Probabilities and Data

Digging into Data: Jordan Boyd-Graber

University of Maryland

February 3, 2014

Slides adapted from Dave Blei and Lauren Hannah
Roadmap

- What are probabilities
  - Discrete
  - Continuous
- How to manipulate probabilities
- Properties of probabilities
Probabilities are the language we use to describe data

A reasonable (but geeky) definition of data science is how to get probabilities we care about from data

Later classes will be about how to do this for different probability models and different types of data

But first, we need key definitions of probability
Probabilities are the language we use to describe data

A reasonable (but geeky) definition of data science is how to get probabilities we care about from data

Later classes will be about how to do this for different probability models and different types of data

But first, we need key definitions of probability

So pay attention!
Preface: Why make us do this?

- Probabilities are the language we use to describe data
- A reasonable (but geeky) definition of data science is how to get probabilities we care about from data
- Later classes will be about how to do this for different probability models and different types of data
- But first, we need key definitions of probability
- So pay attention!
- Also, ya’ll need to get your environments set up
Outline

1. Properties of Probability Distributions
2. Working with probability distributions
3. Combining Probability Distributions
4. Continuous Distributions
5. Expectation and Entropy
Random variable

- Probability is about *random variables*.
- A random variable is any “probabilistic” outcome.
- For example,
  - The flip of a coin
  - The height of someone chosen randomly from a population
- We’ll see that it’s sometimes useful to think of quantities that are not strictly probabilistic as random variables.
  - The temperature on 11/12/2013
  - The temperature on 03/04/1905
  - The number of times “streetlight” appears in a document
Random variable

- Random variables take on values in a *sample space*.
- They can be *discrete* or *continuous*:
  - Coin flip: \( \{H, T\} \)
  - Height: positive real values \((0, \infty)\)
  - Temperature: real values \((-\infty, \infty)\)
  - Number of words in a document: Positive integers \(\{1, 2, \ldots\}\)
- We call the outcomes *events*.
- Denote the random variable with a capital letter; denote a realization of the random variable with a lower case letter.
- E.g., \(X\) is a coin flip, \(x\) is the value \((H\text{ or } T)\) of that coin flip.
A discrete distribution assigns a probability to every event in the sample space. For example, if $X$ is an (unfair) coin, then

$$P(X = H) = 0.7$$
$$P(X = T) = 0.3$$

And probabilities have to be greater than 0.

Probabilities of disjunctions are sums over part of the space. E.g., the probability that a die is bigger than 3:

$$P(D > 3) = P(D = 4) + P(D = 5) + P(D = 6)$$

The probabilities over the entire space must sum to one.
Discrete distribution

- A discrete distribution assigns a probability to every event in the sample space.
- For example, if $X$ is an (unfair) coin, then
  \[ P(X = H) = 0.7 \]
  \[ P(X = T) = 0.3 \]
- And probabilities have to be greater than 0.
- Probabilities of disjunctions are sums over part of the space. E.g., the probability that a die is bigger than 3:
  \[ P(D > 3) = P(D = 4) + P(D = 5) + P(D = 6) \]
- The probabilities over the entire space must sum to one.
A discrete distribution assigns a probability to every event in the sample space.

For example, if $X$ is an (unfair) coin, then

\[
\begin{align*}
P(X = H) &= 0.7 \\
P(X = T) &= 0.3
\end{align*}
\]

And probabilities have to be greater than 0.

Probabilities of disjunctions are sums over part of the space. E.g., the probability that a die is bigger than 3:

\[
P(D > 3) = P(D = 4) + P(D = 5) + P(D = 6)
\]

The probabilities over the entire space must sum to one.

\[
\sum P(X = x) = 1
\]
A discrete distribution assigns a probability to every event in the sample space.

For example, if $X$ is an (unfair) coin, then

$$P(X = H) = 0.7$$
$$P(X = T) = 0.3$$

And probabilities have to be greater than 0.

Probabilities of disjunctions are sums over part of the space. E.g., the probability that a die is bigger than 3:

$$P(D > 3) = P(D = 4) + P(D = 5) + P(D = 6)$$

The probabilities over the entire space must sum to one.

$$\sum_{x} P(X = x) = 1$$
Outline

1. Properties of Probability Distributions
2. Working with probability distributions
3. Combining Probability Distributions
4. Continuous Distributions
5. Expectation and Entropy
Events

An *event* is a set of outcomes to which a probability is assigned

- drawing a black card from a deck of cards
- drawing a King of Hearts

Intersections and unions:

- Intersection: drawing a red and a King

\[ P(A \cap B) \]

- Union: drawing a spade or a King

\[ P(A \cup B) = P(A) + P(B) - P(A \cap B) \]
An event is a set of outcomes to which a probability is assigned

- drawing a black card from a deck of cards
- drawing a King of Hearts

Intersections and unions:

- Intersection: drawing a red and a King
  \[ P(A \cap B) \]  

- Union: drawing a spade or a King
  \[ P(A \cup B) = P(A) + P(B) - P(A \cap B) \]
An *event* is a set of outcomes to which a probability is assigned

- drawing a black card from a deck of cards
- drawing a King of Hearts

Intersections and unions:

- Intersection: drawing a red and a King
  
  \[ P(A \cap B) \] (1)

- Union: drawing a spade or a King
  
  \[ P(A \cup B) = P(A) + P(B) - P(A \cap B) \] (2)
Typically, we consider collections of random variables.

The joint distribution is a distribution over the configuration of all the random variables in the ensemble.

For example, imagine flipping 4 coins. The joint distribution is over the space of all possible outcomes of the four coins.

\[
\begin{align*}
  P(HHHH) &= 0.0625 \\
  P(HHHT) &= 0.0625 \\
  P(HHTH) &= 0.0625 \\
  \vdots
\end{align*}
\]

You can think of it as a single random variable with 16 values.
Visualizing a joint distribution

\[ \sim x \]

\[ \sim x, \sim y \]

\[ \sim x, y \quad x, y \quad x, \sim y \]
Marginalization

If we are given a joint distribution, what if we are only interested in the distribution of one of the variables?

We can compute the distribution of $P(X)$ from $P(X, Y, Z)$ through marginalization:

$$
\sum_y \sum_z P(X, Y = y, Z = z) = \sum_y \sum_z P(X)P(Y = y, Z = z | X)
$$

$$
= P(X) \sum_y \sum_z P(Y = y, Z = z | X)
$$

$$
= P(X)
$$
Marginalization (from Leyton-Brown)

Marginalization allows us to compute distributions over smaller sets of variables:

\[ P(X, Y) = \sum_z P(X, Y, Z = z) \]

- Marginalize out weather
- Marginalize out temperature

<table>
<thead>
<tr>
<th>temperature (T) and weather (W)</th>
<th>T=Hot</th>
<th>T=Mild</th>
<th>T=Cold</th>
</tr>
</thead>
<tbody>
<tr>
<td>W=Sunny</td>
<td>.10</td>
<td>.20</td>
<td>.10</td>
</tr>
<tr>
<td>W=Cloudy</td>
<td>.05</td>
<td>.35</td>
<td>.20</td>
</tr>
</tbody>
</table>

Corresponds to summing out a table dimension

New table still sums to 1
Marginalization (from Leyton-Brown)

### Joint distribution

<table>
<thead>
<tr>
<th>temperature (T) and weather (W)</th>
<th>T=Hot</th>
<th>T=Mild</th>
<th>T=Cold</th>
</tr>
</thead>
<tbody>
<tr>
<td>W=Sunny</td>
<td>.10</td>
<td>.20</td>
<td>.10</td>
</tr>
<tr>
<td>W=Cloudy</td>
<td>.05</td>
<td>.35</td>
<td>.20</td>
</tr>
</tbody>
</table>

Marginalization allows us to compute distributions over smaller sets of variables:

- $P(X, Y) = \sum_z P(X, Y, Z = z)$
- Corresponds to summing out a table dimension
- New table still sums to 1

- Marginalize out weather

<table>
<thead>
<tr>
<th>T=Hot</th>
<th>T=Mild</th>
<th>T=Cold</th>
</tr>
</thead>
</table>

- Marginalize out temperature
Marginalization (from Leyton-Brown)

<table>
<thead>
<tr>
<th>Temperature (T) and Weather (W)</th>
<th>T=Hot</th>
<th>T=Mild</th>
<th>T=Cold</th>
</tr>
</thead>
<tbody>
<tr>
<td>W=Sunny</td>
<td>.10</td>
<td>.20</td>
<td>.10</td>
</tr>
<tr>
<td>W=Cloudy</td>
<td>.05</td>
<td>.35</td>
<td>.20</td>
</tr>
</tbody>
</table>

Marginalization allows us to compute distributions over smaller sets of variables:

\[ P(X, Y) = \sum_z P(X, Y, Z = z) \]

- Marginalize out weather
- Marginalize out temperature

New table still sums to 1
# Marginalization (from Leyton-Brown)

## Joint distribution

<table>
<thead>
<tr>
<th>temperature (T) and weather (W)</th>
<th>T=Hot</th>
<th>T=Mild</th>
<th>T=Cold</th>
</tr>
</thead>
<tbody>
<tr>
<td>W=Sunny</td>
<td>.10</td>
<td>.20</td>
<td>.10</td>
</tr>
<tr>
<td>W=Cloudy</td>
<td>.05</td>
<td>.35</td>
<td>.20</td>
</tr>
</tbody>
</table>

Marginalization allows us to compute distributions over smaller sets of variables:

- \( P(X, Y) = \sum_z P(X, Y, Z = z) \)
- Corresponds to summing out a table dimension
- New table still sums to 1

- Marginalize out weather

<table>
<thead>
<tr>
<th>T=Hot</th>
<th>T=Mild</th>
<th>T=Cold</th>
</tr>
</thead>
<tbody>
<tr>
<td>.15</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Marginalize out temperature
Marginalization allows us to compute distributions over smaller sets of variables:

\[ P(X, Y) = \sum_z P(X, Y, Z = z) \]

- Marginalize out weather

\[
\begin{array}{ccc}
T=Hot & T=Mild & T=Cold \\
W=Sunny & .15 & .55 & .30 \\
W=Cloudy & .10 & .20 & .10 \\
\end{array}
\]

- Marginalize out temperature

\[
\begin{array}{ccc}
T=Hot & T=Mild & T=Cold \\
.15 & .55 & .30 \\
\end{array}
\]

New table still sums to 1
Marginalization (from Leyton-Brown)

Joint distribution

<table>
<thead>
<tr>
<th>temperature (T) and weather (W)</th>
<th>T=Hot</th>
<th>T=Mild</th>
<th>T=Cold</th>
</tr>
</thead>
<tbody>
<tr>
<td>W=Sunny</td>
<td>.10</td>
<td>.20</td>
<td>.10</td>
</tr>
<tr>
<td>W=Cloudy</td>
<td>.05</td>
<td>.35</td>
<td>.20</td>
</tr>
</tbody>
</table>

Marginalization allows us to compute distributions over smaller sets of variables:

- \( P(X, Y) = \sum_z P(X, Y, Z = z) \)
- Corresponds to summing out a table dimension
- New table still sums to 1

- Marginalize out weather
  - T=Hot  | T=Mild | T=Cold
  - .15    | .55    | .30

- Marginalize out temperature
  - W=Sunny
  - W=Cloudy
Marginalization (from Leyton-Brown)

**Joint distribution**

<table>
<thead>
<tr>
<th>temperature (T) and weather (W)</th>
<th>T=Hot</th>
<th>T=Mild</th>
<th>T=Cold</th>
</tr>
</thead>
<tbody>
<tr>
<td>W=Sunny</td>
<td>.10</td>
<td>.20</td>
<td>.10</td>
</tr>
<tr>
<td>W=Cloudy</td>
<td>.05</td>
<td>.35</td>
<td>.20</td>
</tr>
</tbody>
</table>

Marginalization allows us to compute distributions over smaller sets of variables:

- \( P(X, Y) = \sum_z P(X, Y, Z = z) \)
- Corresponds to summing out a table dimension
- New table still sums to 1

- **Marginalize out weather**
<table>
<thead>
<tr>
<th>T=Hot</th>
<th>T=Mild</th>
<th>T=Cold</th>
</tr>
</thead>
<tbody>
<tr>
<td>.15</td>
<td>.55</td>
<td>.30</td>
</tr>
</tbody>
</table>

- **Marginalize out temperature**
<table>
<thead>
<tr>
<th>W=Sunny</th>
<th>W=Cloudy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Marginalization (from Leyton-Brown)

**Joint distribution**

<table>
<thead>
<tr>
<th>temperature (T) and weather (W)</th>
<th>T=Hot</th>
<th>T=Mild</th>
<th>T=Cold</th>
</tr>
</thead>
<tbody>
<tr>
<td>W=Sunny</td>
<td>.10</td>
<td>.20</td>
<td>.10</td>
</tr>
<tr>
<td>W=Cloudy</td>
<td>.05</td>
<td>.35</td>
<td>.20</td>
</tr>
</tbody>
</table>

Marginalization allows us to compute distributions over smaller sets of variables:

- \( P(X, Y) = \sum_z P(X, Y, Z = z) \)
- Corresponds to summing out a table dimension
- New table still sums to 1

Marginalize out weather

<table>
<thead>
<tr>
<th>T=Hot</th>
<th>T=Mild</th>
<th>T=Cold</th>
</tr>
</thead>
<tbody>
<tr>
<td>.15</td>
<td>.55</td>
<td>.30</td>
</tr>
</tbody>
</table>

Marginalize out temperature

<table>
<thead>
<tr>
<th>W=Sunny</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>.40</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>W=Cloudy</th>
</tr>
</thead>
</table>
Marginalization (from Leyton-Brown)

**Joint distribution**

<table>
<thead>
<tr>
<th>temperature (T)</th>
<th>T=Hot</th>
<th>T=Mild</th>
<th>T=Cold</th>
</tr>
</thead>
<tbody>
<tr>
<td>W=Sunny</td>
<td>.10</td>
<td>.20</td>
<td>.10</td>
</tr>
<tr>
<td>W=Cloudy</td>
<td>.05</td>
<td>.35</td>
<td>.20</td>
</tr>
</tbody>
</table>

Marginalization allows us to compute distributions over smaller sets of variables:

- \( P(X, Y) = \sum_z P(X, Y, Z = z) \)
- Corresponds to summing out a table dimension
- New table still sums to 1

- **Marginalize out weather**
<table>
<thead>
<tr>
<th>T=Hot</th>
<th>T=Mild</th>
<th>T=Cold</th>
</tr>
</thead>
<tbody>
<tr>
<td>.15</td>
<td>.55</td>
<td>.30</td>
</tr>
</tbody>
</table>

- **Marginalize out temperature**
<table>
<thead>
<tr>
<th>W=Sunny</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>.40</td>
<td></td>
</tr>
<tr>
<td>W=Cloudy</td>
<td>.60</td>
</tr>
</tbody>
</table>
The *conditional probability* of event $A$ given event $B$ is the probability of $A$ when $B$ is known to occur,

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

The *conditional probability* of event $A$ given event $B$ is the probability of $A$ when $B$ is known to occur,

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$
The conditional probability of event $A$ given event $B$ is the probability of $A$ when $B$ is known to occur,

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$
The **conditional probability** of event $A$ given event $B$ is the probability of $A$ when $B$ is known to occur,

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

The *conditional probability* of event $A$ given event $B$ is the probability of $A$ when $B$ is known to occur,

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$
The *conditional probability* of event $A$ given event $B$ is the probability of $A$ when $B$ is known to occur,

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

What is the probability that the sum of two dice is six given that the first is greater than three?
Conditional Probabilities

Example

What is the probability that the sum of two dice is six given that the first is greater than three?

- $A \equiv$ First die
- $B \equiv$ Second die

<table>
<thead>
<tr>
<th></th>
<th>B=1</th>
<th>B=2</th>
<th>B=3</th>
<th>B=4</th>
<th>B=5</th>
<th>B=6</th>
</tr>
</thead>
<tbody>
<tr>
<td>A=1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>A=2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>A=3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>A=4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>A=5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>A=6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
</tr>
</tbody>
</table>
### Conditional Probabilities

#### Example

What is the probability that the sum of two dice is six given that the first is greater than three?

- **A** ≡ First die
- **B** ≡ Second die

<table>
<thead>
<tr>
<th></th>
<th>B=1</th>
<th>B=2</th>
<th>B=3</th>
<th>B=4</th>
<th>B=5</th>
<th>B=6</th>
</tr>
</thead>
<tbody>
<tr>
<td>A=1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>A=2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>A=3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>A=4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>A=5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>A=6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
</tr>
</tbody>
</table>
Conditional Probabilities

**Example**

What is the probability that the sum of two dice is six given that the first is greater than three?

- $A$ $\equiv$ First die
- $B$ $\equiv$ Second die

<table>
<thead>
<tr>
<th></th>
<th>B=1</th>
<th>B=2</th>
<th>B=3</th>
<th>B=4</th>
<th>B=5</th>
<th>B=6</th>
</tr>
</thead>
<tbody>
<tr>
<td>A=1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>A=2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>A=3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>A=4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>A=5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>A=6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
</tr>
</tbody>
</table>

$P(A > 3 \cap B + A = 6) =$

$P(A > 3) =$

$P(A > 3 | B + A = 6) =$
**Example**

What is the probability that the sum of two dice is six given that the first is greater than three?

- $A \equiv$ First die
- $B \equiv$ Second die

<table>
<thead>
<tr>
<th></th>
<th>B=1</th>
<th>B=2</th>
<th>B=3</th>
<th>B=4</th>
<th>B=5</th>
<th>B=6</th>
</tr>
</thead>
<tbody>
<tr>
<td>A=1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>A=2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>A=3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>A=4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>A=5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>A=6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
</tr>
</tbody>
</table>

\[
P(A > 3 \cap B + A = 6) = \frac{2}{36}
\]

\[
P(A > 3) =
\]

\[
P(A > 3 | B + A = 6) =
\]
Conditional Probabilities

Example

What is the probability that the sum of two dice is six given that the first is greater than three?

- \( A \equiv \) First die
- \( B \equiv \) Second die

\[
P(A > 3 \cap B + A = 6) = \frac{2}{36}
\]

\[
P(A > 3) = \frac{3}{6}
\]

\[
P(A > 3 \mid B + A = 6) =
\]
Example
What is the probability that the sum of two dice is six given that the first is greater than three?

- A ≡ First die
- B ≡ Second die

\[
P(A > 3 \cap B + A = 6) = \frac{2}{36}
\]

\[
P(A > 3) = \frac{3}{6}
\]

\[
P(A > 3 \mid B + A = 6) = \frac{\frac{2}{36}}{\frac{3}{36}} = \frac{2}{6} \times \frac{3}{36} = \frac{2}{36} = \frac{6}{36}
\]
Conditional Probabilities

Example

What is the probability that the sum of two dice is six given that the first is greater than three?

- $A \equiv$ First die
- $B \equiv$ Second die

<table>
<thead>
<tr>
<th></th>
<th>B=1</th>
<th>B=2</th>
<th>B=3</th>
<th>B=4</th>
<th>B=5</th>
<th>B=6</th>
</tr>
</thead>
<tbody>
<tr>
<td>A=1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>A=2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>A=3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>A=4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>A=5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>A=6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
</tr>
</tbody>
</table>

$P(A > 3 \cap B + A = 6) = \frac{2}{36}$

$P(A > 3) = \frac{3}{6}$

$P(A > 3 | B + A = 6) = \frac{\frac{2}{36}}{\frac{3}{6}} = \frac{2}{6} \div \frac{36}{3} = \frac{1}{9}$
Outline

1. Properties of Probability Distributions
2. Working with probability distributions
3. Combining Probability Distributions
4. Continuous Distributions
5. Expectation and Entropy
The definition of conditional probability lets us derive the *chain rule*, which lets us define the joint distribution as a product of conditionals:

\[
P(X, Y) = P(X, Y) \frac{P(Y)}{P(Y)}
\]
The chain rule

The definition of conditional probability lets us derive the chain rule, which lets us define the joint distribution as a product of conditionals:

\[ P(X, Y) = P(X, Y) \frac{P(Y)}{P(Y)} = P(X|Y)P(Y) \]
The chain rule

The definition of conditional probability lets us derive the chain rule, which lets us define the joint distribution as a product of conditionals:

\[
P(X, Y) = P(X, Y) \frac{P(Y)}{P(Y)} = P(X|Y)P(Y)
\]

For example, let \( Y \) be a disease and \( X \) be a symptom. We may know \( P(X|Y) \) and \( P(Y) \) from data. Use the chain rule to obtain the probability of having the disease and the symptom.

In general, for any set of \( N \) variables

\[
P(X_1, \ldots, X_N) = \prod_{n=1}^{N} P(X_n|X_1, \ldots, X_{n-1})
\]
What is the relationship between $P(A|B)$ and $P(B|A)$?

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

1. Start with $P(A|B)$
2. Change outcome space from $B$ to $\Omega$
3. Change outcome space again from $\Omega$ to $A$
Bayes’ Rule

What is the relationship between $P(A|B)$ and $P(B|A)$?

\[ P(B|A) = \frac{P(A|B)P(B)}{P(A)} \]

1. Start with $P(A|B)$
2. Change outcome space from $B$ to $\Omega$
3. Change outcome space again from $\Omega$ to $A$
Bayes’ Rule

What is the relationship between $P(A|B)$ and $P(B|A)$?

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

1. Start with $P(A|B)$
2. Change outcome space from $B$ to $\Omega$: $P(A|B)P(B)$
3. Change outcome space again from $\Omega$ to $A$

![Diagram showing the relationship between $P(A|B)$ and $P(B|A)$]
Bayes’ Rule

What is the relationship between $P(A|B)$ and $P(B|A)$?

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

1. Start with $P(A|B)$
2. Change outcome space from $B$ to $\Omega$: $P(A|B)P(B)$
3. Change outcome space again from $\Omega$ to $A$: $\frac{P(A|B)P(B)}{P(A)}$
Random variables $X$ and $Y$ are independent if and only if
\[ P(X = x, Y = y) = P(X = x)P(Y = y). \]

Conditional probabilities equal unconditional probabilities with independence:
- $P(X = x | Y) = P(X = x)$
- *Knowing $Y$ tells us nothing about $X$*
Random variables $X$ and $Y$ are independent if and only if

$$P(X = x, Y = y) = P(X = x)P(Y = y).$$

Conditional probabilities equal unconditional probabilities with independence:

- $P(X = x | Y) = P(X = x)$
- *Knowing Y tells us nothing about X*

Mathematical examples:

- If I draw two socks from my (multicolored) laundry, is the color of the first sock independent from the color of the second sock?
Independence

Random variables $X$ and $Y$ are independent if and only if

$$P(X = x, Y = y) = P(X = x)P(Y = y).$$

Conditional probabilities equal unconditional probabilities with independence:

- $P(X = x \mid Y) = P(X = x)$
- *Knowing $Y$ tells us nothing about $X$*

Mathematical examples:

- If I draw two socks from my (multicolored) laundry, is the color of the first sock independent from the color of the second sock?

- If I flip a coin twice, is the first outcome independent from the second outcome?
Independence

Intuitive Examples:

- Independent:
  - you use a Mac / the Green line is on schedule
  - snowfall in the Himalayas / your favorite color is blue

- Not independent:
  - you vote for Mitt Romney / you are a Republican
  - there is a traffic jam on the Beltway / the Redskins are playing
Intuitive Examples:

- Independent:
  - you use a Mac / the Green line is on schedule
  - snowfall in the Himalayas / your favorite color is blue

- Not independent:
  - you vote for Mitt Romney / you are a Republican
  - there is a traffic jam on the Beltway / the Redskins are playing
Independence

Sometimes we make convenient assumptions.

- the values of two dice
- the value of the first die and the sum of the values
- whether it is raining and the number of taxi cabs
- whether it is raining and the amount of time it takes me to hail a cab
- the first two words in a sentence
Continuous random variables

- We’ve only used discrete random variables so far (e.g., dice)
- Random variables can be continuous.
- We need a \( p(x) \), which integrates to one.
  
  E.g., if \( x \in \mathbb{R} \) then
  \[
  \int_{-\infty}^{\infty} p(x) \, dx = 1
  \]

- Probabilities are integrals over smaller intervals. E.g.,
  \[
  P(X \in (-2.4, 6.5)) = \int_{-2.4}^{6.5} p(x) \, dx
  \]

- Notice when we use \( P, p, X, \) and \( x \).
Continuous random variables

- We’ve only used discrete random variables so far (e.g., dice)
- Random variables can be continuous.
- We need a density $p(x)$, which integrates to one.
  E.g., if $x \in \mathbb{R}$ then
  $$\int_{-\infty}^{\infty} p(x) \, dx = 1$$
- Probabilities are integrals over smaller intervals. E.g.,
  $$P(X \in (-2.4, 6.5)) = \int_{-2.4}^{6.5} p(x) \, dx$$
- Notice when we use $P$, $p$, $X$, and $x$.
- Integrals? I didn’t sign up for this!
Integrals?
Integrals?
Integrals?
The Gaussian (or Normal) is a continuous distribution.

\[ p(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\} \]

- The density of a point \( x \) is proportional to the negative exponentiated half distance to \( \mu \) scaled by \( \sigma^2 \).
- \( \mu \) is called the mean; \( \sigma^2 \) is called the variance.
The mean $\mu$ controls the location of the bump.

The variance $\sigma^2$ controls the spread of the bump.
Outline

1. Properties of Probability Distributions
2. Working with probability distributions
3. Combining Probability Distributions
4. Continuous Distributions
5. Expectation and Entropy
An *expectation* of a random variable is a weighted average:

\[
E[f(X)] = \sum_{x=1}^{\infty} f(x) p(x) \quad \text{(discrete)}
\]

\[
= \int_{-\infty}^{\infty} f(x) p(x) \, dx \quad \text{(continuous)}
\]
Expectations of constants or known values:

- $E[a] = a$
- $E[Y \mid Y = y] = y$
Example: Gaussian distribution $X \sim N(\mu, \sigma^2)$

$$E[X] = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \, dx$$
Expectation

Example: Gaussian distribution $X \sim N(\mu, \sigma^2)$

$$E[X] = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \, dx$$

$$= \mu$$
Expectation Intuition

- Average or outcome (might not be an event: 2.4 children)
- Center of mass

"Fair Price" of a wager
What is the expectation of the roll of die?

\[
\text{One die: } 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5.
\]

What is the expectation of the sum of two dice?

\[
\text{Two die: } 2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} + 4 \cdot \frac{3}{36} + 5 \cdot \frac{4}{36} + 6 \cdot \frac{5}{36} + 7 \cdot \frac{6}{36} + 8 \cdot \frac{5}{36} + 9 \cdot \frac{4}{36} + 10 \cdot \frac{3}{36} + 11 \cdot \frac{2}{36} + 12 \cdot \frac{1}{36} = 7.
\]
What is the expectation of the roll of die?

**One die**

\[
1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} =
\]
What is the expectation of the roll of die?

**One die**

\[
1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5
\]
What is the expectation of the roll of die?

**One die**

$$1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$$

What is the expectation of the sum of two dice?
What is the expectation of the roll of die?

**One die**

\[
1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5
\]

What is the expectation of the sum of two dice?

**Two die**

\[
2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} + 4 \cdot \frac{3}{36} + 5 \cdot \frac{4}{36} + 6 \cdot \frac{5}{36} + 7 \cdot \frac{6}{36} + 8 \cdot \frac{5}{36} + 9 \cdot \frac{4}{36} + 10 \cdot \frac{3}{36} + 11 \cdot \frac{2}{36} + 12 \cdot \frac{1}{36} = \]
### Expectation of die / dice

What is the expectation of the roll of die?

#### One die

$$1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$$

What is the expectation of the sum of two dice?

#### Two die

$$2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} + 4 \cdot \frac{3}{36} + 5 \cdot \frac{4}{36} + 6 \cdot \frac{5}{36} + 7 \cdot \frac{6}{36} + 8 \cdot \frac{5}{36} + 9 \cdot \frac{4}{36} + 10 \cdot \frac{3}{36} + 11 \cdot \frac{2}{36} + 12 \cdot \frac{1}{36} = 7$$
Entropy

- Measure of disorder in a system
- In the real world, entropy in a system tends to increase
- Can also be applied to probabilities:
  - Is one (or a few) outcomes certain (low entropy)
  - Are things equiprobable (high entropy)
- In data science
  - We look for features that allow us to reduce entropy (decision trees)
  - All else being equal, we seek models that have maximum entropy (Occam’s razor)
Aside: Logarithms

- \( \log(x) = b \iff 2^b = x \)
- Makes big numbers small
- Way to think about them: cutting a carrot

\[
\begin{align*}
\log(1) &= 0 \\
\log(2) &= 1 \\
\log(4) &= 2 \\
\log(8) &= 3
\end{align*}
\]
Aside: Logarithms

\[ \lg(x) = b \iff 2^b = x \]

- Makes big numbers small
- Way to think about them: cutting a carrot
- Negative numbers?

\[ \begin{align*}
\lg(1) &= 0 \\
\lg(2) &= 1 \\
\lg(4) &= 2 \\
\lg(8) &= 3
\end{align*} \]
Aside: Logarithms

- \( \log(x) = b \iff 2^b = x \)
- Makes big numbers small
- Way to think about them: cutting a carrot
- Negative numbers?
- Non-integers?

\[
\begin{align*}
\log(1) &= 0 \\
\log(2) &= 1 \\
\log(4) &= 2 \\
\log(8) &= 3 \\
\end{align*}
\]
Entropy

*Entropy* is a measure of uncertainty that is associated with the distribution of a random variable:

\[
H(X) = -E[\lg(p(X))]
\]

\[
= - \sum_x p(x) \lg(p(x)) \quad \text{(discrete)}
\]

\[
= - \int_{-\infty}^{\infty} p(x) \lg(p(x)) \, dx \quad \text{(continuous)}
\]
Entropy

Entropy is a measure of uncertainty that is associated with the distribution of a random variable:

\[
H(X) = -E[\log(p(X))]
\]

\[
= - \sum_x p(x) \log(p(x)) \quad \text{(discrete)}
\]

\[
= - \int_{-\infty}^{\infty} p(x) \log(p(x)) \, dx \quad \text{(continuous)}
\]

Does not account for the values of the random variable, only the spread of the distribution.

- \( H(X) \geq 0 \)
- uniform distribution \( \equiv \) highest entropy, point mass \( \equiv \) lowest
- suppose \( P(X = 1) = p, \ P(X = 0) = 1 - p \) and
- \( P(Y = 100) = p, \ P(Y = 0) = 1 - p \): \( X \) and \( Y \) have the same entropy
Entropy

Entropy is a measure of uncertainty that is associated with the distribution of a random variable:

\[
H(X) = -E[\lg(p(X))]
\]

\[
= -\sum_x p(x) \lg(p(x)) \quad \text{(discrete)}
\]

\[
= -\int_{-\infty}^{\infty} p(x) \lg(p(x)) \, dx \quad \text{(continuous)}
\]

Does not account for the values of the random variable, only the spread of the distribution.

- \(H(X) \geq 0\)
- uniform distribution = highest entropy, point mass = lowest
- suppose \(P(X = 1) = p, P(X = 0) = 1 - p\) and \(P(Y = 100) = p, P(Y = 0) = 1 - p\): \(X\) and \(Y\) have the same entropy

Examples (in class)

Entropy of one die, two dice.
That’s it for now

You don’t have to be an expert on this stuff (there are other classes for that)

This is to get your feet wet and to know the concepts when you see the math
Next Time

- Technological foundations
- Dealing with messy data
- Telling stories with data