Probability and Conditional Probability

Emile Latour, Nicky Wakim, Meike Niederhausen

January 12, 2026

Where we are

On Monday we worked with:

- frequency tables
- row and column percentages
- contingency tables

Today we formalize those ideas using **probability notation**.

Learning goals

By the end of class, you should be able to:

- Interpret probabilities using notation
- Distinguish joint, marginal, and conditional probabilities
- Compute conditional probabilities from tables
- Understand when events are independent

Probability basics

What is probability?

A **probability** is a number between 0 and 1 that describes **how likely an event is**.

- 0 = impossible
- 1 = certain

A more technical definition:

> probability of an outcome is the proportion of times the outcome would occur if the random phenomenon could be observed an infinite number of times.

Probability notation

We write probabilities using:

$$P(A)$$

which means: **the probability that event A occurs**

Examples:

- $P(\text{Heads})$
- $P(\text{Rolling a } 6)$

Probability notation (continued)

If an event *A* occurs *m* times out of a total of *n* identical trials, then

$$P(A) = \frac{m}{n} = \frac{\text{number of times } A \text{ occurs}}{\text{number of trials}}$$

Examples:

- Flip a fair coin 10 times and record the proportion of heads.
- Rolling a six-sided die lots of times, and recording proportion of times a 6 appears.

Is the coin fair?

- We can think of flipping a coin.
  - There are two possible outcomes (heads or tails).
  - The probability of getting heads is 0.5.

- If we flip the coin 10 times, it is not certain that we will get 5 heads.
- However, if we flip it **enough** times, we will get heads 50% of the flips.

- **Fun "Seeing Theory" demonstration!**

Law of Large Numbers (intuition)

- Impractical to conduct “infinitely” many trials to determined probabilities
- Instead estimate probabilities using the proportion (of times an event occurs) from “large” samples

Law of large numbers If we repeat a random experiment many times:

- the long-run proportion of times an event occurs
- approaches the true probability

This is why we interpret probability as a **long-run relative frequency**.

Complement rule

If A is an event, then:

$$P(A^c) = 1 - P(A)$$

Example:

- If $P(\text{Rain}) = 0.3$
- then $P(\text{No rain}) = 0.7$

Addition rules (working with probability)

Disjoint events

Two events are **disjoint** if they cannot occur at the same time.

If A and B are disjoint:

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

- Also use the term **mutually exclusive** to mean disjoint.
- Can see "or" represented by the $\cup$ symbol. So $P(A \cup B)$.

Example: rolling a die

Let:

- A = rolling a 1
- B = rolling a 6

These are disjoint.

$$P(A \text{ or } B) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6}$$

k disjoint outcomes

If there are k disjoint outcomes A_1, ..., A_k, then the probability that either one of these outcomes will occur is

$$P(A_1) + P(A_2) + \cdots + P(A_k)$$

Example: probability of rolling a 1, 2, 3, 4, 5, or 6.

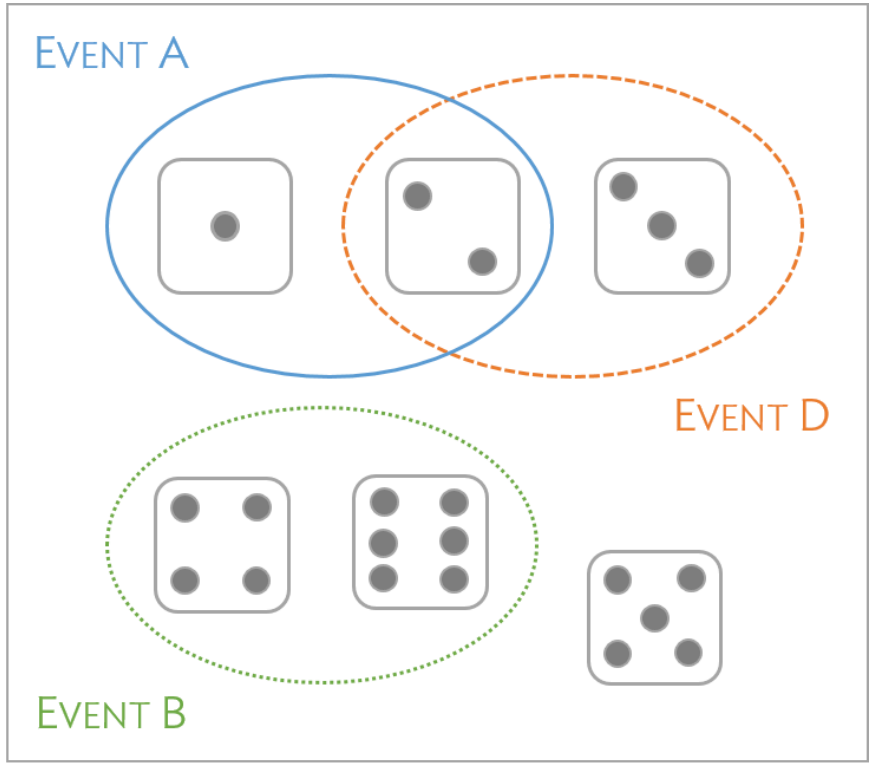
$$\begin{aligned} P(1 \text{ or } 2 \text{ or } 3 \text{ or } 4 \text{ or } 5 \text{ or } 6) &= P(1) + P(2) + P(3) + P(4) + P(5) + P(6) \\ &= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} \\ &= 1 \end{aligned}$$

Textbook's die rolling example

Let:

- A = rolling a 1 or 2
- B = rolling a 4 or 6

$$P(A \text{ or } B) = P(A) + P(B) = \frac{2}{6} + \frac{2}{6} = \frac{2}{6} = \frac{2}{3}$$



Non-disjoint events

What if events CAN occur at the same time?

Example:

- Drawing a card that is a **diamond**
- Drawing a card that is a **face card** (Jack, Queen, King)

These events **overlap** - a card can be both!

General Addition Rule

If A and B are any two events (disjoint or not):

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

We subtract $P(A \text{ and } B)$ because those outcomes were counted twice.

This formula is sometimes written:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Example: Cards

- $P(\text{diamond}) = \frac{13}{52}$
- $P(\text{face card}) = \frac{12}{52}$
- $P(\text{diamond and face card}) = \frac{3}{52}$ (J♦, Q♦, K♦)

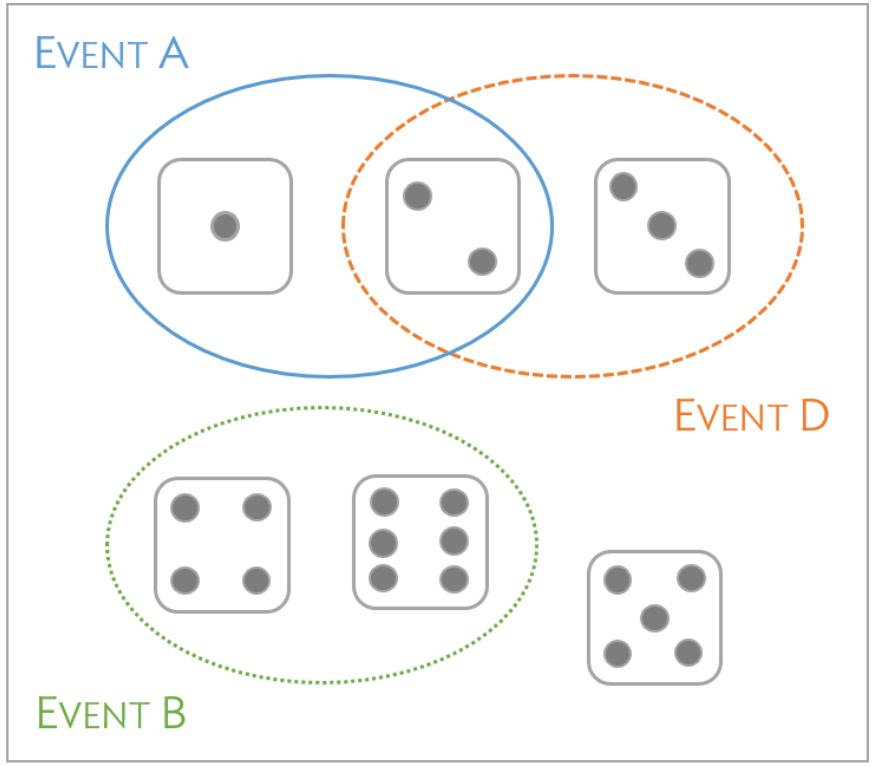
$$\begin{aligned} P(\text{diamond or face card}) &= \frac{13}{52} + \frac{12}{52} - \frac{3}{52} \\ &= \frac{22}{52} \end{aligned}$$

Dice again

Let:

- A = rolling a 1 or 2
- D = rolling a 2 or 3

$$
\begin{aligned}
P(\text{A or D}) &= \frac{2}{6} + \frac{2}{6} - \frac{1}{6} \\
&= \frac{3}{6} \\
&= \frac{1}{2}
\end{aligned}
$$



General Addition Rule when events are disjoint

General Addition Rule again

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

If **events are disjoint**, then $P(A \text{ and } B) = 0$, and we get the same result we saw before for disjoint events.

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

Connecting back to Monday

Remember the contingency tables from Monday?

sex/actn3.r577x	CC	CT	TT	Total
Female	106	149	98	353
Male	67	112	63	242
Total	173	261	161	595

- We calculated **row proportions** and
- Today we're going to give these forma

Contingency tables – tables as probabilities

A simpler example for today

For teaching purposes, let's work with a simpler 2×2 table:

	Disease	No Disease	Total
Test +	90	180	270
Test -	10	720	730
Total	100	900	1000

This represents 1,000 patients who were tested for a disease.

Joint probabilities

Joint probabilities involve **two events happening together**.

Examples:

- $P(\text{Disease and Test } +) = 90/1000$
- $P(\text{No Disease and Test -}) = 720/1000$

	Disease	No Disease	Total
Test +	90	180	270
Test -	10	720	730
Total	100	900	1000

Marginal probabilities

Marginal probabilities ignore the other variable.

Examples:

- $P(\text{Disease}) = 100/1000$
- $P(\text{Test} +) = 270/1000$

	Disease	No Disease	Total
Test +	90	180	270
Test -	10	720	730
Total	100	900	1000

Marginal vs Joint

It can help to remember when looking at a table that:

- **Joint probability**: intersection of row and column
- **Marginal probability**: row or column total

Marginal vs Joint vs Conditional

From the table:

- **Marginal**: $P(\text{Disease}) = 100/1000$ (ignore test results)
- **Joint**: $P(\text{Disease and Test} +) = 90/1000$ (both happening)
- **Conditional**: $P(\text{Disease} \mid \text{Test} +) = ?$ (we'll calculate next)

Key difference: Marginal ignores other variables, conditional focuses within a subgroup.

Conditional probabilities

Conditional probability answers:

Given that B occurred, how likely is A?

Notation:

$$P(A \mid B)$$

Definition of conditional probability

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

This is just a **restricted proportion**.

- We're only looking at cases where B occurred (the denominator)
- Then asking: of those cases, how many also have A?
- We've "restricted" our view to just the B group

- **Example:** $P(\text{Disease} \mid \text{Test} +)$
  - Denominator: only people who tested positive
  - Numerator: of those, how many have disease?
  - We've restricted to the "Test+" subgroup

Example from the table

What is $P(\text{Disease} \mid \text{Test} +)$?

$$
\begin{aligned}
P(\text{Disease} \mid \text{Test} +) &= \frac{P(Disease \text{ and } Test+)}{P(Test+)} \\
&= \frac{90/1000}{270/1000} \\
&= \frac{90}{270} \\
&\approx 0.333
\end{aligned}
$$

	Disease	No Disease	Total
Test +	90	180	270
Test -	10	720	730
Total	100	900	1000

Interpretation: Among those who tested positive, about 33% have the disease.

This is what you did Monday!

When you calculated **row proportions**:

- You were computing $P(\text{genotype} \mid \text{sex})$

sex/actn3.r577x	CC	CT	TT	Total
Female	106	149	98	353
Male	67	112	63	242

sex/actn3.r577x	CC	CT	TT	Total
Female	0.30	0.42	0.28	1.00
Male	0.28	0.46	0.26	1.00

This is what you did Monday!

When you calculated **column proportions**:

- You were computing $P(\text{sex} \mid \text{genotype})$

sex/actn3.r577x	CC	CT	TT
Female	106	149	98
Male	67	112	63
Total	173	261	161

sex/actn3.r577x	CC	CT	TT
Female	0.61	0.57	0.61
Male	0.39	0.43	0.39
Total	1.00	1.00	1.00

The formula just formalizes what you already know how to do.

Independence

What is independence?

Two events are **independent** if knowing one provides no information about the other.

Examples:

- Flipping a coin and rolling a die (independent)
- Your height and the weather tomorrow (independent)
- Having a disease and testing positive (NOT independent!)
- Smoking and lung cancer (NOT independent!)

Checking for independence

Events A and B are independent if:

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

OR equivalently:

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

Why are these equivalent? Starting with the conditional probability formula:

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

Key idea: "Knowing B doesn't change the probability of A"

Example: Independence

Roll two dice. Are the outcomes independent?

- $P(\text{first die} = 1) = \frac{1}{6}$
- $P(\text{second die} = 1) = \frac{1}{6}$
- $P(\text{both} = 1) = \frac{1}{36}$

Check: $\frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$ ✓

The events are independent.

Example: NOT independent

From our medical test example:

- $P(\text{Disease}) = \frac{100}{1000} = 0.10$
- $P(\text{Disease} \mid \text{Test} +) = \frac{90}{270} \approx 0.33$

Since $0.33 \neq 0.10$, the events are **NOT independent**.

Testing positive changes the probability of disease! This is why tests are useful.

Multiplication rule

General multiplication rule

The General multiplication rule

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

which follows from rearranging the definition of conditional probability:

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

This connects **joint** and **conditional** probabilities.

Example: Multiplication rule

What is the probability of randomly selecting a person who is male AND has hypertension?

Given:

- $P(\text{male}) = 0.50$
- $P(\text{hypertension} \mid \text{male}) = 0.18$

$$\begin{aligned} P(\text{male and hypertension}) &= P(\text{hypertension} \mid \text{male}) \times P(\text{male}) \\ &= 0.18 \times 0.50 \\ &= 0.09 \end{aligned}$$

Multiplication rule for independent events

When events are **independent**, the multiplication rule simplifies:

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

Why? Because when events are independent, $P(A \mid B) = P(A)$.

Example: Rolling two dice

$$P(\text{both are } 6) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$

Positive Predictive Value (PPV)

A medical testing scenario

Given information:

- Prevalence (disease rate in population) = 1%
- Sensitivity (correct positive test result when disease present) = 90%
- Specificity (correct negative test result when disease absent) = 90%

Question: If someone tests positive, what's the probability they actually have the disease?

This is asking for: $P(\text{Disease} \mid \text{Test} +)$

Build a contingency table

Let's think about 10,000 people:

	Disease	No Disease	Total
Test +	?	?	?
Test -	?	?	?
Total	?	?	10,000

Let's fill this in step by step...

Building the table using what we learned

We'll use:

- **Marginal probabilities** (prevalence)
- **Conditional probabilities** (sensitivity, specificity)
- **Multiplication rule**: $P(A \text{ and } B) = P(A \mid B) \times P(B)$

This shows how all the concepts work together!

Step 1: How many have the disease?

Prevalence = 1%, so out of 10,000 people:

	Disease	No Disease	Total
Test +	?	?	?
Test -	?	?	?
Total	100	9,900	10,000

- 1% of 10,000 = 100 people have disease
- 99% of 10,000 = 9,900 people don't have disease

Step 2: Among those WITH disease, how many test positive?

Sensitivity = 90%, so of the 100 with disease:

	Disease	No Disease	Total
Test +	90	?	?
Test -	10	?	?
Total	100	9,900	10,000

- 90% of 100 = 90 test positive (true positives)
  - Using multiplication rule: $P(\text{Test} + \mid \text{Disease}) \times 100$
- 10% of 100 = 10 test negative (false negatives)

Step 3: Among those WITHOUT disease, how many test positive?

Specificity = 90% (so 10% test positive when they don't have disease):

	Disease	No Disease	Total
Test +	90	990	?
Test -	10	8,910	?
Total	100	9,900	10,000

- 10% of 9,900 = 990 test positive (false positives)
  - Using multiplication rule: $P(\text{Test} + \mid \text{No Disease}) \times 9,900$
- 90% of 9,900 = 8,910 test negative (true negatives)

Step 4: Fill in the row totals

Now we can complete the table:

	Disease	No Disease	Total
Test +	90	990	1,080
Test -	10	8,910	8,920
Total	100	9,900	10,000

- Total positive tests: 90 + 990 = 1,080
- Total negative tests: 10 + 8,910 = 8,920

Calculate the PPV

$$
\begin{aligned}
P(\text{Disease} \mid \text{Test} +) &= \frac{90}{1,080} \\
&\approx 0.083 \\
&= 8.3\%
\end{aligned}
$$

	Disease	No Disease	Total
Test +	90	990	1,080
Test -	10	8,910	8,920
Total	100	9,900	10,000

Only about **8%** of positive tests indicate disease!

Key takeaway

Even a good test (90% sensitivity, 90% specificity) can have a low PPV when the condition is rare.

Why?

- Only 100 people have the disease
- But 990 people test positive incorrectly
- The false positives outnumber the true positives!

Base rates matter.

Next class: We'll see how Bayes' theorem formalizes this calculation.

Recap

What to remember

- Probability is a **long-run relative frequency** (between 0 and 1)
- **Addition rules**: disjoint vs non-disjoint events
- **Contingency tables** show joint and marginal probabilities
- **Conditional probability**: $P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}$
  - This formalizes the row/column proportions you did Monday!
- **Independence**: $P(A \mid B) = P(A)$ (knowing B doesn't change A)
- **Multiplication rule**: $P(A \text{ and } B) = P(A \mid B) \times P(B)$
- **Base rates matter** for interpreting test results

Next class (after MLK Day): Bayes' theorem and doing this in R with tidyverse!