Bernstein–von Mises Theorem & Why Sampling Matters

Chapter 3 wrap-up $\rightarrow$ Chapters 4–5 motivation

(Lecture slides)

Where we are (end of Chapter 3)

- We have learned: **Bayesian inference** = prior $\times$ likelihood $\rightarrow$ posterior.
- We can often compute posteriors analytically when we have **conjugacy**.
- Today's last big idea in Chapter 3:

Bayesian inference and frequentist inference agree (as data $\rightarrow \infty$).

Where we are (end of Chapter 3)

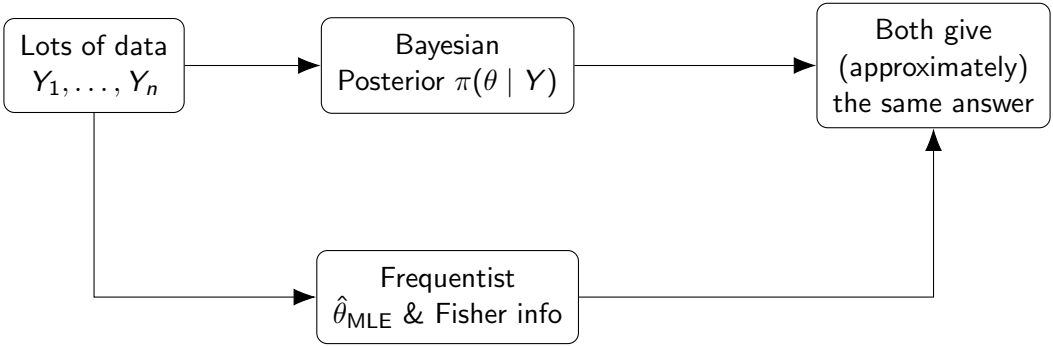
- We have learned: **Bayesian inference** = prior $\times$ likelihood $\rightarrow$ posterior.
- We can often compute posteriors analytically when we have **conjugacy**.
- Today's last big idea in Chapter 3:

Bayesian inference and frequentist inference agree (as data $\rightarrow \infty$).

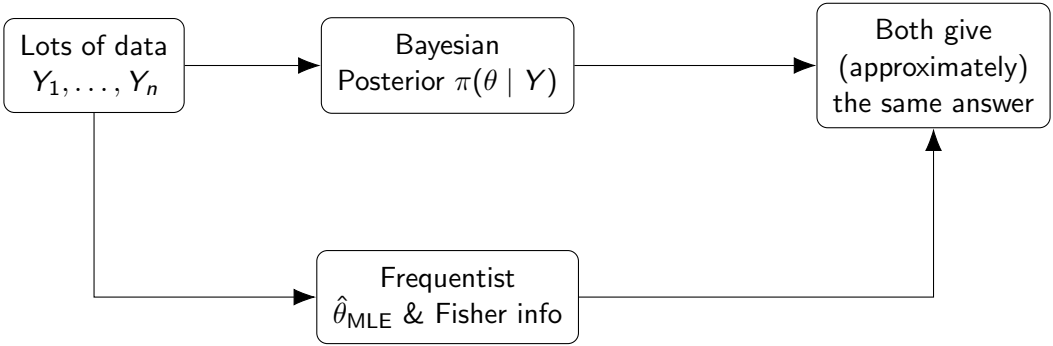
The bridge

Bernstein–von Mises says the posterior distribution becomes approximately Normal, centered at the **MLE**, with covariance determined by the **Fisher information**.

Big picture (what you should remember)



Big picture (what you should remember)



Informal slogan

With enough data and a well-specified model, the prior "washes out" and the posterior looks like a Normal distribution around the MLE.

Set-up (model, prior, posterior)

Data model (parametric, fixed dimension).

$$Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} p(\cdot \mid \theta), \qquad \theta \in \Theta \subset \mathbb{R}^d$$

Prior. A density $\pi(\theta)$ on Θ.

Set-up (model, prior, posterior)

Data model (parametric, fixed dimension).

$$Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} p(\cdot \mid \theta), \qquad \theta \in \Theta \subset \mathbb{R}^d$$

Prior. A density $\pi(\theta)$ on Θ.
Posterior.

$$\pi(\theta \mid Y) \propto \left(\prod_{i=1}^{n} p(Y_i \mid \theta) \right) \pi(\theta).$$

Set-up (model, prior, posterior)

Data model (parametric, fixed dimension).

$$Y_1, \dots, Y_n \overset{\text{i.i.d.}}{\sim} p(\cdot \mid \theta), \qquad \theta \in \Theta \subset \mathbb{R}^d$$

Prior. A density $\pi(\theta)$ on Θ.
Posterior.

$$\pi(\theta \mid Y) \propto \left(\prod_{i=1}^{n} p(Y_i \mid \theta) \right) \pi(\theta).$$

MLE.

$$\hat{\theta}_n := \arg\max_{\theta \in \Theta} \ell_n(\theta), \qquad \ell_n(\theta) := \sum_{i=1}^{n} \log p(Y_i \mid \theta).$$

The Bernstein–von Mises theorem (informal statement)

Theorem (Bernstein–von Mises, informal)

Under regularity conditions (well-specified model, fixed d, smooth prior with positive mass near the truth), the posterior distribution is approximately Normal:

$$\pi(\theta \mid Y) \approx \mathcal{N}\Big(\hat{\theta}_n,\ \big(n\, I(\hat{\theta}_n)\big)^{-1}\Big),$$

and the approximation error goes to 0 as $n \to \infty$ in **total variation distance**.

The Bernstein–von Mises theorem (informal statement)

Theorem (Bernstein–von Mises, informal)

Under regularity conditions (well-specified model, fixed d, smooth prior with positive mass near the truth), the posterior distribution is approximately Normal:

$$\pi(\theta \mid Y) \approx \mathcal{N}\Big(\hat{\theta}_n,\ \big(n\, I(\hat{\theta}_n)\big)^{-1}\Big),$$

and the approximation error goes to 0 as $n \to \infty$ in **total variation distance**.

Interpretation

Bayesian uncertainty (posterior spread) matches frequentist uncertainty (Fisher information / MLE asymptotics).

What does “distance between distributions” mean?

One common distance used in the theorem is **total variation**:

$$\mathrm{TV}(P, Q) = \sup_{A} |P(A) - Q(A)|.$$

What does “distance between distributions” mean?

One common distance used in the theorem is **total variation**:

$$\mathrm{TV}(P, Q) = \sup_{A} |P(A) - Q(A)|.$$

If P and Q have densities p and q,

$$\mathrm{TV}(P, Q) = \frac{1}{2} \int_{\mathbb{R}^d} |p(x) - q(x)| \, dx.$$

What does "distance between distributions" mean?

One common distance used in the theorem is **total variation**:

$$\mathrm{TV}(P, Q) = \sup_A |P(A) - Q(A)|.$$

If P and Q have densities p and q,

$$\mathrm{TV}(P, Q) = \frac{1}{2} \int_{\mathbb{R}^d} |p(x) - q(x)| \, dx.$$

How to think about it

- $\mathrm{TV}(P, Q) = 0$ means the distributions are identical.
- $\mathrm{TV}(P, Q)$ small means the distributions assign almost the same probability to *every* event.
- Here: posterior and its Normal approximation become indistinguishable as $n \to \infty$.

All the qualifiers (why they matter)

1. **Well-specified model:** the model family contains the true data-generating distribution.
   - If data are Normal and you fit Normal $\Rightarrow$ good.
   - If data are Normal and you fit Exponential $\Rightarrow$ the theorem can fail badly.

All the qualifiers (why they matter)

1. **Well-specified model:** the model family contains the true data-generating distribution.
   - If data are Normal and you fit Normal $\Rightarrow$ good.
   - If data are Normal and you fit Exponential $\Rightarrow$ the theorem can fail badly.
2. **Fixed number of parameters d:** d does not grow with n.
   - Many modern "high-dimensional" settings need extra work (and BvM can fail).

All the qualifiers (why they matter)

1. **Well-specified model:** the model family contains the true data-generating distribution.
   - If data are Normal and you fit Normal $\Rightarrow$ good.
   - If data are Normal and you fit Exponential $\Rightarrow$ the theorem can fail badly.
2. **Fixed number of parameters d:** d does not grow with n.
   - Many modern "high-dimensional" settings need extra work (and BvM can fail).
3. **Smooth prior:** no nasty discontinuities; typically continuous density with derivatives.
   - Avoid point-mass priors like $\pi(\theta) = \delta_{\theta_0}$ if you want BvM behaviour.

All the qualifiers (why they matter)

1. **Well-specified model:** the model family contains the true data-generating distribution.
   - If data are Normal and you fit Normal $\Rightarrow$ good.
   - If data are Normal and you fit Exponential $\Rightarrow$ the theorem can fail badly.
2. **Fixed number of parameters d:** d does not grow with n.
   - Many modern "high-dimensional" settings need extra work (and BvM can fail).
3. **Smooth prior:** no nasty discontinuities; typically continuous density with derivatives.
   - Avoid point-mass priors like $\pi(\theta) = \delta_{\theta_0}$ if you want BvM behaviour.
4. **Prior positive near MLE / truth:** $\pi(\theta)$ should not be (essentially) zero where the likelihood concentrates.

Why "prior non-zero near the MLE" is essential

Cautionary example (setting yourself up to fail)

Suppose the likelihood pushes the estimate near $\hat{\theta}_n \approx 5$, but your prior is supported only on $(1, 4)$.

Why "prior non-zero near the MLE" is essential

Cautionary example (setting yourself up to fail)

Suppose the likelihood pushes the estimate near $\hat{\theta}_n \approx 5$, but your prior is supported only on $(1, 4)$.

Then:

- The posterior *cannot* concentrate near 5 (it has zero mass there).
- No matter how much data you get, the prior prevents the posterior from matching the MLE asymptotics.

Why "prior non-zero near the MLE" is essential

Cautionary example (setting yourself up to fail)

Suppose the likelihood pushes the estimate near $\hat{\theta}_n \approx 5$, but your prior is supported only on $(1, 4)$.

Then:

- The posterior *cannot* concentrate near 5 (it has zero mass there).
- No matter how much data you get, the prior prevents the posterior from matching the MLE asymptotics.

Moral

A prior can be "weak", but it must not *rule out* what the likelihood wants.

Frequentist asymptotics you already know (MLE is Normal)

Under standard regularity conditions:

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} \mathcal{N}\big(0,\ I(\theta_0)^{-1}\big),$$

where $I(\theta_0)$ is the Fisher information (per observation).

Frequentist asymptotics you already know (MLE is Normal)

Under standard regularity conditions:

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} \mathcal{N}\big(0,\ I(\theta_0)^{-1}\big),$$

where $I(\theta_0)$ is the Fisher information (per observation).
Equivalently, for large n:

$$\hat{\theta}_n \approx \mathcal{N}\Big(\theta_0,\ \frac{1}{n} I(\theta_0)^{-1}\Big).$$

Frequentist asymptotics you already know (MLE is Normal)

Under standard regularity conditions:

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} \mathcal{N}\big(0,\ I(\theta_0)^{-1}\big),$$

where $I(\theta_0)$ is the Fisher information (per observation).
Equivalently, for large n:

$$\hat{\theta}_n \approx \mathcal{N}\Big(\theta_0,\ \frac{1}{n} I(\theta_0)^{-1}\Big).$$

BvM connection

BvM says the **posterior** itself becomes approximately

$$\theta \mid Y \approx \mathcal{N}\Big(\hat{\theta}_n,\ \frac{1}{n} I(\hat{\theta}_n)^{-1}\Big),$$

so Bayesian uncertainty matches frequentist uncertainty.

Where does the Normal shape come from? (Laplace approximation intuition)

Key idea: Taylor expand the log-posterior near the mode.

Where does the Normal shape come from? (Laplace approximation intuition)

Key idea: Taylor expand the log-posterior near the mode.
Let

$$\log \pi(\theta \mid Y) = \ell_n(\theta) + \log \pi(\theta) + \text{constant}.$$

Where does the Normal shape come from? (Laplace approximation intuition)

Key idea: Taylor expand the log-posterior near the mode.
Let

$$\log \pi(\theta \mid Y) = \ell_n(\theta) + \log \pi(\theta) + \text{constant}.$$

Expand around $\hat{\theta}_n$ (the likelihood mode; under mild conditions also near the posterior mode):

$$\ell_n(\theta) \approx \ell_n(\hat{\theta}_n) - \frac{1}{2}(\theta - \hat{\theta}_n)^\top \left[-\nabla^2 \ell_n(\hat{\theta}_n)\right] (\theta - \hat{\theta}_n).$$

Where does the Normal shape come from? (Laplace approximation intuition)

Key idea: Taylor expand the log-posterior near the mode.
Let

$$\log \pi(\theta \mid Y) = \ell_n(\theta) + \log \pi(\theta) + \text{constant}.$$

Expand around $\hat{\theta}_n$ (the likelihood mode; under mild conditions also near the posterior mode):

$$\ell_n(\theta) \approx \ell_n(\hat{\theta}_n) - \frac{1}{2}(\theta - \hat{\theta}_n)^\top \left[-\nabla^2 \ell_n(\hat{\theta}_n)\right] (\theta - \hat{\theta}_n).$$

If $-\nabla^2 \ell_n(\hat{\theta}_n) \approx n\, I(\hat{\theta}_n)$, then

$$\pi(\theta \mid Y) \text{ looks like } \exp\left(-\frac{1}{2}(\theta - \hat{\theta}_n)^\top (nI(\hat{\theta}_n))(\theta - \hat{\theta}_n)\right),$$

i.e. a Normal density.

So what does BvM *actually* buy us?

- **Posterior mean** $\mathbb{E}[\theta \mid Y]$ is close to $\hat{\theta}_n$ for large n.
- **Posterior credible intervals** are close to **frequentist confidence intervals**.
- **Prior sensitivity decreases** as n grows (under the regularity conditions).

So what does BvM *actually* buy us?

- **Posterior mean** $\mathbb{E}[\theta \mid Y]$ is close to $\hat{\theta}_n$ for large n.
- **Posterior credible intervals** are close to **frequentist confidence intervals**.
- **Prior sensitivity decreases** as n grows (under the regularity conditions).

But keep the caveats in mind

BvM can fail when:

- the model is misspecified,
- the dimension grows with n (high-dimensional),
- the prior is too rough / places zero mass near the truth,
- parameters are on boundaries / non-regular settings.

One-slide summary (end of Chapter 3)

Takeaway

If you have **lots of data** and a **well-specified, fixed-dimensional** model with a **reasonable prior**, then

$$\pi(\theta \mid Y) \;\approx\; \mathcal{N}\Big(\hat{\theta}_n,\; (nI(\hat{\theta}_n))^{-1}\Big) \quad \text{and} \quad \mathrm{TV}(\text{posterior}, \text{Normal approx}) \to 0.$$

One-slide summary (end of Chapter 3)

Takeaway

If you have **lots of data** and a **well-specified, fixed-dimensional** model with a **reasonable prior**, then

$$\pi(\theta \mid Y) \;\approx\; \mathcal{N}\Big(\hat{\theta}_n,\; (nI(\hat{\theta}_n))^{-1}\Big) \quad \text{and} \quad \mathrm{TV}(\text{posterior}, \text{Normal approx}) \to 0.$$

Bayes and MLE agree (approximately), eventually.

Motivation for Chapters 4–5: why sampling?

So far, many examples had **closed-form** posteriors (conjugacy).

Motivation for Chapters 4–5: why sampling?

So far, many examples had **closed-form** posteriors (conjugacy).
But in realistic Bayesian models:

- the posterior normalising constant is unknown,
- the posterior has no recognisable distribution family,
- the posterior may be high-dimensional.

Motivation for Chapters 4–5: why sampling?

So far, many examples had **closed-form** posteriors (conjugacy).
But in realistic Bayesian models:

- the posterior normalising constant is unknown,
- the posterior has no recognisable distribution family,
- the posterior may be high-dimensional.

Core problem

We often cannot compute $\mathbb{E}[g(\theta) \mid Y]$ analytically.

$$\mathbb{E}[g(\theta) \mid Y] = \int g(\theta)\, \pi(\theta \mid Y)\, d\theta \quad \text{(hard!)}$$

Motivation for Chapters 4–5: why sampling?

So far, many examples had **closed-form** posteriors (conjugacy).
But in realistic Bayesian models:

- the posterior normalising constant is unknown,
- the posterior has no recognisable distribution family,
- the posterior may be high-dimensional.

Core problem

We often cannot compute $\mathbb{E}[g(\theta) \mid Y]$ analytically.

$$\mathbb{E}[g(\theta) \mid Y] = \int g(\theta)\, \pi(\theta \mid Y)\, d\theta \quad \text{(hard!)}$$

Solution idea: approximate the integral by **sampling**.

Example: hierarchical model (posterior gets messy)

Typical hierarchical structure:

$$Y_i \mid \lambda \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda), \qquad \lambda \mid \gamma \sim \text{Gamma}(\text{shape},\ \gamma), \qquad \gamma \sim \text{Exponential}(\nu).$$

Example: hierarchical model (posterior gets messy)

Typical hierarchical structure:

$$Y_i \mid \lambda \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda), \qquad \lambda \mid \gamma \sim \text{Gamma}(\text{shape},\ \gamma), \qquad \gamma \sim \text{Exponential}(\nu).$$

- You might derive *conditional* distributions (Gibbs steps).
- But the **joint posterior** $\pi(\lambda, \gamma \mid Y)$ can be unpleasant.
- You may not be able to sample directly from the joint in one go.

Example: hierarchical model (posterior gets messy)

Typical hierarchical structure:

$$Y_i \mid \lambda \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda), \qquad \lambda \mid \gamma \sim \text{Gamma}(\text{shape},\ \gamma), \qquad \gamma \sim \text{Exponential}(\nu).$$

- You might derive *conditional* distributions (Gibbs steps).
- But the **joint posterior** $\pi(\lambda, \gamma \mid Y)$ can be unpleasant.
- You may not be able to sample directly from the joint in one go.

This is exactly where Chapter 5 lives

We will use **Markov chain Monte Carlo (MCMC)** to sample from complicated posteriors.

Example: Pareto likelihood + Exponential prior (no nice posterior)

Pareto likelihood (shape parameter β):

$$p(y \mid \beta) = \frac{\beta}{y^{\beta+1}} \qquad (y \geq 1,\ \beta > 0 \text{ in a standard Pareto set-up}).$$

Example: Pareto likelihood + Exponential prior (no nice posterior)

Pareto likelihood (shape parameter β):

$$p(y \mid \beta) = \frac{\beta}{y^{\beta+1}} \qquad (y \geq 1,\ \beta > 0 \text{ in a standard Pareto set-up}).$$

Prior: $\beta \sim \text{Exponential}(0.01)$:

$$\pi(\beta) = 0.01\, e^{-0.01\beta}, \qquad \beta > 0.$$

Example: Pareto likelihood + Exponential prior (no nice posterior)

Pareto likelihood (shape parameter β):

$$p(y \mid \beta) = \frac{\beta}{y^{\beta+1}} \qquad (y \geq 1,\ \beta > 0 \text{ in a standard Pareto set-up}).$$

Prior: $\beta \sim \text{Exponential}(0.01)$:

$$\pi(\beta) = 0.01\, e^{-0.01\beta}, \qquad \beta > 0.$$

Posterior kernel (up to proportionality):

$$\pi(\beta \mid y) \propto p(y \mid \beta)\, \pi(\beta) \propto \beta\, y^{-(\beta+1)}\, e^{-0.01\beta}.$$

Example: Pareto likelihood + Exponential prior (no nice posterior)

Pareto likelihood (shape parameter β):

$$p(y \mid \beta) = \frac{\beta}{y^{\beta+1}} \qquad (y \geq 1,\ \beta > 0 \text{ in a standard Pareto set-up}).$$

Prior: $\beta \sim \text{Exponential}(0.01)$:

$$\pi(\beta) = 0.01\, e^{-0.01\beta}, \qquad \beta > 0.$$

Posterior kernel (up to proportionality):

$$\pi(\beta \mid y) \propto p(y \mid \beta)\, \pi(\beta) \propto \beta\, y^{-(\beta+1)}\, e^{-0.01\beta}.$$

Point

This is a perfectly valid posterior density, but it is not a standard named family.
Sampling-based methods become the practical tool.

Chapter 4 vs Chapter 5 (what is coming next)

Chapter 4: Sampling basics

- What is random number generation?
- How do we sample from simple distributions?
- Inverse transform sampling, rejection sampling, etc.

Chapter 4 vs Chapter 5 (what is coming next)

Chapter 4: Sampling basics

- What is random number generation?
- How do we sample from simple distributions?
- Inverse transform sampling, rejection sampling, etc.

Chapter 5: MCMC (Bayesian workhorse)

- Construct a Markov chain whose stationary distribution is $\pi(\theta \mid Y)$.
- Use the chain to generate (dependent) samples.
- Compute Monte Carlo approximations of posterior means/intervals.

Random numbers: the first step (Uniform(0,1))

Almost every sampling method starts here:

$$U \sim \mathrm{Uniform}(0, 1).$$

Random numbers: the first step (Uniform(0,1))

Almost every sampling method starts here:

$$U \sim \mathrm{Uniform}(0, 1).$$

If we can generate Uniform$(0, 1)$ samples, we can transform them into samples from many other distributions.

Random numbers: the first step (Uniform(0,1))

Almost every sampling method starts here:

$$U \sim \text{Uniform}(0, 1).$$

If we can generate Uniform$(0, 1)$ samples, we can transform them into samples from many other distributions.

But how do computers generate Uniform(0,1)?

Two broad philosophies:

- **Truly random:** based on a physical process (hardware randomness).
- **Pseudo-random:** deterministic algorithm that *looks* random.

Truly random vs pseudo-random (conceptual)

Truly random

- Comes from **physical noise**:
  - radioactive decay timing,
  - thermal noise,
  - atmospheric noise,
  - hardware entropy sources.
- Hard to model/predict.
- Harder to build, validate, and deploy at scale.

Pseudo-random

- Generated by a **deterministic algorithm**.
- Given the same seed $\Rightarrow$ same sequence.
- Designed to pass statistical tests for "random-looking" behaviour.
- Extremely convenient for science:
  - reproducibility,
  - debugging,
  - controlled experiments.

What R typically does (pseudo-random numbers)

In R, functions like `runif()`, `rnorm()`, `rexp()` generate **pseudo-random** numbers.

What R typically does (pseudo-random numbers)

In R, functions like `runif()`, `rnorm()`, `rexp()` generate **pseudo-random** numbers.

Key idea: seed $\rightarrow$ sequence

- A seed initialises the generator.
- Same seed $\Rightarrow$ identical output sequence.
- Different seed $\Rightarrow$ different (but still deterministic) sequence.

What R typically does (pseudo-random numbers)

In R, functions like `runif()`, `rnorm()`, `rexp()` generate **pseudo-random** numbers.

Key idea: seed $\rightarrow$ sequence

- A seed initialises the generator.
- Same seed $\Rightarrow$ identical output sequence.
- Different seed $\Rightarrow$ different (but still deterministic) sequence.

Period (why the repetition is not an issue in practice)

Pseudo-random generators eventually repeat (they are periodic), but modern generators have an *enormous* period, so you will not cycle in any realistic computation.

Why “truly random” is hard (and a fun example)

Generating true randomness needs physics

A computer is deterministic. To get true randomness you typically need:

- noise from hardware circuits,
- quantum/atomic processes,
- other physical entropy sources.

Why “truly random” is hard (and a fun example)

Generating true randomness needs physics

A computer is deterministic. To get true randomness you typically need:

- noise from hardware circuits,
- quantum/atomic processes,
- other physical entropy sources.

LavaRand (the lava lamp story)

A famous (and delightfully dramatic) idea:

- Keep a wall of lava lamps running.
- Take images of the blob patterns.
- Convert pixel intensities / blob locations into random bits.

Because the physical motion is hard to predict precisely, it provides a source of entropy.

What we do in this course (practical stance)

- We will treat the output of a good pseudo-random generator as "random enough" for statistics.
- This is standard practice in simulation, Monte Carlo methods, and Bayesian computation.

What we do in this course (practical stance)

- We will treat the output of a good pseudo-random generator as "random enough" for statistics.
- This is standard practice in simulation, Monte Carlo methods, and Bayesian computation.

Two reasons we like pseudo-randomness

1. **Reproducibility:** set a seed and rerun the exact experiment.
2. **Debugging:** if something goes wrong, you can reproduce the same random stream.

Preview: inverse transform sampling (next lecture)

Suppose we can generate $U \sim \mathrm{Uniform}(0, 1)$.

Preview: inverse transform sampling (next lecture)

Suppose we can generate $U \sim \mathrm{Uniform}(0, 1)$.
Let F be the CDF of a target distribution, and assume F is invertible.

Preview: inverse transform sampling (next lecture)

Suppose we can generate $U \sim \mathrm{Uniform}(0, 1)$.
Let F be the CDF of a target distribution, and assume F is invertible.

Inverse transform sampling

$$X = F^{-1}(U) \quad \Longrightarrow \quad X \sim \text{target distribution with CDF } F.$$

Preview: inverse transform sampling (next lecture)

Suppose we can generate $U \sim \mathrm{Uniform}(0, 1)$.
Let F be the CDF of a target distribution, and assume F is invertible.

Inverse transform sampling

$$X = F^{-1}(U) \quad \Longrightarrow \quad X \sim \text{target distribution with CDF } F.$$

- This gives a generic method to sample from many distributions.
- It is the starting point for more advanced methods when F^{-1} is not available.

(Optional mini-example) Inverse transform for Exponential

If $X \sim \text{Exponential}(\lambda)$, then

$$F(x) = 1 - e^{-\lambda x}, \qquad x \geq 0.$$

(Optional mini-example) Inverse transform for Exponential

If $X \sim \text{Exponential}(\lambda)$, then

$$F(x) = 1 - e^{-\lambda x}, \qquad x \geq 0.$$

Set $U \sim \text{Uniform}(0, 1)$ and solve $U = 1 - e^{-\lambda X}$:

$$e^{-\lambda X} = 1 - U \quad \Rightarrow \quad X = -\frac{1}{\lambda} \log(1 - U).$$

(Optional mini-example) Inverse transform for Exponential

If $X \sim \text{Exponential}(\lambda)$, then

$$F(x) = 1 - e^{-\lambda x}, \qquad x \geq 0.$$

Set $U \sim \text{Uniform}(0, 1)$ and solve $U = 1 - e^{-\lambda X}$:

$$e^{-\lambda X} = 1 - U \quad \Rightarrow \quad X = -\frac{1}{\lambda} \log(1 - U).$$

Since $1 - U \overset{d}{=} U$:

$$X = -\frac{1}{\lambda} \log U \quad \sim \quad \text{Exponential}(\lambda).$$

Wrap-up

Today

- Bernstein–von Mises: posterior $\approx$ Normal around the MLE (with Fisher information covariance).
- Big message: with lots of data and good modelling, Bayes and frequentist methods align.

Wrap-up

Today

- Bernstein–von Mises: posterior $\approx$ Normal around the MLE (with Fisher information covariance).
- Big message: with lots of data and good modelling, Bayes and frequentist methods align.

Next

- Chapter 4: how to generate samples (starting from Uniform$(0, 1)$).
- Chapter 5: MCMC for sampling from complicated posteriors.

Thanks everyone!