Today: Chapter 2 (continued)

- Generalizing the “or” rule to non-disjoint events
- Joint, marginal and conditional probabilities
- Generalizing the “and” rule to non-independent events
- Reading these probabilities in contingency tables
- Encode natural language into probabilistic statements
Recap of Last Lecture

• **“or” rule:** If two events $A$ and $B$ are disjoint,
  \[ P(A \text{ or } B) = P(A) + P(B). \]

Generalization: If $A_1, \ldots, A_n$ are disjoint,
  \[ P(A_1 \text{ or } \ldots \text{ or } A_n) = P(A_1) + \ldots + P(A_n). \]

• **“and” rule:** If two events $A$ and $B$ are independent,
  \[ P(A \text{ and } B) = P(A) \times P(B). \]

Generalization: If $A_1, \ldots, A_n$ are mutually independent,
  \[ P(A_1 \text{ and } \ldots \text{ and } A_n) = P(A_1) \times \ldots \times P(A_n). \]
“or” rule: Non-Disjoint Events

In general, for any two events $A$ and $B$,

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B).$$

Caution: “or” is inclusive!

“A or B” = “A but not B, B but not A, A and B simultaneously”.
Non-Disjoint Events: Example

80% of college students like learning. 70% of college student like video games. 62% like both learning and video games. What percent like learning or video games?

Let $L$ be the event that a college student likes learning.
Let $V$ be the event that a college student likes video games.

$$P(L \text{ or } V) = P(L) + P(V) - P(L \text{ and } V)$$
$$= 0.8 + 0.7 - 0.62$$
$$= 0.88.$$

The picture on the right is called a Venn diagram.
Describe in words the zone given by:

- **0.18** People who like learning but not video games.
- **0.08 + 0.62** People who like video games.
- **0.12** People who dislike learning and video games.
- **0.18 + 0.08** People who like learning only or video games only
Contingency Table

<table>
<thead>
<tr>
<th></th>
<th>Like video games</th>
<th>Dislike video games</th>
<th>Margin totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Like Learning</td>
<td>0.62</td>
<td>0.18</td>
<td>0.8</td>
</tr>
<tr>
<td>Dislike Learning</td>
<td>0.08</td>
<td>0.12</td>
<td>0.2</td>
</tr>
<tr>
<td>Margin Totals</td>
<td>0.7</td>
<td>0.3</td>
<td>1</td>
</tr>
</tbody>
</table>

- **Joint Probabilities** are probabilities corresponding to two things happening simultaneously.
  Here: 0.62, 0.18, 0.08, 0.12

- **Marginal Probabilities** are probabilities corresponding to the outcome of one variable.
  Here: 0.8, 0.2 (for $L$) and 0.7, 0.3 (for $V$).
Contingency Table: Example

Among the students enrolled in the class,
- 41 are Junior with CS major
- 87 are neither Junior nor with CS major
- 131 have Major other than CS
- 107 are not Junior

What is the probability that a randomly chosen student is CS Major?

<table>
<thead>
<tr>
<th></th>
<th>Junior</th>
<th>Other Levels</th>
<th>Margin totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS Major</td>
<td>41</td>
<td>20</td>
<td>61</td>
</tr>
<tr>
<td>Other Major</td>
<td></td>
<td>87</td>
<td>131</td>
</tr>
<tr>
<td>Margin Totals</td>
<td></td>
<td>107</td>
<td>192</td>
</tr>
</tbody>
</table>

\[ p = \frac{61}{192} \approx 31.77\% \]
Conditional Probability is a tool to handle events that are not independent.

Idea: When computing a probability, you actually have some extra information that you know is true.

Example:
- $A = \{ \text{It will rain today in San Diego} \}$
- $B = \{ \text{You see dark storm clouds in the sky} \}$

Although $P(A) \simeq 11.2\%$ is small, you have a much higher chance to see $A$ happen if you know already that $B$ occurred.
Conditional Probability

A card is drawn from a deck. What is the probability that the card is a heart, given that the card is a king?

Intuition says 1/4.

\[
P(A \text{ given that } B \text{ occurred}) = \frac{P(A \text{ and } B)}{P(B)}
\]
For two events $A, B$ the conditional probability of $A$ given $B$ is

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

Computing $P(A|B)$ amounts to do as if $B$ was the sample space.
Notion of Independence Revisited

By definition, $A$ and $B$ are independent when

$$P(A \text{ and } B) = P(A) \times P(B).$$

But for any two events $A$ and $B$, we have

$$P(A) = P(A|B) \times P(B).$$

Therefore,

$A$ and $B$ are independent $\iff P(A|B) = P(A)$
$\iff P(B|A) = P(B).$
Independent or not?

1. Find $P(A)$
2. Find $P(A$ assuming you know event $B$ has occurred) = $P(A|B)$

Do you get the same answer? = Check if $P(A|B) = P(A)$

- **Yes:** Events are independent
- **No:** Events are NOT independent (= dependent)
Checking for Independence: Example

A poll led on randomly picked North Carolina residents yields the following figures:

\[ P(\text{resident says gun ownership protects citizen}) = 0.58 \]

\[ P(\text{says guns protect citizens } | \text{ is White}) = 0.67 \]

\[ P(\text{says guns protect citizens } | \text{ is Black}) = 0.28 \]

\[ P(\text{says guns protect citizens } | \text{ is Hispanic}) = 0.64 \]

The opinion on gun ownership varies by ethnicity, therefore the variables Opinion-on-guns and Ethnicity seem to be dependent.

\textbf{Caution:} Mind sample size!
Checking for Independence on Sample Data

• If conditional probabilities computed based on sample data suggest dependence between two variables, the next step is to conduct a **hypothesis test** to determine if the observed difference is **significant** or not
  ( = to determine that this difference is likely to have been created by the sampling procedure or not)

• If the observed difference between the conditional probabilities is large, then there is stronger evidence that the difference is real.

• If the sample is large, then even a small difference can provide strong evidence of a real difference.

→ See Chapter 4 later in the course.
email contains data on 3921 emails sent to one user over 3 months. For each case:

- Write an expression for the given probability
- Say if this probability is marginal, joint, or conditional

- What percent of messages are spam with no number? $P(\text{spam and no number})$, joint probability

$$P(\text{spam and no number}) = \frac{149}{3921} \approx 3.8\%$$
Natural Language and Probabilities

<table>
<thead>
<tr>
<th>none</th>
<th>small</th>
<th>big</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>400</td>
<td>2659</td>
</tr>
<tr>
<td>1</td>
<td>149</td>
<td>168</td>
</tr>
</tbody>
</table>

• What is the probability that a spam message will have a small number?
  \[ P(\text{small number} \mid \text{spam}) \], conditional probability

  \[ P(\text{small number} \mid \text{spam}) = \frac{168}{149 + 168 + 50} \approx 45.8\% \]

• What is the likelihood that a randomly-selected message with a big number will not be spam?
  \[ P(\text{not spam} \mid \text{big number}) \], conditional probability

  \[ P(\text{not spam} \mid \text{big number}) = \frac{495}{495 + 50} \approx 90.8\% \]
Natural Language and Probabilities

```
data("email")
table(email$spam, email$number)
none small big
0  400  2659  495
1  149  168   50
```

- What percent of messages do not contain a small number, ignoring the categorization of spam?\
  \[ P(\text{not small}) \], marginal probability\
  \[
P(\text{not small}) = \frac{400 + 149 + 495 + 50}{3921} \approx 27.9\%\
  \]

- What fraction of emails would we expect to be non-spam with small or big numbers?\
  \[ P(\text{not spam and not none}) \], joint probability\
  \[
P(\text{not spam and not none}) = \frac{2659 + 495}{3921} \approx 80.4\%\
  \]
What You Should do Now

• Turn in Homework 1!
• Start Homework 2 (due Friday, 13th 12:50pm)
• Finish reading Chapter 2.