Probability, Statistics, and Bayes' Theorem
Session 2

1 Conditional Probability

When dealing with finite probability, we saw that the most natural way of assigning a probability to an event A is with the following formula:

$$P(A) = \frac{\text{number of } \omega \in A}{\text{total size of } \Omega} = \frac{|A|}{|\Omega|}.$$

We see here explicitly the role of the sample space Ω in the calculation of finite probabilities. It is the denominator, which is "normalizing" all of our probabilities so that they satisfy the first axiom we listed last time ($0 \leq P(A) \leq 1$). Conditional probability is a way of using information previously obtained to change the size of the sample space, and thereby change the probability associated to any event. Before looking at the formal definition, let's consider some intuitive examples.

1.1 Conditional Probability: More Intuitively

Example Say you're watching the Vancouver Canucks. They have a power-play and both Henrik and Daniel Sedin are on the ice. As much as you want to watch, you're really thirsty so you go into the kitchen to get a drink. While you're in there, the Canucks score. Assuming that any player on the ice for the Canucks, except Luongo, is equally likely to score (which isn't really that believable, but let's assume it is), what's the probability that Daniel Sedin scored? Well, since there are five non-goalie players on the ice for Vancouver, and we're assuming that any one of them is equally likely to score, we'd say that the probability that it was Daniel Sedin is 1/5. Using the formula from above, this is

$$P(\text{Daniel scored}) = \frac{\text{number of Daniel Sedin's on ice}}{\text{number of non-goalie players on ice for Vancouver}} = \frac{1}{5}.$$

Suppose now that one of your friends who you're watching the game with shouts to you (while you're still in the kitchen), "Sedin scored!" Now what is the probability that it was

Daniel Sedin? Let's think about what's happening in this situation. You now know that it was either Henrik or Daniel who scored. Since our underlying assumption was that all the players are equally likely to have scored, this applies to the situation where the only possibilities are Henrik or Daniel. This means that the probability that it was Daniel is now 1/2. Again, looking at the formula in this situation, we have

$$P(\text{Daniel scored}) = \frac{\text{number of Daniel Sedin's on ice}}{\text{number of Sedin's on ice}} = \frac{1}{2}.$$

Here we see that all that changed in the two situations was the denominator, the size of the sample space, and this change was caused by the availability of additional information.

The next two examples shift the focus a little, but it is a good introduction to some of the themes that we will be concerned with in what follows. They have more to do with thought-experiments to get you thinking about the interaction between probability and data, and less to do with the actual calculation of values.

Example You work for RIM and you go to a factory where the newest BlackBerry phones are manufactured. The foreman of the factory tells you that out of the last 100 BlackBerry units that a specific machine has produced, 95 of them have passed inspection. What do you think is the probability that the next BlackBerry that the machine produces will pass inspection? Based on only this information, it would seem reasonable to say that the probability would be .95, which intuitively means that you feel 95% certain that the next BlackBerry will pass inspection.

Example You and a friend are going to play a simple gambling game. He's going to flip a coin 10 times and if it lands heads, you pay him a dollar, while if it lands tails he pays you a dollar. He asks you for a coin, and you give him a loonie that you've checked to make sure has both a heads and a tails. He starts flipping the coin and flips 10 heads in a row. You are out $10. At the end of the game he looks at you and says "I wasn't cheating!" How much do you believe him?

1.1.1 Exercises

1. Return to the RIM example. Suppose now that you look at a chronological record of the last 100 BlackBerry units produced by this machine and you see that the first 95 all passed inspection, but the last 5 have all failed. Would you still be 95% certain that the next BlackBerry produced by this machine would pass inspection?

2. Return to the gambling example. Suppose now that the set-up of the game is the same, except this time he takes a loonie out of his pocket and all you see is the heads. He starts flipping and again flips 10 heads in a row. Again, at the end of the game he says "I wasn't cheating!" How much do you believe him this time?

All three of these examples/exercises are in some way related to conditional probability. In the first example we could actually compute a specific value for the probabilities that

interested us. The second example dealt with how we adjust our views of likely future outcomes based on the level of knowledge that we have about the past outcomes. The third example looked at probability as "degree of belief." In order to try to unify these seemingly disparate views of what conditional probability can mean intuitively, we need to start with the formal definition of what conditional probability is.

1.2 Conditional Probability: Less Intuitively

The *conditional probability* of an event A given the event B, which is written $P(A|B)$, is defined to be

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

The left-hand side is often phrased as "the probability of A, given B," or "the probability of A conditioned on B." Looking at the right-hand side, we see that this probability is given by the intersection of the events A and B, divided by the probability of B. Interpreting this according to the formula that we used above, we get

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

Looking at the denominator leads us to believe that we are now treating B as our sample space, while the numerator suggests that we are only concerned with the number of events that lie in both A and B. This is exactly what's happening. If the numerator seems mysterious, think about what we're trying to calculate. We want to know the probability that A occurs given that we know for certain the B occurs. If we were to take an element that was in A but not in B, then this element would be the result of an experiment where B did not occur, and as such we do not want to count it. But every element in $A \cap B$ is the result of an experiment where both A and B occur, and these are exactly what we want to count. Finally, since we are now only interested in situations where we know for certain that B has occurred, we throw away everything in our sample space where B does not occur, and all that we're left with is B itself. This explains the denominator.

Example What is the probability that you roll a 12 with a throw of 2 dice, given that you know the second dice is a 6? We are interested in $P(\text{first dice} = 6|\text{second dice} = 6)$, which by the formula for conditional probability is

$$P(\text{first dice} = 6|\text{second dice} = 6) = \frac{P(\text{first and second dice} = 6)}{P(\text{second dice} = 6)} = \frac{\frac{1}{36}}{\frac{1}{6}} = \frac{1}{6}.$$

Example In a certain school, students were classified by entrance examinations into three categories:

$$25\% \text{ above average}, \quad 50\% \text{ average}, \quad 25\% \text{ below average}.$$

The Admissions Department kept a record of the students overall mark after the first semester, given in the table below:

	A	B	C	D	F
Above Average	20%	50%	20%	10%	0%
Average	10%	20%	35%	25%	10%
Below Average	0%	10%	15%	50%	25%

If a random student received a D, what is the probability that he or she had been rated above average?

The first thing to notice is that the entries of this table (converted to probabilities instead of percentages) don't sum to 1. This is because the grade distribution is done according to entrance exam categories, so that the three rows all sum to 1. This means that we have to adjust the entries so that they do sum to 1 without changing the relative size. Since the above average students make up 25% of the total population, we need to find 20% of 25% for the first entry in the first row. We need to do this to every entry. After we're done, we have the following (writing them as probabilities instead of percentages):

	A	B	C	D	F
Above Average	.050	.125	.050	.025	.000
Average	.050	.100	.175	.125	.050
Below Average	.000	.025	.0375	.125	.0625

You can check that these entries do add up to 1. Now we can compute the probability that an above average student got a D:

$$P(\text{above average}|D) = \frac{P(\text{above average and } D)}{P(D)} = \frac{.025}{.275} = .09.$$

We found the denominator, $P(D)$, by adding the three values in the column for D.

By multiplying the denominator through in the expression for conditional probability, we get that $P(A \cap B) = P(A|B)P(B)$. This might not seem like much, but it allows us to do the following: let B_i for $i = 1, \ldots, n$ be a partition of Ω (recall that this means that $B_i \cap B_j = \emptyset$

for $i \neq j$ and $\cup_{i=1}^n B_i = \Omega$). For any $A \subset \Omega$, $A = \cup_{i=1}^n (A \cap B_i)$, and $(A \cap B_i) \cap (A \cap B_j) = \emptyset$ (since the $B_i's$ are disjoint), we have that

$$P(A) = P(\cup_{i=1}^n (A \cap B_i)) = \sum_{i=1}^n P(A \cap B_i).$$

Substituting in the expression for $P(A \cap B)$ in terms of conditional probability, we get the *law of total proabability*:

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

In the simplest case where the partition is given by $B_1 = B$ and $B_2 = B^c$, this formula becomes

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c).$$

We will see some important uses of this formula later.

1.2.1 Exercises

1. In written English with punctuation and spaces removed, about 10% of the letters are t and about 5% of the letters are h. When you count letter pairs (including pairs linking the end of one word to the beginning of the next) the pair th occurs about 3.8% of the time.

   What is P(next letter is h|this letter is t)?

   What is P(previous letter was t|this letter is h)?

2. When the Maple Leafs are in the playoffs (it will happen again someday),

   $$P(\text{Leafs win game 1}) = .5,$$

   while for the other games

   $$P(\text{Leafs win}|\text{they lost previous game}) = .4,$$
   $$P(\text{Leafs win}|\text{they won previous game}) = .7.$$

   What is P(Leafs win games 1, 2, 4, and 5)?

What is the probability that the Leafs will win one of the first two games (but not both)?

3. You have three urns in which there are colored balls. The balls are identical except for their colors. The contents of the urns are

   Urn 1: 3 green balls, 6 yellow balls, 1 white ball,

   Urn 2: 4 black balls, 3 white balls, 2 green balls, 1 orange ball,

   Urn 3: 6 brown balls, 2 green balls, 2 purple balls.

   The probability of choosing urn 1 is 50%, the probability of choosing urn 2 is 30%, and the probability of choosing urn 3 is 20%. Once a random urn is chosen, a ball is drawn out.

   What is $P(\text{ball is green})$? What is $P(\text{ball is yellow})$? What is $P(\text{ball is white})$?

   There are less yellow balls than green, and all of the yellow balls are in one urn, but the probability of drawing a yellow ball is greater than the probability of drawing a green ball. How do you explain this?

2 Bayes' Theorem

Bayes' Theorem has a strange double life. On the one hand, from a purely formal mathematical point of view, it is an almost trivial consequence of the definition of conditional probability. On the other hand, when considered from a more philosophical (for lack of a better term) point of view, it is a profound statement about the relationship between data (information) and probability (uncertainty). In contrast to the previous section, here we will start with the formal aspect of Bayes' Theorem.

2.1 Bayes' Theorem: Formally

Recall the definition of the conditional probability of A given B:

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

Now consider the related conditional probability of B given A:

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

Since $P(A \cap B) = P(B \cap A)$, if we multiply both denominators through in the expressions above we see

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

which after rearranging a few terms becomes

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

2.1.1 Exercises

1. Let $A_1, A_2, \ldots, A_n$ be a partition of Ω. Use the law of total probability to derive the following generalization of Bayes' Theorem:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \ldots + P(B|A_n)P(A_n)}.$$

2.2 Bayes' Theorem: Intuitively

On a formal level, Bayes' Theorem relates two conditional probabilities:

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

Let the event A be interpreted as a proposition and let the event B be interpreted as evidence. Then $P(A)$ is our initial belief in the truth of proposition A. This is called the *prior* probability for A, or just the prior. $P(A|B)$ represents our belief that A is true taking into account the evidence B. It is called the *posterior* probability of A given B, or just the posterior. The ratio

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

is interpreted as the support that the evidence B provides for A. With these interpretations, Bayes' Theorem gives us a way to update of belief in the truth of proposition A based upon the available evidence B; we start with the prior belief in A, then we gather evidence B, and then we form the posterior belief $P(A|B)$.

Example For this example, use the urns from the previous section:

Urn 1: 3 green balls, 6 yellow balls, 1 white ball,

Urn 2: 4 black balls, 3 white balls, 2 green balls, 1 orange ball,

Urn 3: 6 brown balls, 2 green balls, 2 purple balls.

$P(\text{urn } 1) = .5,\ P(\text{urn } 2) = .3,\ P(\text{urn } 3) = .2.$

Suppose now that you do not know which urn has been selected. Someone tells you that a green ball has been selected. What is the probability that urn 1 was chosen?

We want to use Bayes' Theorem here, so we have to decide what is our A and what is our B. The proposition that we are trying to verify is "urn 1 was chosen," so we let $A = \text{urn } 1$. The evidence that we have is that a green ball was selected, so we have $B = \text{green}$. Using these, we get

$$P(\text{urn } 1|\text{green}) = P(\text{urn } 1)\frac{P(\text{green}|\text{urn } 1)}{P(\text{green})} = \frac{(.5)(.3)}{(.25)} = .6$$

Where did these values come from? We know that $P(\text{urn } 1) = .5$ and we can easily compute $P(\text{green}|\text{urn } 1) = .3$. For the denominator, we want to compute $P(\text{green})$. Our first instinct might be to divide the number of green balls in all three urns by the total number of balls in all of the urns. This isn't going to work because the probability of picking balls out of urn 1 is greater than picking them out of urns 2 and 3 and we have to take this into account. What we do is use the law of total probability.

$$\begin{aligned} P(\text{green}) &= P(\text{green} \cap \text{urn } 1) + P(\text{green} \cap \text{urn } 2) + P(\text{green} \cap \text{urn } 3) \\ &= P(\text{urn } 1)P(\text{green}|\text{urn } 1) + P(\text{urn } 2)P(\text{green}|\text{urn } 2) + P(\text{urn } 3)P(\text{green}|\text{urn } 3) \\ &= (.5)(.3) + (.3)(.2) + (.2)(.2) = .15 + .6 + .4 = .25. \end{aligned}$$

What is this calculation telling us? It tells us that if we know that a green ball was selected then we should change our belief that urn 1 was chosen from 50% to 60%.

2.2.1 Exercises

1. What is the probability that urn 2 was chosen? Urn 3?
2. What if you were told that a white ball had been selected. What is the probability that urn 1 was chosen? Urn 2? Urn 3? Give an intuitive explanation of the value you found for urn 3 being chosen.
3. A coin, which you are not allowed to examine, is either a fair coin ($P(\text{heads}) = .5$) or has two heads. Your initial opinion is $P(\text{fair}) = .9$. The coin is flipped and heads comes up. What is your opinion now?

   The coin is flipped a second time and again heads comes up. What is your opinion now?