Conditional probability

Intro to Data Science

Shawn Santo

Announcements

- Homework 03 is out today -- focus is probability
- Lab 05 on Tuesday -- group lab, encourage your entire group to attend

Today's agenda

- Conditional probability
- Independence
- Bayes' Rule

Conditional probability

The probability an event will occur *given* that another event has already occurred is a **conditional probability**. The conditional probability of event A given event B is:

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

Examples come up all the time in the real world:

- *Given* that it rained yesterday, what is the probability that it will rain today?
- *Given* that a mammogram comes back positive, what is the probability that a woman has breast cancer?
- *Given* the roulette wheel just landing on green, what is the probability it lands on green in the next spin?

Three probabilities

ORIGINAL RESEARCH Annals of Internal Medicine

Coffee Drinking and Mortality in 10 European Countries
A Multinational Cohort Study

	Did not die	Died
Does not drink coffee	5438	1039
Drinks coffee occasionally	29712	4440
Drinks coffee regularly	24934	3601

Define the events A = died and B = non-coffee drinker. Calculate the following probabilities for a randomly selected person in the cohort:

- **Marginal probability**: $P(A), P(B)$
- **Joint probability**: $P(A \cap B)$
- **Conditional probability**: $P(A|B), P(B|A)$

Independence

The multiplicative rule

We can write the definition of condition probability

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

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

What does the multiplicative rule mean in plain English?

Defining independence

Events A and B are said to be **independent** when

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

or

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

That is, when knowing that one event has occurred doesn't cause us to "adjust" the probability we assign to another event.

We can use the multiplicative rule to see that two events are said to be independent when the joint probability of two events exactly equals the marginal probability of their product:

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

Independent vs. disjoint events

Since for two independent events $P(A|B) = P(A)$ and $P(B|A) = P(B)$, knowing that one event has occurred tells us nothing more about the probability of the other occurring.

For two disjoint events A and B, knowing that one has occurred tells us that the other definitely has not occurred: $P(A \cap B) = 0$.

So, two events which are disjoint in general are **not** independent!

Checking independence

ORIGINAL RESEARCH

Annals of Internal Medicine

Coffee Drinking and Mortality in 10 European Countries
A Multinational Cohort Study

	Did not die	Died
Does not drink coffee	5438	1039
Drinks coffee occasionally	29712	4440
Drinks coffee regularly	24934	3601

Are dying and abstaining from coffee independent events? How might we check?

As you take more statistical science courses, you will learn the tools needed to formally assess whether these two events are independent!

Bayes' Rule

The law of total probability

Suppose we partition B into mutually exclusive events $B_1, B_2, \cdots, B_k$ that comprise the entirety of *the entire* sample space.

The **law of total probability** states that the probability of event A is

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + \cdots + P(A \cap B_k)$$

By applying the definition of conditional probability, we can obtain

$$P(A) = P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + \cdots + P(B_k)P(A|B_k)$$

What does this mean in plain English?

An example

In an introductory statistics course, 50% of students were first years, 30% were sophomores, and 20% were upperclassmen.

80% of the first years didn't get enough sleep, 40% of the sophomores didn't get enough sleep, and 10% of the upperclassmen didn't get enough sleep.

What is the probability that a randomly selected student in this class didn't get enough sleep?

Bayes' Rule

As we saw before, the two conditional probabilities $P(A|B)$ and $P(B|A)$ are not the same. But are they related in some way?

We can use **Bayes' rule** to "reverse" the order of condition. By definition, we have:

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{P(B|A)P(A)}{P(B)} \end{aligned}$$

Bayes' Rule (continued)

By using the rules of probability we've learned so far, we have

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{P(B|A)P(A)}{P(B)} \\ &= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} \end{aligned}$$

Note how we used the law of total probability in the denominator

Diagnostic testing

Definitions

Suppose we're interested in the performance of a diagnostic test. Let D be the event that a patient has the disease, and let T be the event that the test is positive for that disease.

- **Prevalence**: $P(D)$
- **Sensitivity**: $P(T|D)$
- **Specificity**: $P(T^c|D^c)$
- **Positive predictive value**: $P(D|T)$
- **Negative predictive value**: $P(D^c|T^c)$

What do these probabilities mean in plain English?

Rapid self-administered HIV tests

From the FDA package insert for the Oraquick ADVANCE Rapid HIV-1/2 Antibody Test,

- Sensitivity, $P(T|D)$, is 99.3%
- Specificity, $P(T^c|D^c)$, is 99.8%

From CDC statistics in 2016, 14.3/100,000 Americans aged 13 or older are HIV+.

Suppose a randomly selected American aged 13+ has a positive test result. What do you think is the probability they have HIV?

Using Bayes' Rule

$$
\begin{aligned}
P(D|T) &= \frac{P(D \cap T)}{P(T)} \\
&= \frac{P(T|D)P(D)}{P(T)} \\
&= \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|D^c)P(D^c)} \\
&= \frac{P(T|D)P(D)}{P(T|D)P(D) + (1 - P(T^c|D^c))(1 - P(D))} \\
&= \frac{sens. \times prev.}{sens. \times prev. + (1 - spec.) \times (1 - prev.)}
\end{aligned}
$$

Using Bayes' Rule

$$P(D|T) = \frac{sens. \times prev.}{sens. \times prev. + (1 - spec.) \times (1 - prev.)}$$

```
sens <- 0.993; spec <- 0.998; prev <- 14.3/100000
prob <- (sens * prev) / ( (sens * prev) + ((1 - spec) * (1 - prev)) )

prob
```

```
#> [1] 0.0663016
```

A discussion

Think about the following questions:

- Is this calculation surprising?
- What is the explanation?
- Was this calculation actually reasonable to perform?
- What if we tested in a different population, such as high-risk individuals?
- The prevalence of HIV in Botswana is approximately 25%. What if we were to test a random individual in Botswana?

Getting some more practice

- Create your personal private repository by clicking https://classroom.github.com/a/9SDlJKyr
- Follow the steps as we have done previously to clone this and create a new RStudio project in the RStudio Docker containers.