Conditional Probability and Bayes' Theorem
XDASI Fall 2021

9/23/2021

Contents

References

- ***Whitlock & Schluter, Chapter 5: Probability***
- ***Aho, Foundational and Applied Statistics for Biologists with R***
    - **Chapter 2: Introduction to Probability**

Conditional probability

When variables are associated, we need additional tools to model **dependencies** between variables.

The **conditional probability** is the probability that event A occurs, given a prior ***condition*** of event B. We can describe this situation as the probability of "A conditioned on B", "A given the prior probability of B", or simply "A given B". We write "given as" using a vertical bar, so the conditional probability is: $Pr[A|B]$ or $P(A|B)$.

Conditional probability is particularly useful when events ar not independent.

Whitlock, Example 5.8

In many species, the mother can alter the relative numbers of male and female offspring depending on the local environment. In this case the "sex of the offspring" and the "environment" are dependent on each other.

The *Nasonia vitripennis* is a parasite that lays eggs on the pupae of flies. If the host has not been already parasitized, than it lays more female eggs, where as if the host has already been parasitized than it lays more male eggs. Therefore the sex of the offspring depends on the status of the pupae.

We can visualize this using a Venn diagram:

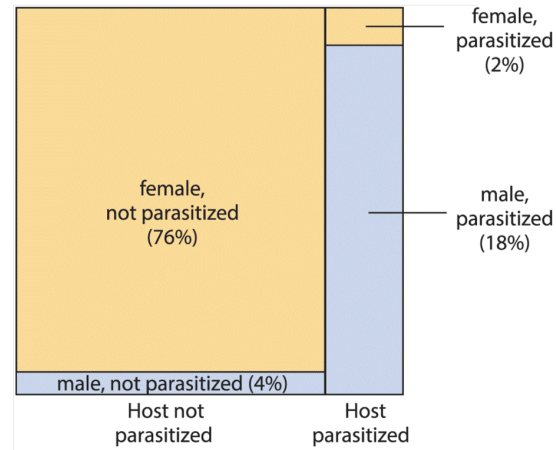We can also use a **decision tree** to lay out this problem visually:

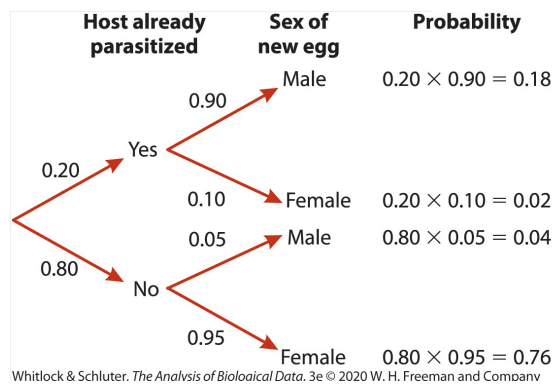Figure 1: Figure 5.8-1: Venn diagram

Figure 2: Figure 5.8-3: Decision Tree

Here we were given a number of **observations**:

The total proportion of the hosts that were parasitized:

- $Pr[par] = P(par) = 0.20$
- $Pr[!\ par] = P(!\ par) = 0.80$

The proportion of male and female eggs found on non-parasitized or parasitized hosts:

- $P(M \cap !\ par) = 0.04$
- $P(F \cap !\ par) = 0.76$
- $P(M \cap par) = 0.18$
- $P(F \cap par) = 0.02$

This allowed us to find the conditional probabilities:

- $P(M|par) = 0.18/0.20 = 0.90$
  - probability that a male egg is laid, **given that** host is parasitized
- $P(F|par) = 0.02/0.2 = 0.10$
  - same as above, but for a female egg
- $P(M|!\ par) = 0.04/0.8 = 0.05$
  - probability that a male egg is laid, **given that** the host is unparasitized.
- $P(F|!\ par) = 0.76/0.8 = 0.95$
  - ditto, for a female egg

The book illustrates finding the probability that a new, randomly chosen egg, is **_male_** by tracing out both possible paths to obtain a male egg, and using the multiplication and addition rules for probabilities to get the **_total probability_** of a male egg:

$$P(M) = P(par) * P(M|par) + P(!\ par) * P(M|!\ par)$$
$$= 0.20 * 0.90 + 0.80 * 0.05 = 0.18 + 0.04$$
$$= 0.22$$

$\Rightarrow$ **_Q: What is the probability that a randomly chosen egg is female?_**

Let's work out the pieces we need and then compute the answer using the given conditonal probabilities.

Answer

$$P(F) = P(par) * P(F|par) + P(!\ par) * P(F|!\ par)$$
$$= 0.20 * 0.10 + 0.80 * 0.95 = 0.02 + 0.76$$
$$= 0.78$$

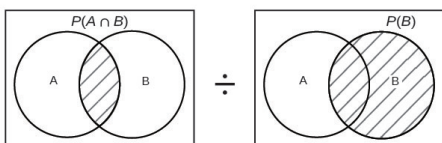How does this line up with the figure we got for male eggs above?

Of course, we could have just computed this using the actual observations:

$$P(F) = P(F \cap !\ par) + P(F \cap par) = 0.76 + 0.02 = 0.78$$

So how does going to all this trouble help us? Because by turning these equations around, we can use them to answer new questions that we don't already know the answer to. Let's see how this could work.

General multiplication rule

We can determine the probability of an event ***A given B*** using the conditional probability, which is simply the fraction of the time that we see A and B together, relative to the total occurrence of B:

$$P(A|B) \ = \ \boxed{\begin{array}{c} P(A \cap B) \\ A \quad B \end{array}} \ \div \ \boxed{\begin{array}{c} P(B) \\ A \quad B \end{array}}$$

We can write this as:

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

If we rearrange this equation (and change our notation slightly), we can say that $P(A \cap B)$ is simply the probability of ***A given B*** times the **total probability of B**:

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

Of course ***the reverse is true as well***:

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

However, keep in mind that usually $P(A|B) \neq P(B|A)$. Why? Consider the relationship between spots and measles!

We actually used this multiplication rule to find the conditional probabilities above. For example:

$$P(M|par) = \frac{P(M \cap par)}{P(par)} = \frac{0.18}{0.20} = 0.95$$

When two events are independent, we know that $P(A \cap B) = P(A) * P(B)$, and so $P(A|B)$ is simply $P(B)$:

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

Law of total probability

When we have multiple possible conditions, or (mutually exclusive) values of B, we can say that B is a **disjoint set** representing multiple ***states of nature***.

The total probability of A will then be **partitioned** among all of the possible values of B. We can visualize this as follows:
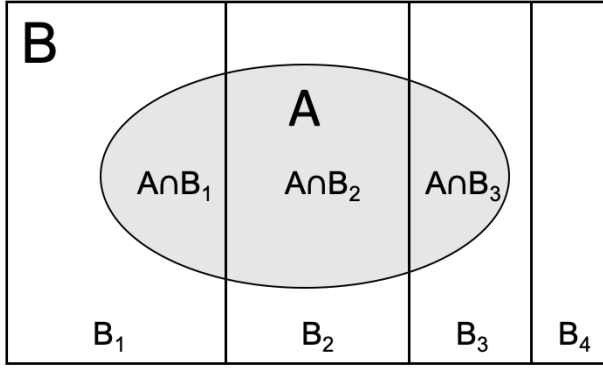
So the total probability of event A, $P(A)$, is the **sum of the intersections of A and each of the partitions of B**. We can write this as:

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + P(A \cap B_3) + P(A \cap B_4) = \sum_{i=1}^{4} P(A \cap B_i)$$

Since each of the intersections can be written as $P(A \cap B_i) = P(A|B_i) * P(B_i)$, we can now write the **law of total probability** in terms of the ***conditional probabilities***:

$$P(A) = \sum_{i=1}^{4} P(A \cap B_i) = \sum_{i=1}^{n} P(A|B_i) * P(B_i)$$

Where n is the total number of ***conditions***, i.e. mutually exclusive states of B. That is, the total probability of A is the conditional probability of (A given B) times the probability of B, for all possible values of B.

For our *Nasonia* example above, this equation gives us the same result as the one we obtained by working our way through the decision tree:

$$P(M) = P(M|par) * P(par) + P(M|!\,par) * P(!\,par)$$

Bayes' theorem

Recall that the intersection of A and B, $P(A \cap B)$, can be written equivalently in terms of conditional probabilities relative to A or relative to B:

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

We can now derive an incredibly useful formula called the **Bayes' Theorem**:

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

The three elements of Bayesian statistics are thus:

- **Priors**: known probabilities for different conditions, e.g. $P(A)$ and $P(B)$
- **Likelihoods**: known conditional probabilities, e.g. $P(B|A)$
- **Posteriors**: unknown conditional probabilities, e.g. $P(A|B)$

This turns out to be a very useful way to frame problems involving conditional probabilities, because it allows us to figure out an unknown quantities for **posterior probabilities**, given **prior probabilities** and **likelihoods**.

Bayes' theorem with multiple priors

More generally, when there are many possible outcomes, we can figure out the conditional probability for a single outcome using the **total law of probability**:

$$P(A_i|B) = \frac{Pr(B|A_i)P(A_i)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{n} P(B|A_j)P(A_j)}$$

Bayes' theorem with binary outcomes

For a situation with binary outcomes, this simplifies to:

$$P(A|B) = \frac{Pr(B|A) * P(A)}{P(B)} = \frac{P(B|A) * P(A)}{P(B|A) * P(A) + P(B|not\ A) * P(not\ A)}$$

Going back to our *Nasonia* example above, we can use this formula to answer a new question like,

⇒ ***"What is the probability that the host has already been parasitized, given that a male egg was laid?"***

Let's work this out together before checking the answer.

Answer

$$\begin{aligned} P(par|M) &= \frac{P(M|par) * P(par)}{P(M)} \\ &= \frac{P(M|par) * P(par)}{P(M|par) * P(par)\ +\ P(M|!\ par) * P(!\ par)} \\ &= \frac{0.90 * 0.20}{0.90 * 0.20\ +\ 0.05 * 0.80} = \frac{0.18}{0.18 + 0.04} \\ &= \frac{0.18}{0.22} = 0.82 \end{aligned}$$

Let's get a better feel for how these ideas apply by working through one more example together.

Example 5.9 - Down's Syndrome

Down's syndrome, or trisomy 21, is a chromosomal condition that occurs in 1/1000 pregnancies. A test called the "triple test", which looks for three hormone concentrations around week 16, has been developed to circumvent the need for amniocentesis, which is a very invasive procedure (but which is also the gold standard for detection, since it provides a definitive diagnosis of the karyotype).

Let's say we want to ask:

If the test on a random fetus gives a *Pos* ***result, what is the probability that fetus actually is*** DS^+***?*** In other words, what is $P(DS^+|Pos)$?

How can we formulate this problem in terms of Bayes' rule?

First, we need the **priors**, or states of nature:

- $P(DS^+) = 0.001$: Rate of Down's syndrome
- $P(DS^-) = 0.999$

The established rates of detection of the test, which give us our **likelihoods**, are as follows:

- $P(Pos|DS^+) = 0.6$
    - **True Positive** (sensitivity): The test was positive, and the fetus has DS
- $P(Neg|DS^+) = 0.4$
    - **False Negative** : The test was negative, however the fetus has DS
- $P(Neg|DS^-) = 0.95$
    - **True Negative** : The test was negative, and the fetus does not have DS
- $P(Pos|DS^-) = 0.05$
    - **False Positive** : Test was positive, but the fetus does not have DS

Using the Bayes' theorem above, let's set A to be the test result, $[Pos|Neg]$, and B to be DS status, $[DS^+|DS^-]$. Now we can formulate our problem as:

$$P(DS^+|Pos) = \frac{P(Pos|DS^+) * P(DS^+)}{P(Pos)}$$

We are given the values for the numerator, but we need the values for the denominator. For this we can use our **law of total probability**.

⇒ ***Q: What is the formula we need to construct the denominator?***

Answer

By the law of total probablity:

$$P(Pos) = P(Pos|DS^+) * P(DS^+) + P(Pos|DS^-) * P(DS^-)$$

⇒ ***Q: What is the final formula and the result after plugging in all of the corresponding values?***

Answer

$$P(DS^+|Pos) = \frac{P(Pos|DS^+) * P(DS^+)}{P(DS^+) * P(Pos|DS^+) + P(DS^-) * P(Pos|DS^-)}$$
$$= \frac{0.6 * 0.001}{0.6 * 0.001 + 0.999 * 0.05}$$

```
result = 0.6*0.001 / (0.6*0.001 + 0.999*0.05)
round(result*100, 2)
```

```
## [1] 1.19
```

The answer is `1.19`%. This means that almost 99% of positive results will be false positives! Even though the FP rate is only 5%, the incidence of Down's is so low that the vast majority of cases that do return a positive test will not actually have Down's syndrome.

In many cases, it's better to err on the side of caution, and any positive test results can be confirmed by further testing.