1 Definition of Probability

A random process is one in which the outcome is unpredictable. We encounter random processes every day: will it rain today? how many minutes will pass until receiving your next text message? will the Seahawks win the Super Bowl? Though the outcome of one particular random process is unpredictable, if we observe the process many many times, the pattern of outcomes, or its probability distribution, can often be modeled mathematically. Though there are several philosophical definitions of probability, we will use the “frequentist” definition of probability, given in section P.3 of our textbook:

**Definition:** The probability of an event is the long-run proportion of times the event would occur if the random process were repeated indefinitely (under identical conditions).

*Example:* Consider the simple example of flipping a fair coin once. What is the probability the coin lands on heads. From its physical properties, we assume the probability of heads is 0.5, but let’s use simulation to examine the probability. The plot below shows the long-run proportion of times a simulated coin flip lands on heads on the y-axis, and the number of tosses on the x-axis. Notice how the long-run proportion starts converging to 0.5 as the number of tosses increases.

2 Probability Tools: Two-way Tables and Tree Diagrams

We can easily solve probability problems without using any equations by either creating a hypothetical two-way table or a tree diagram. These tools are best demonstrated by an example.

*Example:* Suppose your first class on Mondays is in Wilson Hall at 10:00am and you commute to school. From past experience, you know that there is a 20% chance of finding an open parking spot in
the E lot next to Jabs. Otherwise, you have to park in an SB lot. If you find a spot in the E lot, you
only have a 5% chance of being late to class. However, if you have to part in the SB lot, you have a
15% chance of being late to class. What is the probability that you will be late to class this Monday?

2.1 Hypothetical Two-way Tables

Let’s imagine 1000 hypothetical Mondays. Since there is a 20% chance of finding an open spot in the
E lot, on $1000 \times 0.2 = 200$ of these Mondays, you park in E. On the other 800 Mondays, you end up
in SB. We can enter this information into a hypothetical two-way table:

<table>
<thead>
<tr>
<th>Parking lot</th>
<th>Late to class?</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Park in E lot</td>
<td></td>
<td></td>
<td></td>
<td>200</td>
</tr>
<tr>
<td>Park in SB lot</td>
<td></td>
<td></td>
<td></td>
<td>800</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td>1000</td>
</tr>
</tbody>
</table>

Now, of the 200 times you end up in the E lot, you will be late to class on $200 \times 0.05 = 10$ of those
Mondays. Of the 800 Mondays in the SB lot, you will be late to class on $800 \times 0.15 = 120$ of those
days. Let’s enter this in the table:

<table>
<thead>
<tr>
<th>Parking lot</th>
<th>Late to class?</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Park in E lot</td>
<td></td>
<td>10</td>
<td>190</td>
<td>200</td>
</tr>
<tr>
<td>Park in SB lot</td>
<td></td>
<td>120</td>
<td>680</td>
<td>800</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>130</td>
<td>870</td>
<td>1000</td>
</tr>
</tbody>
</table>

We can fill in the rest of the table using addition and subtraction:

Further probability questions can now be answered from the table. Using the table, find the
following probabilities (answers in footnote):

1. What is the probability you are late to class?
2. What is the probability that you park in the E lot and you are not late to class?
3. Given that you were late to class, what is the probability you parked in the SB lot?

Carefully read all probability questions – note the subtle difference between “the probability of
being late to class, given that you parked in SB” (120/800 = 0.15) and “the probability of parking in
SB, given that you were late to class” (120/130 = 0.923). When we are given extra information, this
is called a conditional probability, and the denominator in the probability calculation is a row total
(e.g., 800) or column total (e.g., 130) rather than the overall total.

Answers: 1. 130/1000 = 0.13; 2. 190/1000 = 0.19; 3. 120/130 = 0.923.
2.2 Tree Diagrams

Another probability tool is a **tree diagram**.

1. Grow the first set of tree branches using the given unconditional probabilities.

```
          Park
          /   
0.20     0.80
          /     
        Park   Park
       /     /   
0.20 / 80

Park in E
Park in SB
```

2. Add conditional probabilities for the next set of branches.

```
          Park
          /   
0.05     0.95
          /     
        Park   Park
       /     /   
0.20 / 80

Park in E
Not late to class

Park in E
Late to class
```

3. Calculate the probability of a route on the tree by multiplying the probabilities on the branches.

```
          Park
          /   
0.05     0.95
          /     
        Park   Park
       /     /   
0.20 / 80

Park in E
Not late to class

Park in E
Late to class

Park in E
Park in SB and late
Park in SB
```

```
          Park
          /   
0.15     0.85
          /     
        Park   Park
       /     /   
0.80 / 85

Park in SB
Late to class

Park in SB
Not late to class

Park in SB
Not late to class

Park in SB
Park in SB and not late
Park in SB

Park in SB
```

Park in E and late
(0.20)(0.05) = 0.1

Park in E and not late
(0.20)(0.95) = 0.19

Park in SB and late
(0.80)(0.15) = 0.12

Park in SB and not late
(0.80)(0.85) = 0.68
Note several important features of the tree diagram:

- If you multiply all the end probabilities by 1000, you get the same values as those in your hypothetical two-way table (.01×1000 = 10; .19×1000 = 190; .12×1000 = 120; .68×1000 = 680).
- The sum of all the end probabilities equals 1: .01 + .19 + .12 + .68 = 1.

Answer the same questions using the tree diagram (answers are again in footnote):

1. What is the probability you are late to class?
2. What is the probability that you park in the E lot and you are not late to class?
3. Given that you were late to class, what is the probability you parked in the SB lot?

3 Probability Notation

For ease of translating probability problems into calculations, let’s define some notation. We will denote “events” by upper case letters near the beginning of the alphabet, e.g., A, B, C. The probability of an event A will be denoted by P(A), so P(A) is a number between 0 and 1. The event that A does not happen is called the complement of A and is denoted by P(A^C). Sometimes we have additional information that we would like to condition on, and we denote the conditional probability of A given B by P(A|B), the probability that A happens given that B has already happened.

Example (cont.): In our coin flip example, we could let A be the event that the coin lands on heads. Then we can denote the probability that the coin lands on heads by P(A) = 0.5. We could flip the coin twice and let H_1 be the event that the first flip lands on heads, and H_2 be the event that the second flip lands on heads. Since the coin does not remember its last flip, if the first flip lands on heads, the second flip still has a 50% chance of landing on heads. That is, P(H_2|H_1) = 0.5.

4 Diagnostic Tests

Medical diagnostic tests for diseases spend years in development. Through clinical trials, developers of the diagnostic test are able to determine two important properties of the test:

- The sensitivity of a diagnostic test is the probability the test yields a positive result, given the individual has the disease. In other words, what proportion of the diseased population would test positive?
- The specificity of the diagnostic test is the probability the test yields a negative result, given the individual does not have the disease. That is, what proportion of the non-diseased population would test negative?

A good diagnostic test has very high (near 100%) sensitivity and specificity. However, even for a near-perfect test, the probability that you have the disease given you test positive could still be quite low. To investigate this counter-intuitive result, we need another definition:

- The prevalence of a disease is the proportion of the population that has the disease at a given time, or the probability that a randomly selected individual from this population has the disease.

Answers: 1. 0.01 + 0.12 = 0.13; 2. 0.19; 3. 0.12/(0.01 + 0.12) = 0.923.
We can define these terms in probability notation. If we let $D$ be the event that an individual has the disease and $T$ be the event that an individual tests positive. Then:

- Sensitivity $= P(T|D)$
- Specificity $= P(T^C|D^C)$
- Prevalence $= P(D)$

Note that sensitivity and specificity are conditional probabilities, while prevalence is an unconditional probability. While the above probabilities are useful information, if you test positive on a diagnostic test, the probability you really want to know is $P(D|T)$.

**Example: The Case of Baby Jeff**

A poster in a hospital’s newborn nursery announced that all male newborns would be screened for muscular dystrophy using a heel stick blood test for creatinine phosphokinase (CPK). The test characteristics of the screening tests were nearly perfect: a sensitivity of 100% and a specificity of 99.98%. The prevalence of muscular dystrophy in male newborns ranges from 1 in 3,500 to 1 in 15,000. Baby Jeff had an abnormal CPK test. The parents of the baby wanted to know, "What is the chance that our son has muscular dystrophy?” Doctors informed the parents that though not 100% likely, it was highly probable. First, take a minute and predict this probability – what do you think? 80% chance? 99% chance? Let’s investigate using a two-way table of a hypothetical population of 100,000 male newborns.

For our calculations, let’s use a prevalence of 1 in 10,000. Then out of 100,000 male newborns, we would expect 1 in 10,000 to have muscular dystrophy, or 10: \((1/10000) \times 100000 = 10\). The sensitivity of the test is perfect, so all 10 of the male newborns with muscular dystrophy will test positive. Of the \(100000 – 10 = 99,990\) male newborns that do not have muscular dystrophy, 99.98% will test negative: \((0.9998) \times 99990 = 99,970\) infants. That leaves \(99990 – 99970 = 20\) male newborns that test positive even though they do not have muscular dystrophy. This allows us to fill in the counts in our hypothetical two-way table:

<table>
<thead>
<tr>
<th>Have disease?</th>
<th>Test positive?</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>10</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>No</td>
<td>20</td>
<td>99,970</td>
<td>99,990</td>
</tr>
<tr>
<td>Total</td>
<td>10+20 = 30</td>
<td>0+99970 = 99,970</td>
<td>100,000</td>
</tr>
</tbody>
</table>

Now we can read off the desired probability from the table: of the 30 male newborns we’d expect to test positive, only 10 of them actually have muscular dystrophy. This means the chance Baby Jeff has muscular dystrophy is only about 33%! How would this probability change if the prevalence were 1/3500? 1/15000? Try it.

Let’s create a tree diagram for this scenario. We first split into an event and its complement for which we know the (unconditional) probabilities. In a diagnostic test calculation, we know the prevalence of the disease, so our first split will be into $D$ and $D^C$. We then branch off into the next set of events for which we know the conditional probability, conditioned on the first split, in this case $T$ and $T^C$. Then calculate “and” probabilities by multiplying through the branches:
To calculate the probability of having muscular dystrophy, given a positive test, we need the proportion of babies who have muscular dystrophy out of the subset of babies who test positive. That is,

\[ P(D|T) = \frac{0.0001}{0.0001 + 0.0019998} = 0.333. \]

5 Basic Probability Rules: Not, And, Total, Given

All of the calculations performed using a hypothetical two-way table or a tree diagram can be done using mathematics and some basic probability rules. We’ll demonstrate the basic probability rules in the context of the example from section 2. Recall that you have a 20% chance of parking in an E lot. Otherwise, you park in SB. If you end up in the E lot, there is a 5% chance of being late to class. The probability of being late to class, given you park in SB, is 0.15. Define the following events:

- \( E = \) park in an E lot
- \( L = \) late to class

Start by writing the information given in the problem in terms of these events: \( P(E) = 0.20, P(L|E) = 0.05, P(L|E^C) = 0.15. \)

**Rule 1:** “Not”: \( P(A^C) = 1 - P(A) \)

*Example:* The probability that you do not park in an E lot is \( P(E^C) = 1 - P(E) = 1 - 0.20 = 0.80. \)

**Rule 2:** “And”: \( P(A \text{ and } B) = P(A) \times P(B|A) \)

*Example:* What is the probability that you park in an E lot and you are late to class? We could use the “And” rule by either \( P(E \text{ and } L) = P(E)P(L|E) \) or \( P(L \text{ and } E) = P(L)P(E|L) \). Which equation is easier to use with the given information? We are given \( P(E) \) and \( P(L|E) \) in the problem, but we are not given \( P(L) \) and \( P(E|L) \). Therefore,

\[ P(E \text{ and } L) = P(E)P(L|E) = 0.20(0.05) = 0.01. \]
Rule 3: “Total”: \[ P(A) = P(A \text{ and } B) + P(A \text{ and } B^C) \]

**Example:** What is the unconditional probability that you are late to class, \( P(L) \)? We know conditional probabilities of being late to class if we know where we parked: \( P(L|E) = 0.05 \) and \( P(L|E^C) = 0.15 \). Using the “Total”, “And”, and “Not” rules,

\[
P(L) = P(L \text{ and } E) + P(L \text{ and } E^C) \\
= P(E)P(L|E) + P(E^C)P(L|E^C) \\
= P(E)P(L|E) + (1 - P(E))P(L|E^C) \\
= (0.20)(0.05) + (1 - 0.20)(0.15) = 0.13
\]

Rule 4: “Given”: \[ P(B|A) = \frac{P(A \text{ and } B)}{P(A)} \]

**Example:** Given that you are late to class, what is the probability you parked in an E lot? Using previous results,

\[
P(E|L) = \frac{P(E \text{ and } L)}{P(L)} = \frac{0.01}{0.13} = 0.0769
\]

Note that Rule 4 is just Rule 2 rearranged!

In Stat 216, you will be able to solve any probability problem with a hypothetical table or tree diagram; you do not need to use these mathematical probability rules. (Practice by using the table or tree to answer all the questions above.) If the mathematics helps you, great! Otherwise, stick to the table or tree.

6 References