Module 1 — Probability Foundations

Statistics

hbar.university

Probability Space

Position in Knowledge Graph

System:
Statistics

Structural Domain:
Probability Foundations

Node:
1.1

Formal Definition

A **probability space** is a triple:

$$(\Omega, \mathcal{F}, \mathbb{P})$$

where:

- Ω is the sample space (set of possible outcomes),
- $\mathcal{F} \subseteq 2^\Omega$ is a sigma-algebra of events,
- $\mathbb{P} : \mathcal{F} \to [0, 1]$ is a probability measure satisfying:

1. $\mathbb{P}(\Omega) = 1$
2. Countable additivity:
$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$
for disjoint events A_i.

Intuition

A probability space formalizes uncertainty.

- Ω = what could happen
- $\mathcal{F}$ = what we are allowed to talk about
- $\mathbb{P}$ = how belief mass is distributed

It is the minimal mathematical container for randomness.

Structural Role

Probability spaces are the foundation of all statistical reasoning.

Every statistical model ultimately assumes:

- Data arises from a probability space.
- Parameters index probability measures.
- Inference operates over families of such spaces.

Without this structure, estimation and testing have no formal meaning.

Mathematical Development

From the axioms, we derive:

- $\mathbb{P}(\emptyset) = 0$
- Monotonicity:
$$A \subseteq B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$$
- Complement rule:
$$\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$$

These properties make probability a measure-theoretic object.

Worked Example

Coin Toss

Let:

$$\Omega = \{H, T\}$$

$$\mathcal{F} = 2^{\Omega}$$

$$\mathbb{P}(H) = \mathbb{P}(T) = 0.5$$

This defines a valid probability space.

Cross-Links

- **Linear Algebra:** Measure as linear functional.
- **AI:** Data-generating distribution.
- **Economics:** Uncertainty over states of the world.
- **Quantum:** Born rule defines probability measures over measurement outcomes.

Structural Compression

Invariant:
Total probability mass equals 1.

Mechanism:
Additive measure over disjoint events.

Cross-System Echo:
Conservation principles in physics; normalization in machine learning.

Exercises

1. Prove that $\mathbb{P}(\emptyset) = 0$.
2. Show that $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.
3. Construct a probability space for a fair six-sided die.

Events and Sigma-Algebras

Position in Knowledge Graph

System:
Statistics

Structural Domain:
Probability Foundations

Node:
1.2

Formal Definition

An **event** is a subset of the sample space Ω that we assign a probability to.

A **sigma-algebra** $\mathcal{F} \subseteq 2^{\Omega}$ is a collection of subsets of Ω such that:

1. $\Omega \in \mathcal{F}$
2. If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$
3. If $A_1, A_2, \dots \in \mathcal{F}$, then

$$\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}.$$

From (2) and (3) it follows also that $\mathcal{F}$ is closed under countable intersections.

Intuition

A sigma-algebra is "what questions you are allowed to ask."

Sometimes, the full power set 2^{Ω} is too large or too pathological, so we restrict to a consistent family of events where probability behaves well.

Think:

- Ω: reality-space

- $\mathcal{F}$: the set of measurable questions

- $\mathbb{P}$: the answer-weighting function

Structural Role

Sigma-algebras are the "language layer" of probability.

They ensure:

- probabilities can be consistently assigned,
- limits of events remain events (countable operations),
- random variables can be defined rigorously (measurability depends on $\mathcal{F}$).

This is the gateway to random variables and integration (expectation).

Mathematical Development

Consequences

If $A_1, A_2, \dots \in \mathcal{F}$, then:

- Countable intersections:
$$\bigcap_{i=1}^{\infty} A_i \in \mathcal{F}$$
because
$$\bigcap_{i=1}^{\infty} A_i = \left(\bigcup_{i=1}^{\infty} A_i^c \right)^c$$
and sigma-algebras are closed under complements and countable unions.
- Finite unions and intersections are special cases of the countable rules.

Worked Examples

Example 1: Full power set (finite Ω)

If Ω is finite, then $\mathcal{F} = 2^{\Omega}$ is always a sigma-algebra.

Example 2: Trivial sigma-algebra

$$\mathcal{F} = \{\emptyset, \Omega\}$$

This corresponds to having only one meaningful question: "Did anything happen at all?"

Cross-Links

- **Statistics:** Models restrict which events are relevant (measurable) under a sampling process.
- **AI:** Feature representations define what distinctions the model can express (a "learned sigma-algebra" intuition).
- **Economics:** Information sets correspond to restricted event families (what an agent can distinguish).
- **Quantum:** Measurement choice defines the event structure (what outcomes are distinguishable).

Structural Compression

Invariant:
Closure under complement + countable union keeps probability consistent under limits.

Mechanism:
"Allowed questions" must be stable under repeated refinement.

Cross-System Echo:
In programming: closed interfaces; in physics: observable sets depend on measurement.

Exercises

1. Verify that $\{\emptyset, \Omega\}$ is a sigma-algebra.
2. If $\mathcal{F}$ is a sigma-algebra and $A, B \in \mathcal{F}$, prove that $A \setminus B \in \mathcal{F}$.
3. Let $\Omega = \{1, 2, 3, 4\}$. List all events in the sigma-algebra generated by the partition $\{\{1, 2\}, \{3, 4\}\}$.

Random Variables

Position in Knowledge Graph

System:
Statistics

Structural Domain:
Probability Foundations

Node:
1.3

Formal Definition

A **random variable** is a measurable function:

$$X : \Omega \to \mathbb{R}$$

such that for every Borel set $B \subseteq \mathbb{R}$,

$$X^{-1}(B) \in \mathcal{F}.$$

This measurability condition ensures that probabilities of events of the form $\{X \in B\}$ are well-defined.

Intuition

A random variable is not "a random number."

It is a function that translates outcomes into numbers.

$$\text{Outcome} \longrightarrow \text{Number}$$

It allows us to measure, summarize, and analyze uncertainty quantitatively.

Structural Role

Random variables connect abstract probability spaces to numerical structure.

They enable:

- Expectation
- Variance
- Distribution functions
- Statistical modeling

They are the bridge from pure measure theory to statistics.

Mathematical Development

Induced Distribution

The random variable X induces a probability measure on $\mathbb{R}$:

$$\mathbb{P}_X(B) = \mathbb{P}(X \in B)$$

This is called the **pushforward measure** or distribution of X.

Discrete Example

If:

$$\Omega = \{H, T\}$$

Define:

$$X(H) = 1, \quad X(T) = 0.$$

Then:

$$\mathbb{P}(X = 1) = 0.5, \quad \mathbb{P}(X = 0) = 0.5.$$

Cross-Links

- **Linear Algebra:** Random vectors are vector-valued random variables.
- **AI:** Data features are realizations of random variables.
- **Economics:** Utility functions act on random states.
- **Quantum:** Observables are operator-valued analogues of random variables.

Structural Compression

Invariant:
Random variables preserve measurable structure.

Mechanism:
Function pushes probability measure forward.

Cross-System Echo:
In ML, feature maps transform input space into numeric representation.

Exercises

1. Prove that if X is measurable, then $aX + b$ is measurable.
2. Construct a random variable on a die roll representing parity.
3. Show how a CDF can be defined from the induced distribution.

Distribution Functions

Position in Knowledge Graph

System:
Statistics

Structural Domain:
Probability Foundations

Node:
1.4

Formal Definition

The **cumulative distribution function (CDF)** of a random variable X is the function:

$$F_X(x) = \mathbb{P}(X \leq x).$$

The CDF satisfies:

1. F_X is non-decreasing.
2. $\lim_{x \to -\infty} F_X(x) = 0$.
3. $\lim_{x \to \infty} F_X(x) = 1$.
4. F_X is right-continuous.

Intuition

The CDF accumulates probability from left to right.

At any real number x, it tells us:

> "How much probability mass lies at or below this value?"

It compresses the full distribution of X into a single monotonic function.

Structural Role

The distribution function:

- Fully characterizes the law of a random variable.
- Enables computation of probabilities:

$$\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a).$$

- Bridges discrete and continuous cases into a unified framework.

All density and mass functions are derived from the CDF.

Mathematical Development

Discrete Case

If X takes values x_i with probabilities p_i, then:

$$F_X(x) = \sum_{x_i \le x} p_i.$$

The CDF has jump discontinuities at mass points.

Continuous Case

If X has density $f_X(x)$, then:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt.$$

The density is the derivative of the CDF:

$$f_X(x) = \frac{d}{dx} F_X(x).$$

Worked Example

Fair Die

Let $X \in \{1, 2, 3, 4, 5, 6\}$ with equal probability.

Then:

$$F_X(3) = \frac{3}{6} = 0.5.$$

The CDF increases in steps of $\frac{1}{6}$.

Cross-Links

- **Linear Algebra:** CDFs generalize to multivariate distributions via joint probability measures.
- **AI:** Model outputs often represent CDF or probability mass functions.
- **Economics:** Risk analysis depends on distribution tails.
- **Quantum:** Measurement outcome probabilities form cumulative distributions over eigenvalues.

Structural Compression

Invariant:
CDF is monotone and bounded between 0 and 1.

Mechanism:
Accumulation of probability mass via integration or summation.

Cross-System Echo:
In calculus, integration accumulates area; in learning, loss aggregates over data.

Exercises

1. Prove that every CDF is right-continuous.
2. Show that a CDF uniquely determines a probability measure on $\mathbb{R}$.
3. Sketch the CDF of a Bernoulli(p) random variable.

Expectation

Position in Knowledge Graph

System:
Statistics

Structural Domain:
Probability Foundations

Node:
1.5

Formal Definition

Let X be a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

The **expectation** of X is defined as:

Discrete Case

$$\mathbb{E}[X] = \sum_x x\,\mathbb{P}(X = x),$$

provided the sum converges absolutely.

Continuous Case

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\,dx,$$

provided the integral is finite.

General (Measure-Theoretic Form)

$$\mathbb{E}[X] = \int_\Omega X(\omega)\,d\mathbb{P}(\omega).$$

Intuition

Expectation is not "what usually happens."

It is the **probability-weighted average** of all possible outcomes.

It is the center of mass of the distribution.

If probability is mass, then expectation is balance.

Structural Role

Expectation is the fundamental linear operator of probability theory.

It enables:

- Definition of variance
- Risk evaluation in estimation

- Decision theory
- Optimization of statistical procedures
- Law of Large Numbers

Almost all statistical reasoning is built on properties of expectation.

Mathematical Development

Linearity

For random variables X, Y and constants a, b:

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y].$$

Linearity holds without independence.

Indicator Trick

For an event A, define:

$$\mathbf{1}_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{otherwise} \end{cases}$$

Then:

$$\mathbb{E}[\mathbf{1}_A] = \mathbb{P}(A).$$

This connects expectation directly to probability.

Worked Example

Fair Die

$$\mathbb{E}[X] = \frac{1+2+3+4+5+6}{6} = 3.5.$$

Note: 3.5 is not an achievable outcome. Expectation need not lie in the support of X.

Cross-Links

- **Linear Algebra:** Expectation is a linear functional.
- **AI:** Loss functions are expectations over data distributions.
- **Economics:** Expected utility theory.
- **Quantum:** Expected value of an observable corresponds to $\langle\psi|A|\psi\rangle$.

Structural Compression

Invariant:
Expectation is linear.

Mechanism:
Integration against probability measure.

Cross-System Echo:
In physics, expectation resembles averaging over states; in ML, empirical averages approximate expectations.

Exercises

1. Prove linearity of expectation.
2. Show that $\mathbb{E}[c] = c$ for constant c.
3. Let $X \sim \text{Bernoulli}(p)$. Compute $\mathbb{E}[X]$.
4. Show that expectation of an indicator recovers probability.

Variance and Covariance

Position in Knowledge Graph

System:
Statistics

Structural Domain:
Probability Foundations

Node:
1.6

Formal Definition

Let X be a random variable with finite expectation.

Variance

$$\mathrm{Var}(X) = \mathbb{E}\big[(X - \mathbb{E}[X])^2\big].$$

Equivalently,

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$$

Covariance

For random variables X, Y:

$$\mathrm{Cov}(X, Y) = \mathbb{E}\big[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\big].$$

Intuition

Expectation measures center.

Variance measures spread.

Covariance measures joint variation.

If expectation is "balance,"
variance is "dispersion energy."

Covariance tells us whether two variables move together.

Structural Role

Variance and covariance introduce geometry into probability.

They enable:

- Measurement of uncertainty magnitude
- Correlation analysis
- Construction of covariance matrices

- Linear regression
- Multivariate statistics

They transform probability space into an inner-product structure.

Mathematical Development

Key Properties

1. Non-negativity:
$$\mathrm{Var}(X) \geq 0.$$

2. Scaling:
$$\mathrm{Var}(aX) = a^2 \mathrm{Var}(X).$$

3. Variance of Sum:
$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y).$$

If X and Y are independent:

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

Correlation

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

Correlation is dimensionless and bounded:

$$-1 \leq \rho(X,Y) \leq 1.$$

Worked Example

Bernoulli(p)

If $X \sim \mathrm{Bernoulli}(p)$,

$$\mathbb{E}[X] = p,$$

$$\mathrm{Var}(X) = p(1-p).$$

Maximum variance occurs at $p = 0.5$.

Cross-Links

- **Linear Algebra:** Covariance defines a positive semidefinite matrix.
- **AI:** Covariance matrices appear in PCA and Gaussian models.
- **Economics:** Portfolio variance measures risk.
- **Quantum:** Variance of observables relates to uncertainty principles.

Structural Compression

Invariant:
Variance is always non-negative.

Mechanism:
Quadratic deviation around mean.

Cross-System Echo:
In physics, variance mirrors energy around equilibrium; in ML, variance reflects model sensitivity.

Exercises

1. Prove that variance is non-negative.
2. Show the formula $\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.
3. Compute variance of a fair die.
4. Prove that if X and Y are independent, then $\mathrm{Cov}(X, Y) = 0$.

Conditional Probability and Conditional Expectation

Position in Knowledge Graph

System:
Statistics

Structural Domain:
Probability Foundations

Node:
1.7

Formal Definition

Conditional Probability

For events $A, B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$,

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

Conditional Expectation (Discrete Case)

If Y is a discrete random variable,

$$\mathbb{E}[X \mid Y = y] = \sum_x x \, \mathbb{P}(X = x \mid Y = y).$$

General Definition (Measure-Theoretic Form)

The **conditional expectation** of X given a sigma-algebra $\mathcal{G} \subseteq \mathcal{F}$ is a random variable $\mathbb{E}[X \mid \mathcal{G}]$ such that:

1. $\mathbb{E}[X \mid \mathcal{G}]$ is $\mathcal{G}$-measurable.
2. For all $G \in \mathcal{G}$,

$$\int_G \mathbb{E}[X \mid \mathcal{G}] \, d\mathbb{P} = \int_G X \, d\mathbb{P}.$$

Intuition

Conditional probability updates belief given information.

Conditional expectation updates averages given information.

If expectation is balance, conditional expectation is balance **after learning something**.

It represents the best prediction of X given available information.

Structural Role

Conditional expectation is the backbone of:

- Bayesian inference
- Regression
- Martingale theory
- Filtering
- Sequential decision-making

It formalizes how information changes predictions.

Mathematical Development

Law of Total Probability

If $\{B_i\}$ partitions Ω,

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

Law of Total Expectation

$$\mathbb{E}[X] = \mathbb{E}\big[\mathbb{E}[X \mid Y]\big].$$

This is sometimes called the **tower property**.

Best Prediction Property

Conditional expectation minimizes mean squared error:

$$\mathbb{E}[X \mid Y] = \arg\min_g \mathbb{E}\big[(X - g(Y))^2\big].$$

This connects probability directly to regression.

Worked Example

Let $X \sim \text{Bernoulli}(p)$ and suppose we observe an event B.

Then:

$$\mathbb{P}(X = 1 \mid B) = \frac{\mathbb{P}(B \mid X = 1)\mathbb{P}(X = 1)}{\mathbb{P}(B)}.$$

This is Bayes' rule.

Cross-Links

- **Linear Algebra:** Conditional expectation is an orthogonal projection in L^2.
- **AI:** Bayesian updating and predictive modeling.
- **Economics:** Expectations conditional on information sets.
- **Quantum:** State update after measurement.

Structural Compression

Invariant:
Conditioning preserves total expectation (tower property).

Mechanism:
Projection onto information sigma-algebra.

Cross-System Echo:
In ML, regression projects outcomes onto feature space.

Exercises

1. Prove the law of total probability.
2. Prove the tower property.
3. Show that conditional expectation minimizes mean squared error.
4. Derive Bayes' rule from the definition of conditional probability.

Law of Large Numbers

Position in Knowledge Graph

System:
Statistics

Structural Domain:
Probability Foundations

Node:
1.8

Formal Statement

Let $X_1, X_2, \ldots$ be independent and identically distributed (i.i.d.) random variables with finite expectation $\mu = \mathbb{E}[X_i]$.

Define the sample mean:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Weak Law of Large Numbers (WLLN)

$$\bar{X}_n \xrightarrow{P} \mu.$$

That is, the sample mean converges in probability to the true expectation.

Strong Law of Large Numbers (SLLN)

$$\bar{X}_n \xrightarrow{a.s.} \mu.$$

The convergence occurs almost surely.

Intuition

Averages stabilize.

As we collect more data, random fluctuations cancel out.

The sample mean becomes a reliable estimator of the true mean.

Probability turns into empirical regularity.

Structural Role

The Law of Large Numbers justifies:

- Statistical estimation
- Empirical averages
- Risk minimization
- Machine learning training procedures

Without LLN, statistics would not work.

Mathematical Insight

Variance of the sample mean:

$$\mathrm{Var}(\bar{X}_n) = \frac{\mathrm{Var}(X)}{n}.$$

As $n \to \infty$, variance shrinks to zero.

Concentration drives convergence.

Worked Example

Flip a fair coin.

Let $X_i = 1$ for heads, 0 for tails.

Then:

$$\bar{X}_n \to 0.5.$$

The empirical frequency converges to true probability.

Cross-Links

- **AI:** Empirical risk converges to expected risk.
- **Economics:** Large populations stabilize aggregate behavior.
- **Quantum:** Repeated measurements converge to expectation values.
- **Linear Algebra:** Averaging as projection onto constant vector.

Structural Compression

Invariant:
Sample mean approaches expectation.

Mechanism:
Variance shrinks as $1/n$.

Cross-System Echo:
In ML, more data reduces estimation noise.

Exercises

1. Prove convergence in probability using Chebyshev's inequality.
2. Compute variance of $\bar{X}_n$.
3. Explain difference between weak and strong LLN.

Central Limit Theorem

Position in Knowledge Graph

System:
Statistics

Structural Domain:
Probability Foundations

Node:
1.9

Formal Statement

Let $X_1, X_2, \ldots$ be i.i.d. random variables with:

$$\mathbb{E}[X_i] = \mu, \quad \mathrm{Var}(X_i) = \sigma^2 < \infty.$$

Define:

$$Z_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}.$$

Then:

$$Z_n \xrightarrow{d} \mathcal{N}(0, 1).$$

Intuition

No matter the original distribution (under mild conditions), normalized averages become approximately normal.

The Gaussian distribution emerges universally.

Noise aggregates into structure.

Structural Role

The CLT enables:

- Approximate confidence intervals
- Hypothesis testing
- Asymptotic inference
- Statistical modeling at scale

It connects finite samples to Gaussian structure.

Mathematical Insight

The CLT explains why:

$$\bar{X}_n \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

for large n.

This approximation powers classical statistics.

Worked Example

Let $X_i \sim \text{Bernoulli}(p)$.

Then:

$$\sqrt{n}(\bar{X}_n - p)$$

converges to:

$$\mathcal{N}(0, p(1-p)).$$

Cross-Links

- **AI:** SGD noise approximates Gaussian under large batch regimes.
- **Economics:** Aggregate shocks often modeled as Gaussian.
- **Quantum:** Gaussian approximations in large systems.
- **Linear Algebra:** Multivariate CLT yields multivariate normal with covariance matrix.

Structural Compression

Invariant:
Normal distribution emerges under aggregation.

Mechanism:
Independent fluctuations accumulate additively.

Cross-System Echo:
Universal Gaussian noise in physics and learning systems.

Exercises

1. State Lindeberg condition for CLT.
2. Show how CLT justifies normal approximation to binomial.
3. Explain difference between LLN and CLT.