



Introduction to Data Science

Fundamentals of Probability

15 Sep 2026



Fundamentals of Probability

Today's Route

1. Describe uncertainty

Use outcomes, events, and probabilities to describe a coin or a die.

2. Use information

Change the probability when relevant information is given.

3. Summarize random quantities

Use expectation and variance to understand coin counts and dropout.

Start with outcomes, then summarize what varies.

Events and Probability

Why Do We Need Probability?

FAIR COIN



Two possible results: heads or tails.

Before the toss, we do not know which result will occur.

$$P(\text{heads}) = 1/2$$

FAIR SIX-SIDED DICE



Six possible results: 1, 2, 3, 4, 5, 6.

Before the roll, we do not know which face will appear.

$$P(\text{roll a 5}) = 1/6$$

Probability gives a number to our uncertainty about an outcome.

Events and Probability

Outcome, Sample Space, and Event

Outcome

One result of the experiment, such as rolling a 5.



Sample space S

Every possible result: $S = \{1, 2, 3, 4, 5, 6\}$.

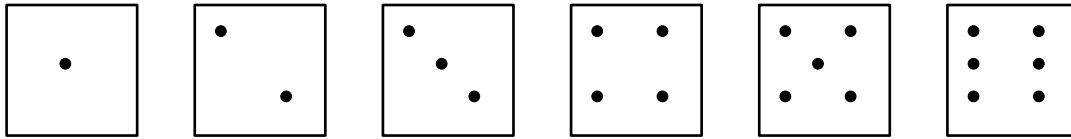
Event E

The results that meet our condition: “at least 3” gives $E = \{3, 4, 5, 6\}$.

An outcome is one result. An event is a set of results.

Events and Probability

Example: At Least a Three



“At least 3” means 3, 4, 5, or 6.

$$P(E) = \frac{4 \text{ favorable outcomes}}{6 \text{ possible outcomes}}$$

The answer is $4/6 = 2/3$. We include 3 because “at least” includes the boundary.

For this fair dice, four of the six equally likely outcomes meet the condition.

Events and Probability

A Probability Distribution

A distribution assigns a probability to every possible outcome.



Face	1	2	3	4	5	6
Probability	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$

$$1/6 + 1/6 + 1/6 + 1/6 + 1/6 + 1/6 = 1$$

Discrete outcomes can be listed: the list may be finite or countably infinite.

All outcome probabilities are nonnegative and sum to one.

Events and Probability

Frequentist and Bayesian Views

FREQUENTIST VIEW

Roll the dice many times.

Use the long-run fraction of rolls that show a particular face.

For a fair dice, that fraction approaches $1/6$.

BAYESIAN VIEW

Start with a belief about the possible outcomes.

Without a reason to favor one face of a fair dice, assign each face $1/6$.

Update beliefs when new data are observed.

The two views explain probability differently; both can describe the same fair die.

Events and Probability

Prior, Data, and Posterior

Prior

Our belief before seeing the new observations.

Data

New outcomes from the experiment, such as additional dice rolls.

Posterior

The updated belief after those observations are taken into account.

Bayesian neural networks place distributions on weights and update them using training data.

A posterior is an updated belief after observing data.

Events and Probability

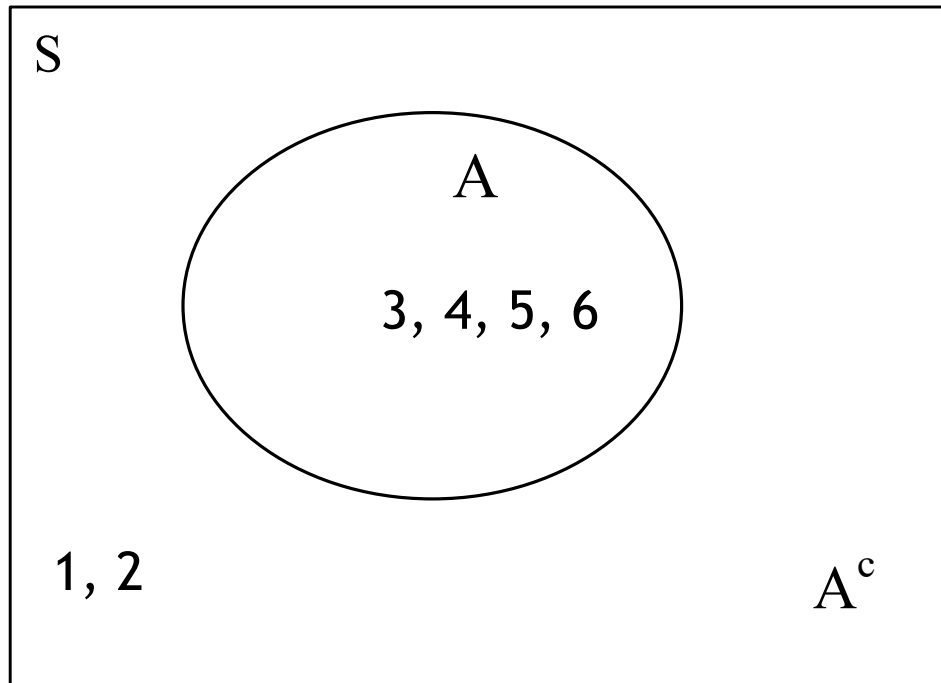
The Four Tenets

Rule	Meaning
Total probability	All outcomes together have probability 1.
Complement	What is outside an event has the remaining probability.
Subset	An event inside another cannot be more probable.
Inclusion and exclusion	Subtract the overlap when combining two events.

These four rules organize probabilities over events.

Events and Probability

An Event and Its Complement



A contains 3, 4, 5, and 6.

The complement, “not A”, contains 1 and 2.

Nothing is left outside these two sets.

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

$$4/6 = 1 - 2/6$$



An event and its complement cover the entire sample space.

Events and Probability

One Event Inside Another

Use the same fair six-sided dice.

SMALLER EVENT

$$E1 = \{6\}$$

Only a 6 satisfies this event.

LARGER EVENT

$$E2 = \{3, 4, 5, 6\}$$

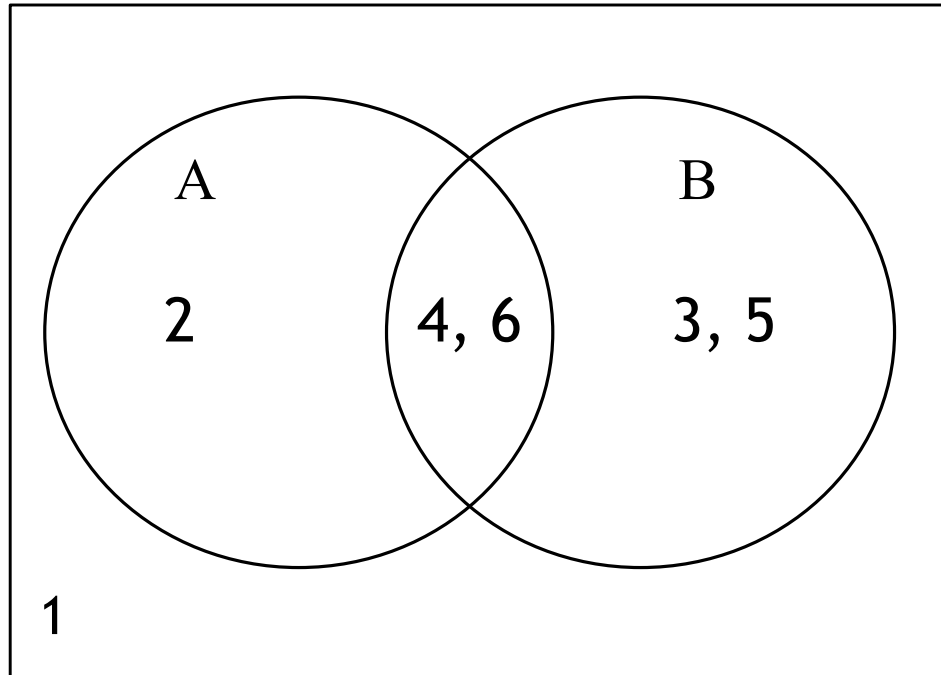
Every outcome in E1 is also in E2.

$$P(E1) = 1/6 \leq P(E2) = 4/6$$

A smaller event cannot have more probability than an event containing it.

Events and Probability

Union and Intersection



A: roll an even number.

B: roll at least 3.



$$A \cap B = \{4, 6\}$$

Both conditions are satisfied.

$$A \cup B = \{2, 3, 4, 5, 6\}$$

At least one condition is satisfied.

The union means A or B. The intersection means both A and B.

Events and Probability

Why Must We Subtract the Overlap?

Keep $A = \{2, 4, 6\}$ and $B = \{3, 4, 5, 6\}$.



$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= 3/6 + 4/6 - 2/6 = 5/6 \end{aligned}$$

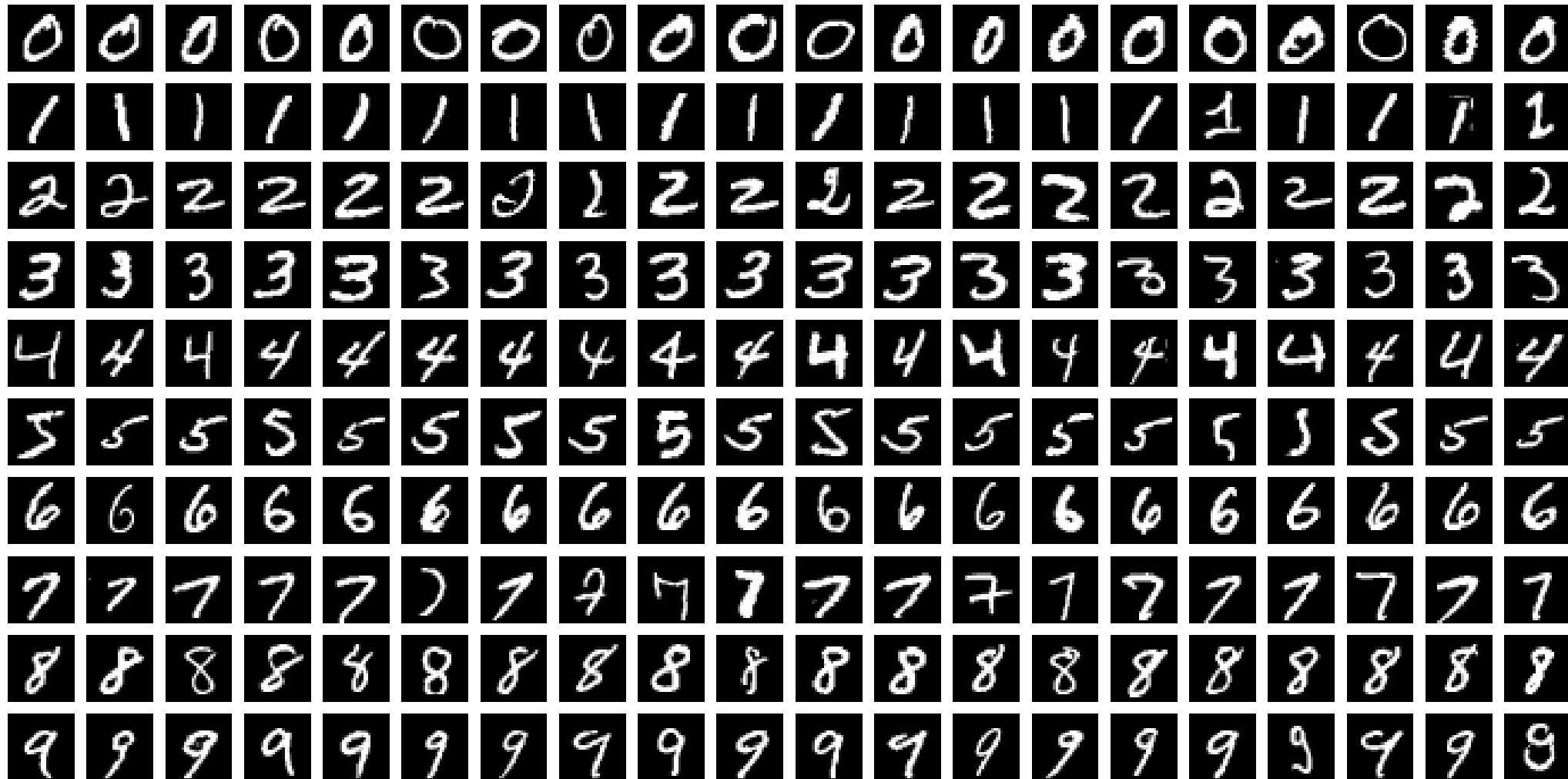
The outcomes 4 and 6 appear in both event lists.

Subtract their probability once, so every outcome in the union is counted once.

Adding $P(A)$ and $P(B)$ counts their shared outcomes twice.

Events and Probability

Deep Learning Connection: MNIST



Events and Probability

Deep Learning Connection: MNIST

For one image, the model distributes probability across the digits 0-9.

Digit	0	1	2	3	4	5	6	7	8	9
P	.01	.01	.02	.03	.02	.80	.04	.02	.03	.02

Illustrative output: digit 5 receives the largest probability, 0.80.

The ten probabilities sum to 1. A high score is not a guarantee of being correct.

A classifier assigns probabilities to candidate labels, then chooses the largest.

Events and Probability

Quick Check: Events

1. Complement

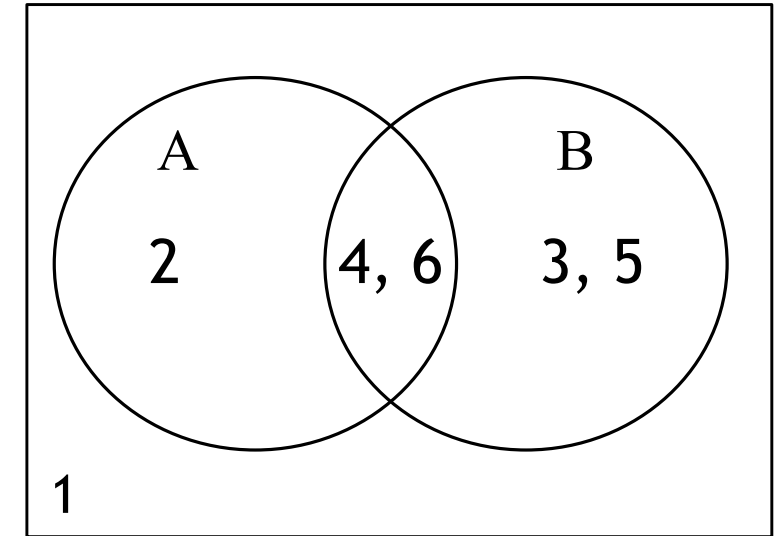
$A = \{2, 4, 6\}$. Which outcomes are not in A?

2. Intersection

$B = \{3, 4, 5, 6\}$. Which outcomes appear in both A and B?

3. Union

What is the probability that at least one of the two conditions is satisfied?



Check the event first, then calculate its probability.

Conditional Probability

Information Can Change Our Belief

The describes a dice with a hidden weight.



After many rolls above 3, we may revise our belief that the dice is fair.

The question becomes: how likely is “biased” given the information observed?

$$P(\text{biased} \mid \text{information})$$

Conditional probability includes information that we already know.

Conditional Probability

A Smaller Set of Possible Outcomes



Now use a fair dice. We are told that the result is greater than 3.

BEFORE THE INFORMATION

Possible results: 1, 2, 3, 4, 5, 6.

A 6 is one of six equally likely results.

$$P(6) = 1/6$$

AFTER THE INFORMATION

Possible results: 4, 5, 6.

A 6 is now one of three equally likely results.

$$P(6 \mid \text{greater than } 3) = 1/3$$

“Given information” means we calculate within the outcomes allowed by information.

Conditional Probability

The Conditional Probability Formula

Use $P(B) > 0$.

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

The denominator selects the B cases. The numerator keeps the cases that also satisfy A.

$$A = \{6\}, B = \{4, 5, 6\}$$

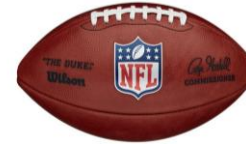
$$P(A \mid B) = (1/6) / (3/6) = 1/3$$

Divide the probability of both events by the probability of the given event.

Conditional Probability

Independence: The Information Does Not Help

Let's imagine a touchdown occurring during the dice experiment.



That unrelated event does not change our probability that the dice is biased.

Independence means that knowing one event does not change the probability of the other.

$$P(E1 \mid E2) = P(E1)$$

The relationship is symmetric when the conditional probabilities are defined.

Independent events do not change each other's probabilities.

Conditional Probability

The Product Rule

$$P(E1 \cap E2) = P(E1 \mid E2) P(E2)$$

Start with E2, then use the probability of E1 within those cases.

$$P(E1 \cap E2) = P(E2 \mid E1) P(E1)$$

Reverse the bookkeeping and the same joint probability is obtained.

Both events must occur. The factorization alone is not a claim about time order.

The same joint event can be factored in either direction.

Conditional Probability

The Independent Special Case

$$P(E1 \cap E2) = P(E1) P(E2)$$

Example: two independent tosses of a fair coin.



E1 is heads on the first toss. E2 is heads on the second toss.

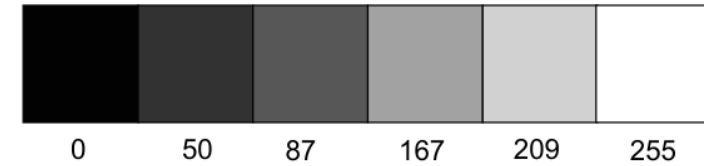
$$P(\text{both heads}) = (1/2)(1/2) = 1/4$$

Independence is the reason we may use the two unconditional probabilities.

For independent events, the joint probability is the product of their probabilities.

Conditional Probability

A Classifier Uses the Input as Information



WITHOUT THE IMAGE

$P(\text{digit } 0)$

This is a prior probability before seeing the image.

WITH THE IMAGE



$P(\text{digit } 0 \text{ given the input})$

The pixel values provide information about the digit.

$$P(0 \mid \text{input})$$

The output distribution can change when a different image is provided.

A prediction is conditional on the particular input image.

Random Variables

One Experiment, Several Questions

Observed sequence: T T H H T



Question	Name	Value for TTHHT
How many heads?	X	2
How many tails?	Y	3
How many tails before the first head?	Z	2

A random variable records a numerical answer to a question about the outcome.

Random Variables

A Random Variable Is a Function

Sequence	X: heads	Y: tails	Z: tails before H
TTHHT	2	3	2
HHTH	3	1	0



The function is fixed; the experimental outcome is random.

Count check: HHTH begins with H, so $Z = 0$.

Once the outcome is known, the value of the random variable is fixed.

Random Variables

Different Outcomes Can Give the Same Value

$$X(\text{HHH}) = 3$$



$$X(\text{HHTH}) = 3$$

These are different sequences, but both contain three heads.

Knowing only $X = 3$ does not tell us the complete sequence.

These examples evaluate the same counting rule on sequences of different lengths.

The random variable keeps the answer we asked for, not the full sequence.



Random Variables

From Outcomes to a Distribution of X

For two independent fair tosses, let X count the heads.

Outcome	TT	TH	HT	HH
X	0	1	1	2
Probability	$1/4$	$1/4$	$1/4$	$1/4$

$$P(X = 1) = P(\text{TH}) + P(\text{HT}) = 1/2$$

The events TH and HT cannot occur together, so there is no overlap to subtract.

Add the probabilities of all outcomes that give the same value of X .

Random Variables

Count Heads by Adding Binary Variables

Let C_i be 1 for heads on toss i and 0 for tails.

$$T T H H T \rightarrow 0, 0, 1, 1, 0$$



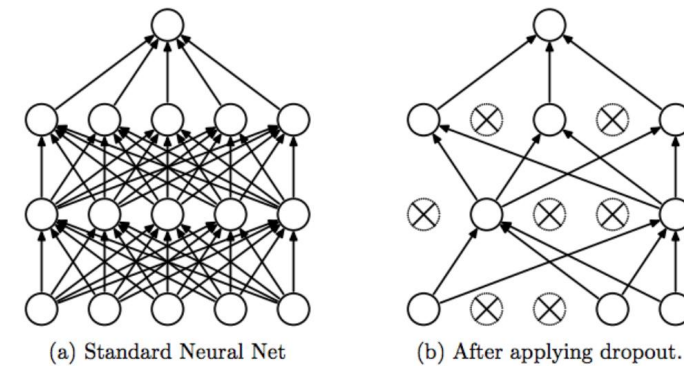
$$X = C_1 + C_2 + \dots + C_n$$

$$\text{For TTHHT: } X = 0 + 0 + 1 + 1 + 0 = 2$$

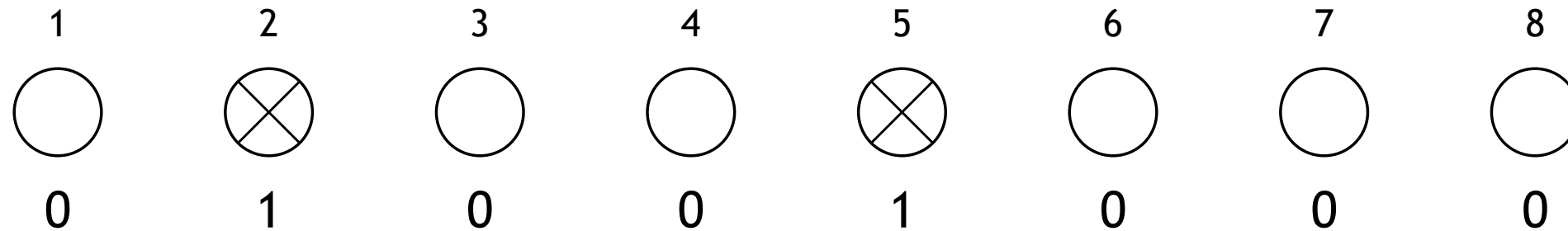
A total count can be written as a sum of 0-or-1 variables.

Random Variables

Deep Learning Connection: Dropout



During training, each neuron is independently masked with probability p .



Use $X = 1$ for masked and $X = 0$ for not masked.

In this example, the eight binary values sum to two masked neurons.

Random masking discourages dependence on a few specific connections.

Each dropout decision is a binary random variable.

Expectation

What Should We Expect on Average?



If we toss a fair coin eight times, how many heads should we expect?

A particular experiment might give three, four, or five heads.

We want the average count across many repetitions of the eight-toss experiment.

One experiment $\neq$ the long-run average

Expectation describes a probability-weighted average, not a promised result.

Expectation

Expectation of One Fair Toss

Let X be 1 for heads and 0 for tails.



$$E[X] = \sum_x x P(X = x)$$

$$E[X] = 0 \times 0.5 + 1 \times 0.5 = 0.5$$

We never observe half a head on one toss.

The value 0.5 describes the average number of heads per toss over repeated trials.

An expected value does not have to be a value observed in one trial.

Expectation

Linearity of Expectation

$$\begin{aligned} E[X_1 + X_2 + \dots + X_n] \\ = E[X_1] + E[X_2] + \dots + E[X_n] \end{aligned}$$



Instead of finding every possible total, add the expected contributions.

This rule also holds when the random variables are dependent.

The expectation of a sum equals the sum of the expectations.

Expectation

Eight Coin Tosses

The total number of heads is $X = X_1 + X_2 + \dots + X_8$.



$$E[X_i] = 0.5$$

$$E[X] = 8 \times 0.5 = 4$$

For n fair tosses, the same reasoning gives $E[X] = 0.5n$.

The number 4 is an average. It does not mean every eight-toss run contains four heads.

Eight fair tosses have an expected total of four heads.

Expectation

Expected Number of Masked Neurons

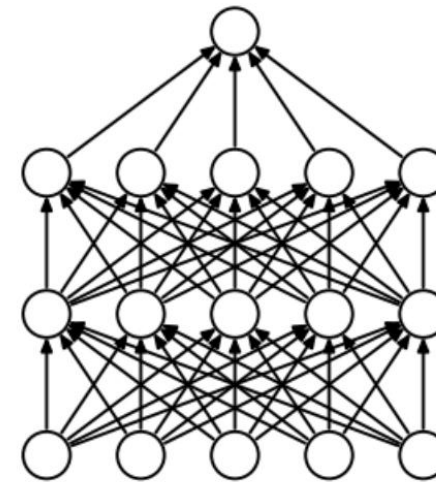
For one neuron, X is 1 with probability p and 0 with probability $1 - p$.

$$E[X] = 1 \times p + 0 \times (1 - p) = p$$

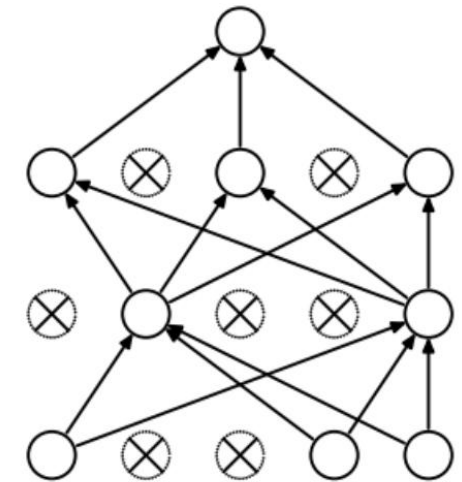
$$E[\text{total masked}] = np$$

Illustrative layer: $n = 8$ and $p = 0.25$.

The expected number masked is $8 \times 0.25 = 2$.



(a) Standard Neural Net



(b) After applying dropout.

For n neurons with mask probability p , the expected masked count is np .

Expectation

Quick Check: What Does the Average Mean?

1. One run

Must every run of eight fair tosses contain exactly four heads?

2. Possible values

Can $E[X]$ equal 0.5 when X can only be 0 or 1?

3. Dependence

Does $E[A + B] = E[A] + E[B]$ require independent A and B ?



Separate an average from the value in a single experiment.

Variance

How Far Do Values Move from the Mean?

Write $\mu = E[X]$. First measure the difference $X - \mu$, then square it.



$$\text{Var}(X) = E[(X - \mu)^2]$$

Fair-toss value X	$X - 0.5$	$(X - 0.5)$ squared
0	-0.5	0.25
1	0.5	0.25

Both outcomes give a squared distance of 0.25, so the variance is 0.25.

Variance is the expected squared distance from the mean.

Variance

Why Not Average the Plain Differences?



PLAIN DIFFERENCES

Tails: $0 - 0.5 = -0.5$

Heads: $1 - 0.5 = 0.5$

Their probability-weighted average is zero.

SQUARED DIFFERENCES

Tails: $(-0.5) \text{ squared} = 0.25$

Heads: $(0.5) \text{ squared} = 0.25$

Their probability-weighted average is 0.25.

A zero average difference does not mean that the random variable never changes.

Positive and negative differences cancel; squared differences do not.

Variance

Does Variance Add Like Expectation?

$$\begin{aligned}\text{Var}(A + B) &= \text{Var}(A) + \text{Var}(B) \\ &\quad + 2 \text{Cov}(A, B)\end{aligned}$$

The extra term is covariance. It describes how the two variables vary together.

If A and B are independent, their covariance is zero.

Then the variance of their sum is the sum of their variances.

For variance, the relationship between the variables matters.

Variance

Covariance: Heads and Tails in the Same Run

Let A count heads and B count tails in the same eight-toss experiment.

$$A + B = 8, \quad B = 8 - A$$



If A increases by 1, B decreases by 1. They do not vary independently.

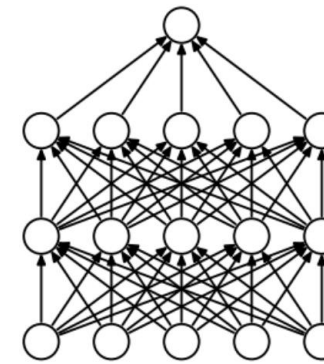
The total $A + B$ is always 8, so its variance is zero.

The covariance term accounts for this relationship between the two counts.

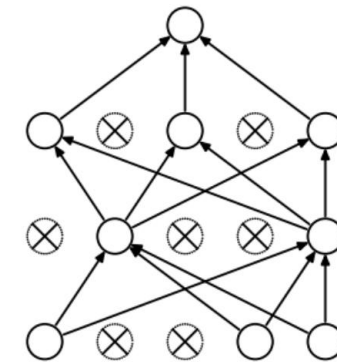
More heads means fewer tails when the total number of tosses is fixed.

Variance

Variance of the Total Dropout Count



(a) Standard Neural Net



(b) After applying dropout.

$$\text{Var}(\text{total masked}) = np(1 - p)$$

Quantity

$n = 8, p = 0.25$

Expected masked count

$$8 \times 0.25 = 2$$

Variance of masked count

$$8 \times 0.25 \times 0.75 = 1.5$$

The average count and its variability answer different questions.

Independent masks let us add the individual variances.

Variance

Expectation and Variance: Keep the Conditions

Quantity	For a sum	Condition used here
Expectation	$E[A + B] = E[A] + E[B]$	No independence required.
Variance	$\text{Var}(A + B) = \text{Var}(A) + \text{Var}(B)$	Independent A and B.

In general, the variance formula also contains $2 \text{Cov}(A, B)$.

Expectation adds without independence. Variance needs an extra check.

Summary

Quick Check: The Eight-Neuron Layer

Assume eight independent mask decisions, each with $p = 0.25$.

1. Average

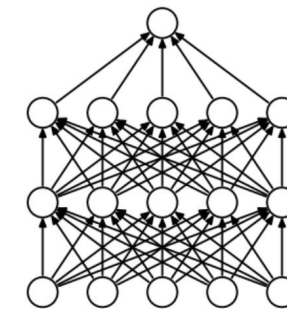
How many masked neurons do we expect?

2. A single realization

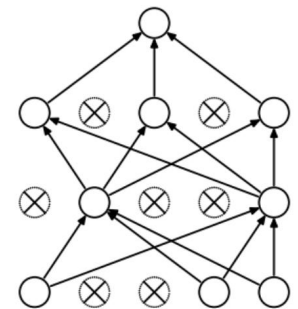
Must every training step mask exactly that many neurons?

3. Variation

What is the variance of the total masked count?



(a) Standard Neural Net



(b) After applying dropout.

Use the meaning of each quantity, not only its formula.

Summary

Quick Check: Answers

1. Expected count = 2

Add eight expectations: $8 \times 0.25 = 2$.

2. The actual count can vary

Independent random decisions do not enforce a fixed count of two.

3. Variance = 1.5

Add eight variances: $8 \times 0.25 \times 0.75 = 1.5$.

An expected count can differ from the count in any particular run.

Summary

What Did We Learn?

Concept	Question it answers
Event and probability	Which outcomes are included, and how likely are they?
Conditional probability	How does known information change the probability?
Random variable	Which numerical summary do we record?
Expectation	What value do we expect on average?
Variance	How much does that value vary around its mean?

Next: infer from evidence, measure uncertainty, and move from sums to integrals.