**PROBABILITY: Quick Review**

Two conceptions:

- **Certainty/Likelihood of particular events**
  - Example: \( \Pr(\text{particular coin flip = heads}) \)
  - Example: \( \Pr(\text{i}^{th} \text{ word in a certain text is a noun}) \)

- **Relative frequency of a certain kind of outcome in an infinitely large number of repetitions of some "repeatable experiment"**
  - Example: proportion of "heads" in 1000000 coin flips
  - Example: proportion of nouns in an 1000000 text corpus

- **In estimating probabilities, we use the latter conception.**

- **In applying (estimated) probabilities, we use the first conception.**

---

**Random Variables**

- A r.v. picks out some observable in a repeatable experiment
- Its value varies from repetition to repetition
  - Example: "Flip" could be \( H \) (for heads), \( T \) (for tails) in repeated coin tosses
  - Example: "CAT" could be \( N, V, P, A, D, t, \ldots \) (i.e., POS) for a word found in some arbitrary location in a corpus

\[
\Pr(\text{Flip} = H) = .497
\]

- says that the prob. of heads in flipping some coin is 49.7%

\[
\Pr(\text{CAT} = N) = .24
\]

- says that the prob. that an arbitrarily selected word in some large corpus (or in some type of corpus) is a noun is 24%

**Abbreviation:**

- \( \Pr(H) = .497 \), \( \Pr(N) = .24 \)
- r.v. is "understood"
Jointly distributed r.v.'s

- We can consider multiple r.v.'s at the same time, coming from one experiment or multiple experiments.
- The values they assume may depend on one another (because there is some causal connection between the r.v.'s), or independent. (We'll formalize this.)

E.g., Let \((X_1, X_2)\) be r.v.'s for the outcomes of 2 successive rolls of a die.
So, values are \((1, 1), (1, 2), \ldots, (6, 6)\)
For a fair die & causally independent, randomized rolls, 
\[ \Pr(X_1 = 1, X_2 = 1) = \Pr(X_1 = 1, X_2 = 2) = \ldots = \Pr(X_1 = 6, X_2 = 6) = \frac{1}{36} \]  "joint probability"

E.g., Let \((C_1, C_2, C_3)\) be r.v.'s for the POS of 3 successive words randomly selected from a large corpus.
Possible values: \((N, N, N)\), \((N, N, V)\), \ldots, \((Adv, Adv, Adv)\)
Interdependent: 
\[ \Pr(N, N, N) \neq (24)^3 \]
\[ \Pr(C_1 = N, C_2 = N, C_3 = N) \]

Probability Axioms

1. The probability of any outcome lies between 0 and 1
\[ 0 \leq \Pr(x = x) \leq 1 \]
where \(x\) is any possible value of r.v. \(X\)

2. The probabilities of all possible alternative outcomes add up to 1. E.g., 
\[ \Pr(\text{FLIP} = H) + \Pr(\text{FLIP} = T) = 1 \]
\[ \sum_{x \in \text{range}(X)} \Pr(x = x) = 1 \]
Informally, 
1. \(0 \leq \Pr(x) \leq 1\)
2. \(\sum_x \Pr(x) = 1\)

Similarly for jointly distributed r.v.'s:
1. \(0 \leq \Pr(x) \leq 1\)
2. \(\sum_x \Pr(x) = 1\)

where \(\Pr(x)\) abbreviates \(\Pr(X = x)\) and 
\[ X = (X_1, \ldots, X_n), \text{ i.e., } n \text{ r.v.'s, and } \]
\[ x = (x_1, \ldots, x_n), \text{ i.e., an } n\text{-tuple of possible values of } X_1, \ldots, X_n. \]
Conditional Probability

For jointly distributed r.v.'s, we may be interested in how probable a certain value is for one of them, **given** the value of another. Write as

\[
\Pr(x=x|y=y), \text{ or briefly, } \Pr(x|y)
\]

E.g., Prob. that \(i\)th word of a text is a noun, \(\text{given}\) that the preceding word is a determiner might be written \(\Pr(C_i = \text{N} | C_{i-1} = \text{Det})\).

Def. \(\Pr(x=x|y=y) = \frac{\Pr(x=x, y=y)}{\Pr(y=y)}\) \(\text{provided}\) \(\Pr(y=y) \neq 0\)

E.g., suppose \(\Pr(C_{i-1} = \text{Det}, C_i = \text{N}) = 0.2\), i.e., when we randomly select a pair of successive words from a corpus, 20% of the time they are a Det, N pair & suppose \(\Pr(C_i = \text{Det}) = 0.25\), i.e., when we randomly select a word from a corpus, 25% of the time it is a Det

Then \(\Pr(C_i = \text{N} | C_{i-1} = \text{Det}) = \frac{\Pr(C_i = \text{N}, C_{i-1} = \text{Det})}{\Pr(C_{i-1} = \text{Det})} = 0.80\)

Chain rule

\[
\Pr(a \& b) = \Pr(a) \Pr(b | a)
\]

from def. of cond. prob

Similarly

\[
\Pr(a \& b \& c) = \Pr(a) \Pr(b | a) \Pr(c | a \& b)
\]

Useful esp. if we can make certain independence assumptions. Here, suppose \(a, c\) are conditionally independent, \(\text{given}\) \(b\):

\[
\Pr(c | a \& b) = \Pr(c | b)
\]

Then

\[
\Pr(a \& b \& c) = \Pr(a) \Pr(b | a) \Pr(c | b)
\]

We can easily generalize to \(n\) events or r.v.'s.

E.g., Suppose \(C_1, \ldots, C_n\) represent the parts of speech of a stretch of \(n\) successive words randomly selected from a text corpus.

\[
\Pr(C_1, \ldots, C_n) = \Pr(C_1) \Pr(C_2 | C_1) \Pr(C_3 | C_2, C_1) \ldots \Pr(C_n | C_{n-1}, \ldots, C_1)
\]

Suppose \(C_i\) is conditionally independent of all earlier parts of speech, given the immediately preceding one:

\[
\Pr(C_i | C_{i-1}, \ldots, C_1) = \Pr(C_i | C_{i-1})
\]

Then

\[
\Pr(C_1, \ldots, C_n) = \Pr(C_1) \prod_{i=2}^{n} \Pr(C_i | C_{i-1})
\]
"Marginal" (absolute) probabilities in terms of joint probabilities

\[ Pr(a) = Pr(a \land b) + Pr(a \land \neg b) \]

More generally, suppose \( b_1, \ldots, b_n \) are mutually exclusive (disjoint), jointly exhaustive events (total prob. 1). Then

\[ Pr(a) = \sum_{i=1}^{n} Pr(a \land b_i) = \sum_{i=1}^{n} Pr(b_i) Pr(a | b_i) \]

"Marginalization"

**Independence**

Intuitively, two events \( A, B \) are independent if the fact that one occurred (or failed to occur) tells us nothing about the probability that the other occurred; i.e.,

- \( Pr(A | B) = Pr(A) \) "\( B \) provides no evidence about \( A \)"
- \( Pr(A | B) = Pr(A) \)
- \( Pr(B | A) = Pr(B) \)
- \( Pr(B | A) = Pr(B) \)
- \( Pr(A \land B) = Pr(A) Pr(B) \)
- \( Pr(A \land B) = Pr(A) Pr(B) \)
- etc.

E.g., Flip a coin twice. Possible outcomes \[ \{ \langle H, H \rangle, \langle H, T \rangle, \langle T, H \rangle, \langle T, T \rangle \} \]

If all 4 probs are \( \frac{1}{4} \), then \( A, B \) are independent.

**Independent r.v.'s**

E.g., For the above prob. space, let

- \( X(\langle H, H \rangle) = X(\langle H, T \rangle) = 1, X(\langle T, H \rangle) = X(\langle T, T \rangle) = 0 \)
- \( Y(\langle H, H \rangle) = Y(\langle H, T \rangle) = 1, Y(\langle T, H \rangle) = Y(\langle T, T \rangle) = 0 \)

i.e., "\( X = 1 \)" if 1st toss is \( H \), & 0 o.w.

"\( Y = 1 \)" if 2nd — \( \cdots \)
Independent r.v.'s, cont'd

So again, if \( P(X=1) = P(X=0) = P(Y=1) = P(Y=0) \),

\( X, Y \) are independent r.v.'s; & we have

\[ P(X=x \land Y=y) = P(X=x) P(Y=y) \text{ for } x,y \in \{0,1\} \]

Also \( P(X=x \mid Y=y) = P(X=x) \text{ for } x,y \in \{0,1\} \), etc.

Often abbreviations are used, such as

\( P(x) \) for \( P(X=x) \), & similarly for \( Y \).

So, for independent \( X, Y \),

\[ P(x,y) = P(x) P(y) \text{, etc.} \]

Another example

Space = CS grad students

- \( X \) = verbal GRE score (up to 800)
- \( Y \) = analytical GRE score (up to 800)
- \( Z \) = 4.44 grade (up to 4.0)

These are probably nearly independent

Conditional independence

\[ P(Z=z \mid X=x, Y=y) = P(Z=z \mid X=x) \]

provides no further evidence about \( Z \), beyond what \( X=x \) provides

Bayes' Rule

\[ P(A \land B) = P(A) P(B \mid A) \]

\[ = P(B) P(A \mid B) \]

\[ \therefore P(B \mid A) \quad \text{prob. of cold, given sore throat} \]

\[ = \frac{P(B) P(A \mid B)}{P(A)} \quad (1) \]

mutually exclusive (disjoint)

\[ = \frac{P(B) P(A \mid B)}{P(A) + P(B) P(A \mid B)} \quad (2) \]

\[ P(B \mid A) = \frac{P(B) P(A \mid B)}{P(B) P(A \mid B) + P(\neg B) P(A \mid \neg B)} \quad (3) \]

E.g., if we know:

\( P(B) = P(\text{cold}) = .01 \) Prior prob of cold hypothesis

\( P(A \mid B) = P(\text{sore throat} \mid \text{cold}) = .9 \) Likelihood of evidence, given the cold hypothesis

\( P(\neg B) = P(\text{sore throat} \mid \text{no cold}) = .005 \) Likelihood of evidence, given falsity of cold hypothesis
Then we can compute $P(\text{hypoth.} \mid \text{new evidence})$:

$$P(B \mid A) = P(\text{cold (sore throat)})$$

**posterior probability of cold hypothesis**

prior

$$\frac{0.1 \cdot 0.9}{0.1 \cdot 0.9 + (1 - 0.1) \cdot 0.005}$$

= 0.645

**NB:** much bigger than prior, 0.1

**Odds - Likelihood form of Bayes' Rule**

Above we needed 3 numbers to compute $P(B \mid A)$, but it really only takes 2 . . .

By analogy with (3):

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

(3')

Divide (3) by (3!) (Note: same denominator!):

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

$$O(B \mid A) = O(B) \cdot L(A \mid B)$$

(4) 

O "odds" 

L "likelihood ratio"

- where the "odds" in favor of an event (or proposition) $A$ is the ratio of its probability of occurrence to its probability of nonoccurrence

$$O(A) = \frac{P(A)}{P(\overline{A})} = \frac{P(A)}{1 - P(A)}$$

Similarly for conditional (or posterior) odds,

$$O(B \mid A) = \frac{P(B \mid A)}{P(\overline{B} \mid A)} = \frac{P(B \mid A)}{1 - P(B \mid A)}$$

- and the likelihood ratio $L(A \mid B)$ is the ratio of the likelihood of the evidence, given the hypothesis, to the likelihood of the evidence, given the denial of the hypothesis

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

How to recover probabilities from (4):

From $O(B) = \frac{P(B)}{1 - P(B)}$, it's easily shown that

$$P(B) = \frac{O(B)}{1 + O(B)}$$

Similarly

$$P(B \mid A) = \frac{O(B \mid A)}{1 + O(B \mid A)}$$
Combining Evidence

E.g., might have 2 pieces of evidence for a "cold" hypothesis:

\[
P(H|E_1 \land E_2) = \frac{P(H|E_1) P(E_2|H)}{P(H|E_1) P(E_2|H) + P(\neg H|E_1) P(E_2|\neg H)}
\]

Write down Bayes rule for \(P(H|E_1)\) again:

\[
P(H|E_1) = \frac{P(H) P(E_1|H)}{P(H) P(E_1|H) + P(\neg H) P(E_1|\neg H)}
\]  

Take this posterior prob. as the prior prob. for the next piece of evidence, \(E_2\):

\[
P(H|E_1 \land E_2) = \frac{P(H|E_1) P(E_2|H \land E_1)}{P(H|E_1) P(E_2|H \land E_1) + P(\neg H|E_1) P(E_2|\neg H \land E_1)}
\]

Now assume that relative to a situation where the hypothesis \(H\) is given (known to be true), \(E_1\) & \(E_2\) are independent (conditional independence):

\[
P(E_2|H \land E_1) = P(E_2|H)
\]

Similarly if \(\neg H\) is given (\(H\) is known to be false):

\[
P(E_2|\neg H \land E_1) = P(E_2|\neg H)
\]

Then (5) simplifies to:

\[
P(H|E_1 \land E_2) = \frac{P(H|E_1) P(E_2|H)}{P(H|E_1) P(E_2|H) + P(\neg H|E_1) P(E_2|\neg H)}
\]

So, we already know how to compute \(P(H|E_1)\) using \(P(H), P(E_1|H),\) & \(P(E_1|\neg H)\)

So (6) shows us how to allow for both pieces of evidence, \(E_1\) & \(E_2\), with the additional data \(P(E_2|H), P(E_2|\neg H)\)

E.g., let's use \(P(\text{cold}) = .01\), \(P(\text{sore throat} | \text{cold}) = .9, \)
\(P(\text{sore throat} | \text{\neg cold}) = .005\) as before,
so we get

\[
P(\text{cold} | \text{sore throat}) = .645
\]
as before.

Now assume new evidence "sniffles", where

\[
P(\text{sniffles} | \text{cold}) = .95
\]
\[P(\text{sniffles} | \text{\neg cold}) = .1\]

From (6):

\[
P(\text{cold} | \text{sore throat} \land \text{sniffles}) = \frac{.645 (.95)}{.645 (.95) + (1-.645) (.1)} = .945\text{ (note the increase!)}
\]
We can get an odds-likelihood version as before: Write an expression for

\[ \frac{P(H|E, \neg E_2)}{P(\neg H|E, \neg E_2)} \]

analogous to (6). Call this (6'). Divide (6)/(6'):

\[
\frac{P(H|E, \neg E_2)}{P(\neg H|E, \neg E_2)} = \frac{P(H|E_1, H)P(E_2|H)}{P(\neg H|E_1)P(E_2|\neg H)}
\]

\[ O(H|E, \neg E_2) = O(H|E_1) \times L(E_2|H) \]

But \[ O(H|E_1) = O(H) \times L(E_1|H) \]

So

\[ O(H|E, \neg E_2) = O(H) \times L(E_1|H) \times L(E_2|H) \]  

(7)

To "update" the odds in favor of \( H \) as new evidence \( E_1, E_2, \ldots \) comes in, we just multiply by the corresponding likelihood ratios \( L(E_1|H), L(E_2|H), \ldots \) (but keep in mind the conditional independence assumption!)