

Lecture 1 — Probability, the Binomial and the Gaussian

This lecture lays the probabilistic foundations used in the whole course. It treats one discrete variable first (Sections 1–7: the rules, average and variance, the Bernoulli and Binomial distributions, and the Gaussian limit of the Binomial), then the Gaussian, densities and entropy (Sections 8–10), and finally several variables (Sections 11–13: sum and product rules, Bayes' theorem, covariance, the multivariate Gaussian). It follows the spirit of Bishop & Bishop, *Deep Learning* (2024), O. Calin, *Deep Learning Architectures: A Mathematical Approach* (2020), L. Susskind's *Theoretical Minimum* lectures, and the *3Blue1Brown* series on probability.

Legend.

-  **Insight:** a key idea, or a connection to machine learning, physics or another part of the course.
-  **Open question:** something to think about; it may be picked up in a later lecture.

Contents

- Lecture 1 — Probability, the Binomial and the Gaussian
 - Contents
 - 1. What is probability?
 - * Two intuitions: counting and frequency
 - * The rules
 - * Where do the numbers come from?
 - * Why the law of large numbers works
 - * Frequencies and beliefs
 - 2. Average, variance and standard deviation
 - * The average
 - * Variance and standard deviation
 - 3. The Bernoulli distribution
 - 4. The Binomial distribution
 - 5. The histogram of the frequency
 - 6. The distribution concentrates
 - 7. From the Binomial to the Gaussian
 - 8. The Gaussian distribution
 - 9. Probability densities
 - 10. Entropy: counting sequences
 - * Large deviations: one count, three lenses
 - 11. The rules of probability
 - 12. Covariance and independence
 - 13. The multivariate Gaussian
 - 14. Summary
 - References

1. What is probability?

Some systems are too complicated to follow in detail, like a gas of 10^{23} molecules. Others are simple but too poorly known, like a die, whose landing face depends on initial conditions we cannot control. In both cases we give up predicting individual outcomes and predict **frequencies** instead: how often each outcome occurs when the experiment is repeated many times. Probability is the language for this.

Definition. A probabilistic model consists of a **random experiment** E (a procedure that can, at least in principle, be repeated), a **sample space** Ω whose elements ω are the possible outcomes of E , and a **probability** P that assigns a number in $[0, 1]$ to each **event**, i.e. to each admissible subset of Ω . A **random variable** is a quantity measured on the outcome: a function $X : \Omega \rightarrow A$, where $X(\omega)$ is the value observed when the outcome is ω .

Example. E = roll two fair dice. The states are the ordered pairs $\omega = (d_1, d_2)$, so Ω has 36 equally likely elements, each with probability $\frac{1}{36}$. Take the sum as random variable, $X(\omega) = d_1 + d_2$, with $A = \{2, \dots, 12\}$.

Several states can give the same value: the event $\{X = 7\} = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$ contains 6 states, while $\{X = 2\} = \{(1, 1)\}$ contains one, so

$$P\{X = 7\} = \frac{6}{36} = \frac{1}{6}, \quad P\{X = 2\} = \frac{1}{36}.$$

The states are equiprobable, but the values of X are not: their probabilities come from **counting states**.

Two intuitions: counting and frequency

This example uses one of the two intuitions we all have about probability. The opening paragraph used the other one.

- **Counting.** Find states that are equally likely. The probability of an event is then the fraction of states it contains, $P(B) = |B|/|\Omega|$ (favourable over possible). This is how we computed $P\{X = 7\}$.
- **Frequency.** Repeat the experiment many times. The probability of an event is then the fraction of trials in which it occurs, $P(B) \approx N_B/N$.

Counting needs a starting point: *why* are the 36 states equally likely? The answer is **symmetry**. The faces of a fair die differ only by their labels, so relabelling them, or rotating the die before the throw, does not change the physics. It therefore cannot change the probabilities, and each face gets $\frac{1}{6}$. For two dice thrown independently, each pair gets $\frac{1}{36}$. Symmetry acts on the states, not on the values of X : no symmetry exchanges a sum of 2 with a sum of 7, so nothing forces these values to be equally likely, and they are not.

The two intuitions answer different questions. Counting is a **prediction**, made before any die is thrown. Frequency is a **measurement**, made after many throws. The rest of this section shows that they agree: symmetry fixes what we count, and the law of large numbers, itself a counting argument, turns the count into a frequency. The same thread runs through the whole lecture: the Binomial counts sequences, the Gaussian describes their fluctuations, and entropy is the logarithm of a count.

The rules

Both intuitions lead to the same rules, because both define a probability as a fraction: of states, or of trials. A fraction is never negative, and the fractions for all possible outcomes add up to one. For a finite state space $\Omega = \{\omega_1, \dots, \omega_n\}$, writing $P(i)$ for $P(\{\omega_i\})$,

$$P(i) \geq 0, \quad \sum_{i=1}^n P(i) = 1$$

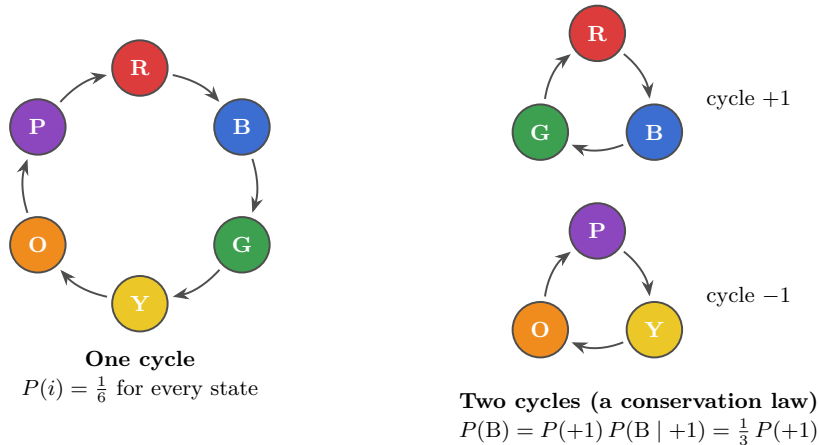
The probability of an event is the sum over its states, $P(B) = \sum_{\omega_i \in B} P(i)$. This is exactly how we computed $P\{X = 7\}$.

Where do the numbers come from?

The rules constrain the numbers $P(i)$ but do not fix them. There are three ways to obtain them.

The first is **symmetry**, which we already used for the dice and which is what makes counting legitimate. Outcomes exchanged by a transformation that leaves the experiment unchanged must be equally likely. This gives $\frac{1}{2}$ per coin face and $\frac{1}{6}$ per die face, and the probabilities of events then follow by counting. When the symmetry is broken, as for a loaded die, counting fails and we need one of the other two routes.

The second is **dynamics**, which needs no symmetry at all. Consider a system with six states (the colours in the figure) that evolves under a deterministic law of motion. Suppose the law takes the system once through all six states before repeating (left). Then a snapshot at a random time finds each state with probability $\frac{1}{6}$, whatever the initial condition. The time spent in each state plays the role that symmetry played for the die. This route sits between the two intuitions: the probability is a **frequency in time**, measured along a single history, but we can compute it by **counting** the states on the cycle, because the cycle treats them all alike.



Suppose instead the law splits the states into two cycles of three (right). Then the system never leaves the cycle it starts on, so the cycle label ± 1 is a **conserved quantity**. The dynamics still fixes probabilities *within* a cycle, by counting its three states: if blue lies on the $+1$ cycle, $P(\text{blue} | +1) = \frac{1}{3}$. But the dynamics says nothing about *which* cycle we are on. $P(+1)$ depends on the initial condition and must come from outside:

$$P(\text{blue}) = P(+1) P(\text{blue} | +1) = \frac{1}{3} P(+1).$$

This is a first instance of **conditional probability** and of the **product rule**, developed in [Section 11](#).

The third is **experiment**, the frequency intuition in its pure form. When neither symmetry nor dynamics is available, as with a loaded die or a real dataset, we measure. Repeat E N times and count the occurrences N_i of state ω_i . The **law of large numbers** guarantees that the measured frequencies converge to the probabilities:

$$\lim_{N \rightarrow \infty} \frac{N_i}{N} = P(i)$$

Why the law of large numbers works

Why should probabilities obtained by counting agree with measured frequencies? The law of large numbers answers this, and it is itself a counting argument, of the same kind as the dice. Flip a fair coin N times. By symmetry all 2^N sequences are equally likely, but the number of heads is not, just as the sum of two dice is not. For $N = 4$:

Number of heads	0	1	2	3	4
Number of sequences	1	4	6	4	1

These are the binomial coefficients $\binom{N}{m}$, the number of sequences with m heads: a row of Pascal's triangle. As N grows, they pile up around $N/2$. For $N = 100$, about 96% of all sequences have between 40 and 60 heads, and for any fixed tolerance around $\frac{1}{2}$, the fraction of sequences within it tends to one, increasing N . Sequences with a frequency of heads far from $\frac{1}{2}$ exist, but they are **rare**, and a rare event specified in advance does not happen in practice. This is the precise sense in which probabilities predict frequencies.

You can watch this happen in [counting_lln.html](#) (open it in a browser). Choose a final number of flips and a window of frequencies, and the animation adds one flip at a time, $N = 1, 2, 3, \dots$, counting the sequences whose frequency of heads falls inside the window. Keep the *frequency* lens for now: Sections 5 and 6 explain its histogram, and we will come back to the same counts with two other lenses in Sections 7 and 10.

💡 This is where the two intuitions merge. The symmetry of a single flip becomes a symmetry of whole sequences, which makes all 2^N of them equally likely. Counting sequences then

shows that almost all of them have the frequency predicted by counting the faces of the coin. Counting and frequency are two faces of the same idea, and symmetry is what connects them.

Frequencies and beliefs

Everything above treats probabilities as frequencies in repeatable experiments. This is the **frequentist** view. Bishop, like much of ML, also uses **Bayesian** probabilities: degrees of belief. These apply to one-off quantities where no repetition is possible, such as the value of a model parameter. The two interpretations obey the same rules, so everything in this course applies to both. Symmetry has a Bayesian version too, Laplace's **principle of indifference**: if nothing we know distinguishes the outcomes, we assign them equal probabilities. Counting therefore makes sense for beliefs as well, even for an experiment that will never be repeated.

2. Average, variance and standard deviation

From now on we work directly with random variables. For a variable X with values $x_1, x_2, \dots$, write

$$p(x_i) \equiv P\{\omega : X(\omega) = x_i\}$$

for the probability that X takes the value x_i : the sum of the probabilities of all the states that give x_i , as for $P\{X = 7\}$. The list of these probabilities is the **distribution** of X . Often we do not need all of it: two numbers, where the values sit and how much they spread, already say a lot.

The average

Repeat the experiment N times and call $X_1, \dots, X_N$ the results. Their arithmetic mean can be computed by grouping the results by value: if X took the value x_i in N_i trials, then $\frac{1}{N} \sum_{k=1}^N X_k = \sum_i x_i N_i / N$. By the law of large numbers, $N_i / N \rightarrow p(x_i)$, so the mean tends to a number that depends only on the distribution. The same holds for any function f of X :

$$\langle f \rangle = \sum_i f(x_i) p(x_i) \approx \frac{1}{N} \sum_{k=1}^N f(X_k)$$

The left side is the **average**. The right side is the **empirical average** over N samples, which becomes exact as $N \rightarrow \infty$. In ML notation, $\langle f \rangle \equiv \mathbb{E}[f]$ ("the expected value of", or "the expectation of", f).

Example. Return to the two dice and the sum $X = d_1 + d_2$. The average can be computed in two ways. Summing over the 36 states, each with weight $\frac{1}{36}$, splits into the two dice separately:

$$\langle X \rangle = \langle d_1 \rangle + \langle d_2 \rangle = \frac{1+2+\dots+6}{6} + \frac{1+2+\dots+6}{6} = 3.5 + 3.5 = 7.$$

Alternatively, group the states by value, as we did for $\{X = 7\}$, and weight each value by its probability:

$$\langle X \rangle = \sum_{x=2}^{12} x p(x) = \frac{2 \cdot 1 + 3 \cdot 2 + 4 \cdot 3 + \dots + 12 \cdot 1}{36} = \frac{252}{36} = 7.$$

Both routes give the same answer because they add the same 36 terms in a different order. The first route shows a general property: **the average of a sum is the sum of the averages**. The average is also linear in a broader sense: constants come out of it, $\langle cf \rangle = c \langle f \rangle$, and the average of a constant is the constant itself, $\langle c \rangle = c \sum_i p(x_i) = c$.

The average of an **indicator**, a quantity that is 1 on an event B and 0 elsewhere, is just the probability of that event: $\langle \mathbf{1}_B \rangle = P(B)$. For instance, $\langle \mathbf{1}[X = 7] \rangle = \frac{1}{6}$. This is one of the many ways in which averages connect to ML. As we will see, the **0–1 loss** of a classifier is the indicator of the event "the prediction is wrong", so its expected loss is its probability of error.

Variance and standard deviation

The average says where the values sit, not how much they spread. A fair die and a variable that always equals 3.5 have the same average. To measure the spread, split each value into the average plus a **fluctuation**:

$$X = \langle X \rangle + \delta X, \quad \delta X = X - \langle X \rangle.$$

The fluctuation averages to zero, $\langle \delta X \rangle = 0$ (we show why after defining the standard deviation): positive and negative fluctuations cancel, so their average says nothing about their size. We average the square instead, which is never negative. This is the **variance**:

$$\text{var}[X] = \langle \delta X^2 \rangle = \langle (X - \langle X \rangle)^2 \rangle = \sum_i (x_i - \langle X \rangle)^2 p(x_i).$$

A second form is often easier to compute. Expand the square, $(X - \langle X \rangle)^2 = X^2 - 2\langle X \rangle X + \langle X \rangle^2$, and average term by term. $\langle X \rangle$ is a number, so it comes out of the average, and the middle term gives $-2\langle X \rangle^2$:

$$\boxed{\text{var}[X] = \langle (X - \langle X \rangle)^2 \rangle = \langle X^2 \rangle - \langle X \rangle^2 \geq 0}$$

The variance has the units of X squared. Its square root, the **standard deviation**,

$$\boxed{\sigma_X = \sqrt{\text{var}[X]}}$$

has the same units as X and is the **typical size of a fluctuation**. Two consequences of the definition are used below. Shifting a variable by a constant does not change its fluctuations, $\text{var}[X + c] = \text{var}[X]$. Rescaling it by c rescales the fluctuations by c and the variance by c^2 : $\text{var}[cX] = c^2 \text{var}[X]$, so $\sigma_{cX} = |c| \sigma_X$.

The average of a fluctuation is zero. This holds for every distribution, with no symmetry and no large N needed: it follows from the definition of the average. Write the average of δX as a sum and split it in two:

$$\boxed{\langle \delta X \rangle = \sum_i (x_i - \langle X \rangle) p(x_i) = \underbrace{\sum_i x_i p(x_i)}_{\langle X \rangle} - \langle X \rangle \underbrace{\sum_i p(x_i)}_1 = 0}$$

💡 Which constant c best predicts X , in the sense of the smallest mean squared error $\langle (X - c)^2 \rangle$? Write $X - c = \delta X + (\langle X \rangle - c)$ and square. The cross term averages to zero because $\langle \delta X \rangle = 0$, so $\langle (X - c)^2 \rangle = \text{var}[X] + (\langle X \rangle - c)^2$. The best constant is the average, and the error left over is the variance, which no constant can remove. This is the first instance of the MSE of [Lecture 2, Section 5](#), where the constant becomes a function $f(x; \theta)$.

3. The Bernoulli distribution

The simplest random variable is binary, $x \in \{0, 1\}$: a coin flip (heads = 1), or whether an input belongs to class 1. If $p(x = 1) = \mu$, then $p(x = 0) = 1 - \mu$, and a single expression covers both cases:

$$\boxed{\text{Bern}(x | \mu) = \mu^x (1 - \mu)^{1-x}}$$

Setting $x = 1$ gives μ , and setting $x = 0$ gives $1 - \mu$. The average and the variance follow from the definitions of Section 2, with only two terms in each sum:

$$\langle x \rangle = 0 \cdot (1 - \mu) + 1 \cdot \mu = \mu, \quad \langle x^2 \rangle = 0^2 \cdot (1 - \mu) + 1^2 \cdot \mu = \mu.$$

The two coincide because $x^2 = x$ for a binary variable. Therefore

$$\boxed{\langle x \rangle = \mu, \quad \text{var}[x] = \mu - \mu^2 = \mu(1 - \mu), \quad \sigma = \sqrt{\mu(1 - \mu)}}$$

The variance vanishes when the outcome is certain, $\mu = 0$ or 1 , and is largest when its derivative $1 - 2\mu$ vanishes, at $\mu = \frac{1}{2}$.

The coin. For a fair coin the symmetry of the faces (Section 1) gives $\mu = \frac{1}{2}$, and $\text{Bern}(x \mid \frac{1}{2}) = (\frac{1}{2})^x (\frac{1}{2})^{1-x} = \frac{1}{2}$ for both faces. Then

$$\langle x \rangle = \frac{1}{2}, \quad \text{var}[x] = \frac{1}{4}, \quad \sigma = \frac{1}{2}.$$

Here the standard deviation is more than a typical size: every flip deviates from the average by exactly $|x - \frac{1}{2}| = \frac{1}{2}$. A biased coin with $\mu = 0.3$ has $\langle x \rangle = 0.3$, $\text{var}[x] = 0.21$ and $\sigma \approx 0.46$: less spread than the fair coin, because one face is more predictable.

💡 In binary classification the label is a Bernoulli variable, and the network outputs its parameter $\mu(x; \theta)$. Maximizing the Bernoulli likelihood of the data gives the binary cross-entropy of [Lecture 2, Section 6](#).

4. The Binomial distribution

Flip the coin N times and count the number m of ones. Each sequence with m ones has probability $\mu^m (1 - \mu)^{N-m}$. This depends only on how many ones there are, not on their order, because i.i.d. trials are **symmetric under permutations**. This symmetry plays the role that the faces' symmetry played for the die: all sequences with m ones are equally likely, so we can count them, even when $\mu \neq \frac{1}{2}$ and a single trial has no symmetry at all. There are $\binom{N}{m}$ of them:

$$\boxed{\text{Bin}(m \mid N, \mu) = \binom{N}{m} \mu^m (1 - \mu)^{N-m}}, \quad \binom{N}{m} = \frac{N!}{m! (N - m)!}$$

For $N = 1$ the count is a single flip, and the Binomial reduces to the Bernoulli.

Normalization follows from the binomial theorem, $(a + b)^N = \sum_{m=0}^N \binom{N}{m} a^m b^{N-m}$. With $a = \mu$ and $b = 1 - \mu$,

$$\sum_{m=0}^N \text{Bin}(m \mid N, \mu) = (\mu + (1 - \mu))^N = 1^N = 1.$$

Mean and variance. Write the count as a sum of N Bernoulli variables, $m = x_1 + \dots + x_N$, where $x_k = 1$ if flip k gives a one. Each term has average μ and variance $\mu(1 - \mu)$ (Section 3).

Mean. The average of a sum is the sum of the averages (Section 2), and each flip contributes μ : $\langle m \rangle = \sum_k \langle x_k \rangle = N\mu$.

Variance. By the definition of Section 2, the variance is the average squared fluctuation, $\text{var}[m] = \langle \delta m^2 \rangle$ with $\delta m = m - \langle m \rangle$. So we need the fluctuation of m , square it, and average.

- **Fluctuation.** Split each flip into its average and its fluctuation, $x_k = \mu + \delta x_k$, with $\delta x_k = x_k - \mu$. By Section 2, each flip has $\langle \delta x_k^2 \rangle = \text{var}[x] = \mu(1 - \mu)$ and, like every fluctuation, $\langle \delta x_k \rangle = 0$. For one flip the boxed sum of Section 2 has two terms: $(1 - \mu)(0 - \mu) + \mu(1 - \mu) = 0$. Summing over the flips, the averages add up to $N\mu = \langle m \rangle$, so what is left is the fluctuation of the count: $\delta m = m - \langle m \rangle = \sum_k \delta x_k$. The fluctuation of a sum is the sum of the fluctuations.

- **Square.** Start with two flips: $\delta m^2 = (\delta x_1 + \delta x_2)^2 = \delta x_1^2 + \delta x_2^2 + 2 \delta x_1 \delta x_2$. With N flips, the square of the sum has N squares δx_k^2 and $N(N-1)$ cross terms $\delta x_k \delta x_l$ with $k \neq l$.
- **Average.** Each square averages to $\text{var}[x]$. Each cross term averages to zero, because the flips are **independent**: the probability of a pair of outcomes (a, b) is the product $p(a)p(b)$, as for the $\frac{1}{36} = \frac{1}{6} \cdot \frac{1}{6}$ of the two dice. The double sum over the pair then splits into two single sums, and each of them is the average of one fluctuation, which vanishes by Section 2:

$$\langle \delta x_k \delta x_l \rangle = \sum_{a, b \in \{0,1\}} p(a) p(b) (a - \mu)(b - \mu) = \underbrace{\left[\sum_a p(a) (a - \mu) \right]}_{\langle \delta x_k \rangle = 0} \underbrace{\left[\sum_b p(b) (b - \mu) \right]}_{\langle \delta x_l \rangle = 0} = 0.$$

Putting the three steps together, the variance of the count is

$$\text{var}[m] = \langle \delta m^2 \rangle = \left\langle \left(\sum_k \delta x_k \right)^2 \right\rangle = \underbrace{\sum_k \langle \delta x_k^2 \rangle}_{N \text{ squares, each } = \text{var}[x]} + \underbrace{\sum_{k \neq l} \langle \delta x_k \delta x_l \rangle}_{N(N-1) \text{ cross terms, each } = 0} = N \text{var}[x].$$

The variance of the count is the sum of the variances of the flips: **for independent terms, variances add.** (Section 11 defines independence in general, and Section 12 calls the cross terms covariances.)

$$\boxed{\langle m \rangle = N\mu, \quad \text{var}[m] = N \text{var}[x] = N\mu(1 - \mu), \quad \sigma_m = \sqrt{N\mu(1 - \mu)}}$$

To see what independence buys, compare with the opposite extreme: N copies of the same flip, $x_1 = x_2 = \dots = x_N$. Then $\delta m = N \delta x_1$, every cross term equals $\text{var}[x]$ instead of zero, and $\text{var}[m] = N \text{var}[x] + N(N-1) \text{var}[x] = N^2 \text{var}[x]$. With independent flips the variance grows like N and the standard deviation like $\sqrt{N}$; with copies, like N^2 and N . The difference is entirely in the $N(N-1)$ cross terms.

The mean grows like N , the standard deviation only like $\sqrt{N}$. Sections 5 and 6 show what this does to the histogram of the animation.

5. The histogram of the frequency

The first panel of the [animation](#), the *frequency* lens, shows the Binomial of a fair coin as a histogram of the frequency of heads, $\nu = m/N$. Its bars are not the probabilities $\text{Bin}(m)$ themselves: each probability is divided by the width of its bin. This section builds the histogram step by step and explains why.

One flip. For $N = 1$ the histogram is the Bernoulli distribution of Section 3: two values, $\nu = 0$ and $\nu = 1$, each with probability $\frac{1}{2}$.

Positions. With N flips the count takes the values $m = 0, \dots, N$, so the frequency takes the $N+1$ values $\nu_m = m/N$. Neighbouring values differ by one head, so they are spaced by

$$\Delta\nu = \nu_{m+1} - \nu_m = \frac{1}{N} :$$

few and far apart for small N ($0, \frac{1}{2}, 1$ for $N = 2$), denser and denser as N grows. This spacing is the natural bin width: one bin per value of m , centred at ν_m .

Areas. Bar m collects all the sequences with m heads, so the probability it represents is $\text{Bin}(m) = \binom{N}{m}/2^N$.

Heights: divide by the width. Why not plot $\text{Bin}(m)$ as the height? As N grows, the $N+1$ probabilities share the same total, one, so each of them gets smaller. A histogram of bare probabilities flattens towards zero, and histograms for different N cannot be compared. Instead, we let the *area* of each bar be its probability and we focus on a quantity that does not depend on N . The height is then the probability per unit length of ν . To write it as a function of ν , invert the position, $m = N\nu_m$: the bar at ν_m is the bar of the count $N\nu_m$. Then

$$p_N(\nu_m) = \frac{\text{Bin}(N\nu_m)}{\Delta\nu} = N \text{Bin}(N\nu_m)$$

The function p_N of the frequency is the Binomial of the count, read at the count that gives that frequency, and multiplied by $1/\Delta\nu = N$.

The total area is one for every N , $\sum_m p_N(\nu_m) \Delta\nu = \sum_m \text{Bin}(m) = 1$, and the probability that ν falls in an interval is the area of the bars above it. This is what the window of the animation measures. A probability per unit length is a **density**: Section 9 defines it for continuous variables, where the bins shrink to zero width.

6. The distribution concentrates

Why does the histogram of Section 5 get narrower? Its centre and its width are the mean and the standard deviation of $\nu = m/N$. Dividing by N divides the mean and the fluctuations by N , and therefore the variance by N^2 ($\text{var}[cX] = c^2 \text{var}[X]$, Section 2). With the results of Section 4,

$$\langle \nu \rangle = \frac{\langle m \rangle}{N} = \mu, \quad \delta\nu = \frac{\delta m}{N}, \quad \text{var}[\nu] = \frac{\text{var}[m]}{N^2} = \frac{\mu(1-\mu)}{N},$$

$$\sigma_\nu = \sqrt{\frac{\mu(1-\mu)}{N}} = \frac{\sigma}{\sqrt{N}}$$

where $\sigma = \sqrt{\mu(1-\mu)}$ is the standard deviation of a single flip. The frequency is centred at the probability, and its width shrinks like $1/\sqrt{N}$ (its variance like $1/N$). The same statement for the count is that its relative width vanishes:

$$\frac{\sigma_m}{\langle m \rangle} = \sqrt{\frac{1-\mu}{\mu}} \frac{1}{\sqrt{N}} \xrightarrow{N \rightarrow \infty} 0.$$

The count grows like N but fluctuates only like $\sqrt{N}$. The mean adds coherently: all N terms push the same way, $\langle m \rangle = N\mu$. The fluctuations add incoherently: their signs are random and they partly cancel, like the steps of a random walk, so only their squares add up, $\text{var}[m] = N \text{var}[x]$. Dividing by N , the frequency fluctuates like $1/\sqrt{N}$.

In short, **the distribution concentrates**. The argument used only two facts: m is a sum of many terms, and the terms are independent, so their (finite) variances add. Whenever this holds, statistical mechanics makes sharp predictions (a macroscopic quantity sums the contributions of $N \sim 10^{23}$ particles, and its relative fluctuations are of order 10^{-12}), and more data makes ML estimates more reliable. When it fails, so does the concentration: strongly correlated parts, as near a critical point, keep the fluctuations large even for large N .

7. From the Binomial to the Gaussian

The delta of Section 6 answers the question "where does the frequency end up?", but it hides the fluctuations around that point. Why a delta? Because of the variable we chose to histogram: the frequency divides the count by N , so its fluctuations shrink like $1/\sqrt{N}$. To see what the fluctuations look like, we have to change variable and zoom in at the right scale. Which one? Since the standard deviation of ν shrinks like $1/\sqrt{N}$, multiplying by $\sqrt{N}$ should undo the shrinking. But multiplying what? Not ν itself: $\sqrt{N}\nu$ has mean $\sqrt{N}\mu$, which runs off to infinity. We keep the centre μ and magnify only the fluctuation $\delta\nu = \nu - \mu$:

$$y = \mu + \sqrt{N} \delta\nu = \mu + \sqrt{N} (\nu - \mu)$$

The fluctuation of the new variable is $\sqrt{N}$ times that of the frequency, and by Section 6 it no longer depends on N :

$$\delta y = y - \mu = \sqrt{N} \delta \nu \sim \sqrt{N} \cdot \frac{\sigma}{\sqrt{N}} = \sigma = \sqrt{\mu(1-\mu)} \quad \text{for every } N.$$

This is the variable of the $\sqrt{N}$ lens: the frequency seen through a magnifying glass of power $\sqrt{N}$ centred at μ . The animation uses a fair coin, so $y = \frac{1}{2} + \sqrt{N}(\nu - \frac{1}{2})$.

Two variables, two distributions. The frequency lens and the $\sqrt{N}$ lens show the *same* sequences, sorted by the *same* number of heads m . What changes is the variable on the horizontal axis, and with it the distribution. The rule is that a change of variable moves probability around but does not create or destroy it: the probability of a bar is the same in every variable. Only its position and width change, and the height adjusts so that height $\times$ width stays equal to the probability.

Step 1: the area of bar m is $\text{Bin}(m)$ in every variable. Bar m collects the sequences with m heads, whatever we call the horizontal axis (Section 5):

$$\text{area of bar } m = \text{Bin}(m) = \binom{N}{m} \mu^m (1-\mu)^{N-m} \quad \left(= \binom{N}{m} \frac{1}{2^N} \text{ for the fair coin} \right).$$

Step 2: the position of bar m . Divide the count by N to get the frequency, then magnify the distance from μ by $\sqrt{N}$:

$$m \rightarrow \nu_m = \frac{m}{N} \rightarrow y_m = \mu + \sqrt{N} \left(\frac{m}{N} - \mu \right) = \mu + \frac{m - N\mu}{\sqrt{N}}.$$

The $\sqrt{N}$ of the zoom multiplies the $1/N$ of the frequency, $\sqrt{N} \cdot \frac{m}{N} = \frac{m}{\sqrt{N}}$.

Step 3: the width of bar m . The width is the distance between neighbouring positions, $\Delta m = 1$, and μ cancels in the difference:

$$\Delta \nu = \nu_{m+1} - \nu_m = \frac{1}{N}, \quad \Delta y = y_{m+1} - y_m = \frac{1}{\sqrt{N}} = \sqrt{N} \Delta \nu.$$

Step 4: the height of bar m . Divide the area of Step 1 by the width of Step 3. To write each height as a function of its own variable, invert the positions of Step 2, $m = N\nu_m$ and $m = N\mu + \sqrt{N}(y_m - \mu)$:

$$p_N(\nu_m) = \frac{\text{Bin}(m)}{\Delta \nu} = N \text{Bin}(N\nu_m), \quad p_N(y_m) = \frac{\text{Bin}(m)}{\Delta y} = \sqrt{N} \text{Bin}(N\mu + \sqrt{N}(y_m - \mu)).$$

The two densities are different functions, of different variables, but at corresponding points they read the same $\text{Bin}(m)$.

The change of distribution. Steps 3 and 4 together are the rule for changing variable in a density. Since y is ν stretched by $\sqrt{N}$ around μ , every length in y is $\sqrt{N}$ times the corresponding length in ν , $\Delta y = \sqrt{N} \Delta \nu$, and every height is divided by $\sqrt{N}$ so that the areas do not change:

$$\boxed{p_N(y) = p_N(\nu) \left| \frac{d\nu}{dy} \right| = \frac{p_N(\nu)}{\sqrt{N}}, \quad \nu = \mu + \frac{y - \mu}{\sqrt{N}}}$$

This is the general rule $p_y = p_\nu |d\nu/dy|$ of Section 9, here with a constant Jacobian $d\nu/dy = 1/\sqrt{N}$. In the animation, switching from the frequency lens to the $\sqrt{N}$ lens does exactly this: the bars are spread out by $\sqrt{N}$ and flattened by $\sqrt{N}$.

Collecting the steps, with $\sigma = \sqrt{\mu(1-\mu)}$:

bar m	count m	frequency ν (frequency lens)	zoom y ($\sqrt{N}$ lens)
area	$\text{Bin}(m)$	$\text{Bin}(m)$	$\text{Bin}(m)$
position	m	m/N	$\mu + (m - N\mu)/\sqrt{N}$
width	1	$1/N$	$1/\sqrt{N}$
height	$\text{Bin}(m)$	$N \text{Bin}(m)$	$\sqrt{N} \text{Bin}(m)$
typical fluctuation	$\sqrt{N} \sigma$	$\sigma/\sqrt{N}$	σ
limit $N \rightarrow \infty$	flattens to 0	$\delta(\nu - \mu)$	$\mathcal{N}(y \mu, \sigma^2)$

💡 The last two rows are the key ones. As N grows, the frequency pins down the mean more and more precisely, $\delta\nu = \sigma/\sqrt{N} \rightarrow 0$, and its distribution collapses to a delta. The fluctuations around it, seen at their own scale, do not shrink: $\delta y = \sigma$ for every N , the standard deviation of a single flip, and their distribution tends to a fixed Gaussian. Same sequences, same probabilities, different variable: a different limiting distribution. More flips sharpen the estimate of μ , not the shape of the noise.

For the fair coin this is the recipe of the animation: for each m , a bar of width $1/\sqrt{N}$ centred at $y_m = \frac{1}{2} + (m - N/2)/\sqrt{N}$, with height $\sqrt{N} \binom{N}{m}/2^N$ (for $N = 4$: $y = -\frac{1}{2}, 0, \frac{1}{2}, \frac{3}{2}$; at $N = 1$, $y = \nu$ and the two lenses coincide). In the y lens the bars get narrower like $1/\sqrt{N}$, but the typical fluctuation stays σ , so the spread of the histogram is fixed. Increase N in the animation: the histogram stops collapsing and settles onto a fixed **bell curve**, the dashed line.

💡 This is the **de Moivre–Laplace theorem**, a special case of the **central limit theorem**: a sum of many independent variables with finite variance tends to a Gaussian. It is why the Gaussian is the default noise model behind the MSE loss.

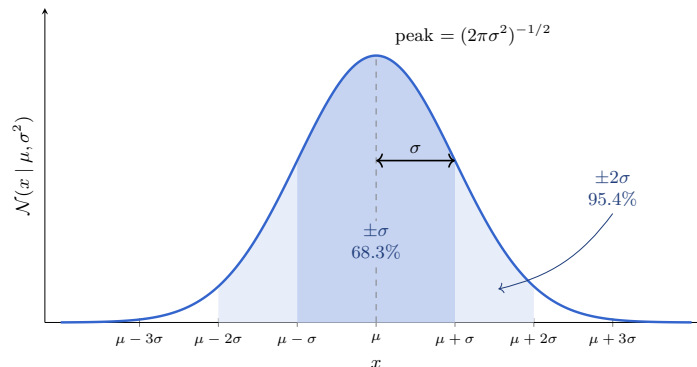
💡 A **Galton board** is the same experiment built in wood. A ball falls through N rows of pegs and bounces left or right with probability $\frac{1}{2}$ at each one, by the left–right symmetry of the peg. Each row is a flip and each path is a sequence: the ball lands in bin m after m right bounces, along any of the $\binom{N}{m}$ paths of Pascal's triangle. The piles of many balls are the $\sqrt{N}$ lens measured as frequencies. Read in time, the path of a ball is a **random walk**: after t steps its typical displacement grows like $\sqrt{t}$, not t , which is the law of **diffusion**, $\langle x^2 \rangle \propto t$. See 3Blue1Brown's [But what is the Central Limit Theorem?](#).

8. The Gaussian distribution

Section 7 found this curve as the limit of the Binomial. It is a **density**, a probability per unit length like the heights of the histograms of Section 5 (Section 9 defines densities in general):

$$\mathcal{N}(x | \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{(x - \mu)^2}{2\sigma^2}\right]$$

σ is the **standard deviation** and $\beta = 1/\sigma^2$ the **precision**. About 68.3% of the probability lies within $\pm\sigma$ and 95.4% within $\pm 2\sigma$.



Normalization ($z = x - \mu$): square the integral and go to polar coordinates,

$$I^2 = \iint e^{-(z_1^2 + z_2^2)/2\sigma^2} dz_1 dz_2 = \int_0^{2\pi} d\phi \int_0^\infty e^{-r^2/2\sigma^2} r dr = 2\pi\sigma^2.$$

Moments. $\mathbb{E}[z] = 0$ by symmetry (odd integrand). Differentiating $\int e^{-z^2/2\sigma^2} dz = (2\pi\sigma^2)^{1/2}$ with respect to σ^2 gives $\int z^2 e^{-z^2/2\sigma^2} dz = \sigma^2(2\pi\sigma^2)^{1/2}$, hence

$$\boxed{\mathbb{E}[x] = \mu, \quad \mathbb{E}[x^2] = \mu^2 + \sigma^2, \quad \text{var}[x] = \sigma^2}$$

💡 Among all distributions with given mean and variance, the Gaussian has **maximum entropy** (Section 10): it is the least committal choice.

9. Probability densities

Section 5 turned the Binomial into a histogram with probability per unit width as height, and Sections 7 and 8 saw it approach a smooth curve, the Gaussian. For a variable that takes a continuum of values, such as a position or a voltage, single values have zero probability, and the histogram with infinitely narrow bins becomes a **density** $p(x)$:

$$p(x \in (a, b)) = \int_a^b p(x) dx, \quad p(x) \geq 0, \quad \int_{-\infty}^{\infty} p(x) dx = 1.$$

A density is not a probability and can be larger than one, as the peak ≈ 8 of the frequency histogram at $N = 100$ shows. The **cumulative distribution** $P(z) = \int_{-\infty}^z p(x) dx$ satisfies $d_z P = p$. Averages become integrals, and the variance keeps the same form:

$$\mathbb{E}[f] = \int p(x) f(x) dx, \quad \text{var}[f] = \mathbb{E}[f^2] - \mathbb{E}[f]^2.$$

Under a change of variables $x = g(y)$, probability is conserved, $p_y |dy| = p_x |dx|$:

$$p_y(y) = p_x(g(y)) \left| \frac{dg}{dy} \right|$$

(Jacobian determinant in several dimensions). This is the rule we used in Section 7 to go from the frequency ν to the zoomed variable y , with $g(y) = \mu + (y - \mu)/\sqrt{N}$ and $|dg/dy| = 1/\sqrt{N}$. Since the density changes by a factor that can depend on y , the location of a density's maximum depends on the choice of variable.

The **Dirac delta** $\delta(x - x_0)$ is a density concentrated at one point: a Gaussian of zero width, like the limit $\delta(\nu - \mu)$ of the frequency histogram (Section 6). It picks the value of a function at that point, $\int \delta(x - x_0) f(x) dx = f(x_0)$.

10. Entropy: counting sequences

Counting configurations has been the thread of this lecture, and entropy is where it leads. In the [animation](#) of Section 1 we counted, for N flips of a fair coin, how many of the 2^N sequences have m heads: $\binom{N}{m}$, for instance 1, 4, 6, 4, 1 out of 16 for $N = 4$. These counts grow exponentially with N , and entropy is the rate of that growth.

How many sequences have a given frequency? Write the number of heads as a frequency, $\nu = m/N$, and use Stirling's formula, $\ln n! \simeq n \ln n - n$ (the neglected terms grow only like $\ln n$):

$$\ln \binom{N}{m} \simeq N \ln N - m \ln m - (N - m) \ln(N - m) = -m \ln \frac{m}{N} - (N - m) \ln \frac{N - m}{N} = N h(\nu),$$

$$\boxed{\binom{N}{m} \simeq e^{N h(\nu)}, \quad h(\nu) = -[\nu \ln \nu + (1 - \nu) \ln(1 - \nu)]}$$

The number of sequences grows exponentially in N , at a rate $h(\nu)$ that depends only on the frequency. It is as if each flip had $e^{h(\nu)}$ choices:

- at $\nu = \frac{1}{2}$, $h = \ln 2$: two choices per flip, every flip is free, and $\binom{N}{N/2} \simeq 2^N$ is, on this exponential scale, almost all the sequences;
- at $\nu = 0$ or 1 , $h = 0$: one choice, and a single sequence (all tails or all heads);
- in between, for instance at $\nu = 0.1$, $h \approx 0.33$ and $e^h \approx 1.38$ choices per flip.

Entropy. The rate $h(\nu)$ depends only on the fractions of heads and tails, ν and $1 - \nu$. These are the probabilities of $\text{Bern}(\nu)$, the distribution of a flip picked at random from the sequence (its **empirical distribution**), with $p(1) = \nu$ and $p(0) = 1 - \nu$.

$$h(\nu) = -[p(1) \ln p(1) + p(0) \ln p(0)].$$

This is the **entropy** of $\text{Bern}(\nu)$, and the definition extends to any distribution:

$$H[p] = -\sum_x p(x) \ln p(x) \quad \text{or} \quad H[p] = -\int p(x) \ln p(x) dx.$$

Entropy is thus the logarithm of the number of microscopic configurations compatible with what we observe, per trial. It is where the two intuitions of Section 1 meet: it **counts** the sequences that show a given **frequency**.

Symmetry reappears too. For n equally likely states, $p(x) = 1/n$, all the terms of the sum are equal:

$$H[\text{uniform}] = -\sum_{x=1}^n \frac{1}{n} \ln \frac{1}{n} = \ln n.$$

The entropy is the logarithm of the number of states, which is counting again: $e^H = n$.

💡 Entropy measures how *spread out* a distribution is, and therefore how much information an experiment gives us: the more spread the distribution, the less we can predict the outcome, and the more we learn by observing it. A high-entropy distribution favours no state over the others; a low-entropy one is biased towards a few. Maximum entropy therefore means minimum bias. This is why many theories, statistical mechanics first of all, start by maximizing the entropy: among all distributions compatible with what we know, we choose the most agnostic one.

For the coin, $n = 2$ gives $\ln 2 = h(\frac{1}{2})$, the maximum of $h(\nu)$ (its derivative, $h'(\nu) = \ln \frac{1-\nu}{\nu}$, vanishes only at $\nu = \frac{1}{2}$). In general, the uniform distribution has the largest entropy among all distributions on n states, and it is exactly the one symmetry would dictate.

Large deviations (advanced): one count, three lenses

We can now answer the question left open in Section 7: why the Gaussian, and why $\sqrt{N}$. We need only the count of sequences and a second-order expansion.

1. Probability = count / total. For a fair coin all $2^N = e^{N \ln 2}$ sequences are equally likely, so the probability of a frequency ν is the fraction of sequences that show it, the ratio plotted in the animation:

$$\boxed{\text{Bin}(m \mid N, \frac{1}{2}) = \frac{\binom{N}{m}}{2^N} \asymp \frac{e^{N h(\nu)}}{e^{N \ln 2}} = e^{-N I(\nu)}, \quad I(\nu) = \ln 2 - h(\nu)}$$

Here $\asymp$ means equal up to factors that change slower than exponentially in N , such as the $\sqrt{2\pi N}$ of Stirling. The **rate function** $I(\nu)$ is how much entropy the frequency ν lacks compared with $\nu = \frac{1}{2}$. Since $h(\nu) \leq \ln 2$, with equality only at $\nu = \frac{1}{2}$, every other frequency is suppressed exponentially in N : the frequency ends up at $\frac{1}{2}$. This is the law of large numbers, and the delta of the first lens.

2. Second-order expansion around the maximum. Differentiate the entropy:

$$h'(\nu) = \ln \frac{1-\nu}{\nu}, \quad h''(\nu) = -\frac{1}{\nu(1-\nu)}.$$

At $\nu = \frac{1}{2}$: $h = \ln 2$, $h' = 0$ (a maximum, so no linear term) and $h'' = -4$. Taylor to second order:

$$h(\nu) \approx \ln 2 - 2(\nu - \frac{1}{2})^2 \implies I(\nu) \approx 2(\nu - \frac{1}{2})^2.$$

3. The Gaussian. Putting this into step 1:

$$\text{Bin}(m | N, \frac{1}{2}) \propto e^{-2N(\nu-1/2)^2} = \exp \left[-\frac{(\nu - \frac{1}{2})^2}{2 \cdot \frac{1}{4N}} \right],$$

a Gaussian in ν centred at $\frac{1}{2}$ with variance $\frac{1}{4N} = \frac{\mu(1-\mu)}{N}$, exactly the σ_ν^2 of Section 6. The $\sqrt{N}$ is now transparent: the exponent is N times a parabola, so it is of order one when $\nu - \frac{1}{2} \sim 1/\sqrt{N}$. In the variable $y = \frac{1}{2} + \sqrt{N}(\nu - \frac{1}{2})$ the exponent becomes $2(y - \frac{1}{2})^2$, the N cancels, and we get $\mathcal{N}(\frac{1}{2}, \frac{1}{4})$, the dashed curve of the $\sqrt{N}$ lens. The Stirling computation of Section 7 was this same expansion, done in m instead of ν .

4. Beyond second order: the tails. The parabola holds only near $\frac{1}{2}$. Typical fluctuations, of size $1/\sqrt{N}$, are close enough; large deviations, of size 1, are not. All N flips come up heads with probability exactly $2^{-N} = e^{-N \ln 2}$, and indeed $I(1) = \ln 2$, while the parabola predicts $I \approx 2 \cdot \frac{1}{4} = \frac{1}{2}$, i.e. $e^{-N/2}$. The Gaussian underestimates how fast the tails die.

A biased coin. For $\mu \neq \frac{1}{2}$ the same steps work. Each sequence with frequency ν now has probability $\mu^m(1-\mu)^{N-m} = e^{N[\nu \ln \mu + (1-\nu) \ln(1-\mu)]}$, and multiplying by the count $e^{Nh(\nu)}$ gives $\text{Bin}(m | N, \mu) \asymp e^{-NI(\nu)}$ with

$$I(\nu) = \nu \ln \frac{\nu}{\mu} + (1-\nu) \ln \frac{1-\nu}{1-\mu}, \quad I(\mu) = I'(\mu) = 0, \quad I''(\mu) = \frac{1}{\mu(1-\mu)},$$

so the second-order expansion gives a Gaussian centred at μ with variance $\mu(1-\mu)/N$.

💡 This $I(\nu)$ is the **Kullback–Leibler divergence** between $\text{Bern}(\nu)$ and $\text{Bern}(\mu)$. It returns in **Lecture 2**, where minimizing it between data and model is the same as maximizing the likelihood.

One function, three readings. The Binomial contains everything, and its logarithm per flip, the rate function $I(\nu) = -\frac{1}{N} \ln \text{Bin}$, built from the entropy, packs it into a single curve. Reading that curve at different depths answers three different questions about what we will observe:

What we read off $I(\nu)$	Question	Scale of $\nu - \mu$	Result	Lens in the animation
its minimum , where $I = 0$	What will I typically observe?	$\rightarrow 0$	the most likely value $\nu = \mu$: law of large numbers	frequency , $\delta(\nu - \mu)$
its curvature at the minimum (second order)	How much does it fluctuate?	$\sim 1/\sqrt{N}$	variance $\mu(1-\mu)/N$, Gaussian shape: central limit theorem	$\sqrt{N}$ zoom , $\mathcal{N}(\mu, \mu(1-\mu))$

What we read off $I(\nu)$	Question	Scale of $\nu - \mu$	Result	Lens in the animation
the whole function , all orders	How rare is an atypical outcome?	~ 1	$P \asymp e^{-NI(\nu)}$: large deviations	log lens , $-\frac{1}{N} \ln \text{Bin} \rightarrow I(\nu)$

Each row is a finer reading of the same function. The minimum is one number, the most likely value. The curvature adds a second one, the variance, and nothing else: this is why the Gaussian of the second row is the same for every distribution with that variance. The whole function keeps everything, including the details that the first two rows forget, and they show up in the tails: 80 heads out of 100, or all heads. The higher terms of the expansion, $\varepsilon^3, \varepsilon^4, \dots$, are the bridge between the second and the third row: they describe how the tails start to depart from the Gaussian, through the asymmetry and the weight of the tails. The first two lenses rescale the *variable*; the third rescales the *logarithm of the probability*, which is why it can still see the tails where the Gaussian lens has nothing left to show.

This is the picture to keep in mind: **the Gaussian is the quadratic part of an entropy**.

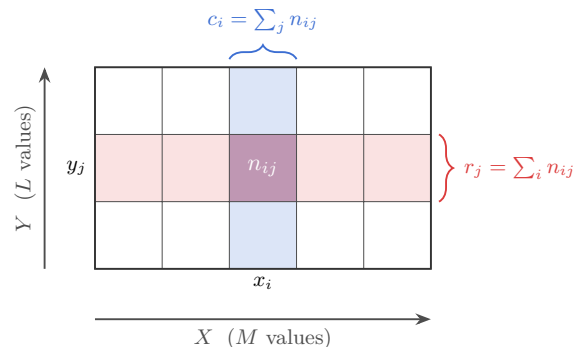
💡 When is the quadratic part not enough? Think of rare events far in the tails, strongly skewed distributions, heavy tails with infinite variance, strongly correlated variables, or bimodal distributions with two minima of I . Be careful about which regime you are in!

💡 The same three readings appear in physics and in ML. In statistical mechanics, $P \propto e^{S/k_B}$: the minimum of the free energy is the equilibrium state, its curvature sets the fluctuations (Einstein), and the whole function the rare events (nucleation, barrier crossing). In ML, the curvature of the expected loss at its minimum is the **Fisher information**: for the coin, $F(\mu) = \frac{1}{\mu(1-\mu)} = I''(\mu)$, and the maximum likelihood estimate $\mu_{\text{ML}} = \nu$ fluctuates with variance $\frac{1}{NF}$ (Lecture 2). A sharp minimum means a well-determined parameter, a flat one a poorly determined one.

11. Several random variables

So far we have studied one random variable at a time. To relate the probabilities of *several* variables, take two random variables measured in the same experiment: X , with values $x_1, \dots, x_M$, and Y , with values $y_1, \dots, y_L$. Repeat the experiment N times and record the outcomes in a table of counts:

- n_{ij} is the number of trials with $X = x_i$ and $Y = y_j$;
- the column total $c_i = \sum_j n_{ij}$ counts the trials with $X = x_i$, whatever Y is;
- the row total $r_j = \sum_i n_{ij}$ counts the trials with $Y = y_j$, whatever X is.



By the law of large numbers, as $N \rightarrow \infty$ these frequencies become probabilities:

- **joint**, $p(x_i, y_j) = n_{ij}/N$: both values occur together;
- **marginal**, $p(x_i) = c_i/N$: $X = x_i$, whatever Y is;
- **conditional**, $p(y_j | x_i) = n_{ij}/c_i$: the fraction with $Y = y_j$ among only the trials with $X = x_i$.

Two identities between the counts become the two rules of probability. Summing a column, $c_i = \sum_j n_{ij}$ and normalizing, gives the sum rule: $\frac{c_i}{N} = \sum_j \frac{n_{ij}}{N}$. Splitting a fraction, $\frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N}$, gives the product rule:

$$\boxed{\text{sum rule: } p(x) = \sum_y p(x, y), \quad \text{product rule: } p(x, y) = p(y | x) p(x)}$$

We derived the rules by counting trials, but the same algebra holds when we count equiprobable states instead, with $|\Omega|$ in place of N : the two intuitions give the same rules. For continuous variables (Section 9) the sums become integrals, $p(x) = \int p(x, y) dy$. A deterministic relation $y = T(x)$ is the conditional density $p(y | x) = \delta(y - T(x))$: in ML we will call this a noiseless supervisor.

Independence. Two variables are **independent** when knowing one tells us nothing about the other, $p(y | x) = p(y)$. By the product rule this is the same as the factorization $p(x, y) = p(x) p(y)$. The two dice of Section 1 are independent: each of the 36 pairs has probability $\frac{1}{36} = \frac{1}{6} \cdot \frac{1}{6}$. Section 4 used this factorization to show that the variances of independent flips add.

The sum rule removes ("marginalizes on") a variable we are not interested in. The product rule builds a joint distribution from a marginal and a conditional. Since nothing distinguishes the roles of X and Y (a symmetry of the table under exchanging rows and columns), the product rule also holds the other way round, $p(x, y) = p(x | y) p(y) = p(y | x) p(x)$, which is the starting point for Bayes' theorem. The latter is

$$p(x | y) = p(y | x) p(x) / p(y)$$

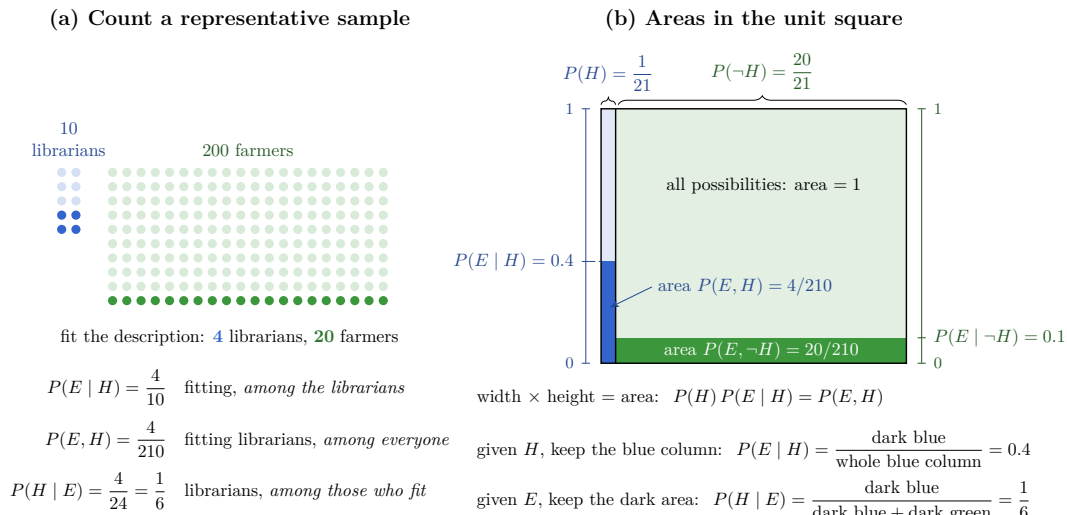
and can also be built using (fixing our) intuition.

Following 3Blue1Brown's *Bayes theorem, the geometry of changing beliefs*, we build it from an example rather than from the formula.

Example (Kahneman & Tversky)(from the book *Thinking Fast and Slow*) Steve is described as shy, withdrawn and tidy, with a passion for detail. Is he more likely a librarian or a farmer? Most people answer "librarian", because the description fits the stereotype. But they forget how many of each there are: assume about 1 librarian for every 20 farmers. Picture a representative sample of 10 librarians and 200 farmers, and suppose the description fits 40% of librarians and 10% of farmers. Then 4 librarians and 20 farmers fit it. Steve is one of these 24 people, so

$$P(\text{librarian} | \text{description}) = \frac{4}{4 + 20} = \frac{1}{6} \approx 17\%.$$

The description is four times more typical of librarians, but farmers are twenty times more numerous, so Steve is still most likely a farmer.



From counting to the formula. Call the claim "Steve is a librarian" the hypothesis H and the description the evidence E . The data of the problem are:

- the **prior** $P(H) = \frac{10}{210} = \frac{1}{21}$, the probability of the hypothesis before seeing the evidence;
- the **likelihood** $P(E | H) = \frac{4}{10} = 0.4$, how probable the evidence is if the hypothesis is true;
- the probability of the evidence if the hypothesis is false, $P(E | \neg H) = \frac{20}{200} = 0.1$.

Now rewrite the count with these numbers, as fractions of all 210 people. The fitting librarians are those who are librarians *and* fit the description: a fraction $P(H) P(E | H) = \frac{1}{21} \cdot 0.4 = \frac{4}{210}$ of everyone (product rule). In the same way, the fitting farmers are $P(\neg H) P(E | \neg H) = \frac{20}{21} \cdot 0.1 = \frac{20}{210}$. Everyone who fits is either a fitting librarian or a fitting farmer, so they are the sum of the two, $\frac{24}{210}$ (sum rule). The ratio "fitting librarians over everyone who fits" does not depend on the total 210, and it becomes

$$P(H | E) = \frac{P(H) P(E | H)}{P(H) P(E | H) + P(\neg H) P(E | \neg H)} = \frac{\frac{1}{21} \cdot 0.4}{\frac{1}{21} \cdot 0.4 + \frac{20}{21} \cdot 0.1} = \frac{1}{6}.$$

The denominator, everyone who fits, is just $P(E)$. In general notation:

$$\underbrace{p(y | x)}_{\text{posterior}} = \frac{\overbrace{p(x | y)}^{\text{likelihood}} \underbrace{p(y)}_{\text{prior}}}{p(x)}, \quad p(x) = \sum_y p(x | y) p(y)$$

In short, the derivation. The joint probability does not care about the order of the variables, $p(x, y) = p(y, x)$. Expanding each side with the product rule gives

$$p(y | x) p(x) = p(x | y) p(y),$$

and dividing by $p(x)$ gives Bayes' theorem. The denominator comes from the sum rule followed by the product rule, $p(x) = \sum_y p(x, y) = \sum_y p(x | y) p(y)$.

Geometric picture. Represent all possibilities as a unit square (panel b).

- Split its width according to the prior: a column of width $P(H)$ and one of width $P(\neg H)$.
- In each column, shade the part where the evidence holds. Its *height*, measured within the column, is $P(E | H)$ or $P(E | \neg H)$.
- The *area* of each shaded rectangle is width $\times$ height, $P(H) P(E | H) = P(E, H)$, which is the product rule. Their total area is $P(E)$, which is the sum rule.

Observing E means throwing away the unshaded part of the square. The posterior is the fraction of what remains that lies in the H column. A strong likelihood cannot compensate for a very thin prior column.

💡 Evidence should not *determine* beliefs from scratch; it should *update* prior beliefs. In Bayesian ML the hypotheses are parameter values θ , the evidence is the dataset $\mathcal{D}$, and $p(\theta | \mathcal{D}) \propto p(\mathcal{D} | \theta) p(\theta)$.

12. Covariance and independence

Averages extend to several variables by weighting each pair of values with the joint probability, $\mathbb{E}[f] = \sum_{x,y} p(x, y) f(x, y)$. A subscript shows which variable is averaged over. For example, $\mathbb{E}_x[f(x, y)]$ is still a function of y . Averaging with a conditional distribution gives the **conditional expectation**, $\mathbb{E}_x[f | y] = \sum_x p(x | y) f(x)$.

The **covariance** measures how two variables vary together. It is positive if they tend to be above their means at the same time, and negative if one tends to be above its mean when the other is below:

$$\text{cov}[x, y] = \mathbb{E}_{x,y}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])] = \mathbb{E}[xy] - \mathbb{E}[x] \mathbb{E}[y].$$

For vectors this becomes the matrix $\text{cov}[\mathbf{x}, \mathbf{y}] = \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{y} - \mathbb{E}[\mathbf{y}])^T]$.

If x and y are independent, the joint distribution factorizes (Section 11), and the double sum splits into a product of two sums, $\mathbb{E}[xy] = \sum_{x,y} p(x)p(y)xy = \mathbb{E}[x]\mathbb{E}[y]$: the covariance vanishes. This is the step used in Section 4 for the cross terms of the Binomial. The converse is false. For example, let $x \in \{-1, 0, 1\}$ with equal probabilities and $y = x^2$. Then $\mathbb{E}[x] = 0$ and $\mathbb{E}[xy] = \mathbb{E}[x^3] = 0$, so $\text{cov}[x, y] = 0$, even though y is completely determined by x .

Expanding the variance of a sum produces a covariance term:

$$\text{var}[x + y] = \text{var}[x] + \text{var}[y] + 2 \text{cov}[x, y].$$

For independent variables this term vanishes, so **variances add**. In particular, the sum of N i.i.d. variables has N times the variance of a single one: this is $\text{var}[m] = N\mu(1 - \mu)$ for the Binomial (Section 4), and $\frac{35}{6} = 2 \cdot \frac{35}{12}$ for the two dice (Section 4).

13. The multivariate Gaussian

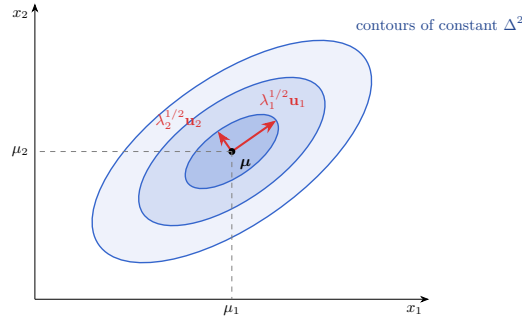
For $\mathbf{x} \in \mathbb{R}^D$, with mean μ and symmetric positive definite covariance Σ :

$$\mathcal{N}(\mathbf{x} \mid \mu, \Sigma) = \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp \left[-\frac{1}{2} (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \right]$$

The exponent contains the **Mahalanobis distance** $\Delta^2 = (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)$. Diagonalizing, $\Sigma \mathbf{u}_i = \lambda_i \mathbf{u}_i$:

$$\Sigma = \sum_{i=1}^D \lambda_i \mathbf{u}_i \mathbf{u}_i^T, \quad \Sigma^{-1} = \sum_{i=1}^D \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^T, \quad \Delta^2 = \sum_i \frac{y_i^2}{\lambda_i}, \quad y_i = \mathbf{u}_i^T (\mathbf{x} - \mu).$$

Contours of constant density are ellipsoids with axes along $\mathbf{u}_i$ and lengths $\propto \lambda_i^{1/2}$.



The change to $\mathbf{y}$ is a rotation (unit Jacobian) and $|\Sigma|^{1/2} = \prod_j \lambda_j^{1/2}$, so the density factorizes into D independent 1D Gaussians:

$$p(\mathbf{y}) = \prod_{j=1}^D \frac{1}{(2\pi\lambda_j)^{1/2}} \exp \left[-\frac{y_j^2}{2\lambda_j} \right].$$

This proves normalization, $\mathbb{E}[\mathbf{x}] = \mu$ and $\text{cov}[\mathbf{x}] = \Sigma$ (and shows why we need $\lambda_i > 0$).

14. Summary

Quantity	Definition	Meaning
Average	$\langle f \rangle = \sum_i f(x_i) p(x_i)$	where the values sit
Variance	$\text{var}[X] = \langle (X - \langle X \rangle)^2 \rangle = \langle X^2 \rangle - \langle X \rangle^2$	mean squared fluctuation
Standard deviation	$\sigma_X = \sqrt{\text{var}[X]}$	typical size of a fluctuation

Distribution	Variable	Parameters	Probability	Mean	Variance
Bernoulli	$x \in \{0, 1\}$	$\mu \in [0, 1]$	$\mu^x (1 - \mu)^{1-x}$	μ	$\mu(1 - \mu)$
Binomial	$m \in \{0, \dots, N\}$	N, μ	$\binom{N}{m} \mu^m (1 - \mu)^{N-m}$	$N\mu$	$N\mu(1 - \mu)$
Gaussian	$x \in \mathbb{R}$	μ, σ^2	$(2\pi\sigma^2)^{-1/2} e^{-(x-\mu)^2/2\sigma^2}$	μ	σ^2
Multivariate Gaussian	$\mathbf{x} \in \mathbb{R}^D$	μ, Σ	$(2\pi)^{-D/2} \Sigma ^{-1/2} e^{-\Delta^2/2}$	μ	Σ

Probability rests on two intuitions: counting equally likely states, which symmetry justifies, and measuring frequencies. The law of large numbers, itself a counting argument, shows that they agree. The average and the variance summarize a distribution by its centre and its spread. The Binomial counts sequences of Bernoulli trials; since independent fluctuations add incoherently, its variance grows like N and the frequency concentrates with a width $1/\sqrt{N}$. Dividing probabilities by the bin width gives densities, and a change of variable changes the density by the Jacobian: the same counts give a delta in the frequency and a Gaussian in the zoomed variable. Several variables obey the sum and product rules, with Bayes' theorem as a direct consequence, and the variances of independent variables add because their covariance vanishes. The same counts, seen through three lenses, give the three limit theorems: at scale $1/N$ the frequency collapses to a delta (law of large numbers), at scale $1/\sqrt{N}$ the independent fluctuations form a Gaussian (central limit theorem), and on a logarithmic scale the tails decay at a rate set by the entropy, the logarithm of the number of sequences per flip (large deviations).

References

- C. M. Bishop and H. Bishop, *Deep Learning: Foundations and Concepts*, Springer, 2024 (chapters on probabilities and standard distributions).
- 3Blue1Brown, *But what is the Central Limit Theorem?*, YouTube.
- L. Susskind, *The Theoretical Minimum*, Stanford Continuing Studies, Statistical Mechanics course; notes on Lesson 1, *Entropy and conservation of information*, [available on GitHub](#).