Week 5: Probability Review

BIOS 20172

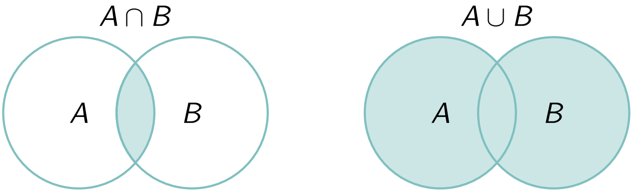
Tiffanie Huang

April 24, 2023

Table of contents

Quick notation+probability review

Notation	Meaning
$P(A \cup B)$	probability of Event A **or** Event B occurring
$P(A \cap B)$	probability of Event A **and** Event B occurring
$P(A \mid B)$	probability of Event A occurring **given** Event B



Quick notation+probability review (cont.)

A couple formulas+interpretations, given Event A and Event B:

- **Additive Law**: probability of Event A **or** Event B occurring

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

  - **if events are* **mutually exclusive**: $P(A \cup B) = P(A) + P(B)$

- **Multiplicative Law**: probability of Event A **and** Event B both (jointly) occurring

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

  - **if events are* **independent**: $P(A \cap B) = P(A)P(B)$

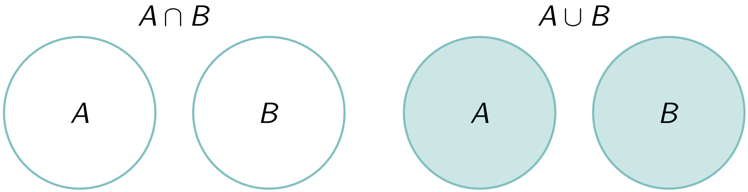
- **Conditional probability**: probability of Event A occurring **given** Event B (rewriting the Multiplicative Law)

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

Quick notation+probability review (cont.)

What if Event A and Event B are mutually exclusive? From the Venn Diagrams of the disjoint sets below you can see:

- why $P(A \cap B) = 0$
  - $(A \cap B) = \varnothing$
- why from the Additive Law, we get $P(A \cup B) = P(A) + P(B)$ (for mutually exclusive events)



Example 1. Evaluating a COVID-19 rapid test

A group of researchers are trying to evaluate a certain brand of COVID-19 rapid test in the US. They know that around 24% of people in the US[1] have COVID. Someone with COVID has a 12.5% chance that their test result is negative. Someone without COVID has an 7.9% chance that their test result is positive.

- What is the prevalence of COVID-19?
- What is the chance that someone with COVID will test positive?
- If someone tests negative, what is the chance that they actually do not have COVID? What is the chance that they do?
- If the percentage of the population with COVID increases, what happens to the positive predictive value, negative predictive value, sensitivity, and specificity? Do they increase, decrease, or remain the same?

[1] I made up these numbers.

Example 1. Evaluating a COVID-19 rapid test

- What is the prevalence of COVID-19?
- What is the chance that someone with COVID will test positive?
- If someone tests negative, what is the chance that they actually do not have COVID? What is the chance that they do?
- If the percentage of the population with COVID increases, what happens to the positive predictive value, negative predictive value, sensitivity, and specificity? Do they increase, decrease, or remain the same?

Prevalence

Prevalence in medicine: a measure of the total number of people in a specific group who have (or had) a certain disease, condition, or risk factor.[2]

Our prevalence in this case is 24%. To rewrite it using probability notation:

$$P(COVID) = 0.24$$

[2] Taken from NCI.

COVID or no COVID, negative or positive?

We have two components to look at:

1. Test results: Is it positive or negative?
2. Actual values: Does the person have COVID-19 or not?

Then we combine them:

- True Positive: test is positive, person does have COVID
- True Negative: test is negative, person does not have COVID
- False Positive: test is positive, person does not have COVID
- False Negative: test is negative, person does have COVID

Sensitivity and specificity

From here, we can find these values using T/F $(+)/(-)$ and also rewrite them in probability notation:

- **sensitivity**: out of all the people w/ COVID, which ended up testing positive? aka true positive rate

$$P(+ \mid COVID) = \frac{TP}{TP + FN}$$

- **specificity**: out of all the people w/o COVID, which ended up testing negative? aka true negative rate

$$P(- \mid no\ COVID) = \frac{TN}{TN + FP}$$

PPV and NPV

- **positive predictive value** (PPV): out of all the people who tested positive, who has COVID?

$$P(COVID \mid +) = \frac{TP}{TP + FP}$$

- **negative predictive value** (NPV): out of all the people who tested negative, who does not have COVID?

$$P(no\ COVID \mid -) = \frac{TN}{TN + FN}$$

Be very careful about conditional probability! You learned that the sum of the probability of each distinct outcome in a sample space is 1. Why won't sensitivity and specificity add up to 1? Why won't PPV and NPV add up to 1?

Example 1. Evaluating a COVID-19 rapid test

- What is the prevalence of COVID-19?
- What is the chance that someone with COVID will test positive?
- If someone tests negative, what is the chance that they actually do not have COVID? What is the chance that they do?
- If the percentage of the population with COVID increases, what happens to the positive predictive value, negative predictive value, sensitivity, and specificity? Do they increase, decrease, or remain the same?

What is the chance that someone with COVID will test positive?

- Rewrite in terms of probability:

$$P(+ \mid COVID)$$

  - Recall the term for this probability: **sensitivity**
- We know that if someone has COVID, there is a 12.5% chance that their test result is negative.

$$P(- \mid COVID) = 0.125$$

- So we can find sensitivity by subtracting that from 1:

$$\begin{aligned} P(+ \mid COVID) &= 1 - P(- \mid COVID) \\ &= 1 - 0.125 \\ &= 0.875 \end{aligned}$$

Example 1. Evaluating a COVID-19 rapid test

- What is the prevalence of COVID-19?
- What is the chance that someone with COVID will test positive?
- If someone tests negative, what is the chance that they actually do not have COVID? What is the chance that they do?
- If the percentage of the population with COVID increases, what happens to the positive predictive value, negative predictive value, sensitivity, and specificity? Do they increase, decrease, or remain the same?

Bayes' Theorem

The formula for Bayes' Theorem to find the probability of Event A given Event B:

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

But where does this come from?

Bayes' Theorem and the multiplicative law

The multiplicative law:

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

Now rearrange for A conditional on B:

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

If someone tests negative, what is the chance that they actually do not have COVID? What is the chance that they do?

- Rewrite in terms of probability:

$$P(\textit{no COVID} \mid -)$$

$$P(\textit{COVID} \mid -)$$

  - Recall the term for the first probability: **NPV**
- Let's start with finding $P(\textit{no COVID} \mid -)$ and use Bayes' Theorem.

$$P(\textit{no COVID} \mid -) = \frac{P(\textit{no COVID})P(- \mid \textit{no COVID})}{P(-)}$$

Written out:

$$P(\textit{no COVID} \mid -) = \frac{P(\textit{no COVID})P(- \mid \textit{no COVID})}{P(-)}$$

- what we need: $P(\textit{no COVID})$, $P(- \mid \textit{no COVID})$, and $P(-)$
- someone w/o COVID has a 7.9% chance that their test result is positive: $P(+ \mid \textit{no COVID}) = 0.079$, and $P(- \mid \textit{no COVID}) = 1 - P(+ \mid \textit{no COVID})$, so $P(- \mid \textit{no COVID}) = 0.921$ (**specificity**)
- from the prevalence we can find $P(\textit{no COVID}) = 1 - P(\textit{COVID}) = 1 - 0.24 = 0.76$
- *how do we find $P(-)$ if we don't know $P(-)$ or $P(+)$?

Law of Total Probability

$$P(A) = \sum_n P(A \cap B_n)$$

You can also write it as:

$$P(A) = \sum_n P(A \mid B_n) P(B_n)$$

All this means is that you can find the probability of Event A by taking the sum of all conditional probabilities of Event A given each Event B_n.

Here, Event A: test $(-)$; Events B_n: COVID, no COVID:

$$P(-) = P(- \mid COVID)P(COVID) + P(- \mid no\ COVID)P(no\ COVID)$$

Aside from our rewritten $P(-)$, we know that:

- $P(COVID) = 0.24$ and $P(no\ COVID) = 0.76$
- $P(- \mid COVID) = 0.125$ and $P(+ \mid COVID) = 0.875$
- $P(- \mid no\ COVID) = 0.921$

Plug these values into your Bayes' Theorem equation!

$$
\begin{aligned}
P(no\ COVID \mid -) &= \frac{P(no\ COVID)P(- \mid no\ COVID)}{P(-)} \\
&= \frac{P(no\ COVID)P(-|no\ COVID)}{P(-|COVID)P(COVID)+P(-|no\ COVID)P(no\ COVID)} \\
&= \frac{0.76 \cdot 0.921}{(0.125 \cdot 0.24) + (0.921 \cdot 0.76)} \\
&= 0.959
\end{aligned}
$$

The probability that someone won't have COVID given a negative test is approximately 0.959. Then the probability that someone has COVID given a negative test is $1 - 0.959 = 0.041$.

Example 1. Evaluating a COVID-19 rapid test

- What is the prevalence of COVID-19?
- What is the chance that someone with COVID will test positive?
- If someone tests negative, what is the chance that they actually do not have COVID? What is the chance that they do?
- If the percentage of the population with COVID increases, what happens to the positive predictive value, negative predictive value, sensitivity, and specificity? Do they increase, decrease, or remain the same?

What happens when prevalence increases?

- Recall that our prevalence $= P(COVID)$.
- An increase in prevalence means an increase in the proportion of COVID cases. Think back to the earlier slides w/ the equations:
- What happens to NPV and PPV?
  - PPV increases – out of all the positive tests, more are going to be true positive/more will actually have COVID
  - NPV decreases – out of all the negative tests, less are going to be true negative/less will actually not have COVID

What happens when prevalence increases?

- What happens to sensitivity and specificity?
  - sensitivity will stay the same – out of all those w/ COVID, whether or not they test positive doesn't have to do with prevalence (it has to do with $P(+)$)
  - specificity will stay the same for the same reason – out of all those w/o COVID, whether or not they test negative is unaffected by prevalence
- What's the difference?
  - given a $(+)$ or $(-)$ test, PPV and NPV evaluate the proportion of people w/ or w/o COVID – which is affected by prevalence
  - given COVID or no COVID, sensitivity+specificity evaluate the proportion of $(+)$ or $(-)$ tests – which is unaffected by prevalence!

Visualizing this with a table

Given any info on the number of individuals, we can use a table showing the count for each category. The true values lie along the diagonal in my table.

COVID Test Outcome

Actual Value		**(−)**	**(+)**
	no COVID	True Negatives	False Positives
	COVID	False Negatives	True Positives

Example 1. A 2x2 table

If the total number of people is 100, the table would look like this:

COVID Test Outcome

		$(-)$	$(+)$	**total**
Actual Value	**no COVID**	TN: 70	FP: 6	76
	COVID	FN: 3	TP: 21	24
	total	73	27	100

Extra practice problems...with solutions!

My answers (with work) will be at the end of the slides.

1. 5.25 from your textbook
2. Can A and B be mutually exclusive if $P(A) = 0.4$ and $P(B) = 0.7$? If $P(A) = 0.4$ and $P(B) = 0.3$? Why?
3. Cards are dealt, one at a time, from a standard 52-card deck.
   a. If the first 2 cards are both spades, what is the probability that the next 3 cards are also spades?
   b. If the first 3 cards are all spades, what is the probability that the next 2 cards are also spades?
   c. If the first 4 cards are all spades, what is the probability that the next card is also a spade?

Extra practice problems (cont.)

4. A football team has a probability of 0.75 of winning when playing any of the other four teams in its conference. If the games are independent, what is the probability the team wins all its conference games?
5. Consider the following events in the toss of a single die:
   A. observe an odd number
   B. observe an even number
   C. observe a 1 or 2

   a. Are A and B independent events?
   b. Are A and C independent events?
6. Of the items produced daily by a factory, 40% come from line I and 60% from line II. Line I has a defect rate of 8%, whereas line II has a defect rate of 10%. If an item is chosen at random from the day's production, find the probability that it will not be defective.

Extra practice problems (cont.)

Bonus question:

7. A diagnostic test for a disease is such that it (correctly) detects the disease in 90% of the individuals who actually have the disease. Also, if a person does not have the disease, the test will report that he or she does not have it with probability 0.9. Only 1% of the population has the disease in question. If a person is chosen at random from the population and the diagnostic test indicates that they have the disease, what is the conditional probability that they do, in fact, have the disease?

Practice problem solutions

1. $2\left(\frac{1}{8} \times \frac{1}{7} \times \frac{1}{6} \times \frac{1}{5} \times \frac{1}{4} \times \frac{1}{3} \times \frac{1}{2} \times 1\right) = 4.96 \times 10^{-5}$
2. No for the first; yes for the second. For the first: $P(A) + P(B) = 1.1$. Since $1.1 > 1$, A and B aren't mutually exclusive for the first question. For the second: $P(A) + P(B) = 0.7$, which is not greater than 1. It's possible for them to be mutually exclusive.
3.
   a. $\frac{11}{50} \times \frac{10}{49} \times \frac{9}{48} = 0.0084$
   b. $\frac{10}{49} \times \frac{9}{48} = 0.0383$
   c. $\frac{9}{48} = 0.1875$
4. $0.75 \times 0.75 \times 0.75 \times 0.75 = 0.75^4 = 0.3164$

5. a. No. These are mutually exclusive (and both have a probability greater than 0), meaning they cannot be independent. A occurring informs you about the probability of B occurring, and vice versa. $P(A)P(B \mid A) \neq P(A)P(B)$ since $P(B \mid A) \neq P(B)$ and $P(A \mid B) \neq P(A)$. Either explanation works:

$$\begin{aligned} P(B \mid A) &\stackrel{?}{=} P(B) \\ P(B \mid A) &= 0 \\ P(B) &= \frac{1}{2} \\ P(B) &\neq P(B \mid A) \end{aligned} \qquad \begin{aligned} P(A \mid B) &\stackrel{?}{=} P(A) \\ P(A \mid B) &= 0 \\ P(A) &= \frac{1}{2} \\ P(A) &\neq P(A \mid B) \end{aligned}$$

b. Yes. $P(A \mid C) = P(A)$ and $P(C \mid A) = P(C)$. Show either:

$$\begin{aligned} P(C \mid A) &= \frac{1}{3} \\ P(C) &= \frac{1}{3} \\ P(C) &= P(C \mid A) \end{aligned} \qquad \begin{aligned} P(A \mid C) &= \frac{1}{2} \\ P(A) &= \frac{1}{2} \\ P(A) &= P(A \mid C) \end{aligned}$$

6. Here I used the law of total probability.

$$
\begin{aligned}
P(\textit{defective}) &= P(\textit{defective} \mid I)P(I) + P(\textit{defective} \mid II)P(II) \\
&= (0.08 \times 0.4) + (0.1 \times 0.6) \\
&= 0.092 \\
P(\textit{not defective}) &= 1 - P(\textit{defective}) \\
&= 1 - 0.092 \\
&= 0.908
\end{aligned}
$$

7. We want to find PPV, written as $P(disease \mid +)$. I did a combination of Bayes' theorem and the law of total probability.

$$\begin{aligned}
P(+ \mid disease) &= 0.9 \\
P(- \mid no\ disease) &= 0.9 \\
P(disease) &= 0.01 \\
P(+) &= P(+|disease)P(disease) + P(+|no\ disease)P(no\ disease) \\
&= 0.9 \times 0.01 + (1 - 0.9) \times (1 - 0.01) \\
&= 0.108 \\
P(disease \mid +) &= \frac{P(+ \mid disease)P(disease)}{P(+)} \\
&= \frac{0.9 \times 0.01}{0.108} \\
&= 0.083
\end{aligned}$$

References

1. Wackerly, D. B. and Mendenhall, W. (2008). Mathematical statistics with applications.
2. Whitlock, C. M. and Schluter, D (2009). The Analysis of Biological Data.