Probability: A Review

Stephen Boyd and Sanjay Lall

EE263
Stanford University

Why probability?

Much of what is left in this course: measurement *noise*, least-squares read as *estimation*, *Gaussians*, and the *covariance matrix* we already met in PCA, rests on a little probability.

This is a self-contained refresher, from the ground up:

- what a probability even *means*
- random variables and their *distributions* (discrete and continuous)
- *expectation*, *variance*, and higher moments
- *covariance*, and where that covariance matrix came from

Outcomes, events, sample space

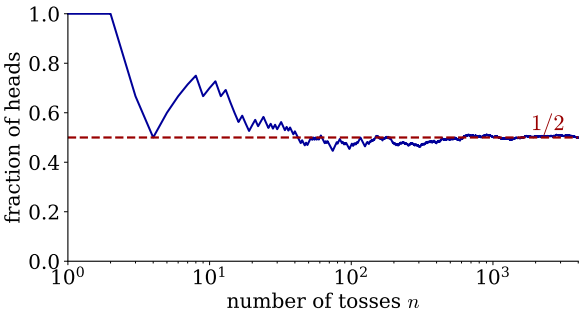
An experiment with an uncertain result is described by three things:

- the *sample space* Ω: the set of all possible outcomes
- an *outcome* $\omega \in \Omega$: one specific thing that happens
- an *event* $A \subseteq \Omega$: a set of outcomes we care about

A *probability* assigns to each event a number $\textbf{Prob}(A) \in [0, 1]$, the chance that it occurs.

example: roll a die. $\Omega = \{1, 2, 3, 4, 5, 6\}$; the event "even" is $A = \{2, 4, 6\}$; for a fair die $\textbf{Prob}(A) = 3/6 = 1/2$.

What does a probability mean?



One concrete meaning: repeat the experiment many times, independently. Then **Prob**(A) is the *long-run fraction* of trials in which A occurs.

For a fair coin, the fraction of heads settles toward $1/2$ as the number of tosses grows.

Probability is the idealization of this fraction. The rules below are what it must obey; the law of large numbers (later) makes the picture precise.

The rules probability obeys

A probability satisfies three axioms:

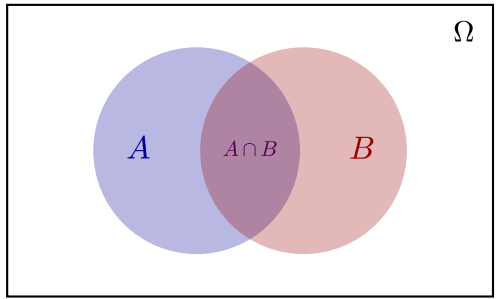
- $\textbf{Prob}(A) \geq 0$ for every event
- $\textbf{Prob}(\Omega) = 1$ (something happens)
- $\textbf{Prob}(A \cup B) = \textbf{Prob}(A) + \textbf{Prob}(B)$ for *disjoint* A, B

Everything else follows. In particular,

$$\textbf{Prob}(A^c) = 1 - \textbf{Prob}(A),$$

and for events that may overlap,

$$\textbf{Prob}(A \cup B) = \textbf{Prob}(A) + \textbf{Prob}(B) - \textbf{Prob}(A \cap B).$$



The overlap $A \cap B$ is counted twice in $\textbf{Prob}(A) + \textbf{Prob}(B)$, so we subtract it once.

Conditioning and independence

Given that B occurred, rescale to the world where B is true:

$$\textbf{Prob}(A \mid B) = \frac{\textbf{Prob}(A \cap B)}{\textbf{Prob}(B)}, \qquad \textbf{Prob}(B) > 0.$$

A and B are *independent* if knowing one says nothing about the other:

$$\textbf{Prob}(A \cap B) = \textbf{Prob}(A)\ \textbf{Prob}(B) \iff \textbf{Prob}(A \mid B) = \textbf{Prob}(A).$$

example: two tosses of a fair coin are independent, so $\textbf{Prob}(\mathrm{HH}) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$.

Bayes' rule

Writing $\textbf{Prob}(A \cap B)$ two ways, $\textbf{Prob}(A \mid B)\,\textbf{Prob}(B) = \textbf{Prob}(B \mid A)\,\textbf{Prob}(A)$, gives a way to *flip the conditioning*:

$$\textbf{Prob}(A \mid B) = \frac{\textbf{Prob}(B \mid A)\ \textbf{Prob}(A)}{\textbf{Prob}(B)}.$$

This turns "probability of the *data* given the cause" into "probability of the *cause* given the data."

That is exactly the move behind estimation: we observe measurements and want the most probable underlying quantity. We return to it in the estimation lecture.

Random variables

A *random variable* X attaches a number to each outcome, $X : \Omega \to \mathbb{R}$. It turns "chance" into something we can add, average, and plot.

- number of heads in 10 tosses (values $0, 1, \ldots, 10$): *discrete*
- the arrival time of a bus (any real ≥ 0): *continuous*

Its *distribution* says how likely each value, or range of values, is. Discrete and continuous variables describe this in parallel ways: a *mass* at each point, or a *density* over the line.

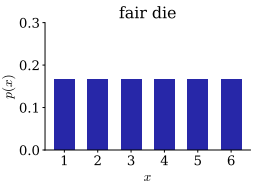
Discrete variables: the mass function

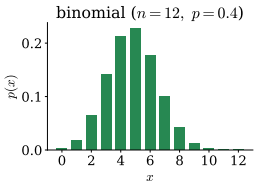
A discrete X takes values in a countable set. Its *probability mass function* gives the weight of each value:

$$p(x) = \textbf{Prob}(X = x), \quad p(x) \geq 0, \quad \sum_x p(x) = 1.$$

Probabilities of events are just sums,

$$\textbf{Prob}(X \in A) = \sum_{x \in A} p(x).$$

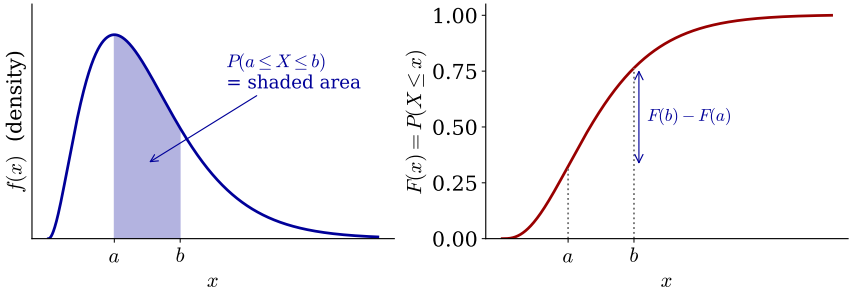




Continuous variables: the density and the CDF

A continuous X has *no* mass at any single point ($\textbf{Prob}(X = x) = 0$). A *probability density* f gives probability as *area*:

$$\textbf{Prob}(a \le X \le b) = \int_a^b f(x)\,dx, \quad f \ge 0, \quad \int_{-\infty}^{\infty} f(x)\,dx = 1.$$



The *cumulative distribution function* $F(x) = \textbf{Prob}(X \le x)$ works for both kinds: sums become integrals, mass becomes density.

A few distributions worth knowing

discrete

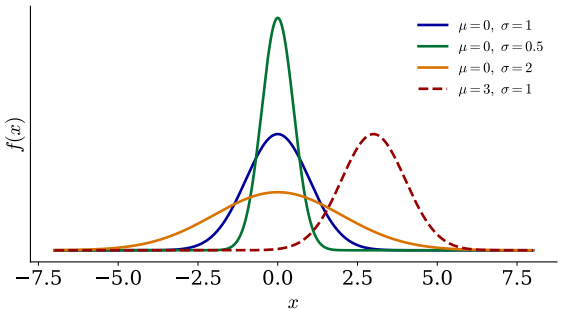
- Bernoulli(p): $X \in \{0, 1\}$, **Prob**$(X = 1) = p$ (one trial)
- binomial(n, p): successes in n trials, $p(k) = \binom{n}{k} p^k (1-p)^{n-k}$
- Poisson(λ): counts of rare events, $p(k) = e^{-\lambda} \lambda^k / k!$

continuous

- uniform$[a, b]$: $f = \dfrac{1}{b-a}$ on $[a, b]$ (no value preferred)
- exponential(λ): $f = \lambda e^{-\lambda x}$, $x \geq 0$ (waiting times)
- Gaussian $\mathcal{N}(\mu, \sigma^2)$: the bell curve (next)

The Gaussian deserves its own two slides: it is the one we lean on most.

The Gaussian: shape and parameters



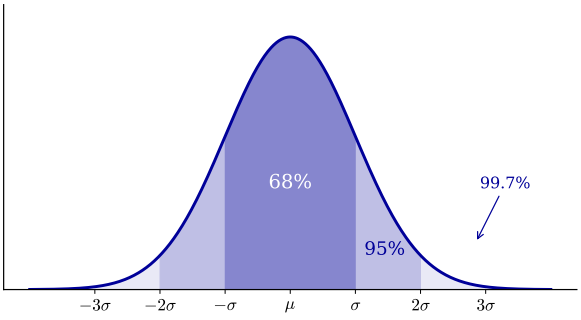
The *Gaussian* (normal) density is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big).$$

We write $X \sim \mathcal{N}(\mu, \sigma^2)$.

The mean μ sets the *center*; the standard deviation σ sets the *width*. These two numbers determine the entire curve.

The Gaussian: the 68–95–99.7 rule

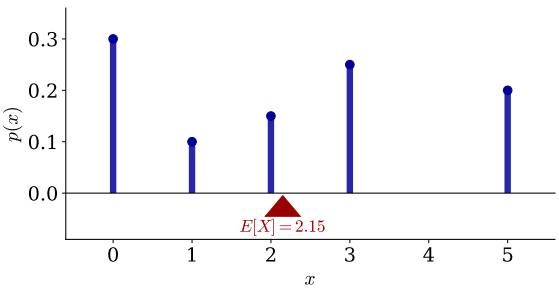


The mass concentrates near the mean, in amounts worth memorizing:

- $\approx 68\%$ within $\mu \pm \sigma$
- $\approx 95\%$ within $\mu \pm 2\sigma$
- $\approx 99.7\%$ within $\mu \pm 3\sigma$

It is the most important distribution here: sums of many small effects look Gaussian (the CLT, at the end), and the estimation story is Gaussian throughout.

Expectation: the average value



The *expectation* (mean) is the probability-weighted average of the values: the *balance point* of the distribution.

$$\mathbf{E}\,X = \sum_x x\,p(x) \quad \text{or} \quad \mathbf{E}\,X = \int x\,f(x)\,dx.$$

Discrete: weight each value by its mass. Continuous: by its density. The mean need not be a value X can actually take.

Working with expectation

Expectation is *linear*, and this needs no independence:

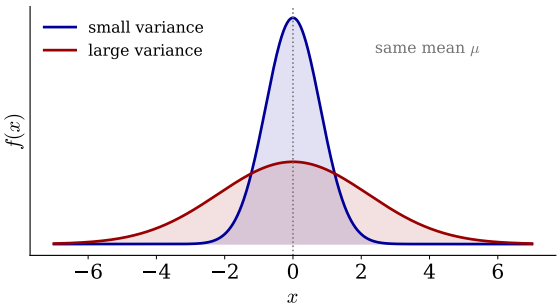
$$\mathbf{E}(aX + b) = a\,\mathbf{E}\,X + b, \qquad \mathbf{E}(X + Y) = \mathbf{E}\,X + \mathbf{E}\,Y.$$

For any function of X,

$$\mathbf{E}\,g(X) = \sum_x g(x)\,p(x) \qquad \text{or} \qquad \mathbf{E}\,g(X) = \int g(x)\,f(x)\,dx.$$

Linearity is the workhorse of the whole subject: it holds even when variables are dependent, and it is what makes least-squares and estimation tractable.

Variance and standard deviation



Variance measures spread about the mean $\mu = \mathbf{E}\, X$:

$$\mathbf{var}\, X = \mathbf{E}\left[(X-\mu)^2\right] = \mathbf{E}\, X^2 - \mu^2.$$

The *standard deviation* $\mathbf{std}\, X = \sqrt{\mathbf{var}\, X}$ puts the spread back in the units of X.

Shifting does not change spread; scaling squares it:

$$\mathbf{var}(aX+b) = a^2\ \mathbf{var}\, X.$$

Moments

Mean and variance are the first two *moments*. In general the k-th moment is $\mathbf{E}\, X^k$, and the k-th *central* moment is $\mathbf{E}(X - \mu)^k$.

- 1st moment $=$ mean (center)
- 2nd central moment $=$ variance (spread)
- 3rd, 4th, ... describe skew and tail weight (shape)

For most of our work the first two are all we need. For a Gaussian they are *everything*: μ and σ^2 pin down the entire distribution.

Mean and variance of the usual suspects

distribution	mean	variance
Bernoulli(p)	p	$p(1-p)$
binomial(n, p)	np	$np(1-p)$
uniform$[a, b]$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
exponential(λ)	$1/\lambda$	$1/\lambda^2$
Gaussian $\mathcal{N}(\mu, \sigma^2)$	μ	σ^2

Worth knowing the first two by heart; the rest you can always look up.

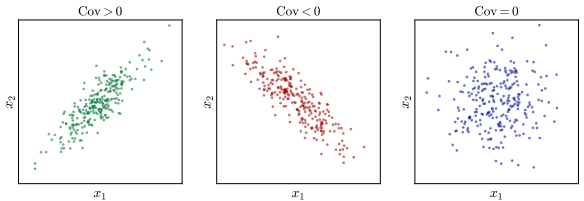
Two random variables at once

Often several quantities vary together. Their *joint* distribution gives probabilities of combinations, e.g. $p(x, y) = \textbf{Prob}(X = x,\, Y = y)$.

- *marginal*: recover one variable by summing (or integrating) out the other
- *independent* if the joint factors, $p(x, y) = p_X(x)\, p_Y(y)$ for all x, y

Independence means X carries no information about Y: the value of one does not shift the distribution of the other.

Covariance and correlation



Covariance measures whether X and Y move together:

$$\operatorname{cov}(X, Y) = \mathbf{E}\left[(X - \mu_X)(Y - \mu_Y)\right].$$

The unit-free version is the *correlation* $\rho = \dfrac{\operatorname{cov}(X, Y)}{\mathbf{std}\, X\ \mathbf{std}\, Y} \in [-1, 1]$.

Independent variables have $\operatorname{cov} = 0$; the converse is false. Stacking n variables, all their covariances form the *covariance matrix* S: exactly the symmetric matrix whose eigenvectors PCA found.

Sums and the sample mean

For the *mean*, sums are effortless (linearity):

$$\mathbf{E}(X_1 + \cdots + X_n) = \mathbf{E}\, X_1 + \cdots + \mathbf{E}\, X_n.$$

For the *variance*, independence matters:

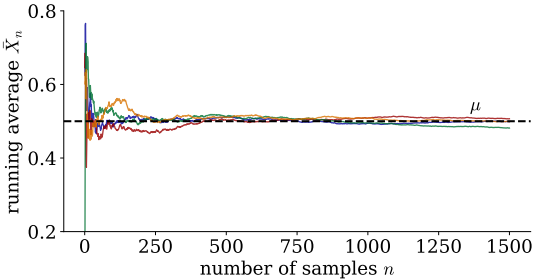
$$\mathbf{var}(X + Y) = \mathbf{var}\, X + \mathbf{var}\, Y \qquad \text{when } X, Y \text{ are independent.}$$

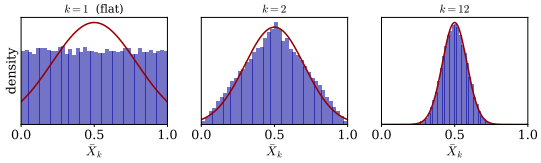
Average n independent copies, each with mean μ and variance σ^2. The *sample mean* $\bar{X}_n = \frac{1}{n}\sum_i X_i$ then has

$$\mathbf{E}\, \bar{X}_n = \mu, \qquad \mathbf{var}\, \bar{X}_n = \frac{\sigma^2}{n}.$$

More data means less spread: averaging cancels noise.

Two theorems that tie it together





Law of large numbers (left): the sample mean $\bar{X}_n$ converges to the true mean μ as n grows. This is the frequency picture from the start of the lecture, made precise.

Central limit theorem (right): for large n the sample mean is approximately *Gaussian*, whatever the original distribution (here the average of k uniforms turns into a bell). Together they explain why averages are trustworthy and why the Gaussian is everywhere.

Summary

- a *probability* obeys three axioms; conditioning and Bayes let us update it
- a *random variable* has a distribution: a *mass function* (discrete) or a *density* (continuous), with the *CDF* common to both
- *expectation* is the weighted average and is *linear*; *variance* is the spread, and moments describe finer shape
- *covariance* measures joint variation; stacked up it is the covariance matrix
- the *law of large numbers* and *central limit theorem* make averages reliable and Gaussian

Two numbers carry most of the weight: the *mean* (center) and the *variance* (spread).