Lecture Notes

BAYESIAN ANALYSIS

DSA 8505

Strathmore University

Lecturer: Dr. Jacob Ong'ala

Contents

2 Prior Distributions & Likelihood Functions

The **prior distribution** represents the initial beliefs or knowledge about the parameters before any data is observed. It is a probability distribution that quantifies our uncertainty about the parameter based on previous information or subjective belief.

2.1 Formulating or Choosing Prior Distribution

The process of selecting or formulating a prior distribution is crucial because the prior can influence the posterior distribution, especially in situations where data is limited.

Choosing an appropriate prior is one of the key decisions in Bayesian analysis. The selection of priors can be guided by domain knowledge, previous research, or computational considerations. Priors can be classified into several types depending on the amount of prior knowledge and the context of the study.

Guidelines for Choosing Priors

Selecting an appropriate prior involves balancing the following factors:

- **Domain Knowledge**: Use prior information that is relevant to the parameter being estimated.
- **Objectivity**: In some cases, non-informative or weakly informative priors are chosen to maintain objectivity and let the data dominate.
- **Computational Simplicity**: For practical reasons, conjugate priors are often chosen to simplify calculations, especially in single-parameter models.
- **Sensitivity Analysis**: It is good practice to perform sensitivity analysis by trying different priors to evaluate how sensitive the posterior is to the choice of prior.

2.2 Types of Priors

2.2.1 Informative Priors

Informative priors are based on substantial prior knowledge or expert opinion. They incorporate real-world information and are chosen when there is enough prior data or experience available about the parameter of interest. These priors contain substantial information about the parameter based on previous studies, expert knowledge, or other

sources. This information can help incorporate domain knowledge into the analysis, particularly when data is scarce or noisy.

For example, if a parameter θ is known to represent a probability that lies between 0 and 1, we might assign a Beta distribution to represent our prior belief. The Beta distribution is flexible and allows us to encode our confidence in different ranges of θ.

Suppose we use Beta$(2, 5)$ as our prior distribution. This choice indicates that, based on prior knowledge, we expect θ to be closer to 0 than to 1, as the shape parameters (2 and 5) cause the distribution to skew toward lower values. The mean of this Beta distribution is calculated as:

$$\text{Mean} = \frac{\alpha}{\alpha + \beta} = \frac{2}{2+5} = \frac{2}{7} \approx 0.286.$$

The prior variance can also be calculated to measure the spread of our beliefs:

$$\text{Variance} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{2 \times 5}{(2+5)^2(2+5+1)} = \frac{10}{392} \approx 0.0255.$$

This prior could arise from expert opinions, where experts believe that the probability of success is more likely to be small but not negligible. Such priors are particularly useful in fields like medicine, engineering, or environmental studies, where prior knowledge can provide critical context for the analysis.

Informative priors help guide the posterior distribution, especially when the observed data is limited. However, care must be taken to ensure that the prior does not overly dominate the results, particularly when it is subjective or based on limited prior evidence.

Example: Suppose a factory has historically produced 95% defect-free products. Based on this historical information, a Beta distribution Beta$(95, 5)$ can be used as a prior for the probability of producing a defect-free product.

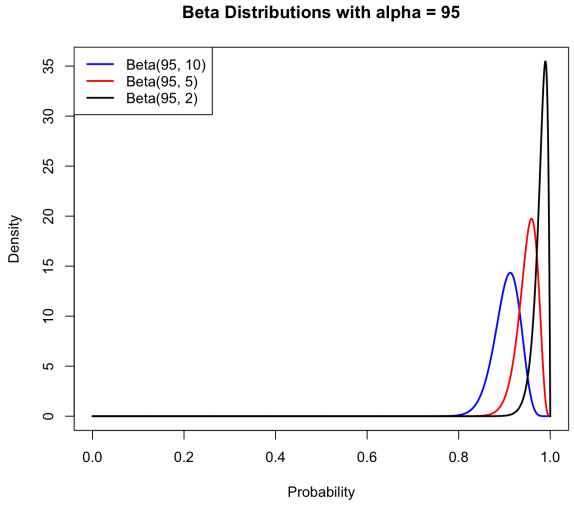

Figure 2.1: prior of Beta(95,β)

2.2.2 Non-informative Priors

Non-informative or vague priors express minimal prior information and are often used when little to no prior knowledge exists about the parameter. They allow the data to

dominate the inference process. Common non-informative priors include uniform distributions over the parameter space.

Example: If we have no strong prior beliefs about the probability of a coin landing heads, we could use a uniform prior over the interval $[0, 1]$:

$$P(\theta) \propto 1 \quad \text{for} \quad \theta \in [0, 1]$$

This prior assumes that all values of θ are equally plausible.

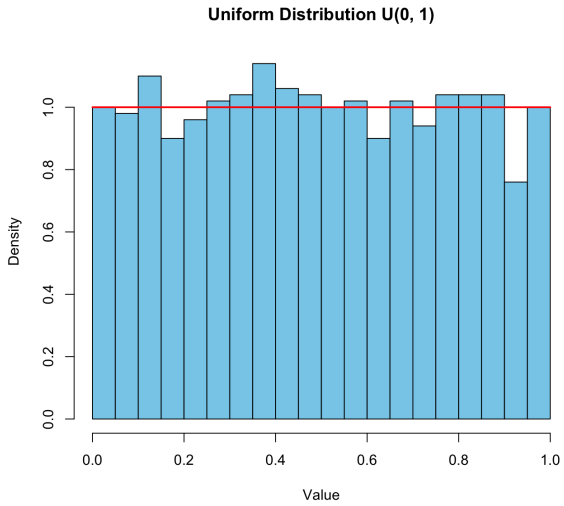

Figure 2.2: Uniform Prior

2.2.3 Weakly Informative Priors

These priors fall between informative and non-informative priors. They reflect some prior knowledge but are designed to have minimal influence on the posterior when sufficient data is available. Weakly informative priors often incorporate conservative assumptions.

Example: In regression modeling, a weakly informative prior might assume that the regression coefficients are near zero without being overly restrictive. A normal prior with a large variance, such as $\beta \sim \mathcal{N}(0, 100^2)$, can serve this purpose.

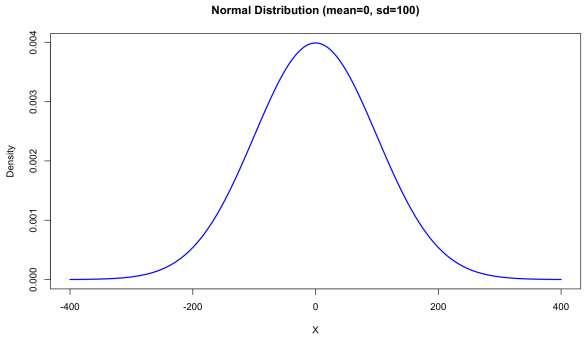

Figure 2.3: Normal prior with hi variance (week informative)

2.2.4 Empirical Priors

Empirical priors are based on previous data or empirical evidence from related studies. These priors are formulated using available data but may not be as subjective as informative priors.

2.2.5 Conjugate Priors

A prior distribution is said to be conjugate to the likelihood function if the posterior distribution is of the same family as the prior. Conjugate priors are mathematically convenient because they simplify the computation of the posterior distribution.

The form of conjugate priors ensures that the posterior distribution can be derived analytically, avoiding the need for complex numerical methods.

2.2.6 Objective vs. Subjective Priors

In Bayesian statistics, the choice of prior can be viewed from two perspectives:

- **Objective Priors**: These are designed to minimize subjectivity. Non-informative or reference priors are examples of objective priors. They are useful in cases where there is little or no prior knowledge.
- **Subjective Priors**: These are based on personal or expert knowledge and are useful in situations where prior information is available and relevant. Informative priors are subjective in nature.

Examples with Beta Priors

In Bayesian inference for binomial or Bernoulli data, the unknown probability parameter $\theta \in (0, 1)$ is commonly assigned a Beta prior distribution due to its **conjugacy** with the Binomial likelihood i.e *when beta prior is chosen in binomial likelihood the beta posterior will be obtained.*

The Beta Distribution as a Prior

Let

$$\theta \sim \text{Beta}(\alpha, \beta),$$

with density

$$p(\theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad 0 < \theta < 1.$$

Key properties of the Beta distribution include:

- **Prior mean:**
$$\mathbb{E}(\theta) = \frac{\alpha}{\alpha + \beta}$$
- **Prior variance:**
$$\text{Var}(\theta) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

- **Prior sample size (strength):**

$$\alpha + \beta$$

The parameters α and β may be interpreted as *prior pseudo-counts*, representing prior successes and failures, respectively.

—

Conjugacy with the Binomial Likelihood

Suppose the data follow a Binomial distribution:

$$Y \mid \theta \sim \text{Binomial}(n, \theta),$$

where Y is the number of observed successes in n trials.

If the prior is

$$\theta \sim \text{Beta}(\alpha, \beta),$$

then the posterior distribution is

$$\theta \mid Y \sim \text{Beta}(\alpha + Y,\ \beta + n - Y).$$

Thus, the posterior parameters are obtained by simply updating the prior parameters with observed data.

—

a) Non-informative Prior: Beta$(1, 1)$

The Beta$(1, 1)$ distribution is equivalent to the Uniform$(0, 1)$ distribution:

$$p(\theta) = 1, \quad 0 < \theta < 1.$$

Properties:

$$\mathbb{E}(\theta) = 0.5, \quad \alpha + \beta = 2$$

Interpretation:

- Represents complete prior ignorance
- All values of θ are equally plausible
- The data dominate the posterior inference

This prior is often used as a baseline for comparison.

—

b) Weakly Informative Prior: Beta$(2, 2)$

The Beta$(2, 2)$ prior is symmetric and mildly concentrated around 0.5.

Properties:

$$\mathbb{E}(\theta) = 0.5, \quad \alpha + \beta = 4$$

Interpretation:

- Slight preference for moderate probabilities

- Extreme values near 0 or 1 are less likely, but still possible
- Provides mild regularization, especially useful for small samples

This prior expresses weak prior belief without overwhelming the data.

—

c) Informative Prior: Beta$(8, 2)$

The Beta$(8, 2)$ prior reflects a belief that success is more likely than failure.

Properties:

$$\mathbb{E}(\theta) = \frac{8}{10} = 0.8, \quad \alpha + \beta = 10$$

Interpretation:

- Equivalent to observing 8 prior successes and 2 prior failures
- Indicates substantial prior knowledge favoring high success probability
- Suitable when historical data or expert opinion is available

This prior will influence the posterior unless the sample size is large.

—

d) Strong Informative Prior: Beta$(95, 5)$

The Beta$(95, 5)$ prior represents very strong prior belief.

Properties:

$$\mathbb{E}(\theta) = 0.95, \quad \alpha + \beta = 100$$

Interpretation:

- Equivalent to 95 prior successes and 5 failures
- Extremely concentrated around $\theta = 0.95$
- Dominates the posterior unless the data sample is very large

Such a prior should only be used when prior information is highly reliable.

—

2.2.7 Illustration: Effect of Different Priors on the Posterior

To illustrate how different prior distributions influence Bayesian inference, we consider a Binomial sampling model.

Model Setup

Let

$$Y \mid \theta \sim \text{Binomial}(n, \theta),$$

where:

- $n = 20$ is the number of trials,
- $y = 12$ is the number of observed successes,
- $\theta \in (0, 1)$ is the probability of success.

The likelihood function is

$$L(\theta \mid y) \propto \theta^{12}(1-\theta)^{8}.$$

We assume a Beta prior:

$$\theta \sim \text{Beta}(\alpha, \beta).$$

Because the Beta distribution is conjugate to the Binomial likelihood, the posterior distribution is:

$$\theta \mid y \sim \text{Beta}(\alpha + y, \beta + n - y).$$

Prior Distributions Considered

We compare four different priors:

1. **Non-informative prior:** $\text{Beta}(1, 1)$
2. **Weakly informative prior:** $\text{Beta}(2, 2)$
3. **Informative prior:** $\text{Beta}(8, 2)$
4. **Strong informative prior:** $\text{Beta}(95, 5)$

Posterior Distributions

Applying Bayes' theorem, the posterior distributions are:

$$\begin{aligned}
\text{Beta}(1 + 12,\ 1 + 8) &= \text{Beta}(13, 9), \\
\text{Beta}(2 + 12,\ 2 + 8) &= \text{Beta}(14, 10), \\
\text{Beta}(8 + 12,\ 2 + 8) &= \text{Beta}(20, 10), \\
\text{Beta}(95 + 12,\ 5 + 8) &= \text{Beta}(107, 13).
\end{aligned}$$

2.2.8 Interpretation

- With non-informative and weakly informative priors, the posterior closely follows the likelihood.
- Informative priors pull the posterior toward prior beliefs.
- Strong priors dominate the posterior, even when the data suggests otherwise.

This example demonstrates that Bayesian inference is a compromise between prior information and observed data.

See Live demonstration HERE

2.3 Likelihood Function

The **likelihood function** is a fundamental concept in statistics, particularly in the context of **Bayesian inference**. It quantifies how well a particular set of observed data supports different possible values of an unknown parameter.

- **Key Concept:** The likelihood function is not the probability of the parameter but the probability of the observed data given a parameter. This distinction is crucial because the likelihood function is a function of the parameter (denoted as θ), and it reflects how plausible different parameter values are, based on the observed data.

Given a set of observed data $D = (x_1, x_2, \ldots, x_n)$ and an unknown parameter θ, the likelihood function is denoted as $P(D|\theta)$. This represents the joint probability of observing the data D, conditioned on the parameter θ.

2.3.1 Mathematical Formulation

If we assume that the data points $x_1, x_2, \ldots, x_n$ are **independent and identically distributed (i.i.d.)**, the likelihood function can be written as the product of the probabilities of each individual observation:

$$P(D|\theta) = P(x_1, x_2, \ldots, x_n|\theta) = \prod_{i=1}^{n} P(x_i|\theta)$$

This formulation assumes that the probability of each data point $P(x_i|\theta)$ is independent of the others. Therefore, the likelihood function aggregates the information from each individual observation to make a judgment about how likely a particular value of θ is, given all of the data.

- **Key Insight:** The likelihood function does not treat θ as a random variable (unlike Bayesian priors or posteriors), but as a fixed, unknown parameter that we aim to infer. It simply quantifies the plausibility of different values of θ based on the observed data.

Likelihood Function vs. Probability Distribution

One common source of confusion is the distinction between a **likelihood function** and a **probability distribution**. While both concepts deal with probabilities, their roles are distinct:

- **Probability Distribution:** Refers to the probability of observing certain data given a fixed parameter. For instance, $P(x_i|\theta)$ represents the probability of observing x_i under the assumption that the parameter θ is known.
- **Likelihood Function:** Refers to the function of the parameter θ given the observed data. In other words, the likelihood function is the probability of observing the given data for different values of θ.

Importantly, while probabilities sum (or integrate) to 1, likelihoods do not necessarily do so. The likelihood function is not constrained to sum to 1 and can be any positive value.

Example 1: Likelihood Function for a Binomial Distribution

Consider a classic example of the likelihood function in the context of a **coin toss experiment**.

- **Scenario:** Suppose we perform 10 coin tosses and observe 7 heads. We want to infer the probability θ that the coin lands heads up. We assume that each toss is independent, and the likelihood of getting heads follows a **Binomial distribution**.

In this case, the likelihood function can be written as:

$$P(D|\theta) = \binom{10}{7}\theta^7(1-\theta)^3$$

This likelihood function shows how different values of θ affect the probability of observing exactly 7 heads out of 10 tosses.

Example 2: Likelihood Function for a Poisson Distribution

Consider an example involving **count data** observed over a fixed period of time.

- **Scenario:** Suppose a call center records the number of customer calls received per hour. Let λ denote the average number of calls per hour. We assume that calls arrive independently and that the number of calls in an hour follows a **Poisson distribution**.

Assume that during a particular hour, the call center receives 6 calls. That is,

$$Y = 6.$$

The probability mass function of the Poisson distribution is:

$$P(Y = y \mid \lambda) = \frac{e^{-\lambda}\lambda^y}{y!}, \quad y = 0, 1, 2, \ldots$$

Treating the observed data $Y = 6$ as fixed, the likelihood function for λ is:

$$P(D \mid \lambda) = \frac{e^{-\lambda}\lambda^6}{6!}.$$

Since the factorial term does not depend on λ, the likelihood function (up to proportionality) can be written as:

$$L(\lambda \mid D) \propto e^{-\lambda}\lambda^6.$$

This likelihood function describes how plausible different values of the rate parameter λ are, given that 6 events were observed in one hour. The likelihood is maximized near $\lambda = 6$, indicating that values of λ close to 6 are most consistent with the observed data.

Example 3: Likelihood Function for a Normal Distribution

Consider an example involving **continuous measurement data**.

- **Scenario:** Suppose a quality control engineer measures the diameter (in millimeters) of a manufactured component. Let μ denote the true mean diameter of the component. Assume that individual measurements are independent and follow a **Normal distribution** with known variance σ^2.

Assume that a single measurement is observed:

$$Y = 52.$$

Further assume that the measurement variability is known and given by:

$$\sigma^2 = 4 \quad (\sigma = 2).$$

The probability density function of the Normal distribution is:

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

Treating the observed data $Y = 52$ as fixed, the likelihood function for μ is:

$$P(D \mid \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(52-\mu)^2}{2\sigma^2}\right).$$

Since the normalizing constant does not depend on μ, the likelihood function (up to proportionality) can be written as:

$$L(\mu \mid D) \propto \exp\left(-\frac{(52-\mu)^2}{2\sigma^2}\right).$$

This likelihood function describes how plausible different values of the mean parameter μ are, given the observed measurement. The likelihood is maximized at $\mu = 52$, indicating that values of μ close to the observed measurement are most consistent with the data.

Click HERE to see illustrations of Likelihood function.

2.3.2 Properties of the Likelihood Function

- **Relative Likelihood:** The likelihood function is primarily useful for comparing the plausibility of different parameter values. For example, if $P(D|\theta_1) > P(D|\theta_2)$, then θ_1 is more plausible than θ_2 given the observed data.
- **Maximum Likelihood Estimate (MLE):** The parameter value that maximizes the likelihood function is known as the **Maximum Likelihood Estimate (MLE)**. This is the value of θ that makes the observed data most probable.
- **Likelihood is not a Probability:** The likelihood function does not sum or integrate to 1 over θ. Instead, it simply represents the relative likelihood of different parameter values.

2.3.3 Marginal Likelihood (Evidence) $P(D)$

The marginal likelihood, or evidence, is the denominator in Bayes' Theorem:

$$P(D) = \int P(D|\theta)P(\theta)d\theta$$

The marginal likelihood serves two purposes:

- **Normalization**: It ensures that the posterior distribution integrates to 1, making it a valid probability distribution.
- **Model Comparison**: In cases where we compare different models, the marginal likelihood can help quantify how well each model explains the observed data. Models with higher marginal likelihood are typically preferred.

Example: Binomial Data with a Beta Prior

Suppose we observe the outcome of $n = 10$ Bernoulli trials and record $y = 7$ successes. Let θ denote the probability of success.

$$Y \mid \theta \sim \text{Binomial}(10, \theta)$$

Assume a Beta prior:

$$\theta \sim \text{Beta}(\alpha, \beta)$$

Step 1: Likelihood The likelihood function is:

$$P(D \mid \theta) = \binom{10}{7}\theta^7(1-\theta)^3.$$

Step 2: Prior The prior density is:

$$P(\theta) = \frac{1}{B(\alpha, \beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

Step 3: Marginal Likelihood The marginal likelihood is:

$$P(D) = \int_0^1 P(D \mid \theta)P(\theta)\, d\theta.$$

Substituting:

$$P(D) = \binom{10}{7}\frac{1}{B(\alpha, \beta)}\int_0^1 \theta^{7+\alpha-1}(1-\theta)^{3+\beta-1}d\theta.$$

Recognizing the Beta integral:

$$\int_0^1 \theta^{a-1}(1-\theta)^{b-1}d\theta = B(a, b),$$

we obtain:

$$P(D) = \binom{10}{7}\frac{B(7+\alpha,\, 3+\beta)}{B(\alpha, \beta)}.$$

Example: Binomial Data with a Beta Prior (Fully Worked)

Suppose we observe $y = 7$ successes from $n = 10$ Bernoulli trials. Let θ denote the probability of success.

$$Y \mid \theta \sim \text{Binomial}(10, \theta)$$

The likelihood function is:

$$P(D \mid \theta) = \binom{10}{7} \theta^7 (1-\theta)^3.$$

We consider three different priors and examine their effects on the posterior distribution and the marginal likelihood.

Case 1: Non-informative Prior

Assume a uniform prior:

$$\theta \sim \text{Beta}(1, 1).$$

Marginal Likelihood Using the marginal likelihood formula:

$$P(D) = \binom{10}{7} \frac{B(8, 4)}{B(1, 1)}.$$

Since $B(1, 1) = 1$, we have:

$$P(D) = \binom{10}{7} B(8, 4) = 0.0909.$$

—

Case 2: Weakly Informative Prior

Assume a weakly informative prior:

$$\theta \sim \text{Beta}(2, 2).$$

Marginal Likelihood The marginal likelihood is:

$$P(D) = \binom{10}{7} \frac{B(9, 5)}{B(2, 2)}.$$

Since:

$$B(2, 2) = \frac{1!1!}{3!} = \frac{1}{6},$$

we obtain:

$$P(D) = 6 \binom{10}{7} B(9, 5).$$

$$\binom{10}{7} = \frac{10!}{7!\,3!} = \frac{10 \times 9 \times 8}{3 \times 2 \times 1} = 120.$$

$$B(a,b) = \frac{(a-1)!(b-1)!}{(a+b-1)!}.$$

Thus,

$$B(9,5) = \frac{(9-1)!(5-1)!}{(9+5-1)!} = \frac{8!\,4!}{13!}.$$

Hence:

$$B(9,5) = \frac{24}{154{,}440} = \frac{1}{6{,}435}.$$

$$P(D) = 6 \times 120 \times \frac{1}{6{,}435}.$$

Thus:

$$P(D) = \frac{720}{6{,}435}.$$

$$P(D) = \frac{48}{429} \approx 0.112.$$

Case 3: Strong Informative Prior

Assume strong prior information suggesting a high probability of success:

$$\theta \sim \text{Beta}(20,5).$$

Marginal Likelihood The marginal likelihood is:

$$P(D) = \binom{10}{7} \frac{B(27,8)}{B(20,5)} = 0.050.$$

This value reflects how well the data are supported by a model that strongly favors high success probabilities.

—

Comparison and Interpretation

Prior Type	Prior	Posterior	Posterior Mean
Non-informative	Beta$(1,1)$	Beta$(8,4)$	0.667
Weakly informative	Beta$(2,2)$	Beta$(9,5)$	0.643
Strong informative	Beta$(20,5)$	Beta$(27,8)$	0.771

Key Observations

- The non-informative prior allows the data to dominate the posterior.
- The weakly informative prior gently shrinks the posterior toward the prior mean.
- The strong prior substantially influences the posterior, pulling it toward prior beliefs.
- The marginal likelihood automatically balances data fit and prior assumptions.