DegreeLive

Bachelor of Commerce (B2001) | Level 2 • Part B Major | Monash University

Statistical thinking

Course Code: ETC2420 • Module 3 Notes

Module 3: Bayesian Statistical Thinking & Decision Analysis

1. The Bayesian Paradigm and Bayes' Theorem for Parameters

In classical (frequentist) statistics, an unknown population parameter θ is treated as a fixed, non-random constant; probability statements apply strictly to how data $\mathbf{y}$ would fluctuate across hypothetical repeated samples. Consequently, a frequentist 95% confidence interval does not express a 95% probability that θ lies inside the observed bounds—once observed, the fixed interval either contains θ or does not. Instead, 95% refers to the long-run coverage rate of the interval-generating procedure across repeated sampling.

Bayesian statistical thinking adopts an epistemic view of probability: probability quantifies rational degrees of belief about uncertain quantities given available information. Because θ is unknown, it is represented by a probability distribution before observing data—the **prior distribution** $p(\theta)$—which is updated coherently upon observing data $\mathbf{y}$ into the **posterior distribution** $p(\theta \mid \mathbf{y})$ via Bayes' theorem.

Dimension	Frequentist Inference	Bayesian Inference
Nature of parameter θ	Fixed, unknown constant	Uncertain quantity governed by probability distribution
Role of probability	Long-run relative frequency over repeated samples	Coherent degree of belief / uncertainty given evidence
Data conditioning	Conditions on θ_0; averages over unobserved samples $\mathbf{Y}$	Conditions on observed $\mathbf{y}$; integrates over θ
Interval summary	95% Confidence Interval (long-run coverage of procedure)	95% Credible Interval (direct posterior probability $P(\theta \in [L, U] \mid \mathbf{y}) = 0.95$)
Prior information	Excluded from formal likelihood inference	Explicitly encoded via $p(\theta)$ and checked via sensitivity analysis

1.1 Continuous Bayes' Theorem, Marginal Likelihood, and Sequential Updating

For a continuous parameter $\theta \in \Theta$ with prior density $p(\theta)$ and parametric sampling model (likelihood) $p(\mathbf{y} \mid \theta)$, the posterior density is:

$$p(\theta \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \theta)\, p(\theta)}{p(\mathbf{y})} = \frac{p(\mathbf{y} \mid \theta)\, p(\theta)}{\int_{\Theta} p(\mathbf{y} \mid \theta)\, p(\theta)\, d\theta}$$

Where: $p(\theta \mid \mathbf{y})$ is the posterior density of θ conditional on observed data $\mathbf{y}$; $p(\mathbf{y} \mid \theta) = L(\theta; \mathbf{y})$ is the likelihood function; $p(\theta)$ is the prior density; and $p(\mathbf{y}) = \int_{\Theta} p(\mathbf{y} \mid \theta) p(\theta)\, d\theta$ is the **marginal likelihood** (also called the *evidence* or prior predictive probability), which serves as the normalising constant ensuring $\int_{\Theta} p(\theta \mid \mathbf{y})\, d\theta = 1$.

Because $p(\mathbf{y})$ does not depend on θ, the core Bayesian updating law is frequently written in unnormalised proportionality form:

$$p(\theta \mid \mathbf{y}) \propto p(\mathbf{y} \mid \theta)\, p(\theta)$$

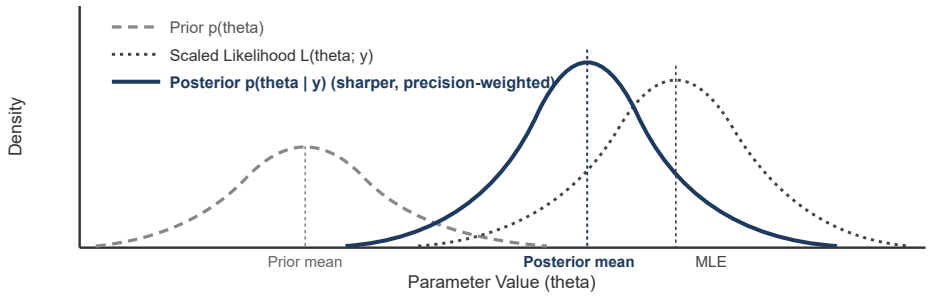
Where: $\propto$ denotes "proportional to" with respect to θ. This expression highlights two foundational principles:

- **The Likelihood Principle:** All experimental evidence relevant to θ resides in the likelihood function $p(\mathbf{y} \mid \theta)$ for the actually observed data $\mathbf{y}$. Unlike frequentist p-values—which depend on unobserved stopping rules (e.g., whether

DegreeLive

an analyst planned fixed $n = 20$ trials or sampled until 5 failures occurred)—Bayesian posterior inference depends solely on the observed likelihood kernel.

- **Sequential Coherence:** If two independent datasets $\mathbf{y}_1$ and $\mathbf{y}_2$ arrive over time, updating $p(\theta)$ with $\mathbf{y}_1$ to obtain $p(\theta \mid \mathbf{y}_1)$ and then using $p(\theta \mid \mathbf{y}_1)$ as the prior for $\mathbf{y}_2$ yields the exact same posterior $p(\theta \mid \mathbf{y}_1, \mathbf{y}_2)$ as analysing both datasets simultaneously. Yesterday's posterior becomes today's prior.
- **Exchangeability and de Finetti's Theorem:** Bayesian models treat sequences of observations $Y_1, Y_2, \ldots$ as *exchangeable* (their joint distribution is invariant under index permutation) rather than strictly independent; de Finetti's representation theorem proves that an infinitely exchangeable sequence behaves mathematically as conditionally i.i.d. given a latent parameter θ endowed with a prior distribution $p(\theta)$.



2. Conjugate Prior Families and Analytical Updating

Computing the marginal likelihood integral $\int_\Theta p(\mathbf{y} \mid \theta)p(\theta)\, d\theta$ is generally impossible in closed form. However, when the prior $p(\theta)$ belongs to a **conjugate family** for a given likelihood $p(\mathbf{y} \mid \theta)$, the posterior $p(\theta \mid \mathbf{y})$ belongs to the exact same parametric family as the prior, reducing Bayesian updating to simple algebraic updates of hyperparameters.

2.1 The Beta-Binomial Model

Suppose we observe y successes in n independent Bernoulli trials with unknown success probability $\theta \in (0, 1)$, so $y \mid \theta \sim \text{Binomial}(n, \theta)$ with likelihood $p(y \mid \theta) \propto \theta^y(1-\theta)^{n-y}$. Assigning a conjugate **Beta prior** $\theta \sim \text{Beta}(\alpha, \beta)$ with shape parameters $\alpha > 0, \beta > 0$:

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}, \qquad \text{E}[\theta] = \frac{\alpha}{\alpha+\beta}, \qquad \text{Var}(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Where: α represents prior pseudo-successes; β represents prior pseudo-failures; and $\alpha + \beta$ represents the **effective prior sample size**. Multiplying the prior by the Binomial likelihood yields:

$$p(\theta \mid y) \propto \theta^y(1-\theta)^{n-y} \cdot \theta^{\alpha-1}(1-\theta)^{\beta-1} = \theta^{(\alpha+y)-1}(1-\theta)^{(\beta+n-y)-1}$$

Where: recognising the kernel of a Beta density gives the exact posterior $\theta \mid y \sim \text{Beta}(\alpha', \beta')$ with updated hyperparameters $\alpha' = \alpha + y$ and $\beta' = \beta + n - y$. Its posterior mean decomposes cleanly into a convex weighted average of the prior mean and the sample proportion $\hat{p} = y/n$:

$$\text{E}[\theta \mid y] = \frac{\alpha+y}{\alpha+\beta+n} = \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)\frac{\alpha}{\alpha+\beta} + \left(\frac{n}{\alpha+\beta+n}\right)\frac{y}{n}$$

Degree Live

Where: $w = \frac{n}{\alpha+\beta+n}$ is the weight placed on the sample data. As $n \to \infty$, $w \to 1$ and the posterior mean converges to the frequentist MLE $\hat{p}$, illustrating how data overwhelm the prior in large samples.

Worked Example 2.1: Bayesian Conversion Rate Estimation for an Australian Retailer

A Melbourne online retailer launches a redesigned mobile product page. Historical launches across similar categories show conversion rates around 10% with moderate uncertainty, encoded as a $\text{Beta}(\alpha = 2, \beta = 18)$ prior (prior mean $2/20 = 0.10$, effective prior sample size 20). During the first day, $n = 80$ visitors view the page and $y = 14$ complete a purchase ($\hat{p} = 14/80 = 0.175$).

Step 1: Update the Beta hyperparameters:

$$\alpha' = 2 + 14 = 16, \qquad \beta' = 18 + (80 - 14) = 84 \implies \theta \mid y \sim \text{Beta}(16, 84)$$

Where: $\alpha' = 16$ and $\beta' = 84$ are the posterior Beta shape parameters summing to 100.

Step 2: Compute posterior mean, posterior mode, and posterior standard deviation:

$$\text{E}[\theta \mid y] = \frac{16}{16 + 84} = 0.160, \qquad \text{Mode}(\theta \mid y) = \frac{16 - 1}{16 + 84 - 2} = \frac{15}{98} \approx 0.1531$$

$$\text{SD}(\theta \mid y) = \sqrt{\frac{16 \times 84}{100^2 \times 101}} = \sqrt{\frac{1{,}344}{1{,}010{,}000}} \approx 0.0365$$

Where: the posterior mean 0.160 shrinks the noisy first-day conversion rate 0.175 slightly toward the prior mean 0.10 via weights $20/100 = 0.20$ on the prior and $80/100 = 0.80$ on the observed sample.

2.2 The Normal-Normal Model (Known and Unknown Variance)

Suppose observations $y_1, \ldots, y_n \mid \mu \overset{\text{i.i.d.}}{\sim} \text{N}(\mu, \sigma^2)$ where the population variance σ^2 is known and we assign a conjugate Normal prior $\mu \sim \text{N}(\mu_0, \tau_0^2)$. Expanding the log-kernel by dropping terms free of μ:

$$\ln p(\mu \mid \mathbf{y}) = -\frac{1}{2}\left[\frac{(\mu - \mu_0)^2}{\tau_0^2} + \frac{\sum_{i=1}^{n}(y_i - \mu)^2}{\sigma^2}\right] + C = -\frac{1}{2}\left[\left(\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}\right)\mu^2 - 2\left(\frac{\mu_0}{\tau_0^2} + \frac{n\bar{y}}{\sigma^2}\right)\mu\right] + C'$$

Where: C and C' denote constants independent of μ. Matching coefficients of μ^2 and μ to the quadratic form $-\frac{1}{2\tau_n^2}(\mu - \mu_n)^2$ yields the Normal posterior $\mu \mid \mathbf{y} \sim \text{N}(\mu_n, \tau_n^2)$ with:

$$\frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}, \qquad \mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} = w\,\bar{y} + (1 - w)\,\mu_0$$

Where: μ_n is the posterior mean; τ_n^2 is the posterior variance; $\kappa_0 = 1/\tau_0^2$ is prior precision; $\kappa_{\text{data}} = n/\sigma^2$ is sample-mean precision; and $w = \kappa_{\text{data}}/(\kappa_0 + \kappa_{\text{data}})$. Thus *posterior precision equals prior precision plus data precision*. When both μ and σ^2 are unknown, assigning a conjugate Normal-Inverse-Gamma prior yields a marginal posterior for μ that follows a scaled Student-t distribution with heavier tails reflecting variance uncertainty.

DegreeLive

Worked Example 2.2: Updating Australian Inflation Expectations

Before an ABS Consumer Price Index release, an economist's prior for underlying annual inflation μ (in %) is $\mathrm{N}\,(\mu_0 = 3.00,\ \tau_0^2 = 0.50^2 = 0.25)$. A survey of $n = 25$ regional price indices yields a sample mean $\bar{y} = 3.45\%$ with known measurement standard deviation $\sigma = 1.00\%$ ($\sigma^2 = 1.00$).

Step 1: Compute prior precision and data precision:

$$\kappa_0 = \frac{1}{0.25} = 4.0, \qquad \kappa_{\text{data}} = \frac{25}{1.00} = 25.0, \qquad \kappa_n = 4.0 + 25.0 = 29.0$$

Where: $\kappa_0 = 4.0$ is prior precision; $\kappa_{\text{data}} = 25.0$ is sample precision; and $\kappa_n = 29.0$ is posterior precision.

Step 2: Compute posterior variance τ_n^2, posterior SD τ_n, and posterior mean μ_n:

$$\tau_n^2 = \frac{1}{29.0} \approx 0.03448 \implies \tau_n \approx 0.1857\%, \qquad \mu_n = \frac{4.0(3.00) + 25.0(3.45)}{29.0} = \frac{12.00 + 86.25}{29.0} \approx 3.388\%$$

Where: the posterior distribution is $\mu \mid \mathbf{y} \sim \mathrm{N}\,(3.388\%,\ 0.1857^2)$. Using $z_{0.975} = 1.96$, the 95% posterior credible interval is $3.388 \pm 1.96(0.1857) = [3.024\%,\ 3.752\%]$.

2.3 Gamma-Poisson and Dirichlet-Multinomial Conjugacy

For count data $y_1, \ldots, y_n \mid \lambda \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda)$ with rate $\lambda > 0$, the conjugate prior is $\lambda \sim \text{Gamma}(\alpha, \beta)$ (shape α, rate β). Multiplying $p(\lambda) \propto \lambda^{\alpha-1} e^{-\beta\lambda}$ by $L(\lambda; \mathbf{y}) \propto \lambda^{\sum y_i} e^{-n\lambda}$ yields the posterior $\lambda \mid \mathbf{y} \sim \text{Gamma}(\alpha + \sum_{i=1}^{n} y_i,\ \beta + n)$. Here α acts as prior total events observed across β prior exposure periods. For instance, if an Australian hospital triage desk starts with $\text{Gamma}(6, 2)$ (prior mean 3.0 cases/hour over 2 pseudo-hours) and records 34 cases over $n = 10$ hours, the posterior is $\text{Gamma}(40, 12)$, giving posterior mean $40/12 \approx 3.33$ cases/hour and posterior variance $40/144 \approx 0.2778$.

When outcomes span $K \geq 3$ mutually exclusive categories—such as market shares across three Australian grocery chains—the Binomial generalises to $\text{Multinomial}(n; \theta_1, \ldots, \theta_K)$ with $\sum_{k=1}^{K} \theta_k = 1$, and the Beta prior generalises to the conjugate **Dirichlet prior** $\boldsymbol{\theta} \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K)$. Observing category counts $(y_1, \ldots, y_K)$ yields the posterior $\boldsymbol{\theta} \mid \mathbf{y} \sim \text{Dirichlet}(\alpha_1 + y_1, \ldots, \alpha_K + y_K)$, where each marginal $\theta_k \mid \mathbf{y}$ follows $\text{Beta}(\alpha_k + y_k,\ \sum_{j \neq k}(\alpha_j + y_j))$.

Worked Example 2.3: Dirichlet-Multinomial Market-Share Updating

A retail analyst models primary grocery-chain shares across three Australian competitors (Chain A, Chain B, and Independent/Other) using a historical prior $\boldsymbol{\theta} \sim \text{Dirichlet}(\alpha_1 = 10,\ \alpha_2 = 10,\ \alpha_3 = 5)$, reflecting effective prior sample size $\alpha_0 = 25$ and prior means $(0.40, 0.40, 0.20)$. A new Melbourne household survey of $n = 100$ shoppers records primary choices $(y_1 = 48,\ y_2 = 34,\ y_3 = 18)$.

Step 1: Update the Dirichlet concentration vector by adding observed counts:

$$\boldsymbol{\theta} \mid \mathbf{y} \sim \text{Dirichlet}(10 + 48,\ \ 10 + 34,\ \ 5 + 18) = \text{Dirichlet}(58,\ 44,\ 23)$$

Where: the posterior total concentration is $\alpha_0' = 58 + 44 + 23 = 125$.

Step 2: Compute posterior mean shares and marginal 95% approximations via $\text{SD}(\theta_k \mid \mathbf{y}) = \sqrt{\frac{\alpha_k'(\alpha_0' - \alpha_k')}{(\alpha_0')^2(\alpha_0' + 1)}}$:

- Chain A: $\mathrm{E}[\theta_1 \mid \mathbf{y}] = 58/125 = 0.464$, with SD $= \sqrt{(58 \times 67)/(125^2 \times 126)} \approx 0.0444$.
- Chain B: $\mathrm{E}[\theta_2 \mid \mathbf{y}] = 44/125 = 0.352$, with SD $= \sqrt{(44 \times 81)/(125^2 \times 126)} \approx 0.0426$.

DegreeLive

- Independent/Other: $\mathrm{E}[\theta_3 \mid \mathbf{y}] = 23/125 = 0.184$, with SD $= \sqrt{(23 \times 102)/(125^2 \times 126)} \approx 0.0345$.

To evaluate the posterior probability $P(\theta_1 > \theta_2 \mid \mathbf{y})$ that Chain A leads Chain B, one simply draws $10{,}000$ vectors from $\mathrm{Dirichlet}(58, 44, 23)$ and counts the fraction of draws satisfying $\theta_1^{(m)} > \theta_2^{(m)}$ (roughly 93.6%).

3. Prior Specification, Credible Intervals, and Computational Bayes

3.1 Informative, Weakly Informative, and Non-Informative Priors

Selecting a prior requires transparency and domain judgement. **Informative priors** encode credible historical evidence or physical constraints (such as the $\mathrm{Beta}(2, 18)$ prior grounded in historical conversion records). **Weakly informative priors** gently regularise extreme or implausible parameter values without dominating moderate datasets (for example, placing $\mathrm{N}(0, 2.5^2)$ on log-odds regression slopes prevents separation explosions in logistic regression). **Non-informative priors** aim for minimal prior influence, such as flat priors $p(\theta) \propto 1$ or **Jeffreys' prior** $p_J(\theta) \propto \sqrt{I(\theta)}$, which is invariant under smooth reparameterisations (for Binomial p, Jeffreys' prior is $\mathrm{Beta}(0.5, 0.5)$).

Regardless of choice, responsible analysts perform **prior sensitivity analysis** by re-computing posteriors under alternative plausible priors to verify whether conclusions remain robust. Returning to Worked Example 2.1 ($y = 14$ conversions out of $n = 80$), the table below illustrates how three distinct priors produce converging posteriors once 80 observations are observed:

Prior Specification	Prior Mean	Effective Prior n_0	Posterior Distribution	Posterior Mean	95% ETI Credible Interval
Uniform / Uninformative $\mathrm{Beta}(1, 1)$	0.500	2	$\mathrm{Beta}(15, 67)$	0.1829	$[0.107,\ 0.273]$
Jeffreys' Prior $\mathrm{Beta}(0.5, 0.5)$	0.500	1	$\mathrm{Beta}(14.5, 66.5)$	0.1790	$[0.104,\ 0.269]$
Historical Informative $\mathrm{Beta}(2, 18)$	0.100	20	$\mathrm{Beta}(16, 84)$	0.1600	$[0.096,\ 0.238]$
Strong Sceptical $\mathrm{Beta}(4, 76)$	0.050	80	$\mathrm{Beta}(18, 142)$	0.1125	$[0.068,\ 0.166]$

3.2 Equal-Tailed vs. Highest Posterior Density (HPD) Credible Intervals

A $(1 - \alpha)$ **credible interval** $[L, U]$ satisfies $P(L \le \theta \le U \mid \mathbf{y}) = 1 - \alpha$ directly under the posterior distribution. Two constructions are standard:

- **Equal-Tailed Interval (ETI):** Places $\alpha/2$ probability in each tail using the $\alpha/2$ and $1 - \alpha/2$ quantiles of the posterior CDF, $[F^{-1}_{\theta|\mathbf{y}}(\alpha/2),\ F^{-1}_{\theta|\mathbf{y}}(1 - \alpha/2)]$. ETIs are easy to compute and invariant under monotone parameter transformations (e.g., transforming σ^2 to σ).
- **Highest Posterior Density (HPD) Interval:** Selects the narrowest possible region $C = \{\theta : p(\theta \mid \mathbf{y}) \ge k_\alpha\}$ containing $1 - \alpha$ posterior mass, so every parameter value inside the HPD interval has strictly higher posterior density than every value outside it.

For symmetric unimodal posteriors such as the Normal-Normal posterior, the ETI and HPD intervals are identical. When the posterior is strongly skewed (such as a $\mathrm{Beta}(2, 18)$ posterior with mode 0.0556 and mean 0.10), the 95% ETI is $[0.0126,\ 0.2565]$ (width 0.2439), whereas the 95% HPD interval is $[0.0040,\ 0.2335]$ (width 0.2295), shifting toward the mode to achieve the shortest interval capturing 95% probability.

DegreeLive

3.3 Posterior Predictive Distribution

Decision-makers rarely care solely about a latent parameter θ; they want to forecast a new, unseen future observation $\tilde{y}$. Rather than plugging in a single point estimate $\hat{\theta}$—which dangerously understates uncertainty by pretending $\theta = \hat{\theta}$ is known with certainty—Bayesian forecasting averages the future sampling distribution $p(\tilde{y} \mid \theta)$ over the entire posterior $p(\theta \mid \mathbf{y})$:

$$p(\tilde{y} \mid \mathbf{y}) = \int_{\Theta} p(\tilde{y} \mid \theta)\, p(\theta \mid \mathbf{y})\, d\theta$$

Where: $p(\tilde{y} \mid \mathbf{y})$ is the **posterior predictive distribution**; $p(\tilde{y} \mid \theta)$ is the sampling density of a new observation given θ; and $p(\theta \mid \mathbf{y})$ is the posterior distribution reflecting remaining parameter uncertainty. In the Normal-Normal model ($\mu \mid \mathbf{y} \sim \mathrm{N}(\mu_n, \tau_n^2)$, $\tilde{y} \mid \mu \sim \mathrm{N}(\mu, \sigma^2)$), the predictive distribution is $\tilde{y} \mid \mathbf{y} \sim \mathrm{N}(\mu_n,\ \sigma^2 + \tau_n^2)$. Using Worked Example 2.2 ($\mu_n = 3.388\%$, $\tau_n^2 = 0.0345$, $\sigma^2 = 1.00$), predictive variance is 1.0345 (SD $= 1.017\%$), yielding a 95% prediction interval $3.388 \pm 1.96(1.017) = [1.39\%,\ 5.38\%]$.

3.4 Grid Approximation and Markov Chain Monte Carlo (MCMC)

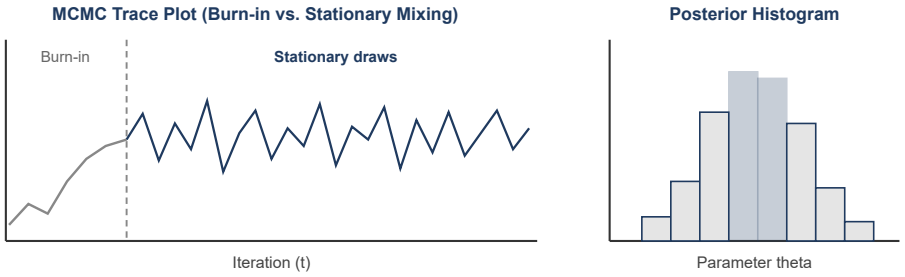
Outside conjugate families (such as Bayesian logistic regression or hierarchical models), the denominator integral $p(\mathbf{y})$ lacks a closed form. In one or two dimensions, **grid approximation** evaluates $p(\mathbf{y} \mid \theta_g)p(\theta_g)$ on a fine discrete grid $\theta_1, \ldots, \theta_G$ and divides by the sum. In higher dimensions, grid size explodes exponentially, so modern Bayesian software uses **Markov Chain Monte Carlo (MCMC)** algorithms—most famously the **Metropolis-Hastings algorithm**:

1. Initialise $\theta^{(0)}$ at a plausible starting value.
2. At iteration t, propose a candidate $\theta^* \sim q(\theta^* \mid \theta^{(t)})$ from a proposal distribution (e.g., symmetric random walk $\mathrm{N}(\theta^{(t)}, s^2)$).
3. Compute the acceptance ratio:

$$\alpha_{\mathrm{MH}} = \min\left(1,\ \frac{p(\mathbf{y} \mid \theta^*)\, p(\theta^*)\, q(\theta^{(t)} \mid \theta^*)}{p(\mathbf{y} \mid \theta^{(t)})\, p(\theta^{(t)})\, q(\theta^* \mid \theta^{(t)})}\right)$$

   Where: the intractable marginal likelihood $p(\mathbf{y})$ cancels identically between numerator and denominator.
4. Draw $u \sim \mathrm{Uniform}(0, 1)$. Set $\theta^{(t+1)} = \theta^*$ if $u \le \alpha_{\mathrm{MH}}$; otherwise retain $\theta^{(t+1)} = \theta^{(t)}$.

After discarding an initial **burn-in** period and verifying chain convergence via trace plots, multiple-chain Gelman-Rubin $\hat{R} \approx 1.00$, and adequate **Effective Sample Size (ESS)** (accounting for autocorrelation), the retained MCMC draws approximate independent samples from $p(\theta \mid \mathbf{y})$.



DegreeLive

4. Bayesian Hypothesis Testing and Bayes Factors

Frequentist p-values measure the probability of obtaining data at least as extreme as observed *assuming* H_0 is true, without evaluating H_1 explicitly. Bayesian hypothesis testing compares two competing models M_0 and M_1 symmetrically by updating prior model odds into posterior model odds via the **Bayes Factor**.

4.1 Posterior Odds, Marginal Likelihood, and Jeffreys' Scale

Assigning prior probabilities $P(M_0)$ and $P(M_1) = 1 - P(M_0)$ to two competing models, applying Bayes' rule to each model and taking their ratio eliminates $p(\mathbf{y})$:

$$\frac{P(M_1 \mid \mathbf{y})}{P(M_0 \mid \mathbf{y})} = \underbrace{\frac{p(\mathbf{y} \mid M_1)}{p(\mathbf{y} \mid M_0)}}_{BF_{10}} \times \frac{P(M_1)}{P(M_0)}, \qquad p(\mathbf{y} \mid M_k) = \int_{\Theta_k} p(\mathbf{y} \mid \theta_k, M_k)\, p(\theta_k \mid M_k)\, d\theta_k$$

Where: $P(M_1 \mid \mathbf{y})/P(M_0 \mid \mathbf{y})$ are the posterior odds in favour of M_1; $P(M_1)/P(M_0)$ are the prior odds; and $BF_{10} = p(\mathbf{y} \mid M_1)/p(\mathbf{y} \mid M_0)$ is the **Bayes Factor** comparing the marginal likelihood (average likelihood across the prior) of M_1 against M_0. Unlike a likelihood ratio evaluated at MLEs, the Bayes factor automatically penalises overly complex or diffuse models: spreading prior mass $p(\theta_1 \mid M_1)$ over parameter regions that poorly predict $\mathbf{y}$ lowers the average likelihood $p(\mathbf{y} \mid M_1)$, implementing a built-in Bayesian **Occam's razor**.

Bayes Factor BF_{10}	$\ln(BF_{10})$	Interpretation (Jeffreys / Kass-Raftery Scale)
1 to 3	0 to 1.1	Barely worth mentioning (anecdotal evidence for M_1)
3 to 20	1.1 to 3.0	Positive / moderate evidence for M_1
20 to 150	3.0 to 5.0	Strong evidence for M_1
> 150	> 5.0	Very strong / decisive evidence for M_1

Worked Example 4.1: Exact Bayes Factor for Quality-Control Randomisation Fairness

An Australian polling auditor tests whether a digital randomisation module selects Option A with exact probability $\theta = 0.5$ ($M_0 : \theta = 0.5$) or exhibits bias ($M_1 : \theta \sim \text{Beta}(1,1) = \text{Uniform}(0,1)$), with equal prior model probabilities $P(M_0) = P(M_1) = 0.5$. Out of $n = 20$ trials, $y = 15$ selections are Option A.

Step 1: Compute the marginal likelihood under the point null M_0:

$$p(y = 15 \mid M_0) = \binom{20}{15}(0.5)^{15}(0.5)^5 = 15{,}504 \times (0.5)^{20} = \frac{15{,}504}{1{,}048{,}576} \approx 0.01479$$

Where: $p(y = 15 \mid M_0)$ is the exact Binomial probability at $\theta = 0.5$.

Step 2: Compute the marginal likelihood under the uniform alternative M_1 using $\int_0^1 \theta^y (1-\theta)^{n-y}\, d\theta = \frac{y!(n-y)!}{(n+1)!}$:

$$p(y = 15 \mid M_1) = \binom{20}{15}\int_0^1 \theta^{15}(1-\theta)^5\, d\theta = \frac{20!}{15!\,5!} \cdot \frac{15!\,5!}{21!} = \frac{1}{21} \approx 0.04762$$

Where: remarkably, under a $\text{Beta}(1,1)$ prior, every outcome $y \in \{0, 1, \ldots, 20\}$ has identical marginal likelihood $1/(n+1) = 1/21$.

DegreeLive

Step 3: Compute the Bayes Factor BF_{10} and posterior model probability $P(M_1 \mid y)$:

$$BF_{10} = \frac{0.04762}{0.01479} \approx 3.22, \qquad P(M_1 \mid y) = \frac{BF_{10}}{1 + BF_{10}} = \frac{3.22}{4.22} \approx 0.763$$

Where: $BF_{10} = 3.22$ provides only moderate evidence for bias, elevating $P(M_1 \mid y)$ from 0.50 to 76.3% while leaving 23.7% posterior probability on M_0. By contrast, a two-sided frequentist Binomial test yields $p = 0.0414 < 0.05$, illustrating how $p < 0.05$ often corresponds to modest Bayes factors—a discrepancy connected to **Lindley's paradox**, where very diffuse priors under M_1 can cause $BF_{10} \to 0$ even when a frequentist p-value is small.

4.2 Savage-Dickey Density Ratio and Region of Practical Equivalence (ROPE)

When $M_0 : \theta = \theta_0$ is nested inside a continuous alternative M_1 with prior $p(\theta \mid M_1)$, the **Savage-Dickey density ratio** simplifies computation dramatically: the Bayes factor in favour of the null equals the ratio of the posterior density to the prior density under M_1 evaluated at θ_0:

$$BF_{01} = \frac{p(\theta = \theta_0 \mid \mathbf{y}, M_1)}{p(\theta = \theta_0 \mid M_1)}$$

Where: if observing data increases the density at θ_0 relative to the prior ($BF_{01} > 1$), the data support M_0; if the posterior moves away from θ_0 so $p(\theta_0 \mid \mathbf{y}, M_1) < p(\theta_0 \mid M_1)$, the data favour M_1 ($BF_{10} = 1/BF_{01} > 1$). Notice in Worked Example 4.1 that $p(\theta = 0.5 \mid M_1) = 1$ and the posterior $\text{Beta}(16, 6)$ density at 0.5 is $\frac{21!}{15!\,5!}(0.5)^{15}(0.5)^5 = 21 \times 0.01479 = 0.3106$, giving $BF_{01} = 0.3106$ and $BF_{10} = 1/0.3106 = 3.22$ identically!

For applied estimation when exact point-null equality $\theta = \theta_0$ is unrealistic, Bayesians frequently specify a **Region of Practical Equivalence (ROPE)**—a narrow interval $[\theta_0 - \epsilon, \theta_0 + \epsilon]$ around zero representing economically negligible effects—and report the posterior probability $P(\theta \in \text{ROPE} \mid \mathbf{y})$ alongside the 95% HPD interval.

5. Bayesian Decision Theory and Optimal Actions Under Uncertainty

Statistical inference ultimately informs action. Whether deciding how much inventory a Sydney retailer should stock, whether a bank should approve a commercial loan, or whether a health authority should deploy a screening intervention, decisions must be made while θ remains uncertain. **Bayesian decision theory** combines the posterior distribution $p(\theta \mid \mathbf{y})$ with an explicit cost or utility structure to identify the optimal action.

5.1 Loss Functions, Posterior Expected Loss, and Bayes Action

A formal decision problem specifies three elements: (1) an uncertain state of nature $\theta \in \Theta$; (2) an action space $a \in \mathrm{A}$; and (3) a **loss function** $L(\theta, a) \geq 0$ quantifying the penalty incurred when action a is taken and the true state turns out to be θ (equivalently, maximising expected utility $U(\theta, a) = -L(\theta, a)$).

Having observed data $\mathbf{y}$, uncertainty about θ is completely summarised by $p(\theta \mid \mathbf{y})$. The **posterior expected loss** of action a is:

$$\rho(a \mid \mathbf{y}) = \mathrm{E}_{\theta \mid \mathbf{y}}[L(\theta, a)] = \int_{\Theta} L(\theta, a)\, p(\theta \mid \mathbf{y})\, d\theta$$

Where: $\rho(a \mid \mathbf{y})$ is the expected loss of choosing action a averaged over the posterior distribution $p(\theta \mid \mathbf{y})$. A **Bayes optimal action** a^* minimises posterior expected loss:

Degreelive

$$a^* = \arg\min_{a \in \mathrm{A}} \rho(a \mid \mathbf{y})$$

Where: a^* is the optimal point estimate or business decision given both the evidence $\mathbf{y}$ and the asymmetric or symmetric consequences $L(\theta, a)$.

5.2 Standard Loss Functions and Their Bayes Estimators

Different loss functions map the same posterior distribution to different optimal point estimates a^*:

- **Squared Error Loss:** $L(\theta, a) = (\theta - a)^2$. Differentiating $\mathrm{E}[(\theta - a)^2 \mid \mathbf{y}]$ with respect to a and setting to zero yields $a^* = \mathrm{E}[\theta \mid \mathbf{y}]$, the **posterior mean**.
- **Absolute Error Loss:** $L(\theta, a) = |\theta - a|$. Penalises errors linearly rather than quadratically, yielding $a^* = \mathrm{Median}(\theta \mid \mathbf{y})$, the **posterior median**.
- **Zero-One Loss (Discrete Classification):** $L(\theta, a) = \mathrm{I}(\theta \neq a)$. Minimising misclassification probability selects the **posterior mode** (maximum a posteriori estimate).
- **Asymmetric Linear (Newsvendor / Quantile) Loss:** In many commercial settings, under-predicting demand costs c_u per unit (lost margin or stockout penalty), whereas over-predicting costs c_o per unit (spoilage or holding cost):

  $$L(\theta, a) = \begin{cases} c_u(\theta - a) & \text{if } \theta > a \\ c_o(a - \theta) & \text{if } \theta \leq a \end{cases}$$

  Where: $c_u > 0$ is the per-unit underage cost; and $c_o > 0$ is the per-unit overage cost. Minimising $\rho(a \mid \mathbf{y})$ yields the **critical ratio quantile** of the posterior:

  $$a^* = F^{-1}_{\theta|\mathbf{y}}\left(\frac{c_u}{c_u + c_o}\right)$$

  Where: $F^{-1}_{\theta|\mathbf{y}}$ is the posterior quantile function. When underage is four times costlier than overage ($c_u = 4c_o$), the Bayes optimal decision is the 80-th percentile of the posterior—deliberately shading upward above the mean or median to insure against costly shortfalls.

Worked Example 5.1: Optimal Action, EVPI, and Value of Sample Information

An Australian electronics importer inspects a shipment of smart meters. Based on current records, the posterior probability that the batch has a high defect rate (θ_1) is $P(\theta_1 \mid \mathbf{y}) = 0.25$, and that it meets compliance (θ_0) is $P(\theta_0 \mid \mathbf{y}) = 0.75$. Management chooses between three actions: a_1 = Ship without rework; a_2 = Full rework before shipping; a_3 = Targeted audit. The financial loss table (in AUD) is:

Action	Loss if Compliant θ_0 ($P = 0.75$)	Loss if High-Defect θ_1 ($P = 0.25$)	Posterior Expected Loss $\rho(a \mid \mathbf{y})$
a_1: Ship as-is	\$0	\$40,000 (recall penalty)	$0.75(0) + 0.25(40{,}000) = \$10{,}000$
a_2: Full rework	\$12,000	\$12,000	$0.75(12{,}000) + 0.25(12{,}000) = \$12{,}000$
a_3: Targeted audit	\$4,000	\$16,000	$0.75(4{,}000) + 0.25(16{,}000) = \$7{,}000$

Degreelive

Step 1 (Immediate Decision): Compare the posterior expected losses: $\rho(a_1) = \$10{,}000$, $\rho(a_2) = \$12{,}000$, and $\rho(a_3) = \$7{,}000$. Thus the Bayes optimal action is $a^* = a_3$ (targeted audit), with minimum expected cost $\$7{,}000$.

Step 2 (EVPI): Compute the **Expected Value of Perfect Information (EVPI)**—the maximum amount the importer should pay for an infallible oracle revealing θ before acting. With perfect information, if θ_0 is revealed we take a_1 (loss $\$0$); if θ_1 is revealed we take a_2 (loss $\$12{,}000$). Expected loss under perfect information is $0.75(0) + 0.25(12{,}000) = \$3{,}000$, giving:

$$\text{EVPI} = \$7{,}000 - \$3{,}000 = \$4{,}000$$

Where: $\text{EVPI} = \$4{,}000$ is the strict upper bound on what any additional diagnostic data can be worth.

Step 3 (EVSI — Pre-Posterior Sample Evaluation): Suppose an imperfect lab screening test costs $\$1{,}200$ and returns either a Positive signal (S^+) or Negative signal (S^-), with sensitivity $P(S^+ \mid \theta_1) = 0.80$ and false-positive rate $P(S^+ \mid \theta_0) = 0.10$. Should the importer purchase the lab test first?

- Marginal signal probabilities: $P(S^+) = 0.80(0.25) + 0.10(0.75) = 0.275$; $P(S^-) = 1 - 0.275 = 0.725$.
- Posterior given S^+: $P(\theta_1 \mid S^+) = \frac{0.20}{0.275} \approx 0.7273$. Expected losses: $\rho(a_1 \mid S^+) = 0.7273(40{,}000) = \$29{,}091$; $\rho(a_2 \mid S^+) = \$12{,}000$; $\rho(a_3 \mid S^+) = 0.2727(4{,}000) + 0.7273(16{,}000) = \$12{,}727$. Optimal given S^+ is a_2 (Full rework, loss $\$12{,}000$).
- Posterior given S^-: $P(\theta_1 \mid S^-) = \frac{0.20(0.25)}{0.725} = \frac{0.05}{0.725} \approx 0.0690$. Expected losses: $\rho(a_1 \mid S^-) = 0.0690(40{,}000) = \$2{,}759$; $\rho(a_2 \mid S^-) = \$12{,}000$; $\rho(a_3 \mid S^-) = 0.9310(4{,}000) + 0.0690(16{,}000) = \$4{,}828$. Optimal given S^- is a_1 (Ship as-is, loss $\$2{,}759$).
- Pre-posterior expected loss before test cost: $0.275(\$12{,}000) + 0.725(\$2{,}759) = \$3{,}300 + \$2{,}000 = \$5{,}300$.

Thus the **Expected Value of Sample Information (EVSI)** is $\$7{,}000 - \$5{,}300 = \$1{,}700$. Because $\text{EVSI} = \$1{,}700$ exceeds the $\$1{,}200$ lab fee, purchasing the test generates a net expected saving of $\$500$.

5.3 Expected Utility, Risk Aversion, and Certainty Equivalents

When outcomes involve large financial stakes relative to a decision-maker's capital—such as hedging currency exposure or insuring catastrophic bushfire risk—linear dollar loss fails to reflect aversion to severe downside shocks. Von Neumann-Morgenstern **Expected Utility Theory** replaces monetary payoffs w with a strictly increasing, concave utility function $U(w)$ (such as logarithmic utility $U(w) = \ln(w)$ or power utility), selecting the action that maximises posterior expected utility $\mathrm{E}_{\theta|\mathbf{y}}[U(W(\theta, a))]$.

Worked Example 5.2: Expected Monetary Value vs. Risk-Averse Expected Utility

An Australian wine exporter holds initial liquid capital of $\$100{,}000$ and faces a single-shipment foreign-exchange exposure. Under the posterior predictive distribution of the AUD/USD rate, leaving the shipment unhedged (a_1) yields a 50% chance of gaining $+\$80{,}000$ (final wealth $\$180{,}000$) and a 50% chance of losing $-\$50{,}000$ (final wealth $\$50{,}000$). Buying a forward currency hedge (a_2) guarantees a certain net gain of $+\$10{,}000$ (final wealth $\$110{,}000$).

- **Risk-neutral (Expected Monetary Value):** $\text{EMV}(a_1) = 0.5(180{,}000) + 0.5(50{,}000) = \$115{,}000$, which exceeds $\text{EMV}(a_2) = \$110{,}000$. A risk-neutral firm leaves the shipment unhedged.
- **Risk-averse (Logarithmic Utility $U(w) = \ln(w)$):**

$$\mathrm{E}[U(a_1)] = 0.5\ln(180{,}000) + 0.5\ln(50{,}000) = 0.5(12.1007) + 0.5(10.8198) = 11.4602$$

$$U(a_2) = \ln(110{,}000) = 11.6082 > 11.4602$$

Where: the **certainty equivalent** of the unhedged gamble is $\mathrm{CE}(a_1) = \exp(11.4602) \approx \$94{,}868$, falling $\$20{,}132$ below its expected monetary value (a *risk premium* of $\$20{,}132$). Under logarithmic utility, the exporter strictly prefers purchasing the $\$10{,}000$-gain hedge a_2 to avoid the severe downside drop to $\$50{,}000$.

DegreeLive