STAD91 Bayesian Statistics

Lecture 1: Organization & Review

Thibault Randrianarisoa

UTSC

January 8, 2026

Computer & Mathematical Sciences
UNIVERSITY OF TORONTO
SCARBOROUGH

Overview of the lecture

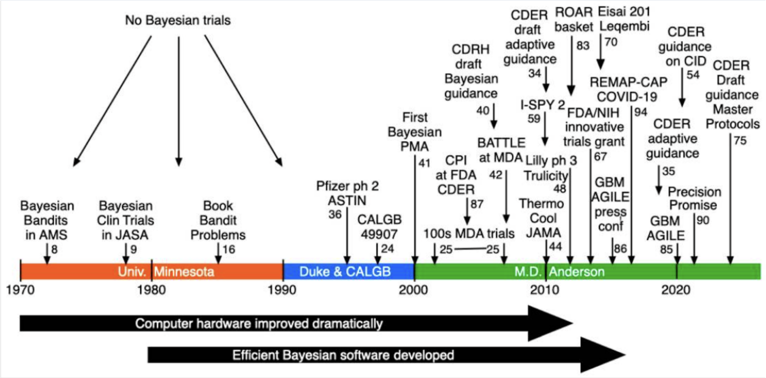
- First part: Organization of the course
  - Motivation
  - Course logistics
  - Assessment

- Second part: Background from Probability and Statistics

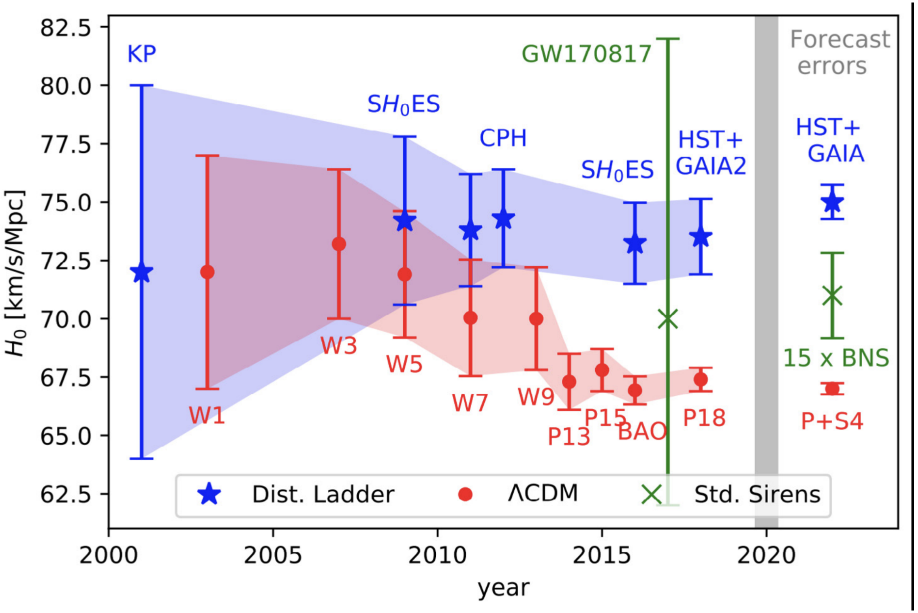
World War II

Search for plane wreck

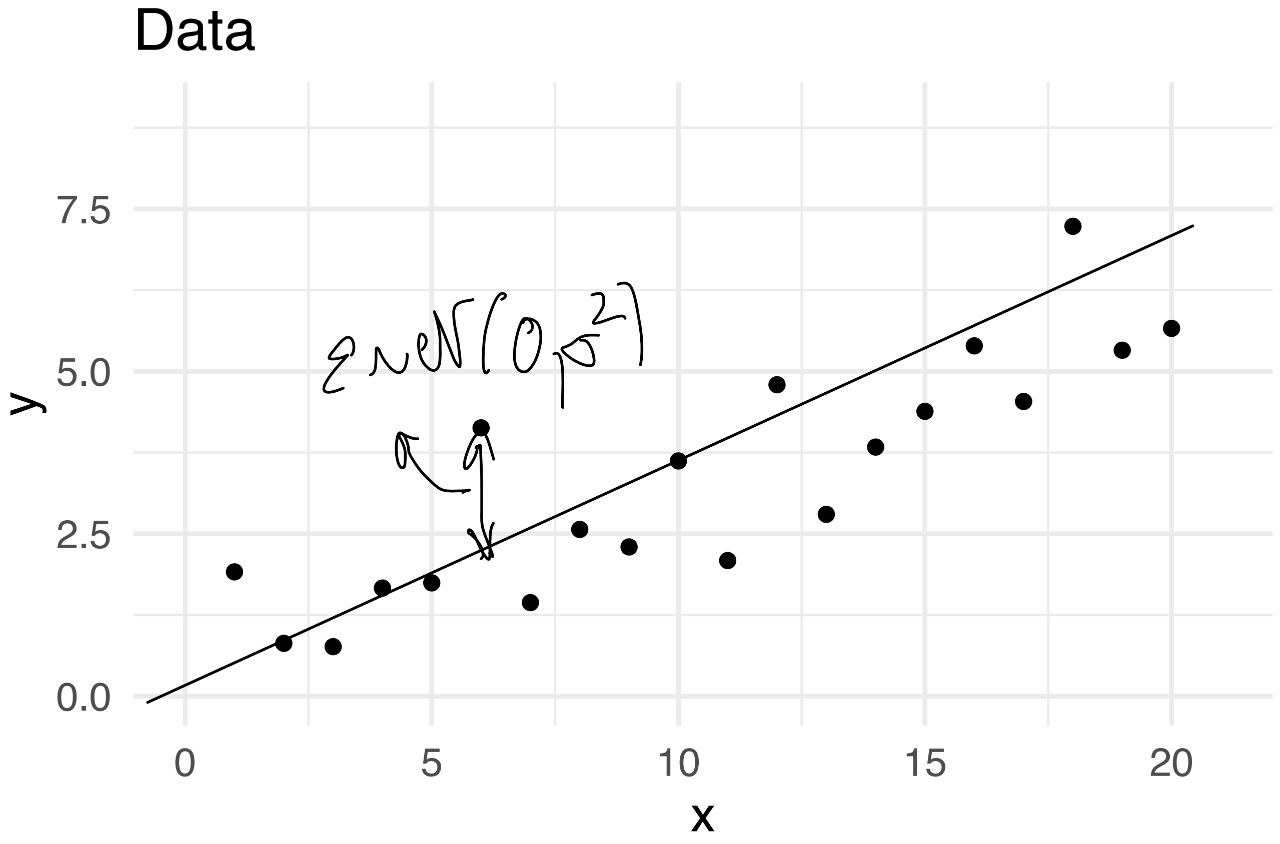
Science



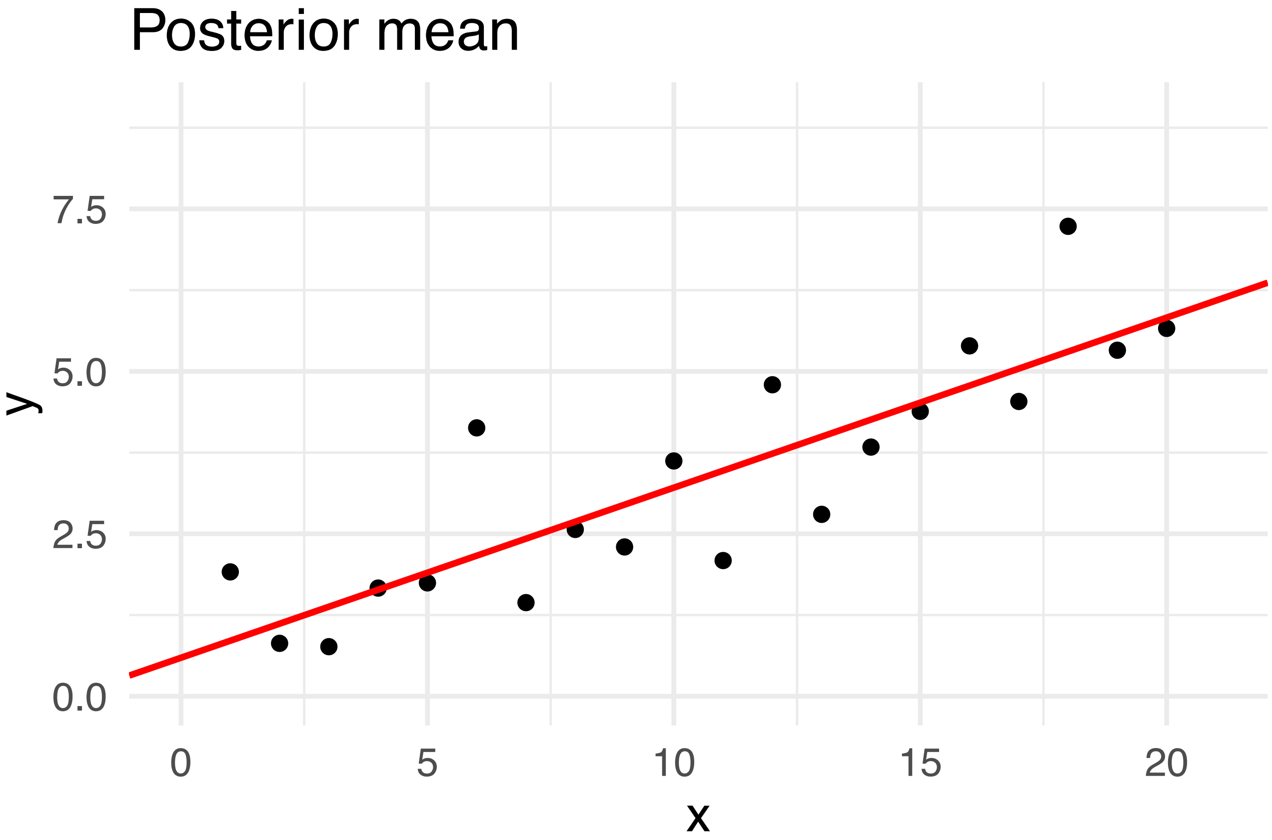
D. Berry, *Adaptive Bayesian Clinical Trials: The Past, Present, and Future of Clinical Research*, 2025



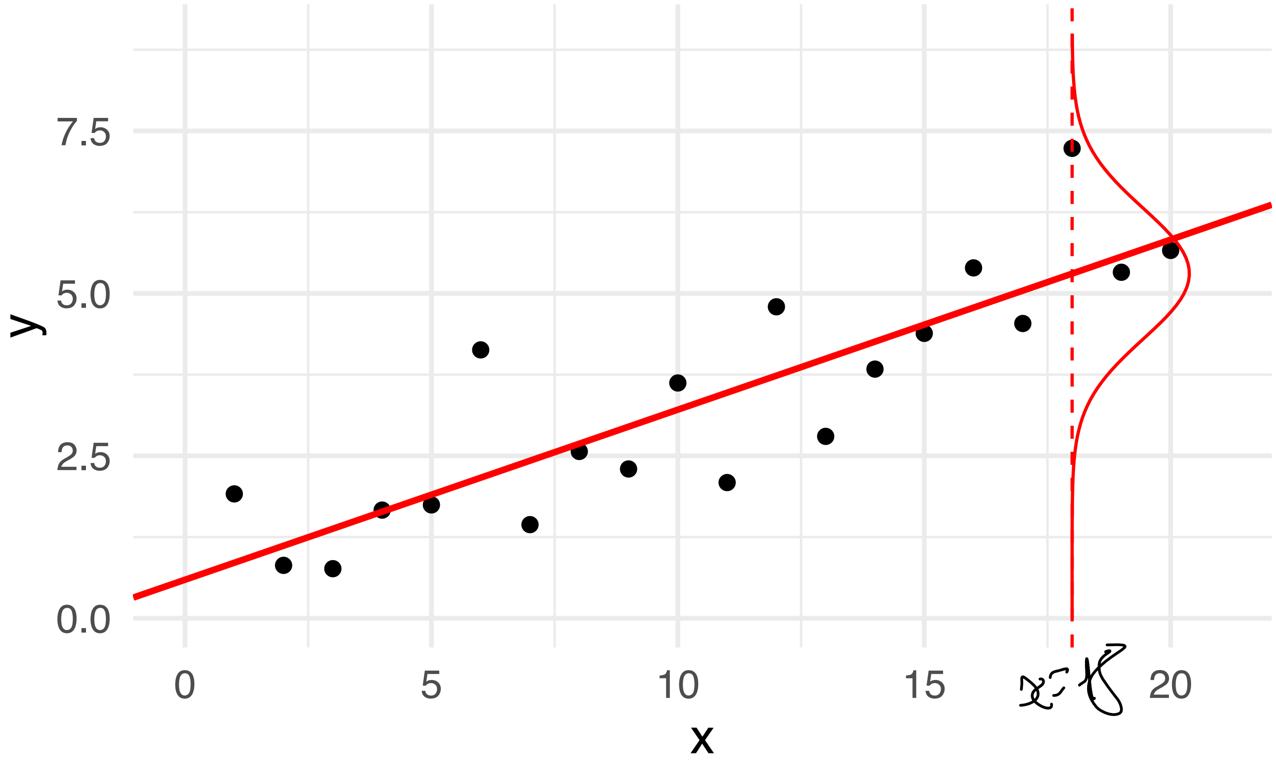
Uncertainty in modeling



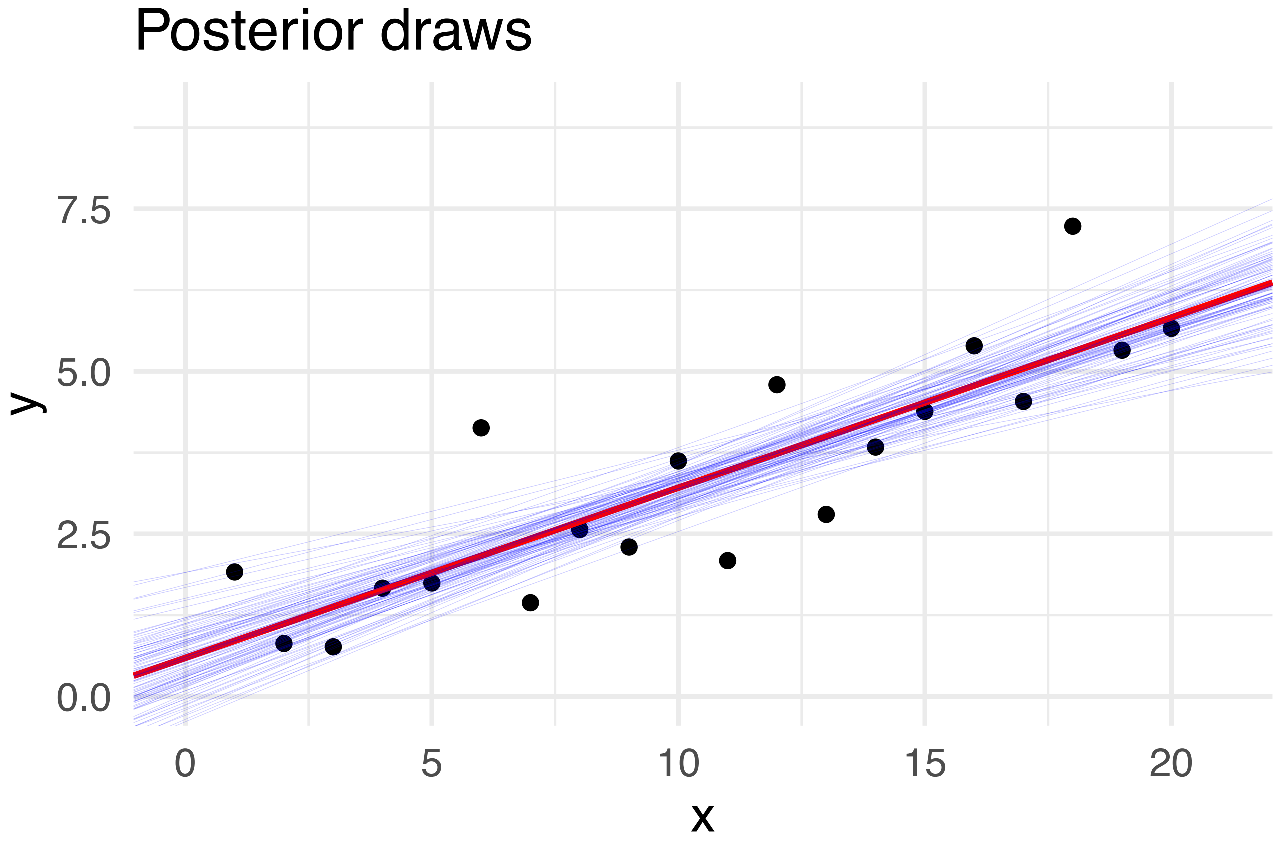
Uncertainty in modeling

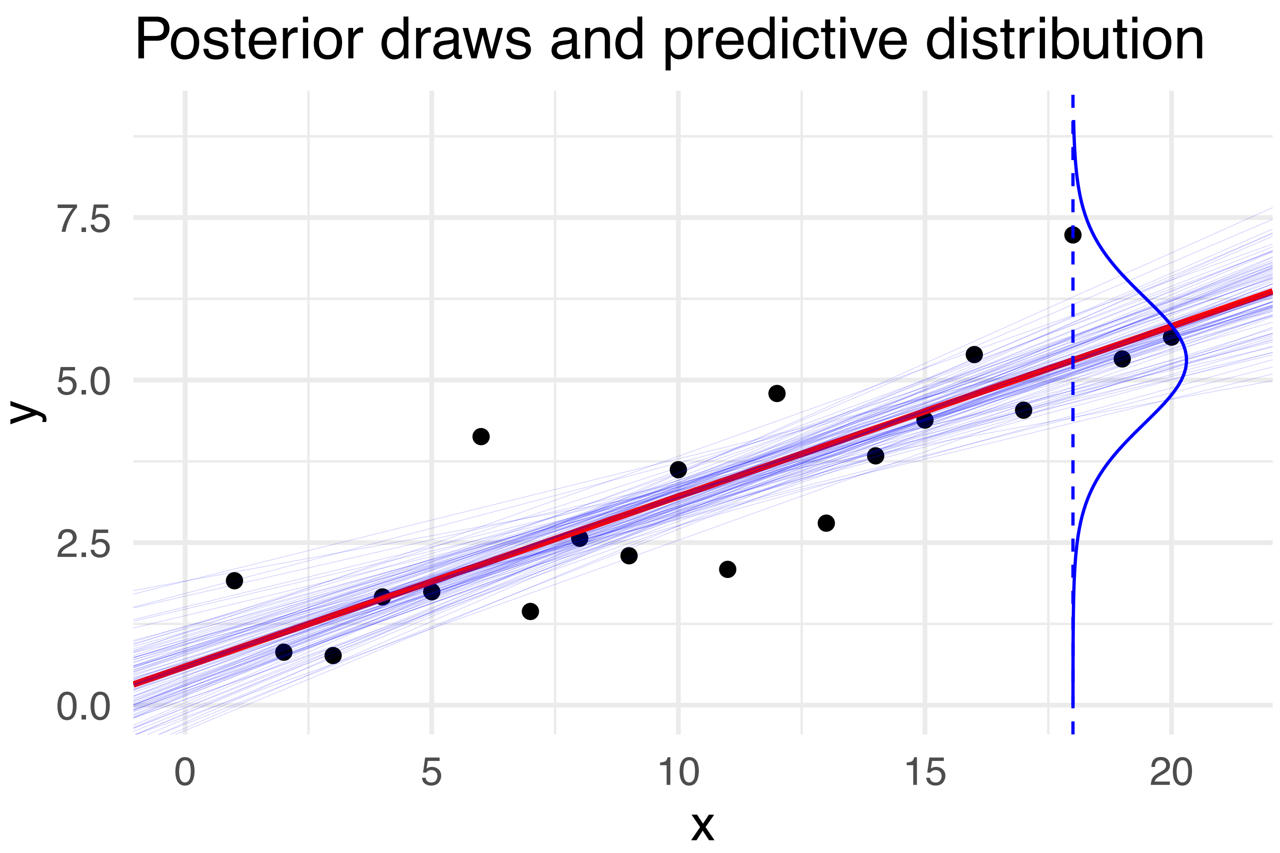


Predictive distribution given posterior mean



Uncertainty in modeling




Posterior draws and predictive distribution
7.5
5.0
2.5
0.0
y
0
5
10
15
20
x

Uncertainty and probabilistic modeling

- Representing uncertainty with probabilities + Updating uncertainty
- Two types of uncertainty: aleatoric and epistemic
- Aleatoric uncertainty due to randomness
  - we are not able to obtain observations which could reduce this uncertainty
- Epistemic uncertainty due to lack of knowledge
  - we are able to obtain observations which can reduce this uncertainty
  - two observers may have different epistemic uncertainty

Impact on society

Better modelling and quantification of uncertainty

$\rightarrow$ better science

$\rightarrow$ better informed decision making
in companies, government, and NGOs

Bayesian probability theory

expert information, previous experiments,...

Bayesian probability theory

expert information, previous experiments,...

$\searrow$

mathematical model

$+$

uncertainty with probabilities

Bayesian probability theory

expert information, previous experiments,... data

↘

mathematical model

+

uncertainty with probabilities

Bayesian probability theory

expert information, previous experiments,... ↘ | data ↙

mathematical model

+

uncertainty with probabilities

+

Bayesian probability theory

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

posterior distribution

$$p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)d\theta$$

Bayesian probability theory

expert information, previous experiments,... $\searrow$

mathematical model
$+$
uncertainty with probabilities

$+$

data $\swarrow$

Bayesian probability theory

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

$$p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)d\theta$$

$\downarrow$

updated uncertainty

Bayesian probability theory

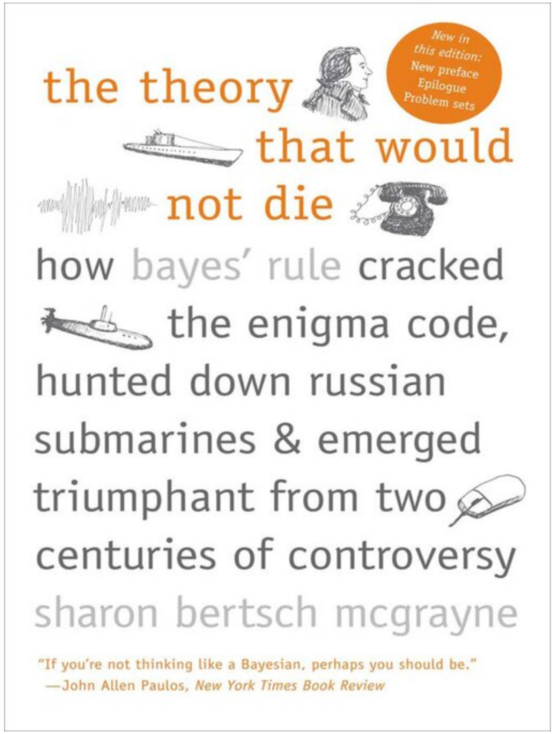
- Based on Bayesian probability theory
  - uncertainty is presented with probabilities
  - probabilities are updated based on new information
- Thomas Bayes (170?–1761)
  - English nonconformist, Presbyterian minister, (amateur) mathematician
  - considered the problem of *inverse probability*

→ simple problem only

- Bayes did not invent all, but was first to solve problem of inverse probability in special case
- Laplace generalized the initial methods and applied it to scientific problems (e.g., astronomy)
- Modern Bayesian theory with rigorous proofs developed in 20th century

Bayesian probability theory

A nice book about history: Sharon Bertsch McGrayne, *The Theory That Would Not Die*, 2012.



Term Bayesian used first time in mid 20th century

- Earlier there was just "probability theory"
  - concept of the probability was not strictly defined, although it was close to modern Bayesian interpretation
  - in the end of 19th century there were increasing demand for more strict definition of probability (mathematical and philosophical problem) → Kolmogorov
- In the beginning of 20th century frequentist view gained popularity
  - accepts definition of probabilities only through frequencies
  - does not accept inverse probability or use of prior
  - gained popularity due to apparent objectivity and "cook book" like reference books
- R. A. Fisher used in 1950 first time term "Bayesian" to emphasize the difference to general term "probability theory"
  - term became quickly popular, because alternative descriptions were longer
- The probabilistic programming revolution started in early 1990's

Bayesian Statistics course

- Probability distributions as model building blocks
  - need to understand the math part (prereq.)
  - continuous vs discrete (prereq.)
  - observation model, likelihood, prior
  - constructing bigger models
- Computation
  - We need to be able to compute expectations

marginalize, conditioning

$$E_{\theta|y}[g(\theta)] = \int p(\theta|y)g(\theta)d\theta$$

  - when analytic solutions are not available, computational approximations with finite number of function evaluations
  - importance sampling, Monte Carlo, Markov chain Monte Carlo, variational Bayes
- **Not in this course:** Diagnostics

Bayesian Statistics course

- *Bayesian inference* : process of statistical learning via Bayes' rule.

$$P(A|E) = \frac{P(E|A)P(A)}{P(E|A)P(A) + P(E|A^c)P(A^c)} = \frac{P(E|A)P(A)}{P(E)}$$

or

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)}$$

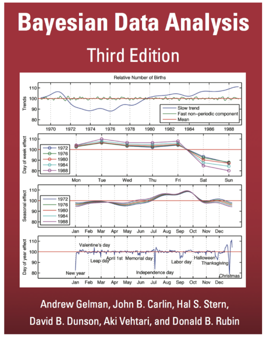
- *Bayesian methods* are data analysis tools that are derived from the principles of Bayesian inference.

- Bayesian methods provide:
  - a rational method for updating beliefs in light of new information;
  - parameter estimates with good statistical properties;
  - predictions for missing data and forecasts of future data;
  - a computational framework for model estimation, selection and validation.

Course Logistics

- Course administered through Quercus
  - Syllabus, Lecture Notes, HW Problems, Quizzes, etc
  - Also look at the course webpage

- Textbook: Bayesian Data Analysis by Gelman, Carlin, Stern, Dunson, Vehtari & Rubin.



- Communication
  - For course content questions use Piazza or OH
  - For personal issues use email (with [STAD91] in the subject) or office hours (TH 10am-1pm @ IA 4064)

Assessment

Evaluation	Weight	Details
Weekly Quizzes (on Kahoot)	20%	• Best 8/10 • Cover previous week's material • At the end of the lecture
Homework Assignments	10%	• Two Homework Assignments (5% each) • Date: After the midterm, TBD • Pen & paper derivations + Coding (Python/Numpy or R)
Term Test	25%	• Covers first 5 weeks • Tentative Date: Feb 26 ~~(wk 10)~~
Final	45%	• Cumulative • Final exam period

No make-up Quizzes/Term Tests; if you miss midterm you must **declare absence on ACORN** & mark will be replaced by final.

Review of Key Concepts

Probability & inference topics you are expected to remember (quick recap)

Definition (absolute continuity)

Let P and μ be two σ-finite measures on a measurable space $(E, \mathcal{E})$. We say that P is absolutely continuous with respect to μ, and write $P \ll \mu$, if

$$\forall A \in \mathcal{E}, \quad \mu(A) = 0 \implies P(A) = 0.$$

Absolute continuity

Definition (absolute continuity)

Let P and μ be two σ-finite measures on a measurable space $(E, \mathcal{E})$. We say that P is absolutely continuous with respect to μ, and write $P \ll \mu$, if

$$\forall A \in \mathcal{E}, \quad \mu(A) = 0 \implies P(A) = 0.$$

Radon–Nikodym theorem

If $P \ll \mu$, then there exists a positive measurable function p such that for every $A \in \mathcal{E}$,

$$P(A) = \int_A p(x)\,\mathrm{d}\mu(x).$$

The function p is called the *Radon-Nikodym derivative* of P with respect to μ and is denoted

$$p = \frac{\mathrm{d}P}{\mathrm{d}\mu}.$$

Notation

One may write, suggestively,

$$P(A) = \int_A \mathrm{d}P(x) = \int_A \frac{\mathrm{d}P(x)}{\mathrm{d}\mu(x)}\,\mathrm{d}\mu(x) = \int_A p(x)\,\mathrm{d}\mu(x).$$

Discrete distributions

- On $E = \{0, 1\}$, the Bernoulli(θ) law has a density with respect to $\mu = \delta_0 + \delta_1$:
$$p(x) = (1 - \theta)\mathbf{1}_{\{x=0\}} + \theta\mathbf{1}_{\{x=1\}}.$$

Notation

One may write, suggestively,

$$P(A) = \int_A \mathrm{d}P(x) = \int_A \frac{\mathrm{d}P(x)}{\mathrm{d}\mu(x)}\,\mathrm{d}\mu(x) = \int_A p(x)\,\mathrm{d}\mu(x).$$

Discrete distributions

- On $E = \{0,1\}$, the Bernoulli(θ) law has a density with respect to $\mu = \delta_0 + \delta_1$:
$$p(x) = (1-\theta)\mathbf{1}_{\{x=0\}} + \theta\mathbf{1}_{\{x=1\}}.$$
- On $E = \{0,1,\ldots,n\}$, the Binomial(n,θ) law is absolutely continuous with respect to the counting measure $\mu = \sum_{k=0}^n \delta_k$, with density
$$p(k) = \binom{n}{k}\theta^k(1-\theta)^{n-k}, \qquad k = 0,1,\ldots,n.$$

Notation

One may write, suggestively,

$$P(A) = \int_A \mathrm{d}P(x) = \int_A \frac{\mathrm{d}P(x)}{\mathrm{d}\mu(x)} \, \mathrm{d}\mu(x) = \int_A p(x) \, \mathrm{d}\mu(x).$$

Discrete distributions

- On $E = \{0, 1\}$, the Bernoulli(θ) law has a density with respect to $\mu = \delta_0 + \delta_1$:
$$p(x) = (1 - \theta)\mathbf{1}_{\{x=0\}} + \theta\mathbf{1}_{\{x=1\}}.$$
- On $E = \{0, 1, \ldots, n\}$, the Binomial(n, θ) law is absolutely continuous with respect to the counting measure $\mu = \sum_{k=0}^{n} \delta_k$, with density
$$p(k) = \binom{n}{k} \theta^k (1 - \theta)^{n-k}, \qquad k = 0, 1, \ldots, n.$$
- On $E = \mathbb{N}^*$, the geometric law with parameter p is absolutely continuous with respect to the counting measure $\mu = \sum_{k \geqslant 1} \delta_k$, with density
$$p(k) = (1 - p)^{k-1} p, \qquad k \geqslant 1.$$

Continuous distributions

- The normal law $\mathcal{N}(\mu, \sigma^2)$ has density with espect to Lebesgue measure on $\mathbb{R}$:

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

- The exponential law with rate $\lambda > 0$ has density (with respect to Lebesgue on $\mathbb{R}$)

$$x \longmapsto \lambda e^{-\lambda x} \mathbf{1}_{\{x \geqslant 0\}}.$$

Markov's inequality

Let X be a non-negative real random variable and $a > 0$. Then

$$\mathbb{P}(X \geqslant a) \leqslant \frac{\mathbb{E}[X]}{a}.$$

In particular, for a real random variable X and $p \in \mathbb{N}^*$, since $x \mapsto x^p$ is increasing on $\mathbb{R}_+$,

$$\mathbb{P}(|X| \geqslant a) = \mathbb{P}(|X|^p \geqslant a^p) \leqslant a^{-p}\,\mathbb{E}[|X|^p].$$

Chebyshev's inequality

Let X be a real random variable and $a > 0$. Then

$$\mathbb{P}\big(|X - \mathbb{E}[X]| \geqslant a\big) \leqslant \frac{\mathrm{Var}(X)}{a^2}, \qquad \mathrm{Var}(X) = \mathbb{E}\big[(X - \mathbb{E}[X])^2\big].$$

Hoeffding's inequality

Let $X_1, \dots, X_n$ be independent random variables, denote $\overline{X}_n = \frac{1}{n}\sum_{k=1}^n X_k$ and suppose $a_i \leqslant X_i \leqslant b_i$ a.s.. Then for all $\varepsilon \geqslant 0$,

$$\mathbb{P}(\bar{X}_n - \underbrace{\mathbb{E}[\bar{X}_n]}_{\mathbb{E}[X_1]} \geqslant \varepsilon) \leqslant \exp\left(-\frac{2\varepsilon^2 n^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}[X_1]| \geqslant \varepsilon) \leqslant 2\exp\left(-\frac{2\varepsilon_n^2 n^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

Classical inequalities

Example: Bernoulli sample mean

Let $X_1, \ldots, X_n$ be i.i.d. with Bernoulli(p) distribution and $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. For every $\varepsilon > 0$,

$\mathrm{Var}(\bar{X}_n) = \frac{1}{n}\mathrm{Var}(X_1)$

Chebyshev's inequality:

$$\mathbb{P}\big(|\bar{X}_n - p| > \varepsilon\big) \leqslant \frac{\mathrm{Var}(\bar{X}_n)}{\varepsilon^2} = \frac{p(1-p)}{n\varepsilon^2}.$$

Since $0 \leqslant X_i \leqslant 1$, we can improve this using **Hoeffding's inequality**:

$$\mathbb{P}\big(|\bar{X}_n - p| > \varepsilon\big) \leqslant 2\exp\big(-2n\varepsilon^2\big).$$

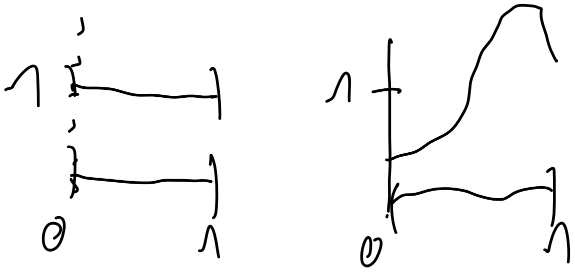
Chebyshev gives a bound of order $1/n$, whereas Hoeffding yields an exponentially small bound in n.

Gamma distribution

For $p, \lambda > 0$, a random variable Z has *Gamma*(p, λ) distribution if it has density

$$f_Z(x) = \frac{\lambda^p}{\Gamma(p)} x^{p-1} e^{-\lambda x} \mathbf{1}_{\{x \geqslant 0\}}, \qquad \Gamma(p) = \int_0^\infty z^{p-1} e^{-z}\, dz.$$

- $\mathbb{E}[Z] = \frac{p}{\lambda}$, $\text{Var}(Z) = \frac{p}{\lambda^2}$.
- Special case: $\Gamma(1, \lambda) = \text{Exp}(\lambda)$.



Beta distribution

For $a, b > 0$, a random variable X has *Beta*(a, b) distribution if it has density

$$f_X(x) = \frac{1}{B(a,b)} x^{a-1} (1-x)^{b-1} \mathbf{1}_{\{0 \leqslant x \leqslant 1\}}, \qquad B(a,b) = \int_0^1 z^{a-1}(1-z)^{b-1}\, dz = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

- $\mathbb{E}[X] = \frac{a}{a+b}$, $\text{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}$.
- Special case: $\text{Beta}(1,1) = \mathcal{U}[0,1]$.

Gamma and Beta: main properties

Additivity of Gamma

If $Y \sim \Gamma(p, \lambda)$ and $Z \sim \Gamma(q, \lambda)$ are independent, then

$$Y + Z \sim \Gamma(p + q, \lambda).$$

In particular, if $E_1, \ldots, E_n$ are i.i.d. $\text{Exp}(\lambda)$, then $\sum_{i=1}^{n} E_i \sim \Gamma(n, \lambda)$.

Scaling of Gamma

If $Y \sim \Gamma(p, \lambda)$ and $t > 0$, then $tY \sim \Gamma\left(p, \frac{\lambda}{t}\right)$.

Gamma–Beta connection

If $X \sim \Gamma(a, \lambda)$ and $Y \sim \Gamma(b, \lambda)$ are independent, then $\frac{X}{X+Y} \sim \text{Beta}(a, b)$. $\quad \frac{Y}{X+Y} \sim \text{Beta}(b, a)$

or $\Gamma(a, \lambda)$

As a special case, if E_1, E_2 are i.i.d. $\text{Exp}(\lambda)$, then

$$\frac{E_1}{E_1 + E_2} \sim \mathcal{U}[0, 1].$$

Dirichlet distribution: definition and properties

Definition (Dirichlet distribution)

Let $K \geqslant 2$ and $\alpha_1, \ldots, \alpha_K > 0$. A random vector $X = (X_1, \ldots, X_K)$ has *Dirichlet*$(\alpha_1, \ldots, \alpha_K)$ distribution if $X_i > 0$, $\sum_{i=1}^K X_i = 1$, and its density on the *simplex* is

$$f_X(x_1, \ldots, x_K) = \frac{\Gamma\left(\sum_{i=1}^K \alpha_i\right)}{\prod_{i=1}^K \Gamma(\alpha_i)} \prod_{i=1}^K x_i^{\alpha_i - 1}, \qquad (x_1, \ldots, x_K) \in S_K = \{x \in [0,1]^K : \sum_i x_i = 1\}$$

Key properties

- **Beta as a special case:** for $K = 2$, $\mathrm{Dir}(\alpha_1, \alpha_2)$ is the $\mathrm{Beta}(\alpha_1, \alpha_2)$ distribution.
- **Marginals are Beta:** if $X \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_K)$, then

$$X_i \sim \mathrm{Beta}\left(\alpha_i, \sum_{k=1}^K \alpha_k - \alpha_i\right), \qquad \mathbb{E}[X_i] = \frac{\alpha_i}{\sum_{k=1}^K \alpha_k}$$

- **Gamma representation:** If $Z_i \sim \Gamma(\alpha_i, \lambda)$ are independent and $Z = \sum_{k=1}^K Z_k$, then

$$\left(\frac{Z_1}{Z}, \ldots, \frac{Z_K}{Z}\right) \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_K).$$

Modes of convergence of random variables

Convergence in probability

Let $X_1, \dots, X_n, \dots$ and X be random variables taking values in $\mathbb{R}^d$, defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The sequence (X_n) *converges in probability* to X, written $X_n \xrightarrow{\mathbb{P}} X$, if

$$\forall \varepsilon > 0, \qquad \mathbb{P}\big(\|X_n - X\| > \varepsilon\big) \xrightarrow[n\to\infty]{} 0.$$

Convergence in L^2

In the same setting, we say that (X_n) *converges in L^2* to X, written $X_n \xrightarrow{L^2} X$, if

$$\mathbb{E}\left[\|X_n - X\|^2\right] \xrightarrow[n\to\infty]{} 0.$$

Modes of convergence of random variables (II)

Almost sure convergence

With the same notation, the sequence (X_n) *converges almost surely* to X, written $X_n \xrightarrow{\text{a.s.}} X$ if

$$\mathbb{P}\left(\left\{\omega \in \Omega : X_n(\omega) \xrightarrow[n\to\infty]{} X(\omega)\right\}\right) = 1.$$

Proposition

We have the implications

$$X_n \xrightarrow{\text{a.s.}} X \implies X_n \xrightarrow{\mathbb{P}} X,$$

and

$$X_n \xrightarrow{L^2} X \implies X_n \xrightarrow{\mathbb{P}} X.$$

Convergence in distribution / in law

Let $(X_n)_{n\geqslant 1}$ and X be random variables with values in $\mathbb{R}^d$. We say that X_n *converges in distribution* (or in law) to X, written $X_n \xrightarrow{\mathcal{L}} X$, if for every bounded continuous function $f : \mathbb{R}^d \to \mathbb{R}$,

$$\mathbb{E}\big[f(X_n)\big] \xrightarrow[n\to\infty]{} \mathbb{E}\big[f(X)\big].$$

Similarly, we say that (X_n) converges in distribution to a probability measure P on $\mathbb{R}^d$ if

$$\mathbb{E}\big[f(X_n)\big] \xrightarrow[n\to\infty]{} \mathbb{E}\big[f(X)\big]$$

for $X \sim P$ and every bounded continuous function f.

Central Limit Theorem in $\mathbb{R}^d$

Multivariate Central Limit Theorem

Let (X_n) be a sequence of i.i.d. random variables with values in $\mathbb{R}^d$, such that $\mathbb{E}\left[\|X_1\|^2\right] < \infty$.

Let

$$\mu = \mathbb{E}[X_1], \qquad \Sigma = \mathbb{E}\left[(X_1 - \mathbb{E}[X_1])(X_1 - \mathbb{E}[X_1])^T\right].$$

Then

$$\sqrt{n}\left(\overline{X}_n - \mu\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma).$$

$$= \frac{1}{n} \sum_{i=1}^{n} X_i$$

Continuous mapping theorem, Slutsky's lemma

Continuous mapping theorem

Let X_n, X be random variables taking values in $\mathbb{R}^d$ and $g : \mathbb{R}^d \to \mathbb{R}^k$ a continuous function.

- If $X_n \xrightarrow{\mathcal{L}} X$, then $g(X_n) \xrightarrow{\mathcal{L}} g(X)$.
- If $X_n \xrightarrow{\mathbb{P}} X$, then $g(X_n) \xrightarrow{\mathbb{P}} g(X)$.
- If $X_n \xrightarrow{\text{a.s.}} X$, then $g(X_n) \xrightarrow{\text{a.s.}} g(X)$.

Slutsky's lemma

Let (X_n) and (Y_n) be sequences of real-valued random variables, X a real-valued random variable, and $a \in \mathbb{R}$.

$$X_n \xrightarrow{\mathcal{L}} X \quad \text{and} \quad Y_n \xrightarrow{\mathbb{P}} a \implies (X_n, Y_n) \xrightarrow{\mathcal{L}} (X, a).$$

Remark

For a constant a, we have

$$Z_n \xrightarrow{\mathcal{L}} a \quad \iff \quad Z_n \xrightarrow{\mathbb{P}} a.$$

Statistical experiment and model

A *statistical experiment* consists of

- a random variable X defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with values in a measurable space $(E, \mathcal{E})$;
- a family of probability measures on $(E, \mathcal{E})$, called a *statistical model*,

$$\mathcal{P} = \{P_\theta : \theta \in \Theta\},$$

$= \mathbb{R}^d$ for instance

where Θ is the *parameter space*.

In the *frequentist approach* one assumes that the law of X belongs to the model:

$$\exists \theta_0 \in \Theta, \qquad X \sim P_{\theta_0}.$$

Statistical inference aims at learning about θ_0 from an observation of X.

Sample model

Also, often X_i are i.i.d.

Example 2: $((X_1, Y_1), \dots, (X_n, Y_n))$

In practice X is often an n-tuple of random variables

$$X = (X_1, \dots, X_n)$$

Then the sample space and the model depend on n.

Example: n-sample model

When $X = (X_1, \dots, X_n)$, one often works with the n-**sample model**

$$\mathcal{P}_n = \left\{ P_\theta^{\otimes n} : \theta \in \Theta \right\},$$

where

$$P_\theta^{\otimes n} = \underbrace{P_\theta \otimes \cdots \otimes P_\theta}_{n \text{ times}}.$$

This corresponds to assuming that $X_1, \dots, X_n$ are i.i.d. with common distribution P_θ.

Identifiability and dominated models

Identifiable model

A statistical model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is *identifiable* if for all $\theta, \theta' \in \Theta$,

$$P_\theta = P_{\theta'} \quad \Longrightarrow \quad \theta = \theta'.$$

Equivalently, the mapping $\theta \mapsto P_\theta$ is injective. This guarantees that each distribution in the model corresponds to a unique parameter value.

Dominated model

The model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is *dominated* if there exists a σ-finite measure μ on E such that, for all $\theta \in \Theta$, $P_\theta \ll \mu$. Then every P_θ admits a density p_θ with respect to μ:

$$\mathrm{d}P_\theta(x) = p_\theta(x)\,\mathrm{d}\mu(x).$$

In what follows we often work with dominated, *parametric models* with $\Theta \subset \mathbb{R}^d$.

I have a family of densities $\{p_\theta : \theta \in \Theta\}$

Example 1: Bernoulli model

Consider $E = \{0, 1\}$ and parameter space $\Theta = (0, 1)$. For $\theta \in \Theta$ let

$$P_\theta(X = 1) = \theta, \qquad P_\theta(X = 0) = 1 - \theta.$$

The model is

$$\mathcal{P} = \{P_\theta : \theta \in (0, 1)\}.$$

- This is a dominated model with respect to the counting measure on $\{0, 1\}$; the density is

$$p_\theta(x) = (1 - \theta)\mathbf{1}_{\{0\}}(x) + \theta\, \mathbf{1}_{\{1\}}(x).$$

- The model is identifiable: $P_\theta = P_{\theta'}$ implies $\theta = P_\theta(X = 1) = P_{\theta'}(X = 1) = \theta'$.

Example 2: Gaussian model with unknown mean

Let $E = \mathbb{R}$, $\Theta = \mathbb{R}$ and fix $\sigma^2 > 0$. For $\theta \in \Theta$ define P_θ as the normal law

$$P_\theta = \mathcal{N}(\theta, \sigma^2).$$

The model is

$$\mathcal{P} = \{\mathcal{N}(\theta, \sigma^2) : \theta \in \mathbb{R}\}.$$

- This model is dominated by Lebesgue measure λ on $\mathbb{R}$ with density

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

- It is identifiable: equality of the densities (or distributions) for all x forces the means to be equal.

if $P_\theta = P_{\theta'}$, $P_\theta = P_{\theta'} \Rightarrow \theta = \theta'$

Estimators in a statistical experiment

Estimator

Consider the statistical experiment $(X, \mathcal{P})$. An *estimator* of the parameter θ is a measurable function

$$\hat{\theta} = \hat{\theta}(X)$$

with values in the parameter space Θ (more precisely, $\hat{\theta}$ is measurable from $(E, \mathcal{E})$ to $(\Theta, \mathcal{B}(\Theta))$, where $\mathcal{B}(\Theta)$ is the Borel σ-algebra).

Sequence of experiments

In practice we often have a sequence of experiments $(X^{(n)}, \mathcal{P}_n)$, $n = 1, 2, \ldots$

↳ e.g., n is the sample size

This leads to a sequence of estimators $(\hat{\theta}_n)$.

Likelihood and maximum likelihood estimator

Assume a dominated model with respect to a measure μ: for each $\theta \in \Theta$,

$$\mathrm{d}P_\theta(x) = p_\theta(x)\,\mathrm{d}\mu(x).$$

Let $X = (X_1, \ldots, X_n) \sim P_\theta^{\otimes n}$. The joint density of X is

$$p_\theta^{\otimes n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} p_\theta(x_i).$$

Viewed as a function of θ for the observed data X, this is the *likelihood function*

$$L_\theta(X) = \prod_{i=1}^{n} p_\theta(X_i).$$

Often we work instead with the *log-likelihood*

$$\ell_\theta(X) = \log L_\theta(X) = \sum_{i=1}^{n} \log p_\theta(X_i).$$

Maximum likelihood estimator (MLE)

Definition (MLE)

In a dominated model, a *maximum likelihood estimator (MLE)* is, when it exists, a value $\hat{\theta}(X) \in \Theta$ such that

$$\hat{\theta}(X) \in \arg\max_{\theta \in \Theta} L_\theta(X), \qquad \text{or equivalently} \qquad \hat{\theta}(X) \in \arg\max_{\theta \in \Theta} \ell_\theta(X).$$

Example / exercise (Bernoulli model). In the Bernoulli model $\mathcal{P} = \{\mathcal{B}(\theta)^{\otimes n} : \theta \in [0,1]\}$, show that the MLE of θ is unique and given by the empirical mean

$$\hat{\theta}(X) = \overline{X}_n.$$

Maximum likelihood estimator (MLE)

$$p_\theta^{\otimes n}(x) = \prod_{i=1}^{N} p_\theta(x_i) = \prod_{i=1}^{N} \theta^{x_i}(1-\theta)^{x_i}$$

$$\log p_\theta^{\otimes m}(x) = \sum_{i=1}^{N}\left[x_i \log\theta + (1-x_i)\log(1-\theta)\right]$$

$$= [\log\theta]\left[\sum_{i=1}^{N} x_i\right] + \left[N - \sum_{i=1}^{N} x_i\right]\log(1-\theta)$$

$$\frac{\partial \cdots}{\partial\theta} = \frac{\sum_{i=1}^{N} x_i}{\theta} - \frac{\left[N-\sum_{i=1}^{N} x_i\right]}{1-\theta} = 0$$

$$\Rightarrow \frac{\theta}{\sum_{i=1}^{N} x_i} = \frac{1-\theta}{N-\sum_{i=1}^{N} x_i} \Rightarrow \frac{\theta}{1-\theta} = \frac{\overline{X}_n}{1-\overline{X}_n}$$

and $x \mapsto \frac{x}{1-x}$ is bijective

Consistency and asymptotic normality

Consistency

Consider a sequence of experiments $(X^{(n)}, \mathcal{P}_n)$ with

$$\mathcal{P}_n = \{P_\theta^{\otimes n} : \theta \in \Theta\}.$$

A sequence of estimators $(\hat{\theta}_n)$ is *consistent* if, for every $\theta \in \Theta$, when $X^{(n)} \sim P_\theta^{\otimes n}$,

$$\hat{\theta}_n(X^{(n)}) \xrightarrow[n\to\infty]{\mathbb{P}} \theta.$$

$\in \mathbb{R}$ $\iff$ $\hat{\Theta}_n(X^{(n)}) \xrightarrow[n \to +\infty]{\mathcal{L}} \Theta$

Asymptotic normality

In the same setting, $(\hat{\theta}_n)$ is *asymptotically normal* if for each $\theta \in \Theta$ there exists a symmetric positive semi-definite matrix Σ_θ such that, when $X^{(n)} \sim P_\theta^{\otimes n}$,

$$\sqrt{n}\left(\hat{\theta}_n(X^{(n)}) - \theta\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \Sigma_\theta).$$

Consistency and asymptotic normality

Exercise. Show that if $(\hat{\theta}_n)$ is asymptotically normal, then it is consistent.

$$\sqrt{n}\left(\hat{\theta}_n(X^{(n)}) - \theta\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \Sigma_\theta).$$

$\longrightarrow (\hat{\theta}_n - \theta) = \frac{1}{\sqrt{n}} \times \sqrt{n}(\hat{\theta}_n - \theta)$, where $\frac{1}{\sqrt{n}} \xrightarrow{\mathbb{P}} 0$ and $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_\theta)$

Slutsky $\Rightarrow \left(\frac{1}{\sqrt{n}}, \sqrt{n}(\hat{\theta}_n - \theta)\right) \xrightarrow{\mathcal{L}} (0, \mathcal{N}(0, \Sigma_\theta))$

Apply $g : \mathbb{R}^2 \to \mathbb{R}$, $(x, y) \mapsto xy$ with continuous mapping theorem:

$\hat{\theta}_n - \theta \xrightarrow{\mathcal{L}} 0 \iff \hat{\theta}_n - \theta \xrightarrow{\mathbb{P}} 0$ as 0 is a constant

Quadratic risk of an estimator

Definition (Quadratic risk)

Let $(X, \mathcal{P})$ be a statistical experiment with $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ and let $\hat{\theta}$ be an estimator. The *quadratic risk* of $\hat{\theta}$ at θ is

$$R(\theta, \hat{\theta}) = \mathbb{E}_\theta\left[\left\|\hat{\theta}(X) - \theta\right\|^2\right] = \int_E \left\|\hat{\theta}(x) - \theta\right\|^2 \mathrm{d}P_\theta(x).$$

Example: Scalar parameter case

When $\Theta \subset \mathbb{R}$, the quadratic risk reduces to

$$R(\theta, \hat{\theta}) = \mathbb{E}_\theta\left[\left(\hat{\theta}(X) - \theta\right)^2\right] = \int_E \left(\hat{\theta}(x) - \theta\right)^2 \mathrm{d}P_\theta(x).$$

A *"good"* estimator typically has small quadratic risk, but remember that $R(\theta, \hat{\theta})$ is a function of θ and may be small for some parameter values and large for others.

Proposition (Bias–variance decomposition)

Let $(X, \mathcal{P})$ be a statistical experiment with $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ and let $\hat{\theta}$ be an estimator. For every $\theta \in \Theta$, if $X \sim P_\theta$,

$$R(\theta, \hat{\theta}) = \mathbb{E}_\theta\left[\left\|\hat{\theta}(X) - \mathbb{E}_\theta[\hat{\theta}(X)]\right\|^2\right] + \left\|\mathbb{E}_\theta[\hat{\theta}(X)] - \theta\right\|^2.$$

The function

$$\theta \longmapsto \mathbb{E}_\theta[\hat{\theta}(X)] - \theta$$

is called the *bias* of $\hat{\theta}$.

Scalar parameter case

If $\Theta \subset \mathbb{R}$, then

$$R(\theta, \hat{\theta}) = \mathrm{Var}_\theta\big(\hat{\theta}(X)\big) + \big(\mathbb{E}_\theta[\hat{\theta}(X)] - \theta\big)^2.$$

Example: Bernoulli model and empirical mean

Setting

Let $(X, \mathcal{P})$ with $\mathcal{P} = \{\mathcal{B}(\theta)^{\otimes n} : \theta \in [0,1]\}$, where $X = (X_1, \dots, X_n)$ and X_i are i.i.d. Bernoulli(θ).

A natural estimator of θ is the empirical mean

$$\hat{\theta}_n(X) = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

- By the (strong) law of large numbers, $\hat{\theta}_n(X) \to \theta$ almost surely, hence $\hat{\theta}_n$ is consistent.
- By the central limit theorem, $\sqrt{n}(\hat{\theta}_n(X) - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \theta(1-\theta))$, so $\hat{\theta}_n$ is asymptotically normal.
- Since $\mathbb{E}_\theta[\hat{\theta}_n(X)] = \theta$, the estimator is unbiased and

$$R(\theta, \hat{\theta}_n) = \mathbb{E}_\theta\Big[\big(\hat{\theta}_n(X) - \theta\big)^2\Big] = \mathrm{Var}_\theta(\hat{\theta}_n(X)) = \frac{\theta(1-\theta)}{n}.$$

Risk and probability of large error

For any estimator $\hat{\theta}$ and any $\varepsilon > 0$, the quadratic risk controls the probability of a large error:

$$\mathbb{P}_\theta\left(\left|\hat{\theta}(X) - \theta\right| \geqslant \varepsilon\right) \leqslant \frac{\mathbb{E}_\theta\left[\left(\hat{\theta}(X) - \theta\right)^2\right]}{\varepsilon^2} = \frac{R(\theta, \hat{\theta})}{\varepsilon^2}.$$

This follows from Markov's (or Chebyshev's) inequality.

Thus, a small quadratic risk implies that $\hat{\theta}(X)$ is close to θ with high probability.

Example: Gaussian mean, two estimators

Setting

Let $X_1, \dots, X_n$ be i.i.d. $\mathcal{N}(\theta, 1)$ with $\theta \in \mathbb{R}$.
We compare two estimators:

- a constant estimator $\tilde{\theta}_n = \theta_0$ for some fixed $\theta_0 \in \mathbb{R}$.

  $$R(\theta, \tilde{\theta}_n) = \mathbb{E}_\theta\big[(\theta_0 - \theta)^2\big] = (\theta - \theta_0)^2.$$

  This risk is zero at $\theta = \theta_0$, but positive elsewhere and does not decrease with n.

- the empirical mean $\hat{\theta}_n(X) = \overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, for which $\mathbb{E}_\theta[\hat{\theta}_n(X)] = \theta$ (unbiased) and

  $$R(\theta, \hat{\theta}_n) = \mathrm{Var}_\theta(\hat{\theta}_n(X)) = \frac{1}{n}.$$

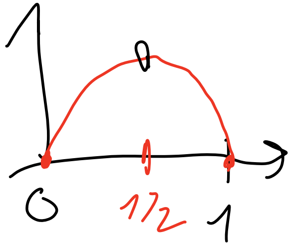
  The risk is independent of θ and decreases at rate $1/n$.

Exercise. For $X \sim \mathrm{Bin}(n, \theta)$ and $\hat{\theta} = X/n$, show that $R(\theta, \hat{\theta}) \leqslant 1/(4n)$ for all $\theta \in [0, 1]$.

$$R(\theta, \hat{\theta}) = \underbrace{\text{bias}}_{=0}{}^2 + \mathrm{Var}(\hat{\theta})$$

$$\mathbb{E}[\hat{\theta}] = n^{-1}\,\mathbb{E}[X] = n^{-1}\, n\,\theta = \theta$$

$$= \frac{n\,\theta(1-\theta)}{n^2} = \frac{\theta(1-\theta)}{n} \leq \frac{1}{4n}$$



Exact confidence intervals and regions

Let $\alpha > 0$.

Definition (exact confidence interval / region)

- **Case $\Theta \subset \mathbb{R}$.** A (random) interval $I(X) = [a(X), b(X)]$ is a *confidence interval of level (at least) $1-\alpha$* if

$$\forall \theta \in \Theta, \quad \mathbb{P}_\theta\big(\theta \in I(X)\big) \geqslant 1-\alpha.$$

- **Case $\Theta \subset \mathbb{R}^d$.** A random subset $\mathcal{R}(X) \subset \Theta$ is a *confidence region of level (at least) $1-\alpha$* if

$$\forall \theta \in \Theta, \quad \mathbb{P}_\theta\big(\theta \in \mathcal{R}(X)\big) \geqslant 1-\alpha.$$

⚠ The interval $I(X)$ cannot depend on the unknown parameter θ; it may only depend on known quantities (such as α, the sample size n, and the data X).

Example: normal mean, exact confidence interval

- <u>Gaussian model</u>
  We observe $X = (X_1, \ldots, X_n)$ i.i.d. with $X_i \sim \mathcal{N}(\theta, 1)$, $\theta \in \mathbb{R}$. Let $\hat{\theta}(X) = \overline{X}_n = n^{-1} \sum_{i=1}^{n} X_i$. Then

$$\sqrt{n}\,(\overline{X}_n - \theta) \sim \mathcal{N}(0, 1).$$

Example: normal mean, exact confidence interval

- <u>Gaussian model</u>
  We observe $X = (X_1, \ldots, X_n)$ i.i.d. with $X_i \sim \mathcal{N}(\theta, 1),\ \theta \in \mathbb{R}$. Let $\hat{\theta}(X) = \overline{X}_n = n^{-1} \sum_{i=1}^{n} X_i$. Then

$$\sqrt{n}\,(\overline{X}_n - \theta) \sim \mathcal{N}(0,1).$$

  Denote by Φ the c.d.f. of $\mathcal{N}(0,1)$ and set $q_\alpha = \Phi^{-1}(1-\alpha/2)$, so that $\mathbb{P}(|\mathcal{N}(0,1)| > q_\alpha) = \alpha$.

Example: normal mean, exact confidence interval

- <u>Gaussian model</u>
  We observe $X = (X_1, \ldots, X_n)$ i.i.d. with $X_i \sim \mathcal{N}(\theta, 1)$, $\theta \in \mathbb{R}$. Let $\hat{\theta}(X) = \overline{X}_n = n^{-1} \sum_{i=1}^n X_i$. Then

$$\sqrt{n}\,(\overline{X}_n - \theta) \sim \mathcal{N}(0, 1).$$

  Denote by Φ the c.d.f. of $\mathcal{N}(0,1)$ and set $q_\alpha = \Phi^{-1}(1 - \alpha/2)$, so that $\mathbb{P}(|\mathcal{N}(0,1)| > q_\alpha) = \alpha$.

- <u>Resulting confidence interval</u>
  We have

$$\iff \hat{\theta}(X) - \frac{q_\alpha}{\sqrt{n}} \leq \theta \leq \hat{\theta}(X) + \frac{q_\alpha}{\sqrt{n}}$$

$$\mathbb{P}_\theta\left(\left|\sqrt{n}\,(\hat{\theta}(X) - \theta)\right| > q_\alpha\right) = \alpha.$$

  Equivalently,

$$I(X) = \left[\hat{\theta}(X) \pm \frac{q_\alpha}{\sqrt{n}}\right]$$

  is an **exact** level $1 - \alpha$ confidence interval for θ.

Asymptotic confidence intervals

Sometimes the finite-sample distribution of an estimator is unknown, but its limiting distribution as $n \to \infty$ is known. This leads to asymptotic confidence intervals/regions.

Definition (asymptotic confidence interval/region)

- **Case $\Theta \subset \mathbb{R}$.** A random interval $I(X^{(n)})$ is an *asymptotic confidence interval of level (at least)* $1 - \alpha$ if

$$\forall \theta \in \Theta, \quad \liminf_{n \to \infty} \mathbb{P}_\theta\big(\theta \in I(X^{(n)})\big) \geqslant 1 - \alpha.$$

- **Case $\Theta \subset \mathbb{R}^d$.** A random set $\mathcal{R}(X^{(n)}) \subset \Theta$ is an *asymptotic confidence region of level (at least)* $1 - \alpha$ if

$$\forall \theta \in \Theta, \quad \liminf_{n \to \infty} \mathbb{P}_\theta\big(\theta \in \mathcal{R}(X^{(n)})\big) \geqslant 1 - \alpha.$$

General construction from an asymptotically normal estimator

Proposition (asymptotic CI from asymptotic normality)

Assume $\Theta \subset \mathbb{R}$ and let $\hat{\theta}_n = \hat{\theta}_n(X)$ be an estimator such that

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0, \sigma^2(\theta)\right),$$

where the function $\theta \mapsto \sigma^2(\theta)$ is continuous.
Let $q_\alpha > 0$ satisfy

$$\mathbb{P}\big(|\mathcal{N}(0,1)| \leqslant q_\alpha\big) = 1 - \alpha \quad (\text{so } q_\alpha = \Phi^{-1}(1-\alpha/2)).$$

Define

$$I(X) = \left[\hat{\theta}_n(X) - \frac{q_\alpha\, \sigma(\hat{\theta}_n(X))}{\sqrt{n}},\ \hat{\theta}_n(X) + \frac{q_\alpha\, \sigma(\hat{\theta}_n(X))}{\sqrt{n}}\right].$$

Then $I(X)$ is an **asymptotic confidence interval of level exactly** $1 - \alpha$.

General construction from an asymptotically normal estimator

proof: $\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0, \sigma^2(\theta)\right)$, → consistency

$$\frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sigma(\hat{\theta}_n)} = \frac{\sigma(\theta)}{\sigma(\hat{\theta}_n)} \times \frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sigma(\theta)} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

$$\frac{\sigma(\theta)}{\sigma(\hat{\theta}_n)} \xrightarrow{\mathbb{P}} 1 \qquad \frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sigma(\theta)} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

$$\lim_{n\to+\infty} \mathbb{P}\left(\left| \frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sigma(\hat{\theta}_n)} \right| > q_\alpha \right) \leq \alpha$$

$$\Longleftrightarrow \theta \in \left[\hat{\theta}_n \pm \frac{q_\alpha \sigma(\hat{\theta}_n)}{\sqrt{n}} \right]$$

Conditional distribution (discrete case)

Definition (discrete conditional law)

Let X and Y be discrete random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with values respectively in sets E and F. For $x \in E$ such that $\mathbb{P}(X = x) > 0$, the conditional distribution of Y given $X = x$, denoted $\mathcal{L}(Y \mid X = x)$, is defined for all $y \in F$ by

$$\mathbb{P}(Y = y \mid X = x) = \frac{\mathbb{P}(Y = y, X = x)}{\mathbb{P}(X = x)}.$$

(Bayes' rule)

This defines, for each fixed x, a probability distribution on F.

Joint densities and marginals

sample space for X

same but for Y

Let

- $(E, \mathcal{E})$ and $(F, \mathcal{F})$ be measurable spaces;
- α a σ-finite positive measure on $(E, \mathcal{E})$, and β a σ-finite positive measure on $(F, \mathcal{F})$;
- X an E-valued random variable and Y an F-valued random variable.

Assume the pair (X, Y) has a joint density $h(x, y)$ with respect to $\alpha \otimes \beta$, i.e.

μ in Radon-Nykodym

$$\mathrm{d}P(x, y) = h(x, y)\,\mathrm{d}\alpha(x)\,\mathrm{d}\beta(y).$$

Note: $A \in \mathcal{E}$, $B \in \mathcal{F}$

$\alpha \otimes \beta(A \times B) = \alpha(A)\beta(B)$

The marginal law of X is the (probability) density

$$f(x) = \int_F h(x, y)\,\mathrm{d}\beta(y),$$

and the marginal law of Y is the (probability) density

$$g(y) = \int_E h(x, y)\,\mathrm{d}\alpha(x).$$

Conditional density (continuous case)

Definition (conditional density for fixed x)

Assume $f(x) > 0$ for some $x \in E$. The conditional law of Y given $X = x$, denoted $\mathcal{L}(Y \mid X = x)$, is the probability measure on F with density (w.r.t. β)

$$g_x(y) = \frac{h(x,y)}{f(x)} = \frac{h(x,y)}{\int_F h(x,y)\,\mathrm{d}\beta(y)}.$$

We may sometimes write $g(y \mid x)$ instead of $g_x(y)$ when there is no risk of confusion.

Remark

For points where $f(x) = 0$ we can define g_x arbitrarily (e.g. 0); these x typically form a set of $\mathcal{L}(X)$-measure zero, so they do not affect integrals.

Random conditional density and joint density factorization

Random conditional density

By extension, we define the conditional density of Y given X as the random density

$$g_X(y) = g(y \mid X) = \begin{cases} \dfrac{h(X,y)}{f(X)}, & f(X) > 0, \\ 0, & f(X) = 0. \end{cases}$$

Since $f(X) > 0$ almost surely, one usually just writes

$$g_X(y) = \frac{h(X,y)}{f(X)}.$$

Conditional expectation via conditional density

Definition (conditional expectation)

Let $\varphi : F \to \mathbb{R}$ be measurable with $\mathbb{E}[\varphi(Y)] < \infty$. The conditional expectation of $\varphi(Y)$ given X is

$$\mathbb{E}[\varphi(Y) \mid X] = \int_F \varphi(y)\, g(y \mid X)\, \mathrm{d}\beta(y).$$

This is a random variable measurable with respect to $\sigma(X)$.

Law of total expectation

For any measurable $\psi : E \times F \to \mathbb{R}$ such that $\psi(X, Y)$ is integrable,

$$\mathbb{E}[\psi(X, Y)] = \mathbb{E}\big[\mathbb{E}[\psi(X, Y) \mid X]\big].$$

In particular, if $\psi(X, Y) = \psi_1(X)\psi_2(Y)$ with integrable $\psi_1(X)$ and $\psi_2(Y)$, then

$$\mathbb{E}[\psi_1(X)\psi_2(Y)] = \mathbb{E}[\psi_1(X)\, \mathbb{E}[\psi_2(Y) \mid X]]\,.$$

Conditional expectation as best L^2 predictor

Projection property (orthogonality)

In the previous setting, let Y be square integrable: $\mathbb{E}[Y^2] < \infty$. Then

$$\inf_{\varphi:E\to\mathbb{R},\ \mathbb{E}[\varphi(X)^2]<\infty} \mathbb{E}\big[(Y - \varphi(X))^2\big] = \mathbb{E}\big[(Y - \mathbb{E}[Y \mid X])^2\big].$$

Thus $\mathbb{E}[Y \mid X]$ is the **best mean-square predictor** of Y among all (square integrable) functions of X.

Conditional expectation as best L^2 predictor

Projection property (orthogonality)

In the previous setting, let Y be square integrable: $\mathbb{E}[Y^2] < \infty$. Then

$$\inf_{\varphi:E\to\mathbb{R},\ \mathbb{E}[\varphi(X)^2]<\infty} \mathbb{E}\big[(Y - \varphi(X))^2\big] = \mathbb{E}\big[(Y - \mathbb{E}[Y \mid X])^2\big].$$

Thus $\mathbb{E}[Y \mid X]$ is the **best mean-square predictor** of Y among all (square integrable) functions of X.

Proof:
For any measurable $\varphi : E \to \mathbb{R}$ with $\mathbb{E}[\varphi(X)^2] < \infty$,

$$\mathbb{E}\big[(Y - \varphi(X))^2\big] = \mathbb{E}\big[(Y - \mathbb{E}[Y \mid X])^2\big] + \mathbb{E}\big[(\mathbb{E}[Y \mid X] - \varphi(X))^2\big].$$

The cross-term is zero because

$$\mathbb{E}[(Y - \mathbb{E}[Y \mid X])(\mathbb{E}[Y \mid X] - \varphi(X))] = \mathbb{E}\big[\mathbb{E}\big[Y - \mathbb{E}[Y \mid X] \mid X\big](\mathbb{E}[Y \mid X] - \varphi(X))\big] = 0.$$

Square integrability of $\mathbb{E}[Y \mid X]$

To justify the previous result we need $\mathbb{E}\big[\mathbb{E}[Y \mid X]^2\big] < \infty$.

By the conditional Jensen inequality,

$$\mathbb{E}\big[\mathbb{E}[Y \mid X]^2\big] \leqslant \mathbb{E}\big[\mathbb{E}[Y^2 \mid X]\big] = \mathbb{E}[Y^2] < \infty.$$

Hence $\mathbb{E}[Y \mid X]$ is square integrable and the projection property makes sense in L^2.

Frequentist approach

In the frequentist approach, we assume that there exists a *true but unknown* parameter value $\theta_0 \in \Theta$ such that the data X follow the law P_{θ_0}:

$$\exists\, \theta_0 \in \Theta \quad \text{s.t.} \quad X \sim P_{\theta_0}.$$

Gaussian model

Let

$$X = (X_1, \ldots, X_n), \qquad \mathcal{P} = \left\{ \mathcal{N}(\theta, 1)^{\otimes n} : \theta \in \mathbb{R} \right\}.$$

The frequentist assumption is that for some $\theta_0 \in \mathbb{R}$, the data are i.i.d. $\mathcal{N}(\theta_0, 1)$. One can then estimate θ_0 by the empirical mean $\bar{X}_n$; by the law of large numbers, $\bar{X}_n \xrightarrow{\mathbb{P}} \theta_0$.

- **Estimation:** construct an estimator $\hat{\theta}(X)$ close to θ_0.
- **Confidence sets:** build random sets $\mathcal{R}(X) \subset \Theta$ with $\theta_0 \in \mathcal{R}(X)$ with high probability under P_{θ_0}.
- **Tests:** answer "true/false" to a property of θ_0 via tests $\varphi(X) \in \{0, 1\}$.

Bayesian approach: intuition

In the Bayesian approach, all unknown quantities are modeled as random variables.

Prior and posterior

- Before observing data, our uncertainty about θ is described by a **prior distribution** Π_0 on Θ.
- After observing X, we update this prior using Bayes' formula to obtain the **posterior distribution** $\Pi(\cdot \mid X)$.

The posterior combines:

- prior knowledge (or belief) about θ;
- the information contained in the data X.

Coin tossing: frequentist vs Bayesian view

Let $\theta \in [0, 1]$ be the probability of "heads".

- Frequentist: θ is fixed; with many tosses, the empirical frequency $\bar{X}_n$ converges to θ (LLN, CLT).
- Bayesian: before any toss, we put a prior on θ (e.g. uniform on $[0, 1]$). Each new observation updates the prior to a posterior that reflects both prior belief and data.