Math for Data Science

Conditional Probability

Joanna Bieri DATA100

Important Information

- Email: joanna_bieri@redlands.edu
- Office Hours take place in Duke 209 unless otherwise noted – Office Hours Schedule

Today's Goals:

- Understand Conditional Probability
- Bayes' Theorem
- Join and Union Conditional Probabilities

(Review) Probability

We express the idea of probability as

$$P(X)$$

where X is the event of interest.

(Review) Probability

Joint Probability of independent events

$$P(A \; and \; B) = P(A) \times P(B)$$

This is also called the **product rule of probability**.

(Review) Probability

Union Probability of independent events.

$$P(A \; or \; B) = P(A) + P(B) - P(A \; and \; B)$$

This is also called **the sum rule of probability**.

Conditional Probabilities

To extend our understanding of probability to the next level we need to understand how to combine probabilities of events that are not assumed independent. *How do I find the probability of event A given that event B occurred?*

This is called a **conditional probability** and the notation we use is:

$$P(A \; given \; B) = P(A|B)$$

Conditional Probabilities

Independent events

If two events are independent then one event does not effect the other.

For example:

What is the probability that it rains in downtown LA this month given that I flipped heads? Unless you are really superstitious about coin flipping you will assume these two events are independent.

Conditional Probabilities

Using the notation from conditional probabilities I would say:

$$P(Rain|Heads) = P(Rain)$$

In other words, the probability of raining does not change given that you flipped heads.

Conditional Probabilities

I can also say

$$P(Heads|Rain) = P(Heads)$$

Does this make sense? What does this mean in words?

Conditional Probabilities

Dependent events

The book has a really nice example:

Say you read a study that says 85% of cancer patients also drank coffee and you want to know how to react. Do I immediately go cold turkey off my morning brew? What really is my risk? What do I really want to know? I want to know the probability that I get cancer given that I drink coffee!

Conditional Probabilities

First there are some other numbers that we might want to know!

- The percentage of people diagnosed with cancer is 0.5% according to cancer.gov.
$$P(Cancer) = .005$$
- The percentage of people who drink coffee is 65% based according to statista.com.
$$P(Coffee) = .65$$

And we want to be able to look at these probabilities and ask the question: Is coffee really the problem? 65% of the population drinks coffee and only 0.5% has cancer at a given time.

Conditional Probabilities

Sometimes it can be a challenge to unpack proportional numbers!! The media often reports numbers/study results out of context. Here is a nice article about this problem:

https://www.fredhutch.org/en/news/center-news/2020/02/spinning-science-overhyped-headlines-snarled-statistics-lead-readers-astray.html

Conditional Probabilities

What would be wrong about an article titled: "New Study Reveals that 85% of Cancer Patients Drink Coffee!"

- Well that number, 85%, seem alarming.
- It is taking a common attribute (drinking coffee) and relating it to a uncommon one (having cancer).
- It is taking advantage of the direction of the condition!
- Our brain said "Holy Crap! I have an 85% chance of getting cancer if I drink coffee!" - but the condition is not actually in that direction.

Conditional Probabilities

What did the study result actually tell me?

- The probability that I drink coffee given that I have cancer is 85%.

$$P(Coffee|Cancer) = .85$$

What does the study result NOT tell me?

- The probability that I have cancer given that I drink coffee.

$$P(Cancer|Coffee) = ?$$

Bayes' Theorem to the rescue!

There is a powerful formula that will allow us to flip these conditional probabilities.

Bayes' Theorem

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

I can use the ratio of the marginal probabilities to change the order of the conditional probability!

Bayes' Theorem

Applying this to our example:

$$P(Cancer|Coffee)$$
$$= P(Coffee|Cancer) * P(Cancer)/P(Coffee)$$

$$= 0.85 * 0.005/0.65 = .0065$$

Bayes' Theorem

```
# Using python to do this calculation:
Pcancer = .005
Pcoffee = .65
Pcoffee_cancer = .85

Pcancer_coffee = Pcoffee_cancer*Pcancer/Pcoffee
print(Pcancer_coffee)
```

```
0.006538461538461539
```

Bayes' Theorem

$$P(Cancer|Coffee) = 0.0065$$

So we could reassess our original fear at reading the news article and realize that the probability that I have cancer given that I drink coffee is 0.65%. This is a much less scary number.

We should be careful here!! There are a lot of other, confounding, variables that could influence these probabilities AND just because to things increase or decrease together (correlate) does not mean that they cause each other!

Joint and Union Probabilities - dependent events

What if now we wanted to ask something like "what is the probability that someone is a coffee drinker AND has cancer"? In other words, what is the probability that someone is simultaneously a coffee drinker and has cancer?

How should I calculate this?

Joint and Union Probabilities - dependent events

Well there are two ways I might try:

- Just use the original formula even though we said it was for independent events:

$$P(Coffee) \times P(Cancer) = .65 \times 0.005 = 0.00325$$

- Use some information about the conditionality between the two:

$$P(Coffee|Cancer) \times P(Cancer)$$
$$= .85 \times 0.005 = 0.00425$$

Joint and Union Probabilities - dependent events

To decide which one makes sense we need to think about what we are calculating, we want to know if someone simultaneously has cancer AND drinks coffee.

- In the first example we use the probability of coffee drinkers (some with cancer and some without) and the probability of having cancer. This seems like mixing the two populations together.

Joint and Union Probabilities - dependent events

- In the second example we use the probability of people who drink coffee given that they have cancer and the group of people who have cancer. This seems more intuitive to get at what we want what is the probability that you simultaneously below to both the coffee drinking cancer group and the cancer group.

Joint and Union Probabilities - dependent events

Update the Joint Probability Formula

$$P(A \; and \; B) = P(B) \times P(A|B)$$

and this is actually a general formula because remember if two events are independent then $P(A|B) = P(A)$.

Joint and Union Probabilities - dependent events

Update the Union Probability Formula

With our new definition of joint probability we can update the union probability formula:

$$P(A \; or \; B) = P(A) + P(B) - P(B) \times P(A|B)$$

Calculating a Conditional Probability from data

When working with data we can calculate conditional probabilities using the formula

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

Does this formula make sense?

Calculating a Conditional Probability from data

- First we find the probability of both events A and B happening (in the data)
- Then we divide out the probability that the event B happened (in the data)
- If the two events are independent then this would just equal the number of times event A happened. There is no event B mingling with my probability of event A
- If the two events are dependent then this will divide out the proportion of the probability of B that does not effect A and leave the overlap.

Calculating a Conditional Probability from data

Let's look at some data.

Imagine you conduct a survey of 100 men and women that asks about their eating habits and you get the following results:

Example from: https://365datascience.com/tutorials/statistics-tutorials/conditional-probability/

Calculating a Conditional Probability from data

	Vegetarian	Not Vegetarian	Total
Women	15	32	47
Men	29	24	53
Total	44	56	100

Example 1

What is the probability of a woman being a vegetarian? AKA what is the probability that of being a vegetarian given that you are a woman?

$$P(Vegetarian|Woman)$$

- Purely looking at the data we would need to count the number of times women said they were vegetarian and divide out the total number of women:

$$\frac{15}{47}$$

Example 1

- Thinking about the formula

$$P(Vegetarian|Woman) = \frac{P(Vegetarian\ and\ Woman)}{P(woman)}$$

- so hopefully the formula makes intuitive sense looking at the data.

Example 1

```
# In python - numerical calculation

P_V_given_W = DF['Vegetarian']['Women']/DF['Total']['Women']
P_V_given_W

# In python - symbolic calculation
P_V_given_W = sp.Rational(DF['Vegetarian']['Women'],DF['Total'
P_V_given_W
```

```
0.3191489361702128
```

$$\frac{15}{47}$$

Example 2

What is the probability of a vegetarian being a woman? AKA what is the probability that you are a woman given that you are a vegetarian?

$$P(Woman|Vegetarian)$$

Well now we can do this two ways! We can do a direct calculation OR we could use Bayes' Theorem!

Example 2

```
# In python - direct symbolic calculation
P_W_given_V = sp.Rational(DF['Vegetarian']['Women'],DF['Vegeta
P_W_given_V

# Using Bayes' Theorem
# P(W|V) = P(V|W)*P(W)/P(V)

P_W_given_V = P_V_given_W * sp.Rational(DF['Total']['Women'],D
P_W_given_V
```

$$\frac{15}{44}$$

$$\frac{15}{44}$$

Example 2

We can check the sum of the probabilities - it should equal one.

```
P_M_given_V = sp.Rational(DF['Vegetarian']['Men'],DF['Vegetari
P_M_given_V+P_W_given_V
```

1

You Try

1. What is the probability that a man is a vegetarian? Write down the symbolic notation.
2. What is the probability that a vegetarian is a man? Write down the symbolic notation.
3. What happens if you sum the probability that a vegetarian is a man and the probability that a vegetarian is a woman?

$$P(V|M) + P(V|W)$$

4. Does this make sense?

The Law of Total Probability

Given a set of all possible events, the total probability of something happening is just the sum of the joint probabilities of each of those events.

Lets say we have events $E_1, E_2, E_3, ...$ and the union of these events is A then the probability of A is given by

$$P(A) = P(A|E_1) \times P(E_1) + P(A|E_2) \times P(E_2) + ...$$

The Law of Total Probability

Lets unpack this using our example!

In our example as a Vegetarian you could have been from one of two possible event groups: male or female. So the total probability that you are a Vegetarian can be given by

$$P(Vegetarian) = P(Vegetarian\ and\ Male)$$
$$+P(Vegetarian\ and\ Female)$$

The Law of Total Probability

But because we know the joint probability formula we can calculate this as

$$\begin{gathered} P(Vegetarian) \\ = P(Vegetarian|Male) \times P(Male) \\ + \\ P(Vegetarian|Female) \times P(Female) \end{gathered}$$

and the numbers from our example give:

$$P(Vegetarian) = \frac{29}{53} \times \frac{53}{100} + \frac{15}{47} \times \frac{47}{100} = 0.44$$

or 44 out of 100 people were vegetarian. Based on our survey there is 44% of someone being a vegetarian.

You Try

Here is some data for you to practice finding conditional probabilities:

You can see the solutions in the lecture notes.

	Outdoor	Indoor	Total
Sunny	15	4	19
Rainy	5	16	21
Cloudy	6	8	14
Total	26	28	54

You Try

You can read the data from the table:

1. What is the probability that it is Sunny, Rainy, Cloudy?
2. What is the probability that you went outdoors or vs stayed indoors?

You try

Do the direct computation using the data:

1. If it was raining, what is the probability that you stayed indoors?
2. If it was sunny, what is the probability that you went outdoors?
3. What is the probability that it was rainy if you stayed indoors?
4. What is the probability that it was sunny if you stayed outdoors?

You try

Now use Bayes' theorem to calculate 3 and 4 from the problem above using the conditional probabilities you found in 1 and 2, respectively. Your answers should match.