3.2 Conditional Probability and Independent Events

Using population-based health studies to *estimate probabilities* relating potential risk factors to a particular disease, evaluate efficacy of medical diagnostic and screening tests, etc.

**Example:** **Events:** $A = \text{“lung cancer”}$  $B = \text{“smoker”}$

<table>
<thead>
<tr>
<th>Disease Status</th>
<th>Lung cancer ($A$)</th>
<th>No lung cancer ($A^C$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes ($B$)</td>
<td>0.12</td>
<td>0.04</td>
</tr>
<tr>
<td>No ($B^C$)</td>
<td>0.03</td>
<td>0.81</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Smoker</th>
<th>0.16</th>
<th>0.84</th>
</tr>
</thead>
</table>

Probabilities: $P(A) = 0.15$  $P(B) = 0.16$  $P(A \cap B) = 0.12$

**Definition:**

**Conditional Probability of Event $A$, given Event $B$**

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

$$= \frac{0.12}{0.16} = 0.75 \gg 0.15 = P(A)$$

**Comments:**

- $P(B \mid A) = \frac{P(B \cap A)}{P(A)} = \frac{0.12}{0.15} = 0.80$, so $P(A \mid B) \neq P(B \mid A)$ in general.

- General formula can be rewritten: $P(A \cap B) = P(A \mid B) \times P(B) \leftarrow \text{IMPORTANT}$

**Example:**

- $P(\text{Angel barks}) = 0.1$
- $P(\text{Brutus barks}) = 0.2$

Therefore…

$P(\text{Angel and Brutus bark}) = 0.06$
Example: Suppose that two balls are to be randomly drawn, one after another, from a container holding four red balls and two green balls. Under the scenario of sampling without replacement, calculate the probabilities of the events $A = \text{First ball is red}$, $B = \text{Second ball is red}$, and $A \cap B = \text{First ball is red AND second ball is red}$. (As an exercise, list the $6 \times 5 = 30$ outcomes in the sample space of this experiment, and use “brute force” to solve this problem.)

This type of problem – known as an “urn model” – can be solved with the use of a tree diagram, where each branch of the “tree” represents a specific event, conditioned on a preceding event. The product of the probabilities of all such events along a particular sequence of branches is equal to the corresponding intersection probability, via the previous formula. In this example, we obtain the following values:

We can calculate the probability $P(B)$ by adding the two “boxed” values above, i.e., $P(B) = P(A \cap B) + P(A^C \cap B) = 12/30 + 8/30 = 20/30$, or $P(B) = 2/3$.

This last formula – which can be written as $P(B) = P(B \mid A) \ P(A) + P(B \mid A^C) \ P(A^C)$ – can be extended to more general situations, where it is known as the Law of Total Probability, and is a useful tool in Bayes’ Theorem (next section).
Suppose event $C = \text{“coffee drinker.”}$

\[
\begin{array}{ccc}
S & A & C \\
\text{0.09} & \text{0.06} & \text{0.34} \\
\text{0.51} & & \\
\end{array}
\]

### Disease Status

<table>
<thead>
<tr>
<th></th>
<th>Lung cancer ($A$)</th>
<th>No lung cancer ($A^C$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes ($C$)</td>
<td>0.06</td>
<td>0.34</td>
</tr>
<tr>
<td>No ($C^C$)</td>
<td>0.09</td>
<td>0.51</td>
</tr>
</tbody>
</table>

|                  | 0.15              | 0.85                   | 1.00 |

**Probabilities:**

$P(A) = 0.15$ \hspace{1cm} $P(C) = 0.40$ \hspace{1cm} $P(A \cap C) = 0.06$

Therefore,

$$P(A \mid C) = \frac{P(A \cap C)}{P(C)} = \frac{0.06}{0.40} = 0.15 = P(A)$$

i.e., the occurrence of event $C$ gives no information about the probability of event $A$.

**Definition:**

Two events $A$ and $B$ are said to be **statistically independent** if either:

1. $P(A \mid B) = P(A)$, i.e., $P(B \mid A) = P(B)$, or equivalently,
2. $P(A \cap B) = P(A) \times P(B)$.

**Exercise:** Are the events $A = \text{“Angel barks”}$ and $B = \text{“Brutus barks”}$ independent?

**Exercise:** Prove mathematically that two events $A$ and $B$ are independent if and only if $P(A \mid B) = P(A \mid B^C)$. [**Hint:** Use the fact that $P(A \cap B^C) = P(A) - P(A \cap B).$]

**Summary**

$A, B$ disjoint $\iff$ If either event occurs, then the other cannot occur: $P(A \cap B) = 0$.

$A, B$ independent $\iff$ If either event occurs, this gives no information about the other: $P(A \cap B) = P(A) \times P(B)$.

**Example:** $A = \text{“Select a 2”}$ and $B = \text{“Select a ♣”}$ are not disjoint events, because $A \cap B = \{2 \clubsuit\} \neq \emptyset$. However, $P(A \cap B) = 1/52 = 1/13 \times 1/4 = P(A) \times P(B)$; hence they are independent events. Can two disjoint events ever be independent? Why?
**Experiment 4 - revisited:** Recall this example from the previous section where, at a party, guests randomly select one pastry from each of two trays. Assuming that their selections are **statistically independent** from one another, characterize the distribution of the **sum** $S = X_1 + X_2$ calories.

### Probability Tables

<table>
<thead>
<tr>
<th>$x$</th>
<th>$f_1(x)$</th>
<th>$+$</th>
<th>$x$</th>
<th>$f_2(x)$</th>
<th>$=$</th>
</tr>
</thead>
<tbody>
<tr>
<td>90</td>
<td>1/3</td>
<td></td>
<td>30</td>
<td>3/6</td>
<td></td>
</tr>
<tr>
<td>120</td>
<td>1/3</td>
<td></td>
<td>60</td>
<td>2/6</td>
<td></td>
</tr>
<tr>
<td>150</td>
<td>1/3</td>
<td></td>
<td>90</td>
<td>1/6</td>
<td></td>
</tr>
</tbody>
</table>

Mean($X_1$) = $\mu_1 = 120$ cals;  
Mean($X_2$) = $\mu_2 = 50$ cals;  
Var($X_1$) = $\sigma_1^2 = 600$ cals$^2$;  
Var($X_2$) = $\sigma_2^2 = 500$ cals$^2$

### Events and Sample Space

<table>
<thead>
<tr>
<th>$S$</th>
<th>Sample Space</th>
</tr>
</thead>
<tbody>
<tr>
<td>120</td>
<td>(90, 30)</td>
</tr>
<tr>
<td>150</td>
<td>(90, 60), (120, 30)</td>
</tr>
<tr>
<td>180</td>
<td>(90, 90), (120, 60), (150, 30)</td>
</tr>
<tr>
<td>210</td>
<td>(120, 90), (150, 60)</td>
</tr>
<tr>
<td>240</td>
<td>(150, 90)</td>
</tr>
</tbody>
</table>

Mean($S$) = $\mu_S = 120\left(\frac{3}{18}\right) + 150\left(\frac{5}{18}\right) + 180\left(\frac{6}{18}\right) + 210\left(\frac{3}{18}\right) + 240\left(\frac{1}{18}\right) = 170$ cals

= $\mu_1 + \mu_2$

Var($S$) = $\sigma_S^2 = (-50)^2\left(\frac{3}{18}\right) + (-20)^2\left(\frac{5}{18}\right) + (10)^2\left(\frac{6}{18}\right) + (40)^2\left(\frac{3}{18}\right) + (70)^2\left(\frac{1}{18}\right) = 1100$ cals$^2$

= $\sigma_1^2 + \sigma_2^2$
Ismor Fischer, 8/11/2008

Same party, same pastries. Again assuming independence between random selections from the two trays, characterize the distribution of the difference \( D = X_1 - X_2 \) calories.

<table>
<thead>
<tr>
<th>Tray 1</th>
<th>Tray 2</th>
<th>Events</th>
<th>Sample Space</th>
</tr>
</thead>
<tbody>
<tr>
<td>90</td>
<td>90</td>
<td>( D = 0 ):</td>
<td>(90, 90)</td>
</tr>
<tr>
<td>120</td>
<td>120</td>
<td>( D = 30 ):</td>
<td>(90, 60), (120, 90)</td>
</tr>
<tr>
<td>150</td>
<td>150</td>
<td>( D = 60 ):</td>
<td>(90, 30), (120, 60), (150, 90)</td>
</tr>
<tr>
<td>120</td>
<td>30</td>
<td>( D = 90 ):</td>
<td>(120, 30), (150, 60)</td>
</tr>
<tr>
<td>150</td>
<td>30</td>
<td>( D = 120 ):</td>
<td>(150, 30)</td>
</tr>
</tbody>
</table>

### Probability Tables

\[
\begin{array}{cc}
  x & f_1(x) \\
  90 & 1/3 \\
  120 & 1/3 \\
  150 & 1/3 \\
\end{array}
\quad
\begin{array}{cc}
  x & f_2(x) \\
  30 & 3/6 \\
  60 & 2/6 \\
  90 & 1/6 \\
\end{array}
\]

- Mean\((X_1) = \mu_1 = 120\) cals;
- Var\((X_1) = \sigma_1^2 = 600\) cals\(^2\);
- Mean\((X_2) = \mu_2 = 50\) cals;
- Var\((X_2) = \sigma_2^2 = 500\) cals\(^2\).

### Events

<table>
<thead>
<tr>
<th>( d )</th>
<th>( f(d) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1/18 = Exercise</td>
</tr>
<tr>
<td>30</td>
<td>3/18 = Exercise</td>
</tr>
<tr>
<td>60</td>
<td>6/18 = Exercise</td>
</tr>
<tr>
<td>90</td>
<td>5/18 = Exercise</td>
</tr>
<tr>
<td>120</td>
<td>3/18 = Exercise</td>
</tr>
</tbody>
</table>

### Exercise:
Sketch the probability histogram of \( D \), and verify the following:

- \( \text{Mean}(D) = \mu_D = 0 \left( \frac{1}{18} \right) + 30 \left( \frac{3}{18} \right) + 60 \left( \frac{6}{18} \right) + 90 \left( \frac{5}{18} \right) + 120 \left( \frac{3}{18} \right) = 70 \text{ cals} = \mu_1 - \mu_2 \)
- \( \text{Var}(D) = \sigma_D^2 = (-70)^2 \left( \frac{1}{18} \right) + (-40)^2 \left( \frac{3}{18} \right) + (-10)^2 \left( \frac{6}{18} \right) + (20)^2 \left( \frac{5}{18} \right) + (50)^2 \left( \frac{3}{18} \right) \)
  \[ = 1100 \text{ cals}^2 = \sigma_1^2 + \sigma_2^2 \]

### GENERAL FACT ~

\[
\text{Mean}(X + Y) = \text{Mean}(X) + \text{Mean}(Y) \quad \text{and} \quad \text{Mean}(X - Y) = \text{Mean}(X) - \text{Mean}(Y)
\]

If \( X \) and \( Y \) are independent random variables,

\[
\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) \quad \text{and} \quad \text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y).
\]

### Comments:

- The difference relations will play an important role in 6.2 – Two Samples inference.
- If \( X \) and \( Y \) are dependent, then the two bottom relations regarding the variance also involve an additional term, Cov\((X, Y)\), the covariance between \( X \) and \( Y \).
**Exercise:** Construct the **probability table** and **probability histogram** for both independent random variables $X, Y$ below, and their difference $D = X - Y$, respectively.

```
<table>
<thead>
<tr>
<th></th>
<th>X</th>
<th></th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>60</td>
<td>40</td>
<td>30</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>40</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

Calculate the means $\mu_X, \mu_Y, \mu_D$, and verify that $\mu_D = \mu_X - \mu_Y$. Also calculate the variances $\sigma_X^2, \sigma_Y^2, \sigma_D^2$, and verify that $\sigma_D^2 = \sigma_X^2 + \sigma_Y^2$.

[Note that the variance relation can be interpreted visually via the Pythagorean Theorem. This is not a superficial coincidence, but illustrates an important geometric connection, expanded upon in the Appendix.]

```
\sigma_D
\sigma_Y
\sigma_X
```

**Optional:** Repeat these calculations with the sum variable $S = X + Y$. Verify that $\mu_S = \mu_X + \mu_Y$ and $\sigma_S^2 = \sigma_X^2 + \sigma_Y^2$. 
More on Conditional Probability and Independent Events

Another example from epidemiology

Suppose that, in a certain study population, we wish to investigate the prevalence of lung cancer \( (A) \), and its associations with obesity \( (B) \) and cigarette smoking \( (C) \), respectively. From the first of the two stylized Venn diagrams above, by comparing the scales drawn, observe that the proportion of the size of the intersection \( A \cap B \) (green) relative to event \( B \) (blue + green), is about equal to the proportion of the size of event \( A \) (yellow + green) relative to the entire population \( S \).

That is,

\[
P(A \cap B) = \frac{P(A)}{P(S)}.
\]

(As an exercise, verify this equality for the following probabilities: yellow = .09, green = .07, blue = .37, white = .47, to two decimals, before reading on.) In other words, the probability that a randomly chosen person from the obese subpopulation has lung cancer, is equal to the probability that a randomly chosen person from the general population has lung cancer (.16). This equation can be equivalently expressed as

\[
P(A | B) = P(A),
\]

since the left side is conditional probability by definition, and \( P(S) = 1 \) in the denominator of the right side. In this form, the equation clearly conveys the interpretation that knowledge of event \( B \) (obesity) yields no information about event \( A \) (lung cancer). In this example, lung cancer is equally probable (.16) among the obese as it is among the general population, so knowing that a person is obese is completely unrevealing with respect to having lung cancer. Events \( A \) and \( B \) that are related in this way are said to be independent. Note that they are not disjoint!

In the second diagram however, the relative size of \( A \cap C \) (orange) to \( C \) (red + orange), is larger than the relative size of \( A \) (yellow + orange) to the whole population \( S \), so \( P(A | C) \neq P(A) \), i.e., events \( A \) and \( C \) are dependent. Here, as is true in general, the probability of lung cancer is indeed influenced by whether a person is randomly selected from among the general population or the smoking subset, where it is much higher. (Statistically, lung cancer would be a rare disease in the U.S., if not for cigarettes (although it is on the rise among nonsmokers for unclear reasons).
Application: “Are Blood Antibodies Independent?”
An example of conditional probability in human genetics
(Adapted from Rick Chappell, Ph.D., UW Dept. of Biostatistics & Medical Informatics)

Background: The surfaces of human red blood cells (“erythrocytes”) are coated with antigens that are classified into four disjoint blood types: O, A, B, and AB. Each type is associated with blood serum antibodies for the other types, that is,

- Type O blood contains both A and B antibodies.
  (This makes Type O the “universal donor”, but capable of receiving only Type O.)
- Type A blood contains only B antibodies.
- Type B blood contains only A antibodies.
- Type AB blood contains neither A nor B antibodies.
  (This makes Type AB the “universal recipient”, but capable of donating only to Type AB.)

In addition, blood is also classified according to the presence (+) or absence (−) of Rh factor (found predominantly in rhesus monkeys, and to varying degree in human populations; they are important in obstetrics). Hence there are eight distinct blood groups corresponding to this joint classification system: O⁺, O⁻, A⁺, A⁻, B⁺, B⁻, AB⁺, AB⁻. According to the American Red Cross, the U.S. population has the following blood group relative frequencies:

<table>
<thead>
<tr>
<th>Rh factor</th>
<th>+</th>
<th>−</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blood Types</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O</td>
<td>.384</td>
<td>.077</td>
<td>.461</td>
</tr>
<tr>
<td>A</td>
<td>.323</td>
<td>.065</td>
<td>.388</td>
</tr>
<tr>
<td>B</td>
<td>.094</td>
<td>.017</td>
<td>.111</td>
</tr>
<tr>
<td>AB</td>
<td>.032</td>
<td>.007</td>
<td>.039</td>
</tr>
<tr>
<td>Totals</td>
<td>.833</td>
<td>.166</td>
<td>.999</td>
</tr>
</tbody>
</table>

From these values (and from the background information above), we can calculate the following probabilities:

\[
P (A \text{ antibodies}) = P (\text{Type O or B}) = P (O) + P (B) = .461 + .111 = .572
\]

\[
P (B \text{ antibodies}) = P (\text{Type O or A}) = P (O) + P (A) = .461 + .388 = .849
\]

\[
P (B \text{ antibodies and Rh}^+) = P (\text{Type O}^+ \text{ or A}^+) = P (O^+) + P (A^+) = .384 + .323 = .707
\]
Using these calculations, we can answer the following.

**Question:** Is having “A antibodies” independent of having “B antibodies”?

**Solution:** We must check whether or not

\[ P(A \text{ and } B \text{ antibodies}) = P(A \text{ antibodies}) \times P(B \text{ antibodies}), \]

i.e.,

\[ P(\text{Type O}) \times .572 \times .849 \]

or

\[ .461 \times .486 \]

This indicates near independence of the two events; there does exist a slight dependence. The dependence would be much stronger if America were composed of two disjoint (i.e., non-interbreeding) groups: Type A (with B antibodies only) and Type B (with A antibodies only), and no Type O (with both A and B antibodies). Since this is evidently not the case, the implication is that either these traits evolved before humans spread out geographically, or they evolved later but the populations became mixed in America.

**Question:** Is having “B antibodies” independent of “Rh\(^+\)”?

**Solution:** We must check whether or not

\[ P(\text{B antibodies and Rh}^+) = P(\text{B antibodies}) \times P(\text{Rh}^+), \]

that is,

\[ .707 = .849 \times .833, \]

which is true, so we have exact independence of these events. These traits probably predate diversification in humans (and were not differentially selected for since).

**Exercises:**

- Is having “A antibodies” independent of “Rh\(^+\)”?
- Find \( P(\text{A antibodies} \mid \text{B antibodies}) \) and \( P(\text{B antibodies} \mid \text{A antibodies}) \). Conclusions?
- Is “Blood Type” independent of “Rh factor”? (Do a separate calculation for each blood type: O, A, B, AB, and each Rh factor: +, −.)