Lesson Plan

1. Bayes Theorem

2. Simpson’s Paradox

3. More worked problems
Why Study Probability?

A probability model describes the random phenomenon that determines how data is produced.

Say we want to know if a coin is fair.

- **Unknown parameter**: probability of heads
- **Population**: infinite number of tosses
- **Sample**: 10 tosses and 4 are heads
- **Estimate**: sample estimate $4/10 = 0.40$

We will later learn how to quantify the uncertainty in our estimate and if it is considerably different from 0.5.
Bayes’ Theorem

Let’s consider the events $A$ and $B$. By the conditional probability formula we have

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

\[ P(A \text{ and } B) = P(B|A)P(A). \]

By combining the two formulas we have

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

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

This is a simplified version of the Bayes’ theorem first proposed by Thomas Bayes in 1763.
Bayes Theorem

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

- We establish the relation between \( P(A|B) \) and \( P(B|A) \).
- The textbook calls this tree reversal.

This is very important in Bayesian analysis which we will see later in the course. Suppose we are interested in a parameter \( \theta \) and we observed some data.

\[ P(\theta|\text{data}) = \frac{P(\text{data}|\theta)P(\theta)}{P(\text{data})} \]

where \( P(\text{data}|\theta) \) describes how the data is generated and \( P(\theta) \) describes our prior information on \( \theta \).
Paternity Suits

Legal cases of disputed paternity in many countries are resolved using blood tests. Laboratories make genetic determinations concerning the mother, child, and alleged father. Suppose you are on a jury considering a paternity suit,

- The mother has blood type O.
- The alleged father has blood type AB.
- The child has blood type B.
Paternity Suits

Here’s some information we need to solve the problem. According to genetics,

- There is a 50% chance that this child will have blood type B if this alleged father is the real father.

- Based on percentage of B genes in the population, there is a 9% chance that this child would have blood type B if this alleged father is not the real father.

- Based on other evidence (e.g., testimonials, physical evidence, records) presented before the DNA test, you believe there is a 75% chance that the alleged father is the real father. This assessment is your prior belief.
Paternity Suits

Given:

- \( P(\text{son} = B | \text{father} = \text{true}) = 0.5 \)
- \( P(\text{son} = B | \text{father} = \text{false}) = 0.09 \)
- \( P(F = \text{true}) = 0.75 \)

Our goal is

\[
P(\text{father} = \text{true} | \text{son} = B) = \frac{P(\text{father} = \text{true} \text{ and } \text{son} = B)}{P(\text{son} = B)}
\]

where

\[
P(\text{son} = B) = P(\text{son} = B \text{ and } \text{father} = \text{true}) + P(\text{son} = B \text{ and } \text{father} = \text{false})
\]

(Draw a tree to see this)
Paternity Suits

Given:
- $P(son = B|father = true) = 0.5$
- $P(son = B|father = false) = 0.09$
- $P(F = true) = 0.75$

Therefore:
- $P(F = false) = 0.25$
- $P(son = B \text{ and } father = false) = 0.5 \times 0.75$
- $P(son = B \text{ and } father = false) = 0.9 \times 0.25$
- $P(son = B) = 0.5 \times 0.75 + 0.9 \times 0.25$

Finally
- $P(father = true|son = B) = \frac{0.50 \times 0.75}{0.5 \times 0.75 + 0.09 \times 0.25} = 0.9434$
Paternity Suits - Impact of Prior

We see that

$$P(\text{father} = \text{true} | \text{son} = B) = \frac{0.50 \times \rho}{0.5 \times \rho + 0.09 \times (1 - \rho)}$$

where $\rho$ is the prior belief that the child belong to the father.
Roommate Example

The simple version of Bayes’ Rule is just an application of the definition of conditional probability.

- Suppose that 90% of statistics majors are looking for a roommate, compared with 10% of econ majors and 50% of bio majors.
- And suppose that Duke has only those three majors, with 25% of the students majoring in stats, 45% majoring in econ, and the rest in bio (with no double majors).

Suppose your friend finds a roommate. What is the probability that it is a statistics major?
Roommate Example

From the definition of conditional probability,

\[ P( \text{stat} \mid \text{roommate} ) = \frac{P( \text{stat and roommate} )}{P( \text{roommate} )}. \]

For the numerator,

\[ P( \text{stat and roommate} ) = P( \text{roommate} \mid \text{stat} ) \times P( \text{stat} ) = 0.9 \times 0.25 \]

For the denominator, (let \( d \) denote looking for roommate),

\[ P(\text{roommate}) = P(\text{s and d} \text{ or e and d} \text{ or b and d}) \]
\[ = P(\text{s and d}) + P(\text{e and d}) + P(\text{b and d}) \]
\[ = P(d|s) \times P(s) + P(d|e) \times P(e) + P(d|b) \times P(b) \]
\[ = 0.9 \times 0.25 + 0.1 \times 0.45 + 0.5 \times 0.3 \]
Roommate Example

Putting everything together we have,

\[
P(\text{stat} \mid \text{roommate}) = \frac{P(\text{stat and roommate})}{P(\text{roommate})} = \frac{0.9 \times 0.25}{0.9 \times 0.25 + 0.1 \times 0.45 + 0.5 \times 0.3} = 0.5357
\]

We can similarly compute:

\[
P(\text{econ} \mid \text{roommate}) = \frac{P(\text{econ and roommate})}{P(\text{roommate})} = \frac{0.1 \times 0.45}{0.9 \times 0.25 + 0.1 \times 0.45 + 0.5 \times 0.3} = 0.107
\]
Roommate Example

\[ P(\text{bio} | \text{roommate}) = \frac{P(\text{econ and roommate})}{P(\text{roommate})} \]

\[ = \frac{0.5 \times 0.35}{0.9 \times 0.25 + 0.1 \times 0.45 + 0.5 \times 0.3} \]

\[ = 0.357 \]

Note that

\[ P(\text{stat}|\text{roommate}) + P(\text{econ}|\text{roommate}) + P(\text{bio}|\text{roommate}) = 1 \]

\[ \text{Reverse conditioning - do you see a pattern yet?} \]
Law of Total Probability

In general, let $A_1, \ldots, A_n$ be mutually exclusive and suppose that $P(A_1 \text{ or } A_2 \text{ or } \ldots \text{ or } A_n) = 1$.

Then the Bayes formula is:

$$P(A_1|B) = \frac{P(B|A_1) \times P(A_1)}{P(B)} = \frac{P(B|A_1) \times P(A_1)}{\sum_{i=1}^{n} P(B|A_i) \times P(A_i)}. \tag{15/24}$$

Note the reversal from $P(B|A_k)$ to $P(A_k|B)$ requires the calculation of $P(B)$ using $P(B|A_k)$.
HIV Test

ELISA is a test for HIV.

► If a person has HIV, ELISA has probability .997 of resulting in a positive test (true positive).
► If a person does not have HIV, then ELISA is negative with probability .985. (true negative)
► About 0.32% of the U.S. population has HIV.

Suppose you get an HIV test (e.g., as part of a marriage license). Your test comes back positive. What is the chance that you have HIV?

We can use Bayes Rule. Let $A_1 = \{\text{have HIV}\}$ and let $A_2 = \{\text{do not have HIV}\}$. 
HIV Test

\[ P[A_1 | \text{pos}] = \frac{P[\text{pos} | A_1] \cdot P[A_1]}{P[\text{pos} | A_1] \cdot P[A_1] + P[\text{pos} | A_2] \cdot P[A_2]} = \frac{.997 \cdot .0032}{(.997) \cdot (.0032) + (1 - .985) \cdot (1 - .0032)} = .1758 \]

So even though the test is positive, you are still unlikely to have HIV. This is because the background rate of HIV is quite low.
Buying a car: Bayes with Tree Diagram

Suppose you need to buy a car.
▶ Based on previous experience you know that 30% of the cars have FAULTY transmission.

To be sure you bring your car to a mechanic. Based on his past experience you know that
▶ he correctly detects cars with good transmission 80% of the cases.
▶ he correctly detects car with bad transmission 90% of the cases.

▶ What is the probability that you buy a faulty car if the mechanic says that is “faulty”?
▶ What is the probability that you buy a faulty car if the mechanic says that is “good”?
Buying a car: Bayes with Tree Diagram
Buying a Car: Bayes with Tree Diagram

It turns out that

- **true positive:** $P(F|"F") = \frac{27\%}{27\%+14\%} = 66\%$
- **false positive:** $P(F|"G") = \frac{3\%}{3\%+56\%} = 5\%$
- **true negative:** $P(G|"G") = \frac{56\%}{3\%+56\%} = 95\%$
- **false negative:** $P(G|"F") = \frac{14\%}{27\%+14\%} = 34\%$

The tree model is easy to be understood, but not practical when you have many cases!! Here we only have 2.
Buying a Car: Bayes without Tree Diagram

\[ P(F) = 30\% \]
\[ P(G) = 70\% \]
\[ P(“F’’|F) = 90\% \]
\[ P(“G’’|F) = 10\% \]
\[ P(“G’’|G) = 80\% \]
\[ P(“F’’|G) = 20\% \]

It turns out that

\[
P(F|“F’’) = \frac{P(“F’’|F)P(F)}{P(“F’’)} = \frac{90\% \times 30\%}{90\% \times 30\% + 20\% \times 70\%} = 66\%
\]
\[
P(F|“G’’) = \frac{P(“G’’|F)P(F)}{P(“G’’)} = \frac{10\% \times 30\%}{10\% \times 30\% + 80\% \times 70\%} = 5\%
\]
The Monty Hall Problem with Bayes

Define

- $O_{xy}$ to be the event such that the TV show presenter opens door $x$ if you pick $y$.
- $W_x$ to be the event such that the door you pick, $x$, is the winning door.

Therefore you are interested in knowing

$$P(W_A | O_{BA}).$$

Conditional probabilities tell us that

$$P(W_A | O_{BA}) = \frac{P(W_A \& O_{BA})}{P(O_{BA})} = \frac{P(O_{BA} | W_A)P(W_A)}{P(O_{BA})}.$$
The Monty Hall Problem with Bayes

- $P(W_A) = \frac{1}{3}$ represents the probability that the door you picked is the winning one!
- $P(O_{BA}|W_A) = \frac{1}{2}$ is the probability that the TV show presenter opens door $B$ given that door you picked and $A$ is the winning one.
- Finally the marginal distribution that the TV show presenter opens door $B$ is

$$P(O_{BA}) = P(W_A \& O_{BA}) + P(W_B \& O_{BA}) + P(W_C \& O_{BA})$$

$$= \frac{1}{6} + 0 + \frac{1}{3} = \frac{3}{6}$$
The Monty Hall Problem with Bayes

- Note that $P(W_A)$ represents our prior distribution (personal probability).

- Opening the door $B$ represents the data that are going to update our personal probability.

- The marginal probability $P(O_{BA})$ re-weight the numerator and again makes the experimental information irrelevant:

$$P(W_A|O_{BA}) = \frac{1/2 \times 1/3}{3/6} = \frac{1}{3}$$