Chapter 8; When One Thing Depends on Another, Conditional Probability

Steve Elston

12/7/2020

Introduction

In the real world, most random variables we observe are dependent on other random variables. For example, the probability of a person contracting an infectious disease is dependent on a number of other random variables. These random variables might include degree of exposure to other people with the infection, precautions taken, genetic disposition to contracting the disease, etc. As a result, to model and understand the probability of contracting the infectious disease, we must include other random variables in our model on which the probability of infection depends. In more technical terms, the probability of contracting the disease is **conditional** on other random variables.

As the foregoing example indicates, statistical models of complex processes invariably require the use of **conditional probability distributions**. In this chapter we will review the key properties of conditional probability distributions we will use in the remainder of this book.

Properties of Conditional Probability

Conditional probability is the probability that event A occurs given that event B has occurred. We can write conditional probability as follow, which we say is the probability of A given B:

$$P(A|B)$$

Start by examining Figure 1, which shows a number of discrete events. The overall **sample space** is the space of all possible events in the set S. This space is divided into several **subspaces** or **subsets**, A, B and C. The subsets A and B are shown as circles, with an **intersection** where the two sets overlap. Events in this intersection occur in both A and B.

We can work out the conditional probability for the intersection between A and B as follows. First, we find the relationship between conditional probability and the intersection between the sets, $P(A \cap B)$. To find this probability notice that it is the product of two probabilities:
1. $P(B)$ since B must be true to be in this intersection. 2. $P(A|B)$ since A must also occur when B is occurring. Given this logic, we can write:

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

Rearranging terms we get the following:

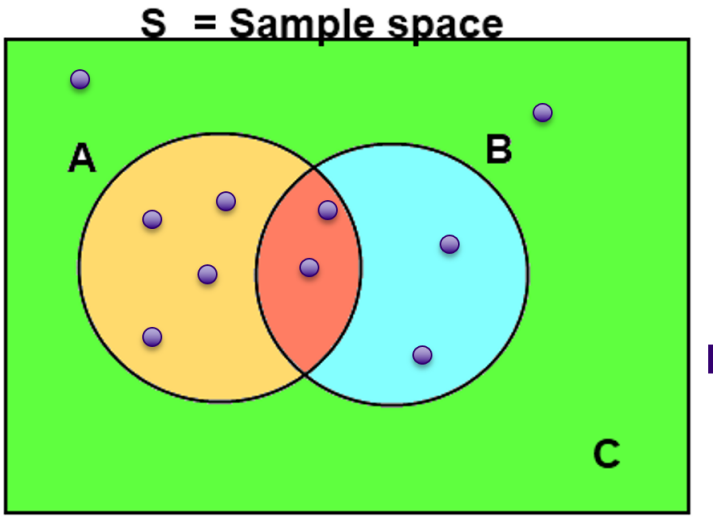

Figure 1: Example of conditional probability of discrete events; credit, Wikipedia commons

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{1}$$

$$= \frac{\frac{2}{10}}{\frac{4}{10}} = \frac{2}{4} = \frac{1}{2} \tag{2}$$

We could have, just as well, written the last equation as:

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

Now, the probability of an identical event in the same intersection must be the same. Therefore we can write:

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

In general, this type of **factorization** of a probability function is a key tool in working with complex conditional probabilities. You can see from the previous equation, that factorization of conditional probability distributions in not unique.

Set operations and probability

Set operations can be readily applied to probability problems. Continuing with our example, we can apply the following common set operations.

1. **Intersection:** We have already discussed intersection of the sets of two events.

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

2. **Union:** The probability of the union of two sets of events is the sum of the probabilities of the sets less the intersection between the sets. Examining Figure 1, you can see that the last term is required to not count the intersection twice. We can express this idea as:

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

Exercise 8-1: Continusing with our example, apply the above equation to compute the probability of the union of A and B.

3. **Negation:** The negation operator, $\neg$, can be applied to sets of events. For example, we can compute the probability of an event being in the subset A but not in the subset B as:

$$P(A \; and \; \neg B) = P(A) - P(B \cap A)$$

Exercise 8-2: Continuing with the running example, compute $P(A \; and \; \neg B)$ using the above relationship.

We can use the combination of basic logical set operations and negation to factor more complex relationships. For example, we can apply **De Morgan's Laws**:

$$P(\neg(A \cup B)) = P(\neg A \; \cap \neg B) \tag{3}$$
$$P(\neg(A \cap B)) = P(\neg A \; \cup \neg B) \tag{4}$$

Exercise 8-3: Apply De Morgan's Laws to compute $P(\neg(A \cup B)$ and $P(\neg(A \cap B))$ for the running example.

Independence and mutual exclusivity

The factorization of probability distributions can be simplified if events are either **independent** or **mutually exclusive**. At first glance, these concepts may seem similar. However, they are quite different, and with very different implications.

To start, we will investigate the concept of independence. As the name implies, the occurrence of one type of event in the set A, does not have any dependency on another type of event in the set B. We can express properties of independent random variables, $A \perp B$, mathematically:

$$P(A \; \cap B) = P(A|B)P(B) = P(A)P(B) \tag{5}$$
$$P(A \; \cup B) = P(A) + P(B) - P(A)P(B) \tag{6}$$
$$P(A|B) = P(A) \tag{7}$$
$$P(A|\neg B) = P(A) \tag{8}$$

But, be careful! Independence of A given B does not imply, independence of B given A:

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

Exercise 8-4: An example helps to make this concept less abstract. Say you have a fair coin and a fair 6-sided dice (numbered 1-6). What is the probability that you will flip the coin and get a head, and roll the dice and get a 6. You should ask yourself if these events have any dependency on each other. Given your answer compute the probability of these two events both ocuring.

But what if the intersection between the events is an empty set, $A \cap B = \emptyset$. In this case we say the events in A are **mutually exclusive** of events in B. In other words, events cannot occur in both the sets A and B. As a result we can write the following:

$$P(A \cup B) = P(A) + P(B) \tag{9}$$
$$P(A|B) = 0 \tag{10}$$
$$P(A|\neg B) = \frac{P(A)}{1 - P(B)} \tag{11}$$

But, just because A is mutually exclusive of B, does not mean B is mutually exclusive of A. Or, in terms of our notation:

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

> **Exercise 8-5:** Consider an example of drawing a playing card randomly from a deck. The deck of cards has equal numbers of suites = $\{hearts, spades, clubs, diamonds\}$. A card can only have one suite. Use the relationships given above to compute the probability that the card you draw is a heart?

Conditional distributions and Bayes' Theorem

Bayes' theorem, also known as **Bayes' rule**, is a powerful tool to think about and analyze conditional probabilities. We will use Bayes theorem in many subsequent parts of this book.

We can derive Bayes Theorem starting with the following relationships:

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

Now:

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

This leads to:

$$P(A|B)P(B) = P(B|A)P(A) \tag{12}$$
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{13}$$

Which is Bayes' theorem!

Interpreting Bayes Theorm

How can we interpret Bayes' theorem in a useful way? Consider the example where we wish to use Bayes Theorem to test a hypothesis given some data or **evidence**. We must make an assertion of our prior probability that the hypothesis is true, $prior(hypothesis)$. We also must choose a likelihood function of the evidence given the hypothesis, $Likelihood(evidence \mid hypothesis)$. Now, we can think of Bayes' theorem in the following terms:

$$Posterior(hypotheis \mid evidence) = \frac{Likelihood(evidence \mid hypothesis)\ prior(hypothesis)}{P(evidence)}$$

We will discuss selection of prior probability distributions and likelihood functions in great detail in subsequent chapters. For now, we will just assume these are known.

A difficult issue here is to come to grips with the denominator, $P(evidence)$. This term is often referred to as the **partition function**, a somewhat confusing reference to statistical mechanics in physics. The denominator is required to normalize the posterior distribution so it is in the range $0 \leq Posterior(hypotheis \mid evidence) \leq 1$ To properly normalize the posterior, the denominator must account for all possible outcomes, or alternative hypotheses, h'. This means that we can write Bayes Theorem as follows:

$$Posterior(hypotheis \mid evidence) = \frac{Likelihood(evidence \mid hypothesis)\ prior(hypothesis)}{\sum_{h' \in\ All\ possible\ hypotheses} Likelihood(evidence \mid h')\ prior(h')}$$

This looks like a formidable problem, and it is! It is often the case that computing this denominator by brute force is simply not feasible. We will address some ways to deal with this problem in subsequent chapters.

Marginal Distributions

In many cases of Bayesian analysis we are interested in the **marginal distribution**. For example, it is often the case that only one or a few parameters of a joint distribution will be of interest. In other words, we are interested in the marginal distribution of these parameters. Further, the denominator of Bayes theorem, $P(data)$, can sometimes be computed as a marginal distribution. For these reasons computing marginal distributions is an important aspect of Bayesian analysis.

Consider a multivariate probability density function with n variables, $p(\theta_1, \theta_2, \ldots, \theta_n)$. A **marginal distribution** is the distribution of one variable with the others integrated out. In other words, if we integrate over all other variables $\{\theta_2, \ldots, \theta_n\}$ the result is the marginal distribution of the variable of interest, $p(\theta_1)$. We can express this idea mathematically as follows:

$$p(\theta_1) = \int_{\theta_2, \ldots, \theta_n} p(\theta_1, \theta_2, \ldots, \theta_n)\ d\theta 2, \ldots, d\theta_n$$

Example, marginal distributions of eye and hair color

Let's consider an example of working with marginal and conditional distributions. This example follows Section 5.1.2 of Kruschke (2015).

A sample population has the following joint probabilities of eye and hair color combinations. These values are the joint probabilities $P(eye, hair) = P(hair, eye)$.

```
eye_hair = pd.DataFrame({
    'black': [0.11, 0.03, 0.03, 0.01],
    'brunette': [0.2, 0.14, 0.09, 0.05],
    'red': [0.04, 0.03, 0.02, 0.02],
    'blond': [0.01, 0.16, 0.02, 0.03],
}, index=['brown', 'blue', 'hazel', 'green'])

eye_hair
```

```
##        black  brunette   red  blond
## brown   0.11      0.20  0.04   0.01
## blue    0.03      0.14  0.03   0.16
## hazel   0.03      0.09  0.02   0.02
## green   0.01      0.05  0.02   0.03
```

How can we understand these joint probabilities in terms of conditional probabilities? We can express this relationship as:

$$p(eye, hair) = p(eye|hair)p(hair) = p(hair|eye)p(eye) = p(hair, eye)$$

At first glance, this relationship can look confusing. But, keep in mind that the table of joint probabilities can be read either row wise or column wise. The joint probability, $p(eye, hair)$, must be the same in either case.

> **Computational note:** Here it is convenient to use a string index with the Pandas data frame for eye color and hair color, rather than the usual numeric zero-based indicies. To access a given (eye, hair) color value, index the data frame like this:

```
eye_hair.loc['hazel', 'red']
```

```
## 0.02
```

Given these joint probabilities, it is easy to compute the marginal distributions of either the eye or hair color by summing over the rows or columns. Since this problem involves discrete values, we can use simple summation. Hair color is in the columns, and the marginal distribution is computed by summing over the rows. Similarly, the eye color is in the rows, so the marginal distribution is computed by summing over the columns. We can express these operations mathematically:

$$p(hair) = \sum_{rows} p(hair|eye)\ p(eye) = \sum_{rows} p(hair, eye) \tag{14}$$

$$p(eye) = \sum_{columns} p(eye|hair)\ p(hair) = \sum_{columns} p(hair, eye) \tag{15}$$

Like all probability distributions, the marginal probability distribution must sum to 1.0. It is always good to check this condition to ensure the sums have been applied correctly.

> **Exercise 8-6:** Use the table of hair and eye color to compute the marginal distributions, $p(eye)$ and $p(hair)$. Make sure you check that these marginal probabilities sum to approximately 1.0. What inferences can you make from these marginal distributions? What is the most common eye color? What is the least common hair color?

Conditional Probability Example

Let's try a simple and widely used example of using conditional probabilities to work out the chance of having a rare disease. The scenario is as follows:

1. Sickle Cell Anemia is a serious, but fairy rare disease. The probability that a given patient, drawn at random from the population of all people in the United States, has the disease is $P(S) = \frac{1}{3200} = 0.0003125$. We can describe the possible events in diagnosing this condition as:

- $S \Rightarrow$ a patient has the disease.
- $S' \Rightarrow$ a patient does not have the disease.
- $\oplus \Rightarrow$ patient tests positive.
- $- \Rightarrow$ a patient tests negative.

2. What if a medical company claims that it has developed a test that is 99% accurate. We can then write:

- $P(S|\oplus) = 0.99$
- $P(S'|-) = 0.99$

On the surface, it seems that a 99% reliable test is rather good. Such a test would ensure that, on average, 99 people out of 100 who have the disease will be identified and treated. But, let's dig into the conditional probabilities and see how things really work out.

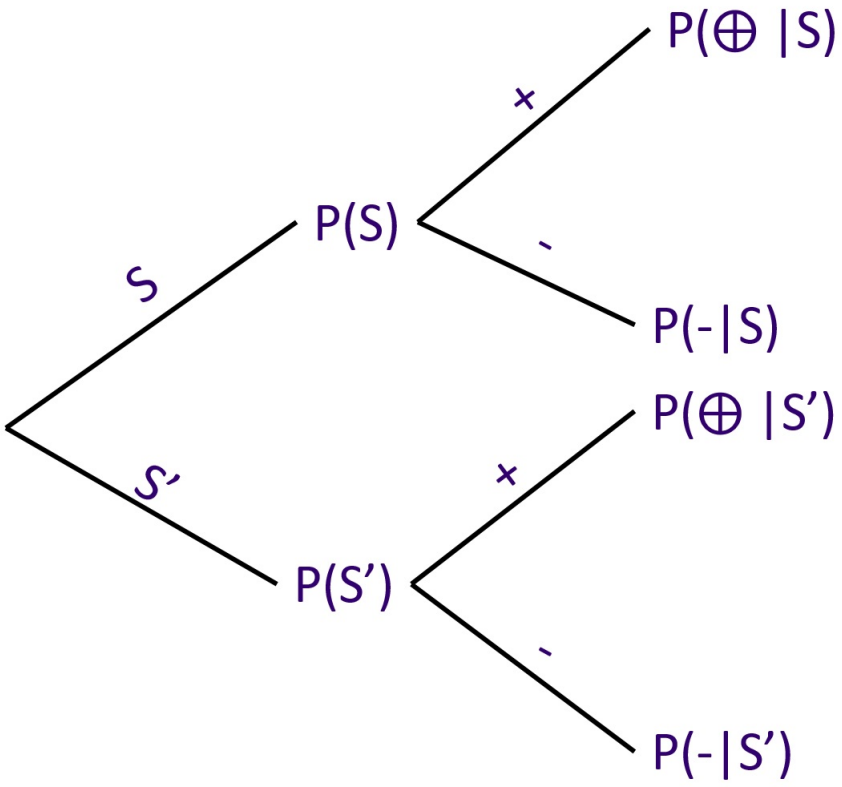

Figure 2: Graph showing dependency of conditional distribution

The Figure 2. shows a **conditional probability tree** for this problem. This tree is technically referred to as a **directed graphical model (DAG)**. Starting at the root the DAG defines a conditional dependency structure of the probability distributions. The goal is to evaluate the medical test as a **decision rule** for providing treatment to patients.

If you follow the tree from the root you can visualize the computation of probabilities for each of the 4 possible outcomes of a test on a patient. Let's summarize the conditional probabilities we need to compute for these outcomes:

- $P(\oplus|S)$ is the conditional probability the test will correctly identify a patient with the disease.
- $P(-|S)$; is the conditional probability of a negative test for a patient with the disease. We call this situation a **Type II Error** or **False Negative**.
- $P(\oplus|S')$ is the conditional probability that a patient with no disease will test positive. We call this situation a **Type I Error** or **False Positive**.
- $P(-|S')$; is the conditional probability of a negative test for a patient who does not have the disease.

We can summarize the four possible outcomes using a **confusion matrix** or **truth table**. The proportion of each case are typically shown in normalized, or decimal, form. An example for this case is shown here:

	Positive Test	Negative Test
Disease	True Positive Rate	False Negative Rate
No Disease	False Positive Rate	True Negative Rate

Exercise 8-7: Using the information provided you will now use conditional probabilities to analyze the effectiveness of the claimed test. Perform the following steps to find out. Your goal with this exercise is to fill in the confusion matrix shown above.
1. Start with the easy cases. You know the probability of a patient having the disease and the accuracy of the test. Create and execute the code to compute the conditional probabilities of a positive test given that a randomly selected patient has the disease, and a negative test given the randomly selected patient does not have the disease. 2. Next, you will compute the conditional probabilities for the cases where the test is in error. Create and execute the code to compute the conditional probabilities of a negative test given the randomly selected patient has the disease, and a positive test given the patient does not have the disease. Compare these results to the conditional probabilities you computed in step 1. 3. Finally, create and execute the code to compute the sum of the probabilities of all the possible outcomes. Does the sum equal to 1.0? 4. Given these results, do you think this test is actually useful? Consider that for many medical conditions, treatment of someone who does not have the condition involves some risk and certainly cost.

Disclaimer!!: The foregoing example is purely hypothetical. The probabilities stated are only to simplify the calculations. None of the information provided should be considered accurate medical information

A Case of Assumptions and Appling Conditional Probability

The results of the foregoing exercise demonstrate that conditional probabilities can lead to unexpected results. Failure to correctly account for conditional probabilities can lead to spectacularly incorrect results.

As an example, the case of Sally Clark is now notorious in the annals of British justice. She was convicted in 1999 of murdering her two young sons. In part, she was convicted based on testimony of Professor Sir Roy Meadow who stated in court that the probability of two such deaths is 1 in 73 million. This analysis was based on an assumption that these death were random independent events. Meadow assumed that the probability of this outcome could be computed as two iid Bernoulli trials, each with $p = 1/8500$, with the 'positive' outcome being a child dying from Sudden Infant Death Syndrome (SIDS). If Meadow's assumptions had been correct, the probability of two such independent deaths is just the square of the probability of a single death.

But, Meadow did not disclose or consider that the second child to die was found to have a serious bacteriological infection. On a second appeal this fact was reveled. A correct analysis would have taken into account the conditional probability of the two deaths given the second child's susceptibility to the infection. Taking account of conditional distributions shows that given these circumstances the probability of both deaths being from natural causes is reasonably high. Further, given this additional condition the Bernoulli distributions of the two childern dying are no iid.

Conditional Probability and the Monte Hall Problem

The long running television game show, *Let's Make A Deal*, originally created and hosted by *Monte Hall*, had its hay-day in the 1970s. At the finale of the show, Monte would tell the winning contestant that they

could pick one of three doors. Behind one door there would be a valuable prize like a car. Worthless items, like a goat, were placed behind the other two doors. The contestant would pick a door. At this point Monte would build suspense by opening one of the other doors. Monte knew which door had the valuable prize, and would always reveal one of the worthless prizes. He would then tell the contestant that they could change their choice of doors. The question is, should the contestant switch or stick with their original choice? Figure 4. illustrates the situation the contestant faces if they pick Door 1:

Figure 3: Illustration of Monte Hall game; Credit, Wikipedia commons

The Monte Hall Problem has a long and convoluted history. In 1975, Steve Selvin published a letter in the *American Statistician* posing the problem of which strategy is optimal. The resulting debate created considerable controversy. This debate was put at full boil when Marilyn vos Savant wrote in her *Ask Marilyn* column in *Parade* magazine that the contestant should definitely switch. She was then ridiculed by several statisticians. But, was ultimately proven to be correct.

Exercise 8-8: What would you do if you were the contestant? Fortunately for you, you know something about conditional probabilities. Further, you know the following probabilities. Your first choice, of one of the three doors, is purely random, since only Monte knows which door hides the car. - There is a probability of 2/3 that your initial pick will be one of the two doors with a goat. At this point, with probability 1 Monte will open the door with the other goat, since he cannot reveal the location of the car. You can only win a car by switching doors.
- There is a probability of 1/3 that your initial pick will be the only door with the car. At this point, with probability 1/2 Monte can open either of the other doors, as they both contain goats. But, if you switch doors at this time, you will win a goat.
- To solve this problem you will create and execute code to simulate the Monte Hall game and analyze the result.
Tip: It will help you understand the dependencies of the conditional probabilities if you draw a graph (tree) showing the relationships.
Now do the following:
1. Create a wrapper function that allows you to run the simulation $n = 1000$ times, with a *switch* or *not switch* strategy.
2. Create a function named `random_door`, which uses numpy.random.choice to Bernoulli sample 1 door randomly from a list of integer door indices (1-3 in this case). Use this function to ramdomly select the door the car is behind and the contestant's initial choice of doors.
3. Create a function `monte_choice`, which chooses the door Monte opens, conditional on the contestant's choice of doors and the door with the car. For the case where the contestant has selected the door with the car, select the door to open by simulating the flip of a fair coin using the scipy.stats.binom.rvs function with $n = 1$.
4. Create a function `win_car`, which determines if the contestant wins the car, conditional on the strategy selected, $\{switch, no_switch\}$, the door the contestant selected, the door with the car, and the door Monte opened.
5. Execute your simulation for each possible strategy. For the two strategies, plot side by side

bar charts showing the numbers of successes and failures for each strategy.
6. Describe the strategy a constestant should adopt for this game. How much will the chosen strategy change the probability of winning a car? Is this result consistent with the conditonal probabilities of this problem.

Bibliography