Comment on the “Tree Diagrams” Section

The reversal of conditional probabilities when using tree diagrams (calculating $P(B \mid A)$ from $P(A \mid B)$ and $P(A \mid B^c)$) is an example of Bayes’ formula, named after the 18th century English clergyman Rev. Thomas Bayes. His formula is the foundation for a whole theory of Statistics called Bayesian statistics.

In common-English discussions of statistics and probability, confusing $P(B \mid A)$ and $P(A \mid B)$ is one of the most common mistakes people make! Bayes’ formula provides a resolution of that issue.
Probability Distributions

Random Variable: A numerical measurement of the outcome of a random phenomenon.

Probability Distribution: All possible values and their probabilities.

Discrete if it's possible to express the possible values as a list.

Not all random variables discrete — main alternative continuous.

If $X$ is a discrete random variable (capital letters for the name of the variable) taking values $x$ (small letters for actual numerical values) then the probability distribution is given by specifying the values $P(x)$, where this has the interpretation “the probability that $X$ is equal to $x$”. For this to be valid, we require that $P(x)$ is between 0 and 1 for each possible $x$, and that the sum of $P(x)$ over all $x$ is 1.
Example 1: Let $X$ be the total number of spots in one throw of a die. Then the possible values of $x$ are 1, 2, 3, 4, 5, 6. In this case $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = \frac{1}{6}$.

Example 2: Let $X$ be the total number of home runs the Red Sox hit in a game. In this case the possible values $x$ are 0, 1, 2, 3, 4, 5, 6, 7, 8, .... — there’s no actual limit to how many home runs they could hit, but we can still arrange the possible values in a list so it’s a discrete random variable. According to some data collected during the 2004 season, the probabilities are:

<table>
<thead>
<tr>
<th>$x$</th>
<th>$P(x)$</th>
<th>$x$</th>
<th>$P(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.23</td>
<td>4</td>
<td>0.03</td>
</tr>
<tr>
<td>1</td>
<td>0.38</td>
<td>5</td>
<td>0.01</td>
</tr>
<tr>
<td>2</td>
<td>0.22</td>
<td>$\geq$ 6</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0.13</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In this case the numbers $P(x)$ are all between 0 and 1 and do add up to 1, so they are a consistent probability distribution.
Example 3: Let $X$ be the number of games that UNC win in their next football season. Since this represents a single outcome rather than some experiment that's repeated many time, any probabilities we might assess for this would have to be subjective probabilities. However the concept of a probability distribution and how we might manipulate the numbers are exactly the same as in the example just discussed.
We’ll use Example 2 to illustrate how probability distributions are used. The general idea is to use given values of \(P(x)\) to calculate probabilities of more complicated events.

- **Question:** “What is the probability the Red Sox score at least 3 home runs in a single game?”

  **Answer:** \(0.13 + 0.03 + 0.01 = 0.17\).

- **Question:** “What is the probability the number of home runs is less than two?”

  **Answer:** \(0.23 + 0.38 = 0.61\).

Be careful about the correct usage of phrases such as “at least 3” (the number 3 is included) or “less than 2” (the number 2 is excluded).
We could also ask “What is the average number of home runs the Red Sox hit in a game?” We could give a “relative frequency” calculation by imagining a hypothetical season of exactly 100 games in which the Sox score home runs in exactly the proportion given by the above probability distribution, i.e. 23 games they score none, 38 games they score 1, and so on. Then the total number of runs scored is

\[(0 \times 23) + (1 \times 38) + (2 \times 22) + (3 \times 13) + (4 \times 3) + (5 \times 1)\]

\[= 38 + 44 + 39 + 12 + 5 = 138,\]

which is an average of 1.38 home runs per game.
Alternatively, we could use the formula

\[ \mu = \sum xP(x) \]

which in practice works out as follows:

<table>
<thead>
<tr>
<th>$x$</th>
<th>$P(x)$</th>
<th>$xP(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.23</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0.38</td>
<td>0.38</td>
</tr>
<tr>
<td>2</td>
<td>0.22</td>
<td>0.44</td>
</tr>
<tr>
<td>3</td>
<td>0.13</td>
<td>0.39</td>
</tr>
<tr>
<td>4</td>
<td>0.03</td>
<td>0.12</td>
</tr>
<tr>
<td>5</td>
<td>0.01</td>
<td>0.05</td>
</tr>
<tr>
<td>$\geq 6$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>1.00</td>
<td>1.38</td>
</tr>
</tbody>
</table>

giving the same answer.

The value $\mu$ (Greek mu) is called the *mean or expected value* of the random variable $X$. 


One place the calculation of $\mu$ is useful is in considering the value of insurance policies.

Example: The actuary for an insurance company determines that you have a 5% chance in any one year of requiring repair damage to your car, which we will simplify by saying the payout is exactly $2,000. Alternatively, with a 0.1% chance, the car will have to be replaced completely, and the payout is $20,000. For this policy, the insurance company charges a premium of $150. Is this fair?
In this case let $X$ be “the payout on the policy in a given year”. The possible values and their probabilities are as follows:

<table>
<thead>
<tr>
<th>$x$</th>
<th>$P(x)$</th>
<th>$xP(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.949</td>
<td>0</td>
</tr>
<tr>
<td>2000</td>
<td>0.05</td>
<td>100</td>
</tr>
<tr>
<td>20000</td>
<td>0.001</td>
<td>20</td>
</tr>
<tr>
<td>Total</td>
<td>1.00</td>
<td>120</td>
</tr>
</tbody>
</table>

So the expected payout per policy is $120$. When allowing for the insurance company’s operating costs and a built-in factor for risk, the premium of $150 does not seem unreasonable.

Real insurance calculations are done a lot like that, but of course they are based on many more possible payout scenarios and also, they take into account a lot more individual information (your age, where you live, record of past claims, etc.)
Spread of a random variable

So far we have only talked about the mean of a random variable, which represents the most natural measure of the center of a distribution. However there are other concepts such as median, mode, ..., reflecting our earlier discussion about summarizing the center of a sample of data.

The spread of a distribution may be measured in numerous ways, of which the standard deviation is the most common. Usually the standard deviation is written $\sigma$ (Greek sigma).

We don’t discuss here the formula for computing $\sigma$. However, it will come up when we deal with specific examples later.
Probability Distribution of a Categorical Variable

In principle, categorical variables lie outside the framework of this discussion since the outcomes are not numerical, but we can make them fit by adopting certain conventions. In particular, for a categorical variable with only two outcomes, we can label them 0 and 1 to turn them into a discrete numerical variable.

Example. In an experiment in which rats are exposed to a toxic substance, 35% of them die. We can label the outcomes as 0 for alive and 1 for dead. The probability distribution is then:

<table>
<thead>
<tr>
<th>$x$</th>
<th>$P(x)$</th>
<th>$xP(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.65</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0.35</td>
<td>0.35</td>
</tr>
</tbody>
</table>

The expected value $\mu$ is equal to the probability of the outcome labelled 1.
Continuous Distributions

A random variable is *continuous* if it takes a continuum of possible values. For example if $X$ is “height of a person in inches”, there’s no requirement that is be a whole number of inches.

- A question like “what is the probability that $X = 64$?” is meaningless because nobody has a height exactly 64 inches.
- However *interval probabilities*, such as “what is the probability that $X$ is between 64 and 65?”, make perfect sense.

We often use histograms to summarize the distribution. As the interval width of the histogram gets smaller and smaller (for a very large sample size), the histogram becomes closer and closer to a continuous curve. We often think of this curve as summarizing the information in a distribution. The technical name for this is *probability density function*. 
The Normal Distribution

The *normal distribution* is said to apply when the shape of the distribution follows a bell-shaped curve that is unimodal and symmetric. It is characterized by the mean $\mu$ and the standard deviation $\sigma$.

*Example:* In a sample of men and women it is found that the women have a mean height $\mu = 65$ inches and a standard deviation $\sigma = 3.5$ inches while the men have a mean height $\mu = 70$ inches and a standard deviation $\sigma = 4.0$ inches.

In what interval do nearly all of the men lie?

*Answer:* according to the empirical rule this is $\mu - 3\sigma$ to $\mu + 3\sigma$, or in other words 58 inches to 82 inches.
A more formal statement of this is *the 68–95–99.7 rule*: If the data follow a normal distribution, then

- 68% of the data lie within one standard deviation of the mean
- 95% of the data lie within two standard deviations of the mean
- 99.7% of the data lie within three standard deviations of the mean
Example of SAT scores.

These are designed to have an overall mean of 500 and a standard deviation of 100.

So...

- 68% of all students score between 400 and 600
- 95% of all students score between 300 and 700
- 16% of all students score more than 600