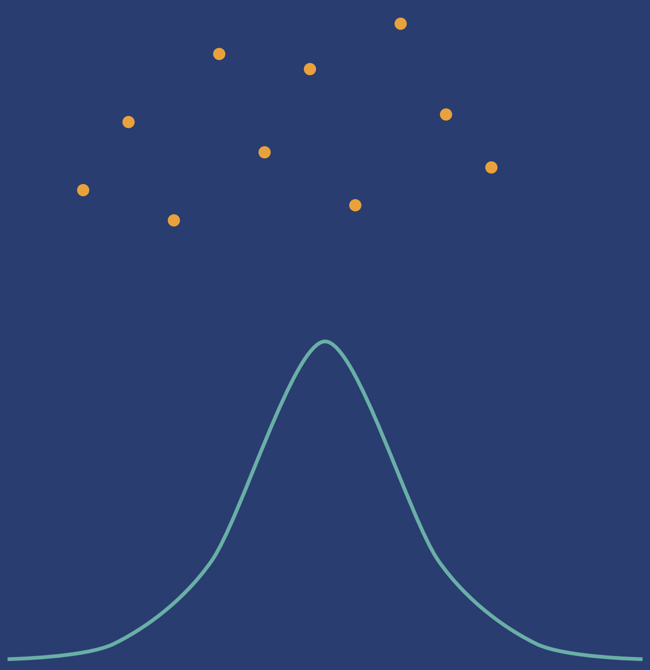
AN INTRODUCTION TO PROBABILITY AND STATISTICS

Chance, Data, and Decisions

A visual, inquiry-driven course with exercises



One-semester course packet

events → random variables → samples → inference

How to use this course

Big idea

Probability starts with a model of what might happen; statistics works backward from what did happen. The course repeatedly follows one cycle:

$$\text{model} \longrightarrow \text{simulate} \longrightarrow \text{observe} \longrightarrow \text{infer} \longrightarrow \text{check.}$$

The point is not merely to calculate a probability or a confidence interval, but to understand which assumptions made that calculation meaningful.

Audience and prerequisites

This packet is designed for a first undergraduate course or a mathematically ambitious secondary course. Students should be comfortable with algebra, functions, summation notation, and basic differentiation and integration. Multivariable calculus is not required. Proofs are included, but the emphasis is on modeling, interpretation, and the disciplined use of computation.

Learning goals

By the end, students should be able to:

1. translate verbal questions into sets, events, and probability models;
2. use counting, conditioning, independence, and Bayes' rule correctly;
3. move fluently among CDFs, probability masses, and densities;
4. compute and interpret expectation, variance, covariance, and correlation;
5. recognize standard distributions and justify an appropriate model;
6. obtain marginal and conditional distributions from a joint model;
7. explain the law of large numbers and central limit theorem through both mathematics and simulation;
8. evaluate estimators by bias, standard error, and mean squared error;
9. construct and interpret basic confidence intervals and hypothesis tests;
10. perform a simple Bayesian update and compare credible and confidence intervals.

A 14-week route

The design assumes two 75-minute meetings per week. “Lab” means that students make a prediction, run an experiment or simulation, and explain a discrepancy.

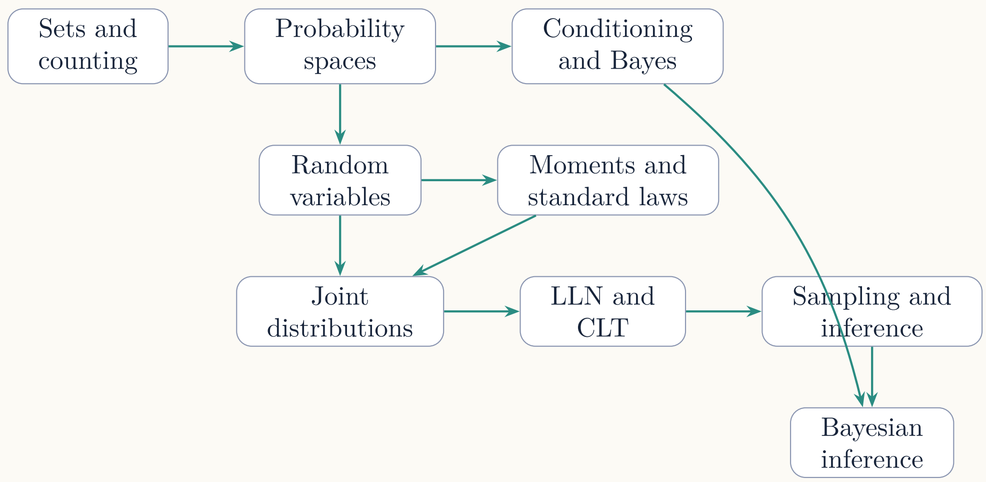
Week	Core mathematics	Suggested lab or assessment
1	Sets, events, Venn diagrams, counting	Inclusion–exclusion gallery
2	Probability spaces and axioms	Build and critique sample spaces
3	Conditional probability and trees	Medical-test simulation
4	Independence, total probability, Bayes	Monty Hall and base-rate lab
5	Random variables, CDFs, PMFs, densities	One distribution in three representations
6	Expectation, variance, moments	Indicator-variable studio
7	Bernoulli, binomial, geometric, Poisson	Rare-event approximation lab
8	Uniform, exponential, normal, beta	Memorylessness and normal areas
9	Joint, marginal, conditional laws	Heat maps and independence
10	Covariance, correlation, conditional expectation	Correlation counterexamples
11	Laws of large numbers; Monte Carlo	Running-average experiment
12	Central limit theorem; sampling distributions	Visual CLT; Problem Set 5
13	Estimation, confidence intervals, tests	Coverage and power simulation
14	Bayesian inference and synthesis	Prior-to-posterior capstone

Assessment design

One workable balance is 35% problem sets, 20% short quizzes, 20% computational labs, and 25% a modeling capstone. Exercises use three labels:

- **Fluency**: calculations and translations that make later reasoning automatic.
- **Reasoning**: derivations, short proofs, and careful interpretation.
- **Exploration**: simulation, visualization, model criticism, or open-ended comparison.

Dependency map



Checkpoint

A good opening prompt. A family has two children and at least one is a girl. What is the probability that both are girls? Ask students to defend not just an answer, but a sample space. Then change how the information was obtained and ask whether the answer changes.

Contents

Chapter 1

The grammar of uncertainty

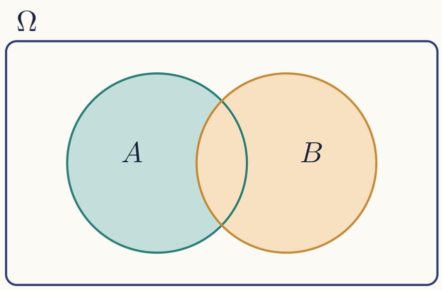
Sets, Venn diagrams, and counting

Guiding question 1.1. Before assigning probabilities, how do we describe precisely what could happen?

A *sample space* is built from possible outcomes; an *event* is a collection of outcomes. Set notation is therefore the grammar of probability.

1.1 Set operations

Let Ω be a universe of outcomes and $A, B \subseteq \Omega$. The union $A \cup B$ means "A or B (or both)," the intersection $A \cap B$ means "both," and the complement A^c means "not A." The difference $A \setminus B$ contains outcomes in A but not B.



Overlap represents $A \cap B$

De Morgan's laws translate negated statements:

$$(A \cup B)^c = A^c \cap B^c, \qquad (A \cap B)^c = A^c \cup B^c.$$

Common trap

In everyday speech, "or" is sometimes exclusive. In probability, $A \cup B$ is inclusive unless explicitly stated otherwise.

1.2 Products and sequences

The Cartesian product $A \times B$ contains ordered pairs (a, b). Repeated trials naturally live in product spaces. Three coin flips have sample space $\{H, T\}^3$, with eight ordered strings. Order matters: HTT and THT are different outcomes.

1.3 Counting without listing

If one choice has m possibilities and a subsequent choice has n, there are mn ordered outcomes. There are

$$n! = n(n-1)\cdots 1$$

orders of n distinct objects, and

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

ways to choose k objects without order.

Proposition 1.2 (Inclusion–exclusion)**.** *For finite sets A and B,*

$$|A \cup B| = |A| + |B| - |A \cap B|.$$

The subtraction repairs the double count of the overlap. For three sets, add the single sizes, subtract pairwise overlaps, then add the triple overlap back.

Example 1.3. A five-card poker hand is an unordered choice from 52 cards, so there are $\binom{52}{5}$ hands. Exactly two aces means choosing two of four aces and three of the remaining 48 cards:

$$\binom{4}{2}\binom{48}{3}.$$

Explore

Draw three overlapping circles for students who take mathematics, computing, and economics. Invent seven region counts, then recover the totals of each course and each pair. Reverse the task: start with totals and explain why the individual regions may not be uniquely determined.

Exercises

1. Write $(A \cap B^c) \cup (A^c \cap B)$ in words. [Fluency]
2. Verify both De Morgan laws using membership tables. [Reasoning]
3. How many eight-character strings contain exactly three letters and five digits? [Fluency]
4. Count five-card hands containing at least one ace. [Reasoning]
5. Prove $\binom{n}{k} = \binom{n}{n-k}$ by a counting argument. [Reasoning]
6. Among integers from 1 to 1,000, how many are divisible by 2 or 3 or 5? [Reasoning]

7. Design two different sampling procedures that produce the same set of possible outcomes but not the same probabilities. [Exploration]

Chapter 2

A measure of possibility

Probability spaces, events, and models

Guiding question 2.1. What rules must any coherent numerical description of uncertainty obey?

2.1 The axioms

Definition 2.2. A *probability space* is a triple $(\Omega, \mathcal{F}, \mathbb{P})$: a sample space Ω, a collection $\mathcal{F}$ of events closed under complements and countable unions, and a function $\mathbb{P} : \mathcal{F} \to [0, 1]$ satisfying

1. $\mathbb{P}(\Omega) = 1$;

2. if $A_1, A_2, \ldots$ are pairwise disjoint, then

$$\mathbb{P}\left(\bigcup_i A_i\right) = \sum_i \mathbb{P}(A_i).$$

For a finite course, $\mathcal{F}$ is usually all subsets of Ω. The more careful notation becomes important for continuous spaces.

Proposition 2.3. *For events A, B,*

$$\mathbb{P}(A^c) = 1 - \mathbb{P}(A), \qquad \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

If $A \subseteq B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$.

Each statement follows by splitting a set into disjoint pieces. This is the basic proof move in probability.

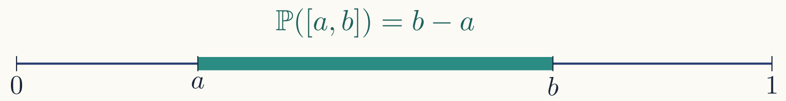
2.2 Equally likely is an assumption

If a finite sample space has N equally likely outcomes and A contains m, then $\mathbb{P}(A) = m/N$. But equality of probabilities is part of the model, not a consequence of counting.

Example 2.4. For two fair dice, the ordered-pair space has 36 equally likely outcomes. The possible sums $2, 3, \ldots, 12$ are not equally likely: sum 7 has six representations, while sum 2 has one.

2.3 Continuity and geometry

Choose a point uniformly from $[0, 1]$. Every individual point has probability zero, yet the entire interval has probability one. Probabilities attach to regions: $\mathbb{P}([a, b]) = b - a$. Zero probability does not mean impossible; it means negligible under the model.



2.4 Simulation and the model–world gap

A probability model is a deliberate simplification. Simulation samples from the model; data come from the world. Agreement between simulation and data supports a model, but never proves it uniquely correct.

Theorem 2.5 (Union bound)**.** *For any events $A_1, \ldots, A_n$,*

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \le \sum_{i=1}^{n} \mathbb{P}(A_i).$$

The events need not be disjoint. Double counting makes the right side larger, which is harmless for an upper bound.

Exercises

1. Derive $\mathbb{P}(\varnothing) = 0$ from the axioms. [Reasoning]
2. Two fair dice are rolled. Find the probability that the maximum is 4. [Fluency]
3. Prove the union bound for three events. [Reasoning]
4. A point is chosen uniformly from the unit square. Find the probability that $x + y < 1$. [Fluency]
5. Give an event with probability zero that is not logically impossible. [Reasoning]
6. Estimate by simulation the probability that two random points in $[0, 1]$ are within 0.1 of each other; then derive it geometrically. [Exploration]

Chapter 3

Information changes the space

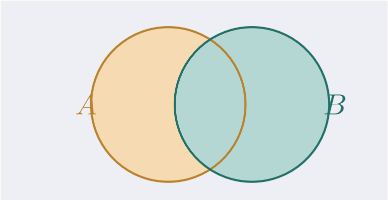
Conditional probability, independence, and Bayes' rule

Guiding question 3.1. When evidence arrives, which probabilities should change—and by how much?

Definition 3.2. If $\mathbb{P}(B) > 0$, the conditional probability of A given B is

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

Conditioning restricts attention to B and renormalizes so that B becomes the new whole.



Within B, measure the part also in A

3.1 Multiplication and trees

Rearranging the definition gives

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

Along a probability tree, multiply down branches and add across disjoint terminal paths.

3.2 Independence

Events A and B are independent when

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

When probabilities are positive, this is equivalent to $\mathbb{P}(A \mid B) = \mathbb{P}(A)$. Disjointness is almost the opposite: two nontrivial disjoint events cannot be independent.

Common trap

Pairwise independence does not imply mutual independence. Three events can be independent in every pair while the third is determined by the first two.

3.3 Total probability and Bayes

If $B_1, \dots, B_k$ partition Ω, then

$$\mathbb{P}(A) = \sum_{i=1}^{k} \mathbb{P}(A \mid B_i)\mathbb{P}(B_i).$$

Bayes' rule reverses the condition:

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

Example 3.3 (A diagnostic test)**.** Suppose prevalence is 1%, sensitivity is 95%, and specificity is 90%. Among 10,000 people, expect 100 cases, of which 95 test positive. Among 9,900 non-cases, 990 test positive. Thus

$$\mathbb{P}(\text{case} \mid +) = \frac{95}{95 + 990} \approx 0.0876.$$

A highly sensitive test can have a modest positive predictive value when the base rate is low.

Explore

Represent the test example with 10,000 dots rather than formulas. Change prevalence while holding sensitivity and specificity fixed. Plot the positive predictive value; locate the prevalence at which a positive result becomes more likely correct than incorrect.

Exercises

1. Draw a tree for two draws without replacement from an urn with three red and two blue balls. [Fluency]
2. Prove that if A and B are independent, then A^c and B are independent. [Reasoning]
3. Construct three pairwise independent events that are not mutually independent. [Reasoning]
4. Solve the diagnostic-test example with prevalence 5%. [Fluency]
5. A randomly selected coin is fair with probability .8 and double-headed with probability .2. It lands heads three times. Find the posterior probability it is double-headed. [Reasoning]
6. Explain precisely why changing the information protocol can change the “two children” answer. [Exploration]
7. Simulate Monty Hall and explain the result using conditioning rather than frequency alone. [Exploration]

Chapter 4

Numbers attached to outcomes

Random variables, distributions, expectation, and moments

Guiding question 4.1. How can one function compress a complicated experiment into a numerical quantity of interest?

Definition 4.2. A random variable is a function $X : \Omega \to \mathbb{R}$. Its cumulative distribution function is

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

Every CDF is nondecreasing, right-continuous, approaches 0 to the left and 1 to the right. For a discrete variable, jumps encode probabilities. For a continuous variable with density f, areas encode probabilities:

$$F_X(x) = \int_{-\infty}^{x} f(t)\,dt, \qquad \mathbb{P}(a < X \le b) = \int_a^b f(t)\,dt.$$

4.1 Expectation

For discrete X with values x_i,

$$\mathbb{E}[X] = \sum_i x_i \mathbb{P}(X = x_i).$$

For continuous X with density f,

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

Expectation is a center of mass, not necessarily a possible value.

Theorem 4.3 (LOTUS)**.** *For a suitable function g,*

$$\mathbb{E}[g(X)] = \sum_x g(x)\mathbb{P}(X = x) \quad \textit{or} \quad \mathbb{E}[g(X)] = \int g(x) f_X(x)\,dx.$$

Theorem 4.4 (Linearity)**.** *For constants a, b and integrable random variables X, Y,*

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

whether or not X and Y are independent.

4.2 Variance and moments

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

Standard deviation has the same units as X. The kth raw moment is $\mathbb{E}[X^k]$; the kth central moment is $\mathbb{E}[(X - \mathbb{E}X)^k]$. The moment-generating function, when finite near zero, is $M_X(t) = \mathbb{E}[e^{tX}]$ and satisfies $M_X^{(k)}(0) = \mathbb{E}[X^k]$.

Proposition 4.5. *For constants a, b,*

$$\operatorname{Var}(aX + b) = a^2 \operatorname{Var}(X).$$

4.3 Indicator variables

For an event A, let $\mathbf{1}_A = 1$ on A and 0 otherwise. Then $\mathbb{E}[\mathbf{1}_A] = \mathbb{P}(A)$. If $N = \sum_i \mathbf{1}_{A_i}$ counts occurrences, linearity gives

$$\mathbb{E}[N] = \sum_i \mathbb{P}(A_i)$$

without independence.

Theorem 4.6 (Markov and Chebyshev)**.** *If $X \geq 0$, then $\mathbb{P}(X \geq a) \leq \mathbb{E}[X]/a$. Consequently,*

$$\mathbb{P}(|X - \mathbb{E}X| \geq k\sigma) \leq \frac{1}{k^2}.$$

Exercises

1. Find the CDF of the number of heads in two fair flips. [Fluency]
2. A density is proportional to x on $[0, 2]$. Normalize it and find its mean. [Fluency]
3. Prove $\operatorname{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}X)^2$. [Reasoning]
4. Find the expected number of fixed points in a random permutation of n objects. [Reasoning]
5. Show that a fair lottery can still have a large variance. Interpret. [Exploration]
6. Use Chebyshev's inequality to bound the chance that a variable lies at least three standard deviations from its mean. Compare with the normal distribution. [Reasoning]
7. Construct two different distributions with the same mean and variance. [Exploration]

Chapter 5

A library of chance mechanisms

Standard discrete and continuous distributions

Guiding question 5.1. Which probability laws recur because the same mechanisms recur?

5.1 Discrete families

Law	Mechanism and mass function	Mean	Variance
Bernoulli(p)	one success indicator; $p^x(1-p)^{1-x}$	p	$p(1-p)$
Binomial(n,p)	successes in n independent trials; $\binom{n}{k}p^k(1-p)^{n-k}$	np	$np(1-p)$
Geometric(p)	trials through first success; $(1-p)^{k-1}p$	$1/p$	$(1-p)/p^2$
Poisson(λ)	rare-event count; $e^{-\lambda}\lambda^k/k!$	λ	λ

The geometric law is memoryless:

$$\mathbb{P}(X > s+t \mid X > s) = \mathbb{P}(X > t).$$

The Poisson law arises as a limit of $\mathrm{Bin}(n,p)$ when n is large, p is small, and $np \approx \lambda$.

5.2 Continuous families

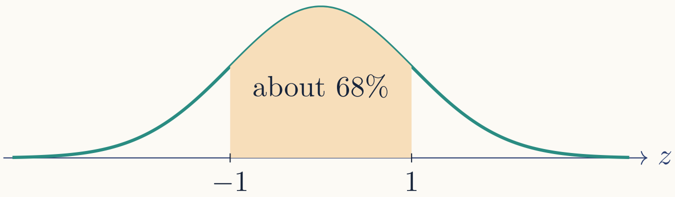
Uniform(a,b) assigns constant density $1/(b-a)$ to an interval. Exponential(λ) has density $\lambda e^{-\lambda x}$ on $x \geq 0$ and is the continuous memoryless law.

The normal density with mean μ and variance σ^2 is

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

Standardization sends $X \sim \mathcal{N}(\mu, \sigma^2)$ to

$$Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1).$$



The beta family on $[0,1]$ has density proportional to $x^{\alpha-1}(1-x)^{\beta-1}$. It is useful for proportions and becomes a natural prior for a Bernoulli probability.

5.3 Choose by mechanism, not by silhouette

A bell-shaped histogram does not prove normality. Ask what generated the observations, what values are possible, whether trials are independent, whether rates are stable, and what approximation error is acceptable.

Exercises

1. Compute $\mathbb{P}(X \geq 2)$ for $X \sim \mathrm{Bin}(5, .3)$. [Fluency]
2. Prove the geometric memoryless property. [Reasoning]
3. Approximate $\mathrm{Bin}(1000, .002)$ by a Poisson law and compare $\mathbb{P}(X = 0)$. [Exploration]
4. Find the median of an exponential distribution. [Fluency]
5. If $X \sim \mathcal{N}(10, 9)$, express $\mathbb{P}(8 < X < 13)$ using the standard normal CDF. [Fluency]
6. Derive the mean and variance of $\mathrm{Uniform}(a, b)$. [Reasoning]
7. For each family in the chapter, invent one plausible and one implausible real-world use. [Exploration]

Chapter 6

Variables together

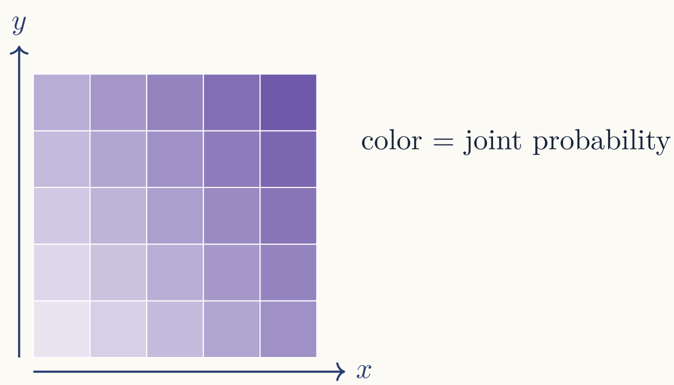
Joint, marginal, and conditional distributions

Guiding question 6.1. How do we describe dependence without confusing association, causation, and prediction?

For discrete variables, a joint mass function $p(x, y) = \mathbb{P}(X = x, Y = y)$ is a table whose entries sum to one. Marginals are obtained by summing:

$$p_X(x) = \sum_y p(x, y), \qquad p_Y(y) = \sum_x p(x, y).$$

For a continuous joint density $f(x, y)$, replace sums with integrals.



6.1 Conditional laws and independence

For $p_Y(y) > 0$,

$$p_{X|Y}(x \mid y) = \frac{p(x, y)}{p_Y(y)}.$$

Variables are independent when their joint law factors:

$$p(x, y) = p_X(x)p_Y(y)$$

or $f(x, y) = f_X(x)f_Y(y)$ throughout the support.

6.2 Covariance and correlation

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

Correlation is the dimensionless quantity

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}.$$

It lies between -1 and 1. Independence implies zero covariance when moments exist, but zero covariance does not generally imply independence.

Theorem 6.2.

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y).$$

For independent variables, the covariance term vanishes.

6.3 Conditional expectation

$\mathbb{E}[X \mid Y]$ is itself a random variable: for each observed value of Y, it gives the mean of the corresponding conditional distribution. The tower property says

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

The conditional mean is also the best mean-square predictor of X using only Y.

Common trap

Correlation measures linear association. A perfect U-shaped relationship can have correlation zero. And neither correlation nor conditioning alone establishes causation.

Exercises

1. Given a 2×3 joint probability table, compute both marginals and one conditional distribution. [Fluency]
2. Prove that independence implies $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$. [Reasoning]
3. Let X be uniform on $\{-1, 0, 1\}$ and $Y = X^2$. Show $\mathrm{Cov}(X,Y) = 0$ although X, Y are dependent. [Reasoning]
4. Derive the variance-of-a-sum identity. [Reasoning]
5. If X, Y are independent Poisson variables, use generating functions to show $X + Y$ is Poisson. [Reasoning]
6. Create three scatterplots with correlations near -1, 0, and 1, then create a nonlinear plot with correlation near zero. [Exploration]
7. Explain the tower property using a population divided into groups. [Exploration]

Chapter 7

From repetition to regularity

Laws of large numbers, the central limit theorem, and sampling

Guiding question 7.1. Why do averages stabilize, and why do bell curves emerge from non-normal data?

Let $X_1, \ldots, X_n$ be independent copies of a variable with mean μ and variance σ^2. Their average is

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \mathbb{E}[\bar{X}_n] = \mu, \qquad \mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n}.$$

Theorem 7.2 (Weak law of large numbers)**.** *For every $\varepsilon > 0$,*

$$\mathbb{P}(|\bar{X}_n - \mu| > \varepsilon) \longrightarrow 0.$$

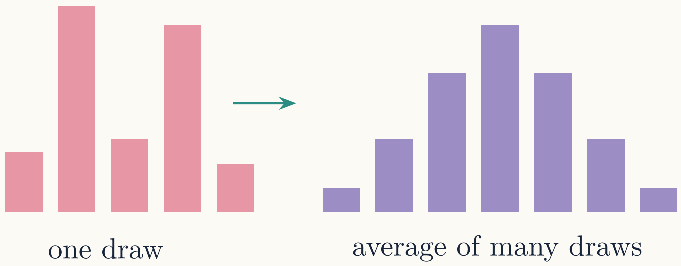
With finite variance, Chebyshev gives the proof immediately:

$$\mathbb{P}(|\bar{X}_n - \mu| > \varepsilon) \le \frac{\sigma^2}{n\varepsilon^2}.$$

Theorem 7.3 (Central limit theorem)**.** *If $0 < \sigma^2 < \infty$, then*

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} \mathcal{N}(0, 1).$$

The LLN says where the average goes; the CLT describes the scale and approximate shape of its remaining fluctuations.



7.1 Sampling distributions

A *statistic* is a function of the sample. Before data are observed it is random, and its probability law is its sampling distribution. The standard deviation of that distribution is the *standard error*.

For a Bernoulli sample, the sample proportion $\hat{p}$ has

$$\mathbb{E}[\hat{p}] = p, qquad \, \mathrm{SE}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}.$$

7.2 Monte Carlo and bootstrap

Monte Carlo approximates an expectation by simulating from a known model. The bootstrap approximates a sampling distribution by repeatedly resampling, with replacement, from observed data. The bootstrap is powerful but cannot repair a badly biased or unrepresentative original sample.

Exercises

1. Compute the standard error of the mean when $\sigma = 12$ and $n = 64$. [Fluency]
2. Prove the weak law using Chebyshev's inequality. [Reasoning]
3. Approximate $\mathbb{P}(480 \leq X \leq 520)$ for $X \sim \mathrm{Bin}(1000, .5)$ using the CLT and a continuity correction. [Fluency]
4. Explain why quadrupling the sample size halves a standard error. [Reasoning]
5. Simulate sample means from a skewed distribution for $n = 1, 2, 5, 30$. Quantify convergence rather than only describing the pictures. [Exploration]
6. Compare Monte Carlo error at $n = 100$, 10^4, and 10^6. Is the $n^{-1/2}$ rate visible? [Exploration]
7. Give an example where bootstrap resampling cannot reveal a sampling bias. [Reasoning]

Chapter 8

Learning from a sample

Estimation, confidence intervals, and hypothesis tests

Guiding question 8.1. How can a random sample support a calibrated statement about an unknown population?

An estimator $T = T(X_1, \ldots, X_n)$ targets a parameter θ. Its bias, variance, and mean squared error are

$$\mathrm{Bias}_\theta(T) = \mathbb{E}_\theta[T] - \theta, \qquad \mathrm{MSE}_\theta(T) = \mathrm{Var}_\theta(T) + \mathrm{Bias}_\theta(T)^2.$$

Unbiasedness is attractive, but MSE captures the bias–variance tradeoff.

8.1 Method of moments and likelihood

The method of moments matches sample moments to model moments. Maximum likelihood chooses the parameter under which the observed data have greatest likelihood.

For Bernoulli observations $x_1, \ldots, x_n$, the likelihood is

$$L(p) = p^{\sum x_i}(1-p)^{n-\sum x_i},$$

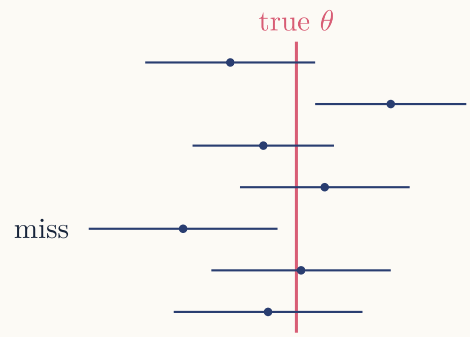
and the maximizer is $\hat{p} = \bar{x}$.

8.2 Confidence intervals

An approximate 95% interval for a mean with known or well-estimated standard error is

$$\text{estimate} \ \pm \ 1.96 \times \text{standard error}.$$

The confidence level refers to the long-run success rate of the procedure, not a posterior probability that a fixed parameter lies in one realized interval.



8.3 Tests and p-values

A hypothesis test specifies a null model H_0, a test statistic, and a rejection rule. The p-value is the probability, assuming H_0, of a result at least as incompatible with H_0 as the observed one.

It is not the probability that H_0 is true. A Type I error rejects a true null; a Type II error fails to reject a false null. Power is the probability of rejecting under a specified alternative.

Common trap

Statistical significance is not practical importance. With enough data, a tiny effect can yield a tiny p-value; with too little data, an important effect can be missed.

8.4 Model checking and design

Random sampling supports generalization; random assignment supports causal comparison. No inferential formula compensates for selection bias, poorly measured outcomes, hidden multiplicity, or a model contradicted by the data.

Exercises

1. Show that the sample mean is unbiased for μ. [Reasoning]
2. Find the Bernoulli maximum-likelihood estimator. [Fluency]
3. Compare two estimators when one has small bias but much smaller variance. [Reasoning]
4. Construct a 95% normal interval for a proportion from 600 observations with 330 successes. [Fluency]
5. Explain the repeated-sampling meaning of that interval without saying there is a 95% chance the fixed parameter is inside. [Reasoning]
6. A test reports $p = .03$. List three claims that do not follow. [Reasoning]
7. Simulate confidence-interval coverage and test power as functions of sample size. [Exploration]
8. Describe how optional stopping or multiple testing can inflate false discoveries. [Exploration]

Chapter 9

Probability for unknowns

An introduction to Bayesian inference

Guiding question 9.1. What changes when an unknown parameter is itself represented by a probability distribution?

Bayesian inference combines a prior distribution $\pi(\theta)$ with a likelihood $p(y \mid \theta)$. Bayes' rule produces the posterior

$$\pi(\theta \mid y) = \frac{p(y \mid \theta)\pi(\theta)}{\int p(y \mid u)\pi(u)\,du} \propto p(y \mid \theta)\pi(\theta).$$

9.1 Beta–binomial updating

Suppose $p \sim \mathrm{Beta}(\alpha, \beta)$ and, conditional on p, $X \sim \mathrm{Bin}(n, p)$. After observing x successes,

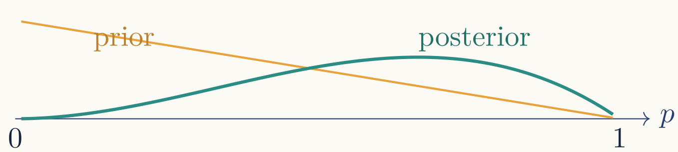
$$p \mid X = x \sim \mathrm{Beta}(\alpha + x, \beta + n - x).$$

The prior parameters behave like pseudocounts: $\alpha - 1$ prior successes and $\beta - 1$ prior failures when both exceed one.

The posterior mean is

$$\mathbb{E}[p \mid x] = \frac{\alpha + x}{\alpha + \beta + n},$$

a weighted compromise between the prior mean and the sample proportion.



9.2 Credible intervals and prediction

A 95% posterior credible interval contains 95% of the posterior probability. This is a conditional probability statement about the parameter under the model and prior. A posterior predictive distribution averages future outcomes over posterior uncertainty:

$$p(y_{\text{new}} \mid y) = \int p(y_{\text{new}} \mid \theta)\pi(\theta \mid y)\, d\theta.$$

9.3 Prior sensitivity and checking

Priors make assumptions visible; they do not make assumptions disappear. Compare reasonable priors, examine posterior predictive simulations, and ask whether the fitted model reproduces important features of the data.

Big idea

Frequentist and Bayesian intervals answer different probability questions. The productive comparison is not "which formula looks similar?" but "what is random, what is conditioned upon, and what operating guarantee or posterior statement is being claimed?"

Exercises

1. Update a $\text{Beta}(2, 2)$ prior after 8 successes in 10 trials. [Fluency]
2. Compute the posterior mean and compare it with the sample proportion. [Fluency]
3. Derive beta–binomial conjugacy by collecting powers of p and $1 - p$. [Reasoning]
4. Under a $\text{Beta}(\alpha, \beta)$ posterior, find the predictive probability that the next Bernoulli trial succeeds. [Reasoning]
5. Compare $\text{Beta}(1, 1)$, $\text{Beta}(2, 2)$, and $\text{Beta}(20, 20)$ priors for the same small dataset. [Exploration]
6. Explain why a credible interval and confidence interval can have the same numerical endpoints but different interpretations. [Reasoning]
7. Design a posterior predictive check for a coin model that might have time-varying bias. [Exploration]

Capstone: one question, two inferential languages

Choose a binomial, count, or measurement dataset. State a generative model, inspect the data, estimate the target, and quantify uncertainty both frequentistically and Bayesianly. End with a model check and a paragraph explaining which conclusions are robust to the modeling choices.

Appendix A

Selected hints and solutions

Chapter 1

For the at-least-one-ace problem, count the complement: $\binom{52}{5} - \binom{48}{5}$. For divisibility by 2, 3, or 5, inclusion–exclusion gives

$$500 + 333 + 200 - 166 - 100 - 66 + 33 = 734.$$

Chapter 2

For the maximum of two dice to equal 4, both dice must be at most 4 but not both at most 3, giving $(4^2 - 3^2)/36 = 7/36$. For two uniform points X, Y, the band $|X - Y| < .1$ has area $1 - 2(.9)^2/2 = .19$.

Chapter 3

For the coin problem,

$$\mathbb{P}(D \mid HHH) = \frac{1 \cdot .2}{1 \cdot .2 + (1/2)^3 \cdot .8} = \frac{2}{3}.$$

A pairwise-not-mutual example: flip two fair coins; let A be first-heads, B second-heads, and C equal-results.

Chapter 4

For fixed points, write $N = \sum_{i=1}^{n} I_i$. Since $\mathbb{P}(I_i = 1) = 1/n$, linearity gives $\mathbb{E}N = 1$. The variance identity follows by expanding $(X - \mu)^2$.

Chapter 5

For an exponential variable, $\mathbb{P}(X > m) = e^{-\lambda m} = 1/2$, hence $m = (\log 2)/\lambda$. For $X \sim \mathcal{N}(10, 9)$,

$$\mathbb{P}(8 < X < 13) = \Phi(1) - \Phi(-2/3).$$

Chapter 6

For X uniform on $\{-1, 0, 1\}$ and $Y = X^2$, symmetry gives $\mathbb{E}X = \mathbb{E}[XY] = 0$, hence covariance zero, but knowing X determines Y. For the tower property, take the weighted average of group means; each observation receives exactly its population weight.

Chapter 7

If $\sigma = 12, n = 64$, then $\mathrm{SE}(\bar{X}) = 12/8 = 1.5$. The weak law follows from $\mathrm{Var}(\bar{X}_n) = \sigma^2/n$ and Chebyshev.

Chapter 8

For 330 successes in 600 trials, $\hat{p} = .55$ and the plug-in standard error is

$$\sqrt{.55(.45)/600} \approx .0203.$$

The approximate 95% interval is $.55 \pm 1.96(.0203)$, or about $(.510, .590)$.

Chapter 9

Beta$(2, 2)$ updated with 8 successes and 2 failures becomes Beta$(10, 4)$, with posterior mean $10/14 \approx .714$. The posterior predictive success probability equals the posterior mean $\alpha'/(\alpha' + \beta')$.

Appendix B

Reference sheets

Core probability identities

$$
\begin{aligned}
\mathbb{P}(A^c) &= 1 - \mathbb{P}(A), \\
\mathbb{P}(A \cup B) &= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B), \\
\mathbb{P}(A \mid B) &= \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, \\
\mathbb{P}(A) &= \sum_i \mathbb{P}(A \mid B_i)\mathbb{P}(B_i), \\
\mathbb{P}(B_j \mid A) &= \frac{\mathbb{P}(A \mid B_j)\mathbb{P}(B_j)}{\sum_i \mathbb{P}(A \mid B_i)\mathbb{P}(B_i)}.
\end{aligned}
$$

Moments and dependence

$$
\begin{aligned}
\mathbb{E}[aX + bY] &= a\mathbb{E}X + b\mathbb{E}Y, \\
\operatorname{Var}(X) &= \mathbb{E}[X^2] - (\mathbb{E}X)^2, \\
\operatorname{Cov}(X, Y) &= \mathbb{E}[XY] - \mathbb{E}X\mathbb{E}Y, \\
\operatorname{Var}(X + Y) &= \operatorname{Var} X + \operatorname{Var} Y + 2\operatorname{Cov}(X, Y), \\
\rho_{X,Y} &= \operatorname{Cov}(X, Y)/(\sigma_X \sigma_Y).
\end{aligned}
$$

Common distributions

Distribution	Support	Mean	Variance
Bernoulli(p)	$\{0, 1\}$	p	$p(1-p)$
Binomial(n, p)	$0, \ldots, n$	np	$np(1-p)$
Geometric(p)	$1, 2, \ldots$	$1/p$	$(1-p)/p^2$
Poisson(λ)	$0, 1, \ldots$	λ	λ
Uniform(a, b)	$[a, b]$	$(a+b)/2$	$(b-a)^2/12$
Exponential(λ)	$[0, \infty)$	$1/\lambda$	$1/\lambda^2$
Normal(μ, σ^2)	$\mathbb{R}$	μ	σ^2
Beta(α, β)	$[0, 1]$	$\alpha/(\alpha+\beta)$	$\alpha\beta/[(\alpha+\beta)^2(\alpha+\beta+1)]$

Inference checklist

1. What is the target population and parameter?

2. How were units sampled or assigned?

3. What model connects observations to the parameter?

4. What is the estimator and its sampling uncertainty?

5. Which assumptions are needed for the interval or test?

6. Is the effect practically meaningful?

7. What diagnostic or predictive check could reveal model failure?

Further reading

- D. Freedman, R. Pisani, and R. Purves, *Statistics*.
- J. Blitzstein and J. Hwang, *Introduction to Probability*.
- L. Wasserman, *All of Statistics* (for a concise next course).
- A. Gelman et al., *Bayesian Data Analysis* (for a more advanced Bayesian treatment).