in the urn before drawing the next ball (this method of drawing balls from the urn is called *sampling with replacement*)
- In this case,

$$P(R_1 \cap R_2) = \frac{1}{3}\left(\frac{1}{3}\right) = \frac{1}{9}$$

- On the surface, then, it would seem that $P(A \cap B) = P(A) \cdot P(B)$
- However, this is not true in general

Sampling without replacement

- Now suppose we don't put the balls back after drawing them (this method of drawing balls from the urn is called *sampling without replacement*)
- Now, it is impossible to draw two red balls; instead of 11%, the probability is 0
- Here, the outcome of the first event changed the random process; after R_1 occurs, $P(R_2)$ is no longer $1/3$, but 0
- When we draw without replacement, $P(R_i)$ depends on what has happened in the earlier draws

Conditional probability

- The notion that the probability of an event may depend on other events is called *conditional probability*
- The conditional probability of event A given event B is written as $P(A|B)$
- For example, in our ball and urn problem, when sampling without replacement:
  - $P(R_2) = \frac{1}{3}$
  - $P(R_2|R_1) = 0$
  - $P(R_2|R_1^C) = \frac{1}{2}$
- Some additional terms (we'll define these more formally later in the course):
  - $P(A)$ is known as the "marginal probability" of A
  - $P(A \cap B)$ is known as the "joint probability" of A and B

The multiplication rule

- To determine $P(A \cap B)$ in general, we need to use the *multiplication rule*
- **Multiplication rule:** For any two events A and B,

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

- Rearranging the formula, we have

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

which allows us to calculate conditional probabilities

Gestational age example

- Recall our earlier example, where the probability that a child's gestational age is less than 37 weeks is 14.2%, the probability that his or her birth weight is less than 2500 grams is 5.1%, and the probability of both is 3.1%
- **Exercise:** What is the probability that a child's birth weight will be less than 2500 grams, given that his/her gestational age is less than 37 weeks?

Independence

- Note that sometimes, event B is completely unaffected by event A, and $P(B|A) = P(B)$
- If this is the case, then events A and B are said to be *independent*
- This works both ways – all the following are equivalent:
  - $P(A) = P(A|B)$
  - $P(B) = P(B|A)$
  - A and B are independent
- Otherwise, if the probability of A depends on B (or vice versa), then A and B are said to be *dependent*

Dependence and independence

Scientific questions often revolve around conditional probability and independence – are two events independent, and if they are dependent, how dependent are they?

Event A	Event B
Patient survives	Patient receives treatment
Student is admitted	Student is male
Person develops lung cancer	Person smokes
Patient will develop disease	Mutation of a certain gene

Independence and the multiplication rule

- Note that if A and B are independent, the multiplication rule reduces to $P(A \cap B) = P(A)P(B)$, which is often much easier to work with, especially when more than two events are involved
- For example, consider an urn with 3 red balls and 2 black balls; what is the probability of drawing three red balls?
- With replacement:

$$P(R_1 \cap R_2 \cap R_3) = \left(\frac{3}{5}\right)^3 = 21.6\%$$

- Without replacement:

$$P(R_1 \cap R_2 \cap R_3) = \frac{3}{5} \cdot \frac{2}{4} \cdot \frac{1}{3} = 10\%$$

Independent versus mutually exclusive

- It is important to keep in mind that "independent" and "mutually exclusive" mean very different things
- For example, consider drawing a random card from a standard deck of playing cards
  - A deck of cards contains 52 cards, with 4 suits of 13 cards each
  - The 4 suits are: hearts, clubs, spades, and diamonds
  - The 13 cards in each suit are: ace, king, queen, jack, and 10 through 2
- If event A is drawing a queen and event B is drawing a heart, then A and B are independent, but not mutually exclusive
- If event A is drawing a queen and event B is drawing a four, then A and B are mutually exclusive, but not independent
- It is impossible for two events to be both mutually exclusive and independent

The Chevalier de Méré, Part II

- We can also use the rules of probability in combination to solve the problem that stumped the Chevalier de Méré
- Recall that we are interested in two probabilities:
  - What is the probability of rolling four dice and getting at least one ace?
  - What is the probability of rolling 24 pairs of dice and getting at least one double-ace?

The Chevalier de Méré, Part II (cont'd)

- First, we can use the complement rule:

$$P(\text{At least one ace}) = 1 - P(\text{No aces})$$

- Next, we can use the multiplication rule:

$$\begin{aligned} P(\text{No aces}) =& P(\text{No aces on roll 1}) \\ & \cdot P(\text{No aces on roll 2} | \text{No aces on roll 1}) \\ & \cdots \end{aligned}$$

- Are rolls of dice independent?
- Yes; therefore,

$$\begin{aligned} P(\text{At least one ace}) &= 1 - \left(\frac{5}{6}\right)^4 \\ &= 51.7\% \end{aligned}$$

The Chevalier de Méré, Part II (cont'd)

- By the same reasoning,

$$P(\text{At least one double-ace}) = 1 - \left(\frac{35}{36}\right)^{24}$$
$$= 49.1\%$$

- Note that this is a little smaller than the first probability, and that both are much smaller than the $\frac{2}{3}$ probability reasoned by the Chevalier

Caution

- With dice, independence is clear and we can multiply probabilities to get the right answer
- However, people often multiply probabilities when events are not independent, leading to incorrect answers
- A dramatic example of misusing the multiplication rule occurred during the 1999 trial of Sally Clark, on trial for the murder of her two children
- Clark had two sons, both of which died of sudden infant death syndrome (SIDS)

The Sally Clark case

- One of the prosecution's key witnesses was the pediatrician Roy Meadow, who calculated that the probability of one of Clark's children dying from SIDS was 1 in 8543, so the probability that both children had died of natural causes was

$$\left(\frac{1}{8543}\right)^2 = \frac{1}{73,000,000}$$

- This figure was portrayed as though it represented the probability that Clark was innocent, and she was sentenced to life imprisonment

The Sally Clark case (cont'd)

- However, this calculation is both inaccurate and misleading
- In a concerned letter to the Lord Chancellor, the president of the Royal Statistical Society wrote:

> *The calculation leading to 1 in 73 million is invalid. It would only be valid if SIDS cases arose independently within families, an assumption that would need to be justified empirically. Not only was no such empirical justification provided in the case, but there are very strong reasons for supposing that the assumption is false. There may well be unknown genetic or environmental factors that predispose families to SIDS, so that a second case within the family becomes much more likely than would be a case in another, apparently similar, family.*

The Sally Clark case (cont'd)

- There are also a number of issues, also mentioned in the letter, with the accuracy of the calculation that produced the "1 in 8543" figure
- Finally, it is completely inappropriate to interpret the probability of two children dying of SIDS as the probability that the defendant is innocent
- The probability that a woman would murder both of her children is also extremely small; one needs to compare the probabilities of the two explanations
- The British court of appeals, recognizing the statistical flaws in the prosecution's argument, overturned Clark's conviction and she was released in 2003, having spent three years in prison

The law of total probability

- A rule related to the addition rule is called *the law of total probability*, which states that if you divide A into the part that intersects B and the part that doesn't, then the sum of the probabilities of the parts equals $P(A)$
- **Theorem (Law of total probability):** For any events A and B,

$$P(A) = P(A \cap B) + P(A \cap B^C)$$

The law of total probability in action

- Again recall the gestational age problem: $P(E) = 0.142$, $P(L) = 0.051$, and $P(E \cap L) = 0.031$
- **Exercise:** What is the probability of low birth weight (L) given that the gestational age was greater than 37 weeks (E^C)? How do the conditional probabilities $P(L|E)$ and $P(L|E^C)$ relate to the unconditional probability $P(L)$?

Introduction

- Conditional probabilities are often easier to reason through (or collect data for) in one direction than the other
- For example, suppose a woman is having twins
- Obviously, if she were having identical twins, the probability that the twins would be the same sex would be 1, and if her twins were fraternal, the probability would be 1/2
- But what if the woman goes to the doctor, has an ultrasound performed, learns that her twins are the same sex, and wants to know the probability that her twins are identical?

Bayes' rule

- So, we know $P(\text{Same sex}|\text{Identical})$, but we want to know $P(\text{Identical}|\text{Same sex})$
- To flip these probabilities around, we can use something called *Bayes' rule*
- **Theorem (Bayes' Rule)**: For any events A and B with nonzero probability,

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

Applying Bayes' rule

- To apply Bayes' rule, we need to know one other piece of information: the unconditional probability that a pair of twins will be identical
- The proportion of all twins that are identical is roughly 1/3
- **Exercise:** For the woman in question, what is the probability that her twins are identical?

Meaning behind Bayes' rule

- Let's think about what happened
- Before the ultrasound, $P(\text{Identical}) = \frac{1}{3}$
- This is called the *prior* probability
- After we learned the results of the ultrasound, $P(\text{Identical}) = \frac{1}{2}$
- This is called the *posterior* probability

Bayesian statistics

- In fact, this prior/posterior way of thinking can be used to establish an entire statistical framework known as *Bayesian statistics*
- In this way of thinking, we start out with an idea of the possible values of some quantity θ (note that this is an example of a non-repeatable event)
- This distribution of possibilities $P(\theta)$ is our prior belief about the unknown; we then observe data D and update those beliefs, arriving at our posterior beliefs about the unknown, $P(\theta|D)$
- Mathematically, this updating uses Bayes' rule, hence the name for this line of inferential reasoning

Bayesian statistics (cont'd)

- As noted earlier, not everyone agrees with the notion of assigning prior probabilities to represent subjective beliefs
- At least for the present, the long-run frequency interpretation of probability has been adopted more widely and represents the most common approach to statistical inference
- Nevertheless, the Bayesian approach offers many advantages – in particular, it offers a natural representation of human thought and allows us to quantify probabilities about non-repeatable events – and has become more widespread in recent decades, although whether this trend will continue or not is anyone's guess
- We will focus primarily on "frequentist" statistics in this course, although we will return to Bayesian statistics again

Testing and screening

- A common application of Bayes' rule in biostatistics is in the area of diagnostic testing
- For example, older women in the United States are recommended to undergo routine X-rays of breast tissue (*mammograms*) to look for cancer
- Even though the vast majority of women will not have developed breast cancer in the year or two since their last mammogram, this routine screening is believed to save lives by catching the cancer while it is relatively treatable
- The application of a diagnostic test to asymptomatic individuals in the hopes of catching a disease in its early stages is called *screening*

Terms involved in screening

- Let D denote the event that an individual has the disease that we are screening for
- Let $+$ denote the event that their screening test is positive, and $-$ denote the event that the test comes back negative
- Ideally, both $P(+|D)$ and $P(-|D^C)$ would equal 1
- However, diagnostic tests are not perfect

Terms involved in screening (cont'd)

- Instead, there are always *false positives*, patients for whom the test comes back positive even though they do not have the disease
- Likewise, there are *false negatives*, patients for whom the test comes back negative even though they really do have the disease
- Suppose we test a person who truly does have the disease:
  - $P(+|D)$ is the probability that we will get the test right
  - This probability is called the *sensitivity* of the test
  - $P(-|D)$ is the probability that the test will be wrong (that it will produce a false negative)

Terms involved in screening (cont'd)

- Alternatively, suppose we test a person who does not have the disease:
  - $P(-|D^C)$ is the probability that we will get the test right
  - This probability is called the *specificity* of the test
  - $P(+|D^C)$ is the probability that the test will be wrong (that the test will produce a false positive)

Terms involved in screening (cont'd)

- The accuracy of a test is determined by these two factors:
  - Sensitivity: $P(+|D)$
  - Specificity: $P(-|D^C)$
- One final important term is the probability that a person has the disease, regardless of testing: $P(D)$
- This is called the *prevalence* of the disease

Values for mammography

- According to an article in *Cancer* (more about this later),
  - The sensitivity of a mammogram is 0.85
  - The specificity of a mammogram is 0.80
  - The prevalence of breast cancer is 0.003
- With these numbers, we can calculate what we really want to know: if a woman has a positive mammogram, what is the probability that she has breast cancer?

Using Bayes' rule for diagnostic testing

- Applying Bayes' rule to this problem,

$$\begin{aligned} P(D|+) &= \frac{P(D)P(+|D)}{P(D)P(+|D) + P(D^C)P(+|D^C)} \\ &= \frac{.003(.85)}{.003(.85) + (1-.003)(1-.8)} \\ &= 0.013 \end{aligned}$$

- In the terminology of Bayes' rule, the prior probability that a woman had breast cancer was 0.3%
- After the new piece of information (the positive mammogram), that probability jumps to 1.3%

Controversy (Part 1)

- So according to our calculations, for every 100 positive mammograms, only one represents an actual case of breast cancer
- Because $P(D|+)$ is so low, screening procedures like mammograms are controversial
- We are delivering scary news to 99 women who are free from breast cancer
- On the other hand, we may be saving that one other woman's life
- These are tough choices for public health organizations

Two studies of mammogram accuracy

- In our example, we calculated that the probability that a woman has breast cancer, given that she has a positive mammogram, is 1.3%
- The numbers we used (sensitivity, specificity, and prevalence) came from the article
  Hulka B (1988). Cancer screening: degrees of proof and practical application. *Cancer*, 62: 1776-1780.
- A more recent study is
  Carney P, et al. (2003). Individual and combined effects of age, breast density, and hormone replacement therapy use on the accuracy of screening mammography. *Annals of Internal Medicine*, 138: 168-175.

Comparing the two studies

	Hulka (1988)	Carney (2003)
Sensitivity	.85	.750
Specificity	.80	.923
Prevalence	.003	.005
$P(D\|+)$	1.3%	4.7%

- It would seem, then, that radiologists have gotten more conservative in calling a mammogram positive, and this has increased $P(D|+)$
- However, the main point remains the same: a woman with a positive mammogram is much more likely *not* to have breast cancer than to have it

Controversy (Part 2)

- Based on these kinds of calculations, in 2009 the US Preventive Services Task Force changed its recommendations:
  - It is no longer recommended for women under 50 to get routine mammograms
  - Women over 50 are recommended to get mammograms every other year, as opposed to every year
- Of course, not everyone agreed with this change, and much debate ensued (my Google search for USPSTF "breast cancer screening" controversy returned over 20,000 hits)

Summary

- Complement rule: $P(A^C) = 1 - P(A)$
- Addition rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Multiplication rule: $P(A \cap B) = P(A)P(B|A)$
- Law of total probability: $P(A) = P(A \cap B) + P(A \cap B^C)$
- Bayes' Rule:

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

- Important terms: mutually exclusive, independent, conditional probability, sensitivity, specificity, prevalence