Probability Lecture

Conditional Probability, Independence, and Beyond

Instructor Name

Motivating Example 1/3

Consider a fair six-sided die.

- Given that we rolled an *even* number, what is the probability that we rolled a 2?

Motivating Example 1/3

Consider a fair six-sided die.

- Given that we rolled an *even* number, what is the probability that we rolled a 2?
- What is the probability that we rolled a 3?

Example 1 Solution

If we know the outcome is even ($\{2, 4, 6\}$), the conditional space contains three equally likely outcomes. Therefore

$$P(\text{rolled } 2 \mid \text{even}) = \frac{1}{3}, \qquad P(\text{rolled } 3 \mid \text{even}) = 0.$$

Motivating Example 2/3

Suppose the same fair die rolled a number *greater than 3*. What is the probability that:

- the outcome was an *even* number?

Motivating Example 2/3

Suppose the same fair die rolled a number *greater than 3*. What is the probability that:

- the outcome was an *even* number?
- it was a 4? A 3?

Example 2 Solution

Given the outcome is > 3 ($\{4,5,6\}$), we again have three equally likely cases. Thus

$$P(\text{even} \mid > 3) = \frac{|\{4,6\}|}{|\{4,5,6\}|} = \frac{2}{3}, \qquad P(4 \mid > 3) = \frac{1}{3}, \quad P(3 \mid > 3) = 0.$$

Motivating Example 3/3

Now roll two fair dice. Knowing that their sum is ≤ 5, determine the probability that:

- one of the dice shows a 5. What about a 4?

Motivating Example 3/3

Now roll two fair dice. Knowing that their sum is ≤ 5, determine the probability that:

- one of the dice shows a 5. What about a 4?
- the *first* die shows a 2?

Motivating Example 3/3

Now roll two fair dice. Knowing that their sum is ≤ 5, determine the probability that:

- one of the dice shows a 5. What about a 4?
- the *first* die shows a 2?
- at least one die shows a 2?

Example 3 Solution

Let $B = \{\text{sum} \leq 5\}$. Enumerating all 10 outcomes in B:

$$(1,1) \mid (1,2) \quad (1,3) \quad (2,1) \quad (2,2) \quad (2,3) \quad (3,1) \quad (3,2) \quad (4,1) \quad (2,2)$$

Among these 10 equally likely outcomes:

- *No* pair shows a 5, so $P(\text{a } 5 \mid B) = 0$ while $P(\text{a } 4 \mid B) = 2/10 = 1/5$.

Example 3 Solution

Let $B = \{\text{sum} \leq 5\}$. Enumerating all 10 outcomes in B:

$$(1,1) \mid (1,2) \quad (1,3) \quad (2,1) \quad (2,2) \quad (2,3) \quad (3,1) \quad (3,2) \quad (4,1) \quad (2,2)$$

Among these 10 equally likely outcomes:

- *No* pair shows a 5, so $P(\text{a } 5 \mid B) = 0$ while $P(\text{a } 4 \mid B) = 2/10 = 1/5$.
- Three outcomes have first dieă$= 2 \Rightarrow P(\text{first die } 2 \mid B) = 3/10$.

Example 3 Solution

Let $B = \{\text{sum} \leq 5\}$. Enumerating all 10 outcomes in B:

$$(1,1) \mid (1,2) \quad (1,3) \quad (2,1) \quad (2,2) \quad (2,3) \quad (3,1) \quad (3,2) \quad (4,1) \quad (2,2)$$

Among these 10 equally likely outcomes:

- *No* pair shows a 5, so $P(\text{a } 5 \mid B) = 0$ while $P(\text{a } 4 \mid B) = 2/10 = 1/5$.
- Three outcomes have first dieă$= 2 \Rightarrow P(\text{first die } 2 \mid B) = 3/10$.
- Five outcomes contain at least oneă2 $\Rightarrow P(\text{someă2} \mid B) = 1/2$.

Definition of Conditional Probability

Given two events A and B with $P(B) > 0$, the **conditional probability** of A *given* B is defined as:

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

In words: $P(A \mid B)$ is the probability that A occurs *under the assumption that B has occurred*.

Equivalently, once we know B happened, we "restrict" our sample space to B and renormalize probabilities so the total within B is 1.

Checking the Examples

Using $P(A \mid B) = P(A \cap B)/P(B)$, we can verify our answers:

- Example 1: $A = \{\text{roll } 2\}, B = \{\text{even}\} = \{2, 4, 6\}$.

$$P(2 \mid \text{even}) = \frac{P(\{2\} \cap \{2,4,6\})}{P(\{2,4,6\})} = \frac{P(\{2\})}{P(\{2,4,6\})} = \frac{1/6}{3/6} = \frac{1}{3}.$$

Similarly, $P(3 \mid \text{even}) = \frac{0}{3/6} = 0$.

Checking the Examples

Using $P(A \mid B) = P(A \cap B)/P(B)$, we can verify our answers:

- Example 1: $A = \{\text{roll } 2\}, B = \{\text{even}\} = \{2, 4, 6\}$.

$$P(2 \mid \text{even}) = \frac{P(\{2\} \cap \{2, 4, 6\})}{P(\{2, 4, 6\})} = \frac{P(\{2\})}{P(\{2, 4, 6\})} = \frac{1/6}{3/6} = \frac{1}{3}.$$

  Similarly, $P(3 \mid \text{even}) = \frac{0}{3/6} = 0$.

- Example 2: $B = \{4, 5, 6\}, A = \{\text{even}\} = \{2, 4, 6\}$.

$$P(\text{even} \mid > 3) = \frac{P(\{4, 6\})}{P(\{4, 5, 6\})} = \frac{2/6}{3/6} = \frac{2}{3}.$$

  And $P(4 \mid > 3) = \frac{1/6}{1/2} = 1/3$, $P(3 \mid > 3) = 0$.

Checking the Examples

Using $P(A \mid B) = P(A \cap B)/P(B)$, we can verify our answers:

- Example 1: $A = \{\text{roll } 2\}, B = \{\text{even}\} = \{2, 4, 6\}$.

$$P(2 \mid \text{even}) = \frac{P(\{2\} \cap \{2,4,6\})}{P(\{2,4,6\})} = \frac{P(\{2\})}{P(\{2,4,6\})} = \frac{1/6}{3/6} = \frac{1}{3}.$$

  Similarly, $P(3 \mid \text{even}) = \frac{0}{3/6} = 0$.

- Example 2: $B = \{4, 5, 6\}, A = \{\text{even}\} = \{2, 4, 6\}$.

$$P(\text{even} \mid > 3) = \frac{P(\{4,6\})}{P(\{4,5,6\})} = \frac{2/6}{3/6} = \frac{2}{3}.$$

  And $P(4 \mid > 3) = \frac{1/6}{1/2} = 1/3$, $P(3 \mid > 3) = 0$.

- Example 3: $B = \{\text{sum} \leq 5\}$ (10 outcomes). $A_1 = \{\text{a die is } 4\}$, $P(A_1 \cap B) = 2/36, P(B) = 10/36 \implies P(A_1 \mid B) = 2/10 = 0.2$. Similarly $P(\text{first die is } 2 \mid B) = 3/10$, and $P(\text{at least one } 2 \mid B) = 5/10$.

Sanity Checks

Some quick sanity checks (for $P(B) > 0$):

- $P(B \mid B) = \frac{P(B \cap B)}{P(B)} = 1$. (If we condition on B, the probability B occurred is 1.)

Sanity Checks

Some quick sanity checks (for $P(B) > 0$):

- $P(B \mid B) = \frac{P(B \cap B)}{P(B)} = 1$. (If we condition on B, the probability B occurred is 1.)
- $P(\emptyset \mid B) = \frac{P(\emptyset)}{P(B)} = 0$. (Given B, an impossible event still has probability 0.)

Sanity Checks

Some quick sanity checks (for $P(B) > 0$):

- $P(B \mid B) = \frac{P(B \cap B)}{P(B)} = 1$. (If we condition on B, the probability B occurred is 1.)
- $P(\emptyset \mid B) = \frac{P(\emptyset)}{P(B)} = 0$. (Given B, an impossible event still has probability 0.)
- $P(\Omega \mid B) = \frac{P(\Omega \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1$. (The whole space event always occurs given B.)

Sanity Checks

Some quick sanity checks (for $P(B) > 0$):

- $P(B \mid B) = \frac{P(B \cap B)}{P(B)} = 1$. (If we condition on B, the probability B occurred is 1.)
- $P(\emptyset \mid B) = \frac{P(\emptyset)}{P(B)} = 0$. (Given B, an impossible event still has probability 0.)
- $P(\Omega \mid B) = \frac{P(\Omega \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1$. (The whole space event always occurs given B.)
- If A^c is the complement of A, then $P(A^c \mid B) = 1 - P(A \mid B)$.

Properties of Conditional Probability

Basic Properties: For any events A, B, C with $P(C) > 0$:

- $0 \leq P(A \mid C) \leq 1$.

Additivity: If $A_1, A_2, \ldots$ are disjoint events, then:

$$P\left(\bigcup_j A_j \,\middle|\, C\right) = \sum_j P(A_j \mid C).$$

(Think of $P(\cdot \mid C)$ as a probability measure on C itself. The additivity axiom still holds under the conditioning.)

Properties of Conditional Probability

Basic Properties: For any events A, B, C with $P(C) > 0$:

- $0 \leq P(A \mid C) \leq 1$.
- If $A \subseteq B$, then $P(A \mid C) \leq P(B \mid C)$.

Additivity: If $A_1, A_2, \ldots$ are disjoint events, then:

$$P\left(\bigcup_j A_j \,\middle|\, C\right) = \sum_j P(A_j \mid C).$$

(Think of $P(\cdot \mid C)$ as a probability measure on C itself. The additivity axiom still holds under the conditioning.)

Properties of Conditional Probability

Basic Properties: For any events A, B, C with $P(C) > 0$:

- $0 \leq P(A \mid C) \leq 1$.
- If $A \subseteq B$, then $P(A \mid C) \leq P(B \mid C)$.
- $P(A \mid C) + P(A^c \mid C) = 1$.

Additivity: If $A_1, A_2, \ldots$ are disjoint events, then:

$$P\left(\bigcup_j A_j \,\middle|\, C\right) = \sum_j P(A_j \mid C).$$

(Think of $P(\cdot \mid C)$ as a probability measure on C itself. The additivity axiom still holds under the conditioning.)

Properties of Conditional Probability

Basic Properties: For any events A, B, C with $P(C) > 0$:

- $0 \leq P(A \mid C) \leq 1$.
- If $A \subseteq B$, then $P(A \mid C) \leq P(B \mid C)$.
- $P(A \mid C) + P(A^c \mid C) = 1$.
- $P(A \setminus B \mid C) = P(A \mid C) - P(A \cap B \mid C)$.

Additivity: If $A_1, A_2, \ldots$ are disjoint events, then:

$$P\left(\bigcup_j A_j \,\middle|\, C\right) = \sum_j P(A_j \mid C).$$

(Think of $P(\cdot \mid C)$ as a probability measure on C itself. The additivity axiom still holds under the conditioning.)

Countable Additivity Proof

Fix C with $P(C) > 0$. Define

$$Q(E) := \frac{P(E \cap C)}{P(C)},$$

for any event E.

Then $Q(\cdot)$ is a probability measure on C. If $\{A_j\}$ are disjoint:

$$\begin{aligned}
Q\Big(\bigcup_j A_j\Big) &= \frac{P\Big((\bigcup_j A_j) \cap C\Big)}{P(C)} = \frac{P\big(\bigcup_j (A_j \cap C)\big)}{P(C)} \\
&= \frac{\sum_j P(A_j \cap C)}{P(C)} = \sum_j \frac{P(A_j \cap C)}{P(C)} \\
&= \sum_j Q(A_j).
\end{aligned}$$

Thus $P(\bigcup_j A_j \mid C) = \sum_j P(A_j \mid C)$. In other words, $P(\cdot \mid C)$ is countably additive.

Example: State of Origin

In a best-of-3 series, suppose each game is equally likely to be won by the Blues (B) or Maroons (M). All $2^3 = 8$ possible win sequences are equally likely (e.g. BBB, BBM, ..., MMM).

Let $B1 = \{\text{Blues win game 1}\} = \{\text{BBB, BBM, BMB, BMM}\}$ (4 outcomes).
Let $BW = \{\text{Blues win the series}\} = \{\text{BBB, BBM, BMB, MBB}\}$ (Blues win ≥ 2 games).

Why do we care?

"Look, winning that *first* game is so importantjust ask any longsuffering Blues fan. Momentum, hoodoo, the vibeyou name it, pundits will invoke it. Queensland bellows Queenslander! while NSW talk themselves into a lather."

We have:

Thus:

$$P(BW \mid B1) = \frac{P(BW \cap B1)}{P(B1)} = \frac{3/8}{1/2} = \frac{3/8}{4/8} = \frac{3}{4} = 75\%.$$

So, if a team wins the first game, their chance of winning the series is 75%. (Sports commentators often cite: "75% of teams who win Game 1 win the series.")

Example: State of Origin

In a best-of-3 series, suppose each game is equally likely to be won by the Blues (B) or Maroons (M). All $2^3 = 8$ possible win sequences are equally likely (e.g. BBB, BBM, ..., MMM).

Let $B1 = \{\text{Blues win game 1}\} = \{\text{BBB, BBM, BMB, BMM}\}$ (4 outcomes).
Let $BW = \{\text{Blues win the series}\} = \{\text{BBB, BBM, BMB, MBB}\}$ (Blues win ≥ 2 games).

Why do we care?

"Look, winning that *first* game is so importantjust ask any longsuffering Blues fan. Momentum, hoodoo, the vibeyou name it, pundits will invoke it. Queensland bellows Queenslander! while NSW talk themselves into a lather."

We have:

- $P(BW \cap B1) = P(\{\text{BBB, BBM, BMB}\}) = 3/8.$

Thus:

$$P(BW \mid B1) = \frac{P(BW \cap B1)}{P(B1)} = \frac{3/8}{1/2} = \frac{3/8}{4/8} = \frac{3}{4} = 75\%.$$

So, if a team wins the first game, their chance of winning the series is 75%. (Sports commentators often cite: "75% of teams who win Game 1 win the series.")

Example: State of Origin

In a best-of-3 series, suppose each game is equally likely to be won by the Blues (B) or Maroons (M). All $2^3 = 8$ possible win sequences are equally likely (e.g. BBB, BBM, ..., MMM).

Let $B1 = \{\text{Blues win game 1}\} = \{\text{BBB, BBM, BMB, BMM}\}$ (4 outcomes).
Let $BW = \{\text{Blues win the series}\} = \{\text{BBB, BBM, BMB, MBB}\}$ (Blues win ≥ 2 games).

Why do we care?

"Look, winning that *first* game is so importantjust ask any longsuffering Blues fan. Momentum, hoodoo, the vibeyou name it, pundits will invoke it. Queensland bellows Queenslander! while NSW talk themselves into a lather."

We have:

- $P(BW \cap B1) = P(\{\text{BBB, BBM, BMB}\}) = 3/8.$
- $P(B1) = 4/8 = 1/2.$

Thus:

$$P(BW \mid B1) = \frac{P(BW \cap B1)}{P(B1)} = \frac{3/8}{1/2} = \frac{3/8}{4/8} = \frac{3}{4} = 75\%.$$

So, if a team wins the first game, their chance of winning the series is 75%. (Sports commentators often cite: "75% of teams who win Game 1 win the series.")

The Monty Hall Problem

Problem: You're on a game show with 3 doors. Behind one door is a car (prize), behind the others are goats. You pick one door (without opening). The host, who knows where the car is, then opens a *different* door, revealing a goat. You are then offered a choice: **stay** with your original door, or **switch** to the other unopened door.

Question: To maximize your chance of winning the car, should you stay or switch?

Monty Hall: Quick Solution

Initially, $P(\text{car behind your door}) = \frac{1}{3}$ and $P(\text{car behind one of the other two}) = \frac{2}{3}$.

The host's action (always revealing a goat behind a door you didn't pick) doesn't change these initial probabilities, but it does give information. In fact:

- If you **stick** with your original door, you win only if you initially picked correctly (probability $1/3$).

Conclusion: Switching doubles your chance of winning (from 33.3% to 66.7%).

Monty Hall: Quick Solution

Initially, $P(\text{car behind your door}) = \frac{1}{3}$ and $P(\text{car behind one of the other two}) = \frac{2}{3}$.

The host's action (always revealing a goat behind a door you didn't pick) doesn't change these initial probabilities, but it does give information. In fact:

- If you **stick** with your original door, you win only if you initially picked correctly (probability $1/3$).
- If you **switch**, you win if your original pick was wrong (probability $2/3$).

Conclusion: Switching doubles your chance of winning (from 33.3% to 66.7%).

Monty Hall: Detailed Analysis

Label the doors 1, 2, 3. Suppose you initially pick Door 1. There are three equally likely scenarios:

In 2 of the 3 cases (when the car is behind #2 or #3), switching wins the car. So $P(\text{win}|\text{switch}) = 2/3$, whereas $P(\text{win}|\text{stay}) = 1/3$.

Monty Hall: Detailed Analysis

Label the doors 1, 2, 3. Suppose you initially pick Door 1. There are three equally likely scenarios:

In 2 of the 3 cases (when the car is behind #2 or #3), switching wins the car. So $P(\text{win}|\text{switch}) = 2/3$, whereas $P(\text{win}|\text{stay}) = 1/3$.

Monty Hall: Detailed Analysis

Label the doors 1, 2, 3. Suppose you initially pick Door 1. There are three equally likely scenarios:

- Car is behind Door 1 (prob $1/3$): Host opens one of #2 or #3 at random (both goats). If you switch, you end up with a goat (lose); if you stay, you win the car.

In 2 of the 3 cases (when the car is behind #2 or #3), switching wins the car. So $P(\text{win}|\text{switch}) = 2/3$, whereas $P(\text{win}|\text{stay}) = 1/3$.

Monty Hall: Detailed Analysis

Label the doors 1, 2, 3. Suppose you initially pick Door 1. There are three equally likely scenarios:

- Car is behind Door 1 (prob $1/3$): Host opens one of #2 or #3 at random (both goats). If you switch, you end up with a goat (lose); if you stay, you win the car.
- Car is behind Door 2 (prob $1/3$): Host must open Door 3 (the only remaining goat door). If you switch (to #2), you win the car; if you stay, you lose.

In 2 of the 3 cases (when the car is behind #2 or #3), switching wins the car. So $P(\text{win}|\text{switch}) = 2/3$, whereas $P(\text{win}|\text{stay}) = 1/3$.

Monty Hall: Detailed Analysis

Label the doors 1, 2, 3. Suppose you initially pick Door 1. There are three equally likely scenarios:

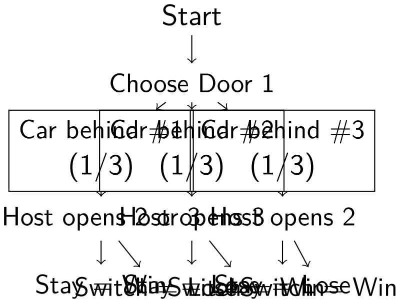
- Car is behind Door 1 (prob $1/3$): Host opens one of #2 or #3 at random (both goats). If you switch, you end up with a goat (lose); if you stay, you win the car.
- Car is behind Door 2 (prob $1/3$): Host must open Door 3 (the only remaining goat door). If you switch (to #2), you win the car; if you stay, you lose.
- Car is behind Door 3 (prob $1/3$): Host opens Door 2. If you switch (to #3), you win; if you stay, you lose.

In 2 of the 3 cases (when the car is behind #2 or #3), switching wins the car. So $P(\text{win}|\text{switch}) = 2/3$, whereas $P(\text{win}|\text{stay}) = 1/3$.

Monty Hall: Detailed Analysis

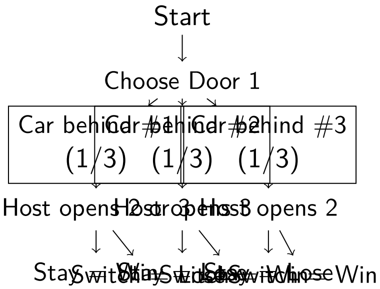
Label the doors 1, 2, 3. Suppose you initially pick Door 1. There are three equally likely scenarios:

- Car is behind Door 1 (prob $1/3$): Host opens one of #2 or #3 at random (both goats). If you switch, you end up with a goat (lose); if you stay, you win the car.
- Car is behind Door 2 (prob $1/3$): Host must open Door 3 (the only remaining goat door). If you switch (to #2), you win the car; if you stay, you lose.
- Car is behind Door 3 (prob $1/3$): Host opens Door 2. If you switch (to #3), you win; if you stay, you lose.



In 2 of the 3 cases (when the car is behind #2 or #3), switching wins the car. So $P(\text{win}|\text{switch}) = 2/3$, whereas $P(\text{win}|\text{stay}) = 1/3$.

Monty Hall Decision Tree (big)



Monty Hall: A Variation

What if the host doesn't pick randomly when faced with a choice? (E.g. if your initial pick is correct, host always opens the lowest-numbered goat door instead of choosing randomly.)

If the host's strategy is known, you can condition on it:

- If the host opens Door 3, that would only happen (under this rule) if the car was behind Door 2. So switching to Door 2 in that case wins with probability 1.

If you always switch regardless, the overall win probability still works out to $2/3$. However, an optimal strategy could be: switch only when the host opens a particular door. In general, the solution to Monty Hall-type problems depends on the host's behavior, but the classic version assumes the host opens a random goat door when able.

Monty Hall: A Variation

What if the host doesn't pick randomly when faced with a choice? (E.g. if your initial pick is correct, host always opens the lowest-numbered goat door instead of choosing randomly.)

If the host's strategy is known, you can condition on it:

- If the host opens Door 3, that would only happen (under this rule) if the car was behind Door 2. So switching to Door 2 in that case wins with probability 1.
- If the host opens Door 2, it could mean the car is either behind Door 1 or Door 3 (each with 50% conditional probability). In that case, switching yields a 50% chance of winning (no better than staying).

If you always switch regardless, the overall win probability still works out to $2/3$. However, an optimal strategy could be: switch only when the host opens a particular door. In general, the solution to Monty Hall-type problems depends on the host's behavior, but the classic version assumes the host opens a random goat door when able.

False Positive Paradox — Setup

Scenario: A rare genetic disorder occurs in 0.1% of births. A prenatal screening test has:

- **Sensitivity:** If the fetus *has* the disorder, the test is positive 100% of the time.

Question: If a randomly selected expectant mother gets a **positive** result, what is the probability that her baby actually has the disorder?

False Positive Paradox — Setup

Scenario: A rare genetic disorder occurs in 0.1% of births. A prenatal screening test has:

- **Sensitivity:** If the fetus *has* the disorder, the test is positive 100% of the time.
- **Specificity:** If the fetus is *healthy*, the test is negative 95% of the time (i.e. a 5% false-positive rate).

Question: If a randomly selected expectant mother gets a **positive** result, what is the probability that her baby actually has the disorder?

False Positive Paradox — Computation

Let us reason with an imaginary cohort of 100 000 pregnancies:

- **Diseased babies**: $0.001 \times 100{,}000 = 100$.

Hence

$$P(\text{Disorder} \mid \text{Positive}) = \frac{100}{5{,}095} \approx 1.96\%.$$

So despite a perfectly sensitive test, a positive result still means only a $\sim 2\%$ chance of disease because of the tiny base rate.

False Positive Paradox — Computation

Let us reason with an imaginary cohort of 100 000 pregnancies:

- **Diseased babies**: $0.001 \times 100{,}000 = 100$.
- **Healthy babies**: 99,900.

Hence

$$P(\text{Disorder} \mid \text{Positive}) = \frac{100}{5{,}095} \approx 1.96\%.$$

So despite a perfectly sensitive test, a positive result still means only a $\sim 2\%$ chance of disease because of the tiny base rate.

False Positive Paradox — Computation

Let us reason with an imaginary cohort of 100 000 pregnancies:

- **Diseased babies**: $0.001 \times 100{,}000 = 100$.
- **Healthy babies**: 99,900.
- **Positive tests** break down as:

Hence

$$P(\text{Disorder} \mid \text{Positive}) = \frac{100}{5{,}095} \approx 1.96\%.$$

So despite a perfectly sensitive test, a positive result still means only a $\sim 2\%$ chance of disease because of the tiny base rate.

False Positive Paradox — Computation

Let us reason with an imaginary cohort of 100 000 pregnancies:

- **Diseased babies**: $0.001 \times 100{,}000 = 100$.
- **Healthy babies**: 99,900.
- **Positive tests** break down as:
  - Diseased & positive (true positives): 100 (sensitivity $= 100\%$).

Hence

$$P(\text{Disorder} \mid \text{Positive}) \;=\; \frac{100}{5{,}095} \;\approx\; 1.96\%.$$

So despite a perfectly sensitive test, a positive result still means only a $\sim 2\%$ chance of disease because of the tiny base rate.

False Positive Paradox — Computation

Let us reason with an imaginary cohort of 100 000 pregnancies:

- **Diseased babies**: $0.001 \times 100{,}000 = 100$.
- **Healthy babies**: 99,900.
- **Positive tests** break down as:
  - Diseased & positive (true positives): 100 (sensitivity $= 100\%$).
  - Healthy & *false* positives: 5% of 99,900 $\approx$ 4,995.

Hence

$$P(\text{Disorder} \mid \text{Positive}) \;=\; \frac{100}{5{,}095} \;\approx\; 1.96\%.$$

So despite a perfectly sensitive test, a positive result still means only a $\sim 2\%$ chance of disease because of the tiny base rate.

False Positive Paradox — Computation

Let us reason with an imaginary cohort of 100 000 pregnancies:

- **Diseased babies**: $0.001 \times 100{,}000 = 100$.
- **Healthy babies**: 99,900.
- **Positive tests** break down as:
  - Diseased & positive (true positives): 100 (sensitivity $= 100\%$).
  - Healthy & *false* positives: 5% of $99{,}900 \approx 4{,}995$.
- **Total positive tests**: $100 + 4{,}995 \approx 5{,}095$.

Hence

$$P(\text{Disorder} \mid \text{Positive}) = \frac{100}{5{,}095} \approx 1.96\%.$$

So despite a perfectly sensitive test, a positive result still means only a $\sim 2\%$ chance of disease because of the tiny base rate.

Base Rate Matters

If instead the disorder was common (say 50% prevalence):

- Diseased babies: 50,000 (all test positive).

Now $P(\text{Disorder} \mid \text{Positive}) \approx 95.2\%$.

This dramatic change (2% vs 95%) illustrates **base rate neglect**: people often ignore prior probabilities. Always account for the base rate when interpreting diagnostic tests!

Other examples:

Base Rate Matters

If instead the disorder was common (say 50% prevalence):

- Diseased babies: 50,000 (all test positive).
- Healthy babies: 50,000 (false positives: $0.05 \times 50{,}000 = 2{,}500$).

Now $P(\text{Disorder} \mid \text{Positive}) \approx 95.2\%$.

This dramatic change (2% vs 95%) illustrates **base rate neglect**: people often ignore prior probabilities. Always account for the base rate when interpreting diagnostic tests!

Other examples:

Base Rate Matters

If instead the disorder was common (say 50% prevalence):

- Diseased babies: 50,000 (all test positive).
- Healthy babies: 50,000 (false positives: $0.05 \times 50{,}000 = 2{,}500$).
- Total positives: 52,500, of which 50,000 are true.

Now $P(\text{Disorder} \mid \text{Positive}) \approx 95.2\%$.

This dramatic change (2% vs 95%) illustrates **base rate neglect**: people often ignore prior probabilities. Always account for the base rate when interpreting diagnostic tests!

Other examples:

Base Rate Matters

If instead the disorder was common (say 50% prevalence):

- Diseased babies: 50,000 (all test positive).
- Healthy babies: 50,000 (false positives: $0.05 \times 50{,}000 = 2{,}500$).
- Total positives: 52,500, of which 50,000 are true.

Now $P(\text{Disorder} \mid \text{Positive}) \approx 95.2\%$.

This dramatic change (2% vs 95%) illustrates **base rate neglect**: people often ignore prior probabilities. Always account for the base rate when interpreting diagnostic tests!

Other examples:

- Security alarms for very rare events will trigger mostly false alarms.

Base Rate Matters

If instead the disorder was common (say 50% prevalence):

- Diseased babies: 50,000 (all test positive).
- Healthy babies: 50,000 (false positives: $0.05 \times 50{,}000 = 2{,}500$).
- Total positives: 52,500, of which 50,000 are true.

Now $P(\text{Disorder} \mid \text{Positive}) \approx 95.2\%$.

This dramatic change (2% vs 95%) illustrates **base rate neglect**: people often ignore prior probabilities. Always account for the base rate when interpreting diagnostic tests!

Other examples:

- Security alarms for very rare events will trigger mostly false alarms.
- Drug tests in low-prevalence populations yield more false positives than true positives.

The Birthday Problem

Question: How many people must be in a room so that the probability of at least two sharing the same birthday exceeds 50%?
Surprisingly, the answer is only 23 people! (Much lower than 183, which would be 50% of 365.)

Assumptions:

- 365 days in a year (no Feb 29).

Solution approach: Compute the complement (no shared birthdays):

For $n = 23$:

$$P(B_{23}) = \frac{365 \times 364 \times \cdots \times 343}{365^{23}} \approx 0.4927,$$

so $P(\text{at least one match}) \approx 1 - 0.4927 = 0.5073$ (just over 50%).

The Birthday Problem

Question: How many people must be in a room so that the probability of at least two sharing the same birthday exceeds 50%?
Surprisingly, the answer is only 23 people! (Much lower than 183, which would be 50% of 365.)

Assumptions:

- 365 days in a year (no Feb 29).
- Each person's birthday is uniformly random in $\{1, \ldots, 365\}$, independent of others.

Solution approach: Compute the complement (no shared birthdays):

For $n = 23$:

$$P(B_{23}) = \frac{365 \times 364 \times \cdots \times 343}{365^{23}} \approx 0.4927,$$

so $P(\text{at least one match}) \approx 1 - 0.4927 = 0.5073$ (just over 50%).

The Birthday Problem

Question: How many people must be in a room so that the probability of at least two sharing the same birthday exceeds 50%?
Surprisingly, the answer is only 23 people! (Much lower than 183, which would be 50% of 365.)

Assumptions:

- 365 days in a year (no Feb 29).
- Each person's birthday is uniformly random in $\{1, \ldots, 365\}$, independent of others.

Solution approach: Compute the complement (no shared birthdays):

- Let B_n = event that all n birthdays are distinct.

For $n = 23$:

$$P(B_{23}) = \frac{365 \times 364 \times \cdots \times 343}{365^{23}} \approx 0.4927,$$

so $P(\text{at least one match}) \approx 1 - 0.4927 = 0.5073$ (just over 50%).

The Birthday Problem

Question: How many people must be in a room so that the probability of at least two sharing the same birthday exceeds 50%?
Surprisingly, the answer is only 23 people! (Much lower than 183, which would be 50% of 365.)

Assumptions:

- 365 days in a year (no Feb 29).
- Each person's birthday is uniformly random in $\{1, \ldots, 365\}$, independent of others.

Solution approach: Compute the complement (no shared birthdays):

- Let B_n = event that all n birthdays are distinct.
- $P(B_n) = \frac{365 \times 364 \times \cdots \times (365-n+1)}{365^n}$.

For $n = 23$:

$$P(B_{23}) = \frac{365 \times 364 \times \cdots \times 343}{365^{23}} \approx 0.4927,$$

so $P(\text{at least one match}) \approx 1 - 0.4927 = 0.5073$ (just over 50%).

Birthday Problem: Intuition

At $n = 23$, there are $\binom{23}{2} = 253$ pairs of people. Each pair has probability $1/365$ of matching birthdays, and 253 pairs create many opportunities for a match.

As n grows, the chance of a match increases rapidly:

- $n = 10$: $\approx 12\%$ chance of a shared birthday.

This counterintuitive result is known as the **Birthday Paradox**. One implication:

Birthday Problem: Intuition

At $n = 23$, there are $\binom{23}{2} = 253$ pairs of people. Each pair has probability $1/365$ of matching birthdays, and 253 pairs create many opportunities for a match.

As n grows, the chance of a match increases rapidly:

- $n = 10$: $\approx 12\%$ chance of a shared birthday.
- $n = 23$: $\approx 50.7\%$ chance.

This counterintuitive result is known as the **Birthday Paradox**. One implication:

Birthday Problem: Intuition

At $n = 23$, there are $\binom{23}{2} = 253$ pairs of people. Each pair has probability $1/365$ of matching birthdays, and 253 pairs create many opportunities for a match.

As n grows, the chance of a match increases rapidly:

- $n = 10$: $\approx 12\%$ chance of a shared birthday.
- $n = 23$: $\approx 50.7\%$ chance.
- $n = 30$: $\approx 70\%$ chance.

This counterintuitive result is known as the **Birthday Paradox**. One implication:

Birthday Problem: Intuition

At $n = 23$, there are $\binom{23}{2} = 253$ pairs of people. Each pair has probability $1/365$ of matching birthdays, and 253 pairs create many opportunities for a match.

As n grows, the chance of a match increases rapidly:

- $n = 10$: $\approx 12\%$ chance of a shared birthday.
- $n = 23$: $\approx 50.7\%$ chance.
- $n = 30$: $\approx 70\%$ chance.
- $n = 50$: $\approx 97\%$ chance.

This counterintuitive result is known as the **Birthday Paradox**. One implication:

Birthday Problem: Intuition

At $n = 23$, there are $\binom{23}{2} = 253$ pairs of people. Each pair has probability $1/365$ of matching birthdays, and 253 pairs create many opportunities for a match.

As n grows, the chance of a match increases rapidly:

- $n = 10$: $\approx 12\%$ chance of a shared birthday.
- $n = 23$: $\approx 50.7\%$ chance.
- $n = 30$: $\approx 70\%$ chance.
- $n = 50$: $\approx 97\%$ chance.

This counterintuitive result is known as the **Birthday Paradox**. One implication:

- In cryptography, finding two inputs with the same hash ("collision") is easier than brute force due to the birthday paradox (birthday attack).

Birthday Problem: Intuition

At $n = 23$, there are $\binom{23}{2} = 253$ pairs of people. Each pair has probability $1/365$ of matching birthdays, and 253 pairs create many opportunities for a match.

As n grows, the chance of a match increases rapidly:

- $n = 10$: $\approx 12\%$ chance of a shared birthday.
- $n = 23$: $\approx 50.7\%$ chance.
- $n = 30$: $\approx 70\%$ chance.
- $n = 50$: $\approx 97\%$ chance.

This counterintuitive result is known as the **Birthday Paradox**. One implication:

- In cryptography, finding two inputs with the same hash ("collision") is easier than brute force due to the birthday paradox (birthday attack).
- In any large group, surprising coincidences (like shared birthdays) are more common than naive intuition suggests.

Definition of Independence

Two events A and B are **independent** if knowing that one occurs does not change the probability of the other. Formally, A and B are independent if

$$P(A \mid B) = P(A),$$

whenever $P(B) > 0$. (Equivalently $P(B \mid A) = P(B)$.)

If $P(A) > 0, P(B) > 0$, this is equivalent to

$$P(A \cap B) = P(A)\,P(B).$$

(This product rule can be taken as the definition of independence.)

Properties of Independence

- If $A \perp\!\!\!\perp B$, then automatically $B \perp\!\!\!\perp A$. (Symmetry: $P(A \cap B) = P(B \cap A)$.)

Properties of Independence

- If $A \perp\!\!\!\perp B$, then automatically $B \perp\!\!\!\perp A$. (Symmetry: $P(A \cap B) = P(B \cap A)$.)
- If $A \perp\!\!\!\perp B$, then A is also independent of B^c (the complement of B).

$$P(A \cap B^c) = P(A) - P(A \cap B) = P(A) - P(A)P(B) = P(A)P(B^c).$$

Properties of Independence

- If $A \perp\!\!\!\perp B$, then automatically $B \perp\!\!\!\perp A$. (Symmetry: $P(A \cap B) = P(B \cap A)$.)
- If $A \perp\!\!\!\perp B$, then A is also independent of B^c (the complement of B).

$$P(A \cap B^c) = P(A) - P(A \cap B) = P(A) - P(A)P(B) = P(A)P(B^c).$$

- If $A \perp\!\!\!\perp B$, then any event defined in terms of A is independent of any event defined in terms of B. (E.g. $A \perp\!\!\!\perp B$ implies $A \perp\!\!\!\perp B^c$, $A^c \perp\!\!\!\perp B$, $A^c \perp\!\!\!\perp B^c$.)

Informal Examples of Independence

- The result of a coin flip and the result of a die roll are independent events.

Informal Examples of Independence

- The result of a coin flip and the result of a die roll are independent events.
- Two successive flips of a fair coin are independent (the first being heads has no effect on the probability of the second being heads).

Informal Examples of Independence

- The result of a coin flip and the result of a die roll are independent events.
- Two successive flips of a fair coin are independent (the first being heads has no effect on the probability of the second being heads).
- The event of rain in Sydney today and the event of rolling a 6 on a die are essentially independent.

Informal Examples of Independence

- The result of a coin flip and the result of a die roll are independent events.
- Two successive flips of a fair coin are independent (the first being heads has no effect on the probability of the second being heads).
- The event of rain in Sydney today and the event of rolling a 6 on a die are essentially independent.
- *(Gambler's Fallacy)*: People often *assume* dependence incorrectly. After 5 reds in a row at roulette, one might think "black is due" next. In reality, each spin is independent; the wheel has no memory.

Formal Examples: Independence

- Draw one card from a well-shuffled deck. Let $A = \{\text{card is Heart}\}$, $B = \{\text{card is a Face card (J,Q,K)}\}$.

  $$P(A) = 1/4, \quad P(B) = 12/52 = 3/13.$$

  $P(A \cap B) = P(\text{Heart and Face}) = 3/52$. And $P(A)P(B) = \frac{1}{4} \cdot \frac{3}{13} = 3/52$. So $P(A \cap B) = P(A)P(B)$: being a Heart and being a Face card are independent (because the deck's composition is uniform across suits).

Formal Examples: Independence

- Draw one card from a well-shuffled deck. Let $A = \{\text{card is Heart}\}$, $B = \{\text{card is a Face card (J,Q,K)}\}$.

  $$P(A) = 1/4, \quad P(B) = 12/52 = 3/13.$$

  $P(A \cap B) = P(\text{Heart and Face}) = 3/52$. And $P(A)P(B) = \frac{1}{4} \cdot \frac{3}{13} = 3/52$. So $P(A \cap B) = P(A)P(B)$: being a Heart and being a Face card are independent (because the deck's composition is uniform across suits).

- Roll a fair die. Let $C = \{\text{roll is even}\}$, $D = \{\text{roll is prime}\}$.

  $$P(C) = 3/6 = 1/2, \quad P(D) = 3/6 = 1/2,$$

  $$P(C \cap D) = P(\{\text{roll} = 2\}) = 1/6.$$

  $P(C)P(D) = 1/4 = 0.25$, whereas $P(C \cap D) \approx 0.1667$. So C and D are **not** independent (knowing the roll is prime changes the chance of it being even).

Mutual Independence (≥ 3 events)

We defined independence for pairs of events. How about three events A, B, C? One might attempt: A, B, C are independent if

$$P(A \cap B \cap C) = P(A)P(B)P(C).$$

But that alone is *not sufficient*. For mutual independence, we require *all* nontrivial intersections factor:

- $A \perp\!\!\!\perp B$, $A \perp\!\!\!\perp C$, $B \perp\!\!\!\perp C$ (all pairs independent), **and**

In general, events $A_1, \ldots, A_n$ are **mutually independent** if every subcollection's intersection probability equals the product of those probabilities. (This includes all pairs, triples, etc.)
Warning: Pairwise independence does *not* imply mutual independence, and vice versa. We'll see examples.

Mutual Independence (≥ 3 events)

We defined independence for pairs of events. How about three events A, B, C? One might attempt: A, B, C are independent if

$$P(A \cap B \cap C) = P(A)P(B)P(C).$$

But that alone is *not sufficient*. For mutual independence, we require *all* nontrivial intersections factor:

- $A \perp\!\!\!\perp B$, $A \perp\!\!\!\perp C$, $B \perp\!\!\!\perp C$ (all pairs independent), **and**
- $P(A \cap B \cap C) = P(A)P(B)P(C)$.

In general, events $A_1, \ldots, A_n$ are **mutually independent** if every subcollection's intersection probability equals the product of those probabilities. (This includes all pairs, triples, etc.)
Warning: Pairwise independence does *not* imply mutual independence, and vice versa. We'll see examples.

Counterexample 1: Triple Product Only

It's possible for $P(A \cap B \cap C) = P(A)P(B)P(C)$ to hold even if some pairs are *not* independent.

Consider 8 equally likely outcomes $\{1, 2, \ldots, 8\}$. Define:

$$A = \{1, 2, 3, 4\}, \quad B = \{1, 5, 6, 7\}, \quad C = \{1, 5, 7, 8\}.$$

Then $P(A) = P(B) = P(C) = 4/8 = 1/2$. Check intersections:

- $A \cap B = \{1\}$, so $P(A \cap B) = 1/8 = 0.125$ vs $P(A)P(B) = 0.25$. (A, B not independent.)

Here the triple product rule holds, but none of the pairs are independent! This is why the definition of mutual independence requires all subsets, not just the full intersection.

Counterexample 1: Triple Product Only

It's possible for $P(A \cap B \cap C) = P(A)P(B)P(C)$ to hold even if some pairs are *not* independent.

Consider 8 equally likely outcomes $\{1, 2, \ldots, 8\}$. Define:

$$A = \{1, 2, 3, 4\}, \quad B = \{1, 5, 6, 7\}, \quad C = \{1, 5, 7, 8\}.$$

Then $P(A) = P(B) = P(C) = 4/8 = 1/2$. Check intersections:

- $A \cap B = \{1\}$, so $P(A \cap B) = 1/8 = 0.125$ vs $P(A)P(B) = 0.25$. (A, B not independent.)
- $A \cap C = \{1\}$, so $P(A \cap C) = 0.125$ vs $0.25 = P(A)P(C)$. (A, C not independent.)

Here the triple product rule holds, but none of the pairs are independent! This is why the definition of mutual independence requires all subsets, not just the full intersection.

Counterexample 1: Triple Product Only

It's possible for $P(A \cap B \cap C) = P(A)P(B)P(C)$ to hold even if some pairs are *not* independent.

Consider 8 equally likely outcomes $\{1, 2, \ldots, 8\}$. Define:

$$A = \{1, 2, 3, 4\}, \quad B = \{1, 5, 6, 7\}, \quad C = \{1, 5, 7, 8\}.$$

Then $P(A) = P(B) = P(C) = 4/8 = 1/2$. Check intersections:

- $A \cap B = \{1\}$, so $P(A \cap B) = 1/8 = 0.125$ vs $P(A)P(B) = 0.25$. (A, B not independent.)
- $A \cap C = \{1\}$, so $P(A \cap C) = 0.125$ vs $0.25 = P(A)P(C)$. (A, C not independent.)
- $B \cap C = \{1, 5, 7\}$, so $P(B \cap C) = 3/8 = 0.375$ vs $0.25 = P(B)P(C)$. (B, C not independent.)

Here the triple product rule holds, but none of the pairs are independent! This is why the definition of mutual independence requires all subsets, not just the full intersection.

Counterexample 1: Triple Product Only

It's possible for $P(A \cap B \cap C) = P(A)P(B)P(C)$ to hold even if some pairs are *not* independent.

Consider 8 equally likely outcomes $\{1, 2, \ldots, 8\}$. Define:

$$A = \{1, 2, 3, 4\}, \quad B = \{1, 5, 6, 7\}, \quad C = \{1, 5, 7, 8\}.$$

Then $P(A) = P(B) = P(C) = 4/8 = 1/2$. Check intersections:

- $A \cap B = \{1\}$, so $P(A \cap B) = 1/8 = 0.125$ vs $P(A)P(B) = 0.25$. (A, B not independent.)
- $A \cap C = \{1\}$, so $P(A \cap C) = 0.125$ vs $0.25 = P(A)P(C)$. (A, C not independent.)
- $B \cap C = \{1, 5, 7\}$, so $P(B \cap C) = 3/8 = 0.375$ vs $0.25 = P(B)P(C)$. (B, C not independent.)
- $A \cap B \cap C = \{1\}$, so $P(A \cap B \cap C) = 0.125$. And $P(A)P(B)P(C) = 0.125$ as well.

Here the triple product rule holds, but none of the pairs are independent! This is why the definition of mutual independence requires all subsets, not just the full intersection.

Counterexample 2: Pairwise but Not Mutual

Roll two fair dice. Define:

- $A = \{\text{sum of dice} = 7\}$.

We have:

All three pairs are independent, yet A, B, C are **not** mutually independent (the triple intersection doesn't factor).

Counterexample 2: Pairwise but Not Mutual

Roll two fair dice. Define:

- $A = \{\text{sum of dice} = 7\}$.
- $B = \{\text{first die} = 3\}$.

We have:

All three pairs are independent, yet A, B, C are **not** mutually independent (the triple intersection doesn't factor).

Counterexample 2: Pairwise but Not Mutual

Roll two fair dice. Define:

- $A = \{\text{sum of dice} = 7\}$.
- $B = \{\text{first die} = 3\}$.
- $C = \{\text{second die} = 4\}$.

We have:

All three pairs are independent, yet A, B, C are **not** mutually independent (the triple intersection doesn't factor).

Counterexample 2: Pairwise but Not Mutual

Roll two fair dice. Define:

- $A = \{\text{sum of dice} = 7\}$.
- $B = \{\text{first die} = 3\}$.
- $C = \{\text{second die} = 4\}$.

We have:

- $P(A) = 6/36 = 1/6$, $P(B) = 1/6$, $P(C) = 1/6$.

All three pairs are independent, yet A, B, C are **not** mutually independent (the triple intersection doesn't factor).

Counterexample 2: Pairwise but Not Mutual

Roll two fair dice. Define:

- $A = \{\text{sum of dice} = 7\}$.
- $B = \{\text{first die} = 3\}$.
- $C = \{\text{second die} = 4\}$.

We have:

- $P(A) = 6/36 = 1/6$, $P(B) = 1/6$, $P(C) = 1/6$.
- $P(A \cap B)$: outcomes where first die 3 and sum 7 $\implies$ only $(3, 4)$, so $P(A \cap B) = 1/36 = (1/6)(1/6) = P(A)P(B)$. Similarly $P(A \cap C) = 1/36 = P(A)P(C)$, and $P(B \cap C) = 1/36 = P(B)P(C)$. (All pairs independent.)

All three pairs are independent, yet A, B, C are **not** mutually independent (the triple intersection doesn't factor).

Counterexample 2: Pairwise but Not Mutual

Roll two fair dice. Define:

- $A = \{\text{sum of dice} = 7\}$.
- $B = \{\text{first die} = 3\}$.
- $C = \{\text{second die} = 4\}$.

We have:

- $P(A) = 6/36 = 1/6$, $P(B) = 1/6$, $P(C) = 1/6$.
- $P(A \cap B)$: outcomes where first die 3 and sum 7 $\implies$ only $(3, 4)$, so $P(A \cap B) = 1/36 = (1/6)(1/6) = P(A)P(B)$. Similarly $P(A \cap C) = 1/36 = P(A)P(C)$, and $P(B \cap C) = 1/36 = P(B)P(C)$. (All pairs independent.)
- $P(A \cap B \cap C) = P(\{(3, 4)\}) = 1/36$. But $P(A)P(B)P(C) = (1/6)^3 = 1/216$.

All three pairs are independent, yet A, B, C are **not** mutually independent (the triple intersection doesn't factor).

Conditional Independence

Sometimes events that are not independent can become independent when conditioning on a third event.

Notation: $A \perp\!\!\!\perp B \mid C$ means "A is independent of B given C". Formally:

$$A \perp\!\!\!\perp B \mid C \iff P(A \mid B \cap C) = P(A \mid C),$$

whenever $P(C) > 0$ (and $P(B \cap C) > 0$).
Equivalently:

$$A \perp\!\!\!\perp B \mid C \iff P(A \cap B \mid C) = P(A \mid C)\, P(B \mid C).$$

It's the same idea as regular independence, but in the probability space restricted to C.

Conditional Independence: Examples

Informal Examples:

- Height and vocabulary in children: Not independent in general (older children tend to be taller and know more words). But given a fixed age, height and vocabulary might be (approximately) independent. Age is a lurking variable creating a dependence that disappears when conditioning on age.

Formal Example: Consider student performance in two subjects. Let $A = \{\text{student passes math}\}$, $B = \{\text{student passes English}\}$. These events are positively correlated in the overall population (a strong student is likely to pass both, a struggling student might fail both). However, let C represent the student's underlying ability level (e.g. $C = \{\text{high ability}\}$ vs $C^c = \{\text{low ability}\}$). Within each ability group, performance in math and English may be independent. Thus A and B can be considered *conditionally independent* given C, even though they are not independent marginally.

Conditional Independence: Examples

Informal Examples:

- Height and vocabulary in children: Not independent in general (older children tend to be taller and know more words). But given a fixed age, height and vocabulary might be (approximately) independent. Age is a lurking variable creating a dependence that disappears when conditioning on age.
- Ice cream sales and shark attacks: These are correlated (both higher in summer). However, given information about the weather/season, the number of ice cream sales and shark attacks become (nearly) independent. The weather "explains" the correlation.

Formal Example: Consider student performance in two subjects. Let $A = \{\text{student passes math}\}$, $B = \{\text{student passes English}\}$. These events are positively correlated in the overall population (a strong student is likely to pass both, a struggling student might fail both). However, let C represent the student's underlying ability level (e.g. $C = \{\text{high ability}\}$ vs $C^c = \{\text{low ability}\}$). Within each ability group, performance in math and English may be independent. Thus A and B can be considered *conditionally independent* given C, even though they are not independent marginally.

Coupon Collector Problem

Problem: There are N different types of coupons (e.g., trading cards). Each new coupon you collect is equally likely to be any of the N types (independent draws). How many coupons do you expect to collect to get at least one of each type?

Small cases:

- $N = 1$: Trivial, one draw gets the unique type.

General result: The expected number of draws T_N to collect all N types is

$$E[T_N] = N\left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{N}\right) = N \cdot H_N,$$

where H_N is the Nth harmonic number.

For large N, $H_N \approx \ln N + \gamma$ ($\gamma \approx 0.577$), so $E[T_N] \approx N \ln N + 0.577N$. For example, with $N = 50$, $E[T_{50}] \approx 50 \ln 50 \approx 196$ draws.

Coupon Collector Problem

Problem: There are N different types of coupons (e.g., trading cards). Each new coupon you collect is equally likely to be any of the N types (independent draws). How many coupons do you expect to collect to get at least one of each type?

Small cases:

- $N = 1$: Trivial, one draw gets the unique type.
- $N = 2$: The first draw gives some type. The chance the next draw is the *other* type is $1/2$. The number of draws needed to get the second type is geometric with $p = 1/2$, with expected value 2. So $E[T_2] = 1 + 2 = 3$.

General result: The expected number of draws T_N to collect all N types is

$$E[T_N] = N\left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{N}\right) = N \cdot H_N,$$

where H_N is the Nth harmonic number.

For large N, $H_N \approx \ln N + \gamma$ ($\gamma \approx 0.577$), so $E[T_N] \approx N \ln N + 0.577N$. For example, with $N = 50$, $E[T_{50}] \approx 50 \ln 50 \approx 196$ draws.

Coupon Collector Problem

Problem: There are N different types of coupons (e.g., trading cards). Each new coupon you collect is equally likely to be any of the N types (independent draws). How many coupons do you expect to collect to get at least one of each type?

Small cases:

- $N = 1$: Trivial, one draw gets the unique type.
- $N = 2$: The first draw gives some type. The chance the next draw is the *other* type is $1/2$. The number of draws needed to get the second type is geometric with $p = 1/2$, with expected value 2. So $E[T_2] = 1 + 2 = 3$.
- $N = 3$: $E[T_3] = 1$ (first type) $+\frac{3}{2}$ (expected draws to get one of the 2 remaining types) $+\frac{3}{1}$ (expected draws to get the last type) $= 1 + 1.5 + 3 = 5.5$.

General result: The expected number of draws T_N to collect all N types is

$$E[T_N] = N\left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{N}\right) = N \cdot H_N,$$

where H_N is the Nth harmonic number.
For large N, $H_N \approx \ln N + \gamma$ ($\gamma \approx 0.577$), so $E[T_N] \approx N \ln N + 0.577N$. For example, with $N = 50$, $E[T_{50}] \approx 50 \ln 50 \approx 196$ draws.

Coupon Collector: Analysis

Why does $E[T_N] = NH_N$? One intuitive derivation:

- Let $T_j = \#$ of extra coupons needed to go from $j-1$ collected types to j collected types. Then $T_N = \sum_{j=1}^{N} T_j$.

The coupon collector problem shows that even for moderate N, collecting all types takes quite a lot of trials (due to the long tail for that last new coupon).

Coupon Collector: Analysis

Why does $E[T_N] = NH_N$? One intuitive derivation:

- Let $T_j = \#$ of extra coupons needed to go from $j-1$ collected types to j collected types. Then $T_N = \sum_{j=1}^{N} T_j$.
- $T_1 = 1$ (first draw always yields a new type).

The coupon collector problem shows that even for moderate N, collecting all types takes quite a lot of trials (due to the long tail for that last new coupon).

Coupon Collector: Analysis

Why does $E[T_N] = NH_N$? One intuitive derivation:

- Let $T_j = \#$ of extra coupons needed to go from $j-1$ collected types to j collected types. Then $T_N = \sum_{j=1}^{N} T_j$.
- $T_1 = 1$ (first draw always yields a new type).
- If you have $j-1$ types, the probability the next coupon is a new type is $\frac{N-(j-1)}{N}$. So T_j is geometric with success probability $(N-j+1)/N$ and $E[T_j] = \frac{1}{(N-j+1)/N} = \frac{N}{N-j+1}$.

The coupon collector problem shows that even for moderate N, collecting all types takes quite a lot of trials (due to the long tail for that last new coupon).

Coupon Collector: Analysis

Why does $E[T_N] = NH_N$? One intuitive derivation:

- Let $T_j = \#$ of extra coupons needed to go from $j-1$ collected types to j collected types. Then $T_N = \sum_{j=1}^{N} T_j$.
- $T_1 = 1$ (first draw always yields a new type).
- If you have $j-1$ types, the probability the next coupon is a new type is $\frac{N-(j-1)}{N}$. So T_j is geometric with success probability $(N-j+1)/N$ and $E[T_j] = \frac{1}{(N-j+1)/N} = \frac{N}{N-j+1}$.
- Thus:

$$E[T_N] = \sum_{j=1}^{N} \frac{N}{N-j+1} = N \sum_{k=1}^{N} \frac{1}{k} = NH_N.$$

The coupon collector problem shows that even for moderate N, collecting all types takes quite a lot of trials (due to the long tail for that last new coupon).

Law of Total Probability (LTP)

Suppose $\{B_1, \ldots, B_n\}$ is a **partition** of Ω (disjoint B_i whose union is Ω).
Then for any event A:

$$A = \bigcup_{i=1}^{n} (A \cap B_i),$$

a union of disjoint events. By additivity:

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i).$$

Using $P(A \cap B_i) = P(A \mid B_i)P(B_i)$:

$$\boxed{P(A) = \sum_{i=1}^{n} P(A \mid B_i)\, P(B_i).}$$

This is the Law of Total Probability.

LTP: Special Case & Bayes' Theorem

For partition $\{B, B^c\}$:

$$P(A) = P(A \mid B)P(B) + P(A \mid B^c)P(B^c).$$

This splits $P(A)$ based on whether B occurred.

Bayes' Theorem: For partition $\{B_i\}$:

$$P(B_j \mid A) = \frac{P(A \mid B_j)P(B_j)}{\sum_i P(A \mid B_i)P(B_i)}.$$

This allows inverting conditional probabilities. It's especially useful in diagnostic scenarios (like our false positive example) to find $P(\text{cause} \mid \text{evidence})$ from $P(\text{evidence} \mid \text{cause})$.

LTP Example: Disease Prevalence

5% of men and 1% of women have dichromacy (color blindness). If 60% of a population is female, what is the probability a random person has the condition?

Let D = person is dichromatic, F = person is female. We have:

$$P(D \mid F^c) = 0.05, \quad P(D \mid F) = 0.01, \quad P(F) = 0.6, \; P(F^c) = 0.4.$$

By LTP (partition on gender):

$$P(D) = P(D \mid F)P(F) + P(D \mid F^c)P(F^c) = 0.01(0.6) + 0.05(0.4).$$

Compute: $= 0.006 + 0.020 = 0.026 = 2.6\%$.

So about 2.6% of the population is dichromatic. Using Bayes' theorem, one can find that a dichromatic person from this population is about 77% likely to be male (since men have higher prevalence).

Another LTP Example

A factory has two machines producing widgets. Machine A produces 70% of widgets with a 2% defect rate. Machine B produces 30% with a 5% defect rate. If we pick a random widget, what's $P(\text{defective})$?

Partition by machine:

$$\begin{aligned} P(\text{defective}) &= P(\text{def}|A)P(A) + P(\text{def}|B)P(B) \\ &= 0.02(0.70) + 0.05(0.30) \\ &= 0.014 + 0.015 = 0.029 = 2.9\%. \end{aligned}$$

So 2.9% of widgets are defective.

If a widget is defective, Bayes' theorem tells us it's more likely from Machine B (despite B's lower output share) because B has a higher defect probability. That is:

$$P(B \mid \text{def}) = \frac{0.05(0.30)}{0.029} \approx 0.517 \ (\approx 51.7\%).$$

Moments of a Random Variable

If X is a random variable, the nth **moment** of X is $E[X^n]$ (assuming it exists).

- $E[X]$ is the first moment (the mean).

Moments describe aspects of a distribution (variance is based on second moment, skewness on third, etc.). But finding $E[X^n]$ directly can be tedious.

Idea: Collect all moments into a single function. This leads to the **moment generating function**.

Moments of a Random Variable

If X is a random variable, the nth **moment** of X is $E[X^n]$ (assuming it exists).

- $E[X]$ is the first moment (the mean).
- $E[X^2]$ is the second moment, etc.

Moments describe aspects of a distribution (variance is based on second moment, skewness on third, etc.). But finding $E[X^n]$ directly can be tedious.

Idea: Collect all moments into a single function. This leads to the **moment generating function**.

Definition: Moment Generating Function

For a random variable X, the **moment generating function (MGF)** is defined as:

$$M_X(t) = E[e^{tX}],$$

for all t where this expectation exists.

In particular:

$$M_X(t) = \begin{cases} \sum_x e^{tx} \, P(X = x), & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} e^{tx} f_X(x) \, dx, & \text{if } X \text{ is continuous with PDF } f_X, \end{cases}$$

within the t-range of convergence.

If $M_X(t)$ exists in an interval around 0, it uniquely determines the distribution of X. It's called moment generating because its derivatives yield the moments.

MGF and Moments

Expand e^{tX} as a power series:

$$e^{tX} = 1 + tX + \frac{t^2X^2}{2!} + \frac{t^3X^3}{3!} + \cdots .$$

Taking expectation:

$$M_X(t) = E[e^{tX}] = 1 + t\,E[X] + \frac{t^2}{2!}E[X^2] + \frac{t^3}{3!}E[X^3] + \cdots .$$

Thus, the coefficients of the power series for $M_X(t)$ give the moments:

- $M_X(0) = 1$.

So if we find a nice formula for $M_X(t)$, we can differentiate to obtain moments instead of computing integrals/sums for each $E[X^n]$ separately.

MGF and Moments

Expand e^{tX} as a power series:

$$e^{tX} = 1 + tX + \frac{t^2X^2}{2!} + \frac{t^3X^3}{3!} + \cdots .$$

Taking expectation:

$$M_X(t) = E[e^{tX}] = 1 + t\,E[X] + \frac{t^2}{2!}E[X^2] + \frac{t^3}{3!}E[X^3] + \cdots .$$

Thus, the coefficients of the power series for $M_X(t)$ give the moments:

- $M_X(0) = 1.$
- $M'_X(0) = E[X].$

So if we find a nice formula for $M_X(t)$, we can differentiate to obtain moments instead of computing integrals/sums for each $E[X^n]$ separately.

MGF and Moments

Expand e^{tX} as a power series:

$$e^{tX} = 1 + tX + \frac{t^2X^2}{2!} + \frac{t^3X^3}{3!} + \cdots .$$

Taking expectation:

$$M_X(t) = E[e^{tX}] = 1 + t\,E[X] + \frac{t^2}{2!}E[X^2] + \frac{t^3}{3!}E[X^3] + \cdots .$$

Thus, the coefficients of the power series for $M_X(t)$ give the moments:

- $M_X(0) = 1.$
- $M'_X(0) = E[X].$
- $M''_X(0) = E[X^2].$

So if we find a nice formula for $M_X(t)$, we can differentiate to obtain moments instead of computing integrals/sums for each $E[X^n]$ separately.

MGF and Moments

Expand e^{tX} as a power series:

$$e^{tX} = 1 + tX + \frac{t^2X^2}{2!} + \frac{t^3X^3}{3!} + \cdots .$$

Taking expectation:

$$M_X(t) = E[e^{tX}] = 1 + t\,E[X] + \frac{t^2}{2!}E[X^2] + \frac{t^3}{3!}E[X^3] + \cdots .$$

Thus, the coefficients of the power series for $M_X(t)$ give the moments:

- $M_X(0) = 1$.
- $M_X'(0) = E[X]$.
- $M_X''(0) = E[X^2]$.
- In general, $M_X^{(n)}(0) = E[X^n]$.

So if we find a nice formula for $M_X(t)$, we can differentiate to obtain moments instead of computing integrals/sums for each $E[X^n]$ separately.

Why Use MGFs?

1. They make finding moments easier: $E[X^n] = M_X^{(n)}(0)$.

Why Use MGFs?

1. They make finding moments easier: $E[X^n] = M_X^{(n)}(0)$.
2. MGFs (when they exist) uniquely determine the distribution. If two random variables have the same MGF (in a neighborhood of 0), they have the same distribution.

Why Use MGFs?

1. They make finding moments easier: $E[X^n] = M_X^{(n)}(0)$.
2. MGFs (when they exist) uniquely determine the distribution. If two random variables have the same MGF (in a neighborhood of 0), they have the same distribution.
3. MGFs convert convolutions into products: If X and Y are independent,

$$M_{X+Y}(t) = E[e^{t(X+Y)}] = E[e^{tX} e^{tY}] = E[e^{tX}]\, E[e^{tY}] = M_X(t)\, M_Y(t).$$

   Thus the MGF of a sum is the product of MGFs. This is extremely useful for finding distributions of sums of independent variables (e.g. sum of independent Poissons, etc.).

Example: Bernoulli MGF

Let $X \sim \text{Bernoulli}(p)$ (takes value 1 with prob p, 0 with prob $1-p$).
The MGF is
$$M_X(t) = E[e^{tX}] = e^{t\cdot 0}(1-p) + e^{t\cdot 1}(p) = (1-p) + p\,e^t.$$

Using it:

- $M_X(0) = (1-p) + p = 1.$

Example: Bernoulli MGF

Let $X \sim \text{Bernoulli}(p)$ (takes value 1 with prob p, 0 with prob $1-p$).
The MGF is

$$M_X(t) = E[e^{tX}] = e^{t\cdot 0}(1-p) + e^{t\cdot 1}(p) = (1-p) + p\,e^t.$$

Using it:

- $M_X(0) = (1-p) + p = 1$.
- $M_X'(t) = p\,e^t$, so $M_X'(0) = p = E[X]$ (correct).

Example: Bernoulli MGF

Let $X \sim \text{Bernoulli}(p)$ (takes value 1 with prob p, 0 with prob $1-p$).
The MGF is

$$M_X(t) = E[e^{tX}] = e^{t\cdot 0}(1-p) + e^{t\cdot 1}(p) = (1-p) + p\,e^t.$$

Using it:

- $M_X(0) = (1-p) + p = 1$.
- $M_X'(t) = p\,e^t$, so $M_X'(0) = p = E[X]$ (correct).
- $M_X''(t) = p\,e^t$ as well, so $M_X''(0) = p$. But $E[X^2] = p$ for a Bernoulli (since $X^2 = X$ when $X \in \{0, 1\}$).

Example: Bernoulli MGF

Let $X \sim \text{Bernoulli}(p)$ (takes value 1 with prob p, 0 with prob $1-p$).
The MGF is

$$M_X(t) = E[e^{tX}] = e^{t\cdot 0}(1-p) + e^{t\cdot 1}(p) = (1-p) + p\,e^t.$$

Using it:

- $M_X(0) = (1-p) + p = 1$.
- $M_X'(t) = p\,e^t$, so $M_X'(0) = p = E[X]$ (correct).
- $M_X''(t) = p\,e^t$ as well, so $M_X''(0) = p$. But $E[X^2] = p$ for a Bernoulli (since $X^2 = X$ when $X \in \{0,1\}$).
- We recover $\text{Var}(X) = E[X^2] - (E[X])^2 = p - p^2 = p(1-p)$.

Example: Binomial MGF

If $X \sim \text{Binomial}(n, p)$, think of $X = X_1 + \cdots + X_n$ as sum of n independent Bernoulli(p) trials.

We have $M_{X_i}(t) = (1 - p) + pe^t$ for each trial. By independence:

$$M_X(t) = [M_{X_1}(t)]^n = [(1 - p) + p\, e^t]^n.$$

For example:

- $M'_X(t) = n[(1 - p) + pe^t]^{\,n-1} \cdot pe^t$. So $M'_X(0) = n(1 - p + p)^{\,n-1}p = np$. Thus $E[X] = np$.

But the MGF derivation nicely shows a Binomial is a sum of n i.i.d. Bernoullis by the factorization $[(1 - p) + pe^t]^n$.

Example: Binomial MGF

If $X \sim \text{Binomial}(n, p)$, think of $X = X_1 + \cdots + X_n$ as sum of n independent Bernoulli(p) trials.

We have $M_{X_i}(t) = (1 - p) + pe^t$ for each trial. By independence:

$$M_X(t) = [M_{X_1}(t)]^n = [(1 - p) + p\, e^t]^n.$$

For example:

- $M'_X(t) = n[(1 - p) + pe^t]^{\,n-1} \cdot pe^t$. So $M'_X(0) = n(1 - p + p)^{\,n-1} p = np$. Thus $E[X] = np$.
- $M''_X(0)$ would give $E[X^2]$, etc. It's easier to use known formulas (like $\text{Var}(X) = np(1 - p)$).

But the MGF derivation nicely shows a Binomial is a sum of n i.i.d. Bernoullis by the factorization $[(1 - p) + pe^t]^n$.

Example: Poisson MGF

If $X \sim \text{Poisson}(\lambda)$:

$$P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots$$

Then the MGF is

$$\begin{aligned} M_X(t) &= \sum_{k=0}^{\infty} e^{tk} e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda e^t)^k}{k!} \\ &= e^{-\lambda}\exp(\lambda e^t) = \exp(\lambda(e^t - 1)). \end{aligned}$$

From this:

- $M_X'(0) = \lambda e^0 (= \lambda)$, so $E[X] = \lambda$.

Also, if $X \sim \text{Poisson}(\lambda_1)$ and $Y \sim \text{Poisson}(\lambda_2)$ independent, $M_{X+Y}(t) = \exp(\lambda_1(e^t - 1))\exp(\lambda_2(e^t - 1)) = \exp((\lambda_1 + \lambda_2)(e^t - 1))$. This is the MGF of $\text{Poisson}(\lambda_1 + \lambda_2)$. (Sums of independent Poissons are Poisson.)

Example: Poisson MGF

If $X \sim \text{Poisson}(\lambda)$:

$$P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots$$

Then the MGF is

$$\begin{aligned} M_X(t) &= \sum_{k=0}^{\infty} e^{tk} e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda e^t)^k}{k!} \\ &= e^{-\lambda}\exp(\lambda e^t) = \exp(\lambda(e^t - 1)). \end{aligned}$$

From this:

- $M_X'(0) = \lambda e^0 (= \lambda)$, so $E[X] = \lambda$.
- $M_X''(0) = \lambda e^0 + \lambda^2 e^0 = \lambda + \lambda^2$, so $E[X^2] = \lambda + \lambda^2$. Then $\text{Var}(X) = \lambda$.

Also, if $X \sim \text{Poisson}(\lambda_1)$ and $Y \sim \text{Poisson}(\lambda_2)$ independent, $M_{X+Y}(t) = \exp(\lambda_1(e^t - 1))\exp(\lambda_2(e^t - 1)) = \exp((\lambda_1 + \lambda_2)(e^t - 1))$. This is the MGF of $\text{Poisson}(\lambda_1 + \lambda_2)$. (Sums of independent Poissons are Poisson.)

Example: Exponential MGF

If $X \sim \text{Exponential}(\beta)$ (rate β, mean $1/\beta$), PDF $f(x) = \beta e^{-\beta x}$ for $x \geq 0$.
The MGF:

$$M_X(t) = \int_0^\infty e^{tx} \beta e^{-\beta x} dx = \beta \int_0^\infty e^{-(\beta - t)x} dx, \quad (\text{for } t < \beta)$$
$$= \beta \left[\frac{1}{\beta - t} \right]_{x=0}^\infty = \frac{\beta}{\beta - t}, \qquad (t < \beta).$$

Now:

- $M_X'(t) = \frac{\beta}{(\beta - t)^2}$. So $M_X'(0) = \frac{\beta}{\beta^2} = \frac{1}{\beta} = E[X]$.

As expected, an $\text{Exp}(\beta)$ has variance $1/\beta^2$.

Example: Exponential MGF

If $X \sim \text{Exponential}(\beta)$ (rate β, mean $1/\beta$), PDF $f(x) = \beta e^{-\beta x}$ for $x \geq 0$.
The MGF:

$$\begin{aligned} M_X(t) &= \int_0^\infty e^{tx} \beta e^{-\beta x} dx = \beta \int_0^\infty e^{-(\beta - t)x} dx, \quad (\text{for } t < \beta) \\ &= \beta \left[\frac{1}{\beta - t} \right]_{x=0}^{\infty} = \frac{\beta}{\beta - t}, \qquad (t < \beta). \end{aligned}$$

Now:

- $M_X'(t) = \frac{\beta}{(\beta - t)^2}$. So $M_X'(0) = \frac{\beta}{\beta^2} = \frac{1}{\beta} = E[X]$.
- $M_X''(0) = \frac{2\beta}{\beta^3} = \frac{2}{\beta^2}$. Then $E[X^2] = 2/\beta^2$, giving $\text{Var}(X) = 2/\beta^2 - (1/\beta)^2 = 1/\beta^2$.

As expected, an $\text{Exp}(\beta)$ has variance $1/\beta^2$.