Lecture Notes

BAYESIAN ANALYSIS

DSA 8505

Strathmore University

Lecturer: Dr. Jacob Ong'ala

Contents

1 Introduction to Bayesian Inference

Bayesian inference offers a probabilistic framework for incorporating prior knowledge and updating beliefs in light of new evidence. This approach fundamentally differs from the classical (frequentist) paradigm, which treats parameters as fixed and unknown quantities. Bayesian methods, in contrast, view parameters as random variables described by probability distributions. This distinction provides several advantages:

- **Incorporation of prior knowledge:** Bayesian inference allows analysts to incorporate existing knowledge or expert opinions into the analysis through prior distributions. This is particularly useful in situations where data are scarce or expensive to collect.
- **Unified framework:** Bayesian methods offer a cohesive approach to inference, prediction, and decision-making. The posterior distribution contains all relevant information about the parameters, facilitating direct probabilistic statements about their values.
- **Flexibility:** Bayesian models can easily accommodate complex structures, hierarchical relationships, and uncertainty quantification.

Connections with the classical approach include:

- **Likelihood functions:** Both Bayesian and frequentist methods rely heavily on the likelihood function to capture the relationship between the data and the model parameters. In Bayesian inference, the likelihood is combined with a prior to compute the posterior.
- **Large-sample behavior:** Under certain conditions, Bayesian posterior distributions converge to frequentist estimators as sample sizes increase. For example, the mean of the posterior distribution often approximates the maximum likelihood estimate (MLE), and credible intervals can align closely with confidence intervals.

1.1 Comparison of Bayesian and Classical Approaches

Parameter Treatment:

- Frequentist methods treat parameters as fixed and unknown constants, focusing on sampling variability of the data.
- Bayesian methods treat parameters as random variables with a prior distribution that reflects their uncertainty before observing data.

Uncertainty Quantification:

- Frequentist confidence intervals provide a range of values that, under repeated sampling, would contain the true parameter value with a certain frequency (e.g., 95%).
- Bayesian credible intervals directly express the probability that the parameter lies within a given range based on the observed data and the prior.

Incorporation of Prior Knowledge:

- Frequentist methods rely solely on the data and do not incorporate prior information.
- Bayesian methods combine prior distributions with the likelihood, enabling analysts to use external knowledge or past research.

Interpretation of Results:

- Frequentist p-values and confidence intervals can be challenging to interpret and may not provide direct probabilistic statements about parameters.
- Bayesian inference directly quantifies uncertainty, offering intuitive probabilistic interpretations of parameters and predictions.

Applications:

- Frequentist methods are well-suited for scenarios with large datasets and minimal prior information.
- Bayesian methods excel in fields such as medicine, engineering, and finance, where prior knowledge is critical and data may be limited.

This comparison highlights the complementary nature of Bayesian and frequentist approaches, with each offering distinct strengths depending on the context of the analysis.

1.2 Subjective Interpretation of Probability

Bayesian probability is interpreted subjectively as a degree of belief, representing an individual's uncertainty about a proposition or parameter. This contrasts with the frequentist interpretation, which defines probability as the long-run relative frequency of an event occurring in repeated experiments. The subjective interpretation offers several key advantages:

- **A coherent framework for updating beliefs:** By using Bayes' theorem, prior beliefs can be updated in light of new data to reflect current knowledge.
- **Quantifying uncertainty in unique events:** Unlike the frequentist view, Bayesian methods allow for probability statements about single events or unique situations (e.g., the likelihood of rain tomorrow).

Real-World Applications

The subjective interpretation of probability is particularly valuable in practical settings where uncertainty plays a significant role:

- **Risk assessment in financial markets:** Analysts can use Bayesian methods to estimate the probability of market crashes or evaluate investment risks based on historical data and expert opinions.
- **Medical decision-making under uncertainty:** Physicians can integrate prior clinical experience and trial data to determine the probability of treatment success for individual patients.
- **Engineering reliability analysis:** Bayesian methods help assess the probability of system failures or estimate the remaining lifespan of critical components, combining field data with expert knowledge.

Updating Beliefs Using Bayes' Theorem

Consider a practical scenario where subjective probabilities evolve with new information:

- **Example: Diagnosing a Disease**
  - **Prior belief:** Based on population data, a physician estimates that 5% of patients presenting with certain symptoms have Disease A.
  - **Likelihood:** A diagnostic test has a 95% sensitivity (true positive rate) and a 90% specificity (true negative rate).
  - **Data:** The patient tests positive for the disease.
  - **Posterior belief:** Using Bayes' theorem, the physician updates the probability of Disease A for the patient to approximately 34%.

1.3 Introduction to Bayes' Theorem and Its Use in Updating Information

Bayes' Theorem is a fundamental concept in probability theory and statistics. It provides a way to update the probability of a hypothesis (θ) given new data (D). In Bayesian inference, this theorem allows us to revise our beliefs about a parameter θ after observing new evidence D. This theorem is crucial for decision-making in various fields, such as medicine, machine learning, and scientific research.

1.3.1 Bayes' Theorem Formula and Conditional Distributions

Bayes' Theorem is rooted in the concept of conditional probability. It describes how to update the probability of a hypothesis θ based on new data D. This is done by utilizing prior knowledge and the likelihood of the observed data.

The general form of Bayes' Theorem is given as:

$$P(\theta|D) = \frac{P(D|\theta) \cdot P(\theta)}{P(D)} \tag{1.1}$$

Where:

- $P(\theta|D)$ is the **posterior probability**: the probability of the hypothesis θ being true after observing the data D. It represents the updated belief about the hypothesis after accounting for the data.
- $P(D|\theta)$ is the **likelihood**: the probability of observing the data D, given that the hypothesis θ is true. It reflects how well the hypothesis explains the observed data.
- $P(\theta)$ is the **prior probability**: the probability assigned to the hypothesis θ before observing any data. This prior encapsulates the initial belief or background knowledge about θ.
- $P(D)$ is the **marginal likelihood** or **evidence**: the total probability of observing the data D under all possible hypotheses. It normalizes the posterior probability to ensure it is a valid probability distribution.

1.3.2 Conditional Probability and Bayes' Theorem

Bayes' Theorem is a direct result of the definition of conditional probability. The conditional probability of an event A given another event B is defined as:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{1.2}$$

Similarly, the conditional probability of B given A is:

$$\begin{aligned} P(B|A) &= \frac{P(A \cap B)}{P(A)} \\ P(B|A) \cdot P(A) &= P(A \cap B) \end{aligned} \tag{1.3}$$

substituting $P(A \cap B)$ of Equation 1.3 in Equation 1.2, we derive the following Bayes' Theorem:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

In the context of Bayes' Theorem applied to statistical inference, we substitute the hypothesis θ for event A and the data D for event B, yielding the formula:

$$P(\theta|D) = \frac{P(D|\theta) \cdot P(\theta)}{P(D)}$$

1.3.3 Explanation of Terms

1. Prior Probability $P(\theta)$

The prior probability represents the initial belief about the hypothesis θ before any data is observed. It can be subjective, based on expert knowledge or past experience, or it can be derived from prior data. In the Bayesian framework, this is where any existing information about θ is incorporated.

For example, in a medical context, the prior probability $P(\theta)$ could represent the prevalence of a disease in the population before any diagnostic tests are performed. If a certain disease affects 2% of the population, the prior probability that a randomly chosen person has the disease is 0.02.

2. Likelihood $P(D|\theta)$

The likelihood is the probability of observing the data D, assuming that the hypothesis θ is true. The likelihood function plays a crucial role in Bayesian inference as it measures how well the hypothesis θ explains the data.

In many practical cases, the likelihood is determined from a statistical model. For example, in a clinical trial, the likelihood $P(D|\theta)$ might be based on a binomial or normal distribution, depending on the type of data collected (e.g., success/failure outcomes or continuous measurements).

3. Marginal Likelihood $P(D)$

The marginal likelihood (also called the evidence) is the total probability of observing the data D, regardless of which hypothesis θ is true. It is computed by summing (or integrating) over all possible hypotheses:

$$P(D) = \sum_{\theta} P(D|\theta) \cdot P(\theta) \tag{1.4}$$

in discrete case in $\theta = \{\theta_1 \texttt{ and } \theta_2\}$ then

$$P(D) = P(D|\theta_1) \cdot P(\theta_1) + P(D|\theta_2) \cdot P(\theta_2) \tag{1.5}$$

For continuous distributions, this becomes an integral:

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

The marginal likelihood ensures that the posterior probability $P(\theta|D)$ is properly normalized. It also plays a critical role in model comparison in Bayesian inference, where models are compared based on their ability to explain the observed data.

4. Posterior Probability $P(\theta|D)$

The posterior probability is the main quantity of interest in Bayesian inference. It represents the updated probability of the hypothesis θ after considering the observed data D. The posterior combines both the prior belief $P(\theta)$ and the likelihood $P(D|\theta)$, providing a new probability distribution over the hypotheses.In decision-making contexts, the posterior probability helps in making informed decisions based on updated beliefs. For instance, after observing a positive test result, the posterior probability tells us the likelihood that a patient has a disease.

Formally, if we treat θ as a random variable with prior distribution $P(\theta)$, and D as the observed data, the posterior distribution $P(\theta|D)$ is updated via Bayes' Theorem as:

$$P(\theta|D) \propto P(D|\theta) \cdot P(\theta)$$

1.4 Examples and Applications of Bayes' Theorem

Example 1: Medical Diagnosis

Consider a scenario where we want to calculate the probability that a patient has a certain disease after receiving a positive test result.

- **Prior Probability ($P(\theta)$):** The prevalence of the disease in the population is 1%, meaning $P(\theta) = 0.01$.
- **Likelihood ($P(D|\theta)$):** The probability of testing positive, given that the patient has the disease, is 90%, meaning $P(D|\theta) = 0.9$.
- **False Positive Rate ($P(D|\neg\theta)$):** The probability of testing positive when the patient does not have the disease is 5%, meaning $P(D|\neg\theta) = 0.05$.
- **Marginal Probability ($P(D)$):** The total probability of a positive test result, regardless of whether the patient has the disease or not.

We now apply Bayes' Theorem to calculate the **posterior probability** $P(\theta|D)$, the probability that the patient has the disease given a positive test result:

$$P(\theta|D) = \frac{P(D|\theta) \cdot P(\theta)}{P(D)}$$

First, we compute $P(D)$, the marginal probability:

$$P(D) = P(D|\theta) \cdot P(\theta) + P(D|\neg\theta) \cdot P(\neg\theta)$$

$$P(D) = (0.9 \times 0.01) + (0.05 \times 0.99) = 0.009 + 0.0495 = 0.0585$$

Now we can compute the posterior:

$$P(\theta|D) = \frac{0.9 \times 0.01}{0.0585} = \frac{0.009}{0.0585} \approx 0.154$$

So, even after a positive test result, the probability that the patient actually has the disease is about 15.4%, highlighting the importance of prior probabilities in making decisions.

Example 2: Spam Filtering

Bayes' Theorem is widely used in email spam filters. The goal is to determine whether an email is spam (θ) given the data (D), such as the words contained in the email.

- **Prior Probability $P(\theta)$:** This could represent the proportion of spam emails in your inbox, based on past observations.
- **Likelihood $P(D|\theta)$:** This is the probability of observing certain words or phrases in spam emails, such as "prize," "win," or "money."
- **Posterior Probability $P(\theta|D)$:** After observing the words in a new email, the posterior probability gives the updated belief that the message is spam.

1.5 Concept of Belief Updating

Belief updating is a fundamental concept in probability and statistics, often used in Bayesian inference. It refers to the process of adjusting or revising one's beliefs (probability estimates) in light of new evidence.

In statistical terms:

- **Prior belief**: The initial probability or belief before observing any data.
- **Likelihood**: The probability of observing the new data given the current state of belief.
- **Posterior belief**: The updated belief after incorporating the new evidence.

The process of belief updating helps us make more accurate predictions and decisions by continuously refining our understanding based on observed data.

1.5.1 Bayes' Theorem

The mathematical foundation for belief updating is based on **Bayes' Theorem**:

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)}$$

Where:

- $P(H|E)$ is the **posterior probability** (updated belief) of hypothesis H given evidence E.
- $P(E|H)$ is the **likelihood**, the probability of observing the evidence E given that the hypothesis H is true.
- $P(H)$ is the **prior probability** of hypothesis H before observing the evidence.
- $P(E)$ is the **marginal likelihood**, the overall probability of observing the evidence E, which can be computed by summing over all possible hypotheses.

1.6 Step-by-Step Inference Examples

1.6.1 Example 1: Medical Diagnosis

Imagine a scenario where a doctor wants to update the belief that a patient has a certain disease based on a test result. Let's walk through the steps:

1. **Define Prior Belief**
   The doctor knows from past experience that 1% of patients have the disease. This is the prior belief:
   $$P(\text{Disease}) = 0.01$$

2. **Define the Likelihood**
   Suppose a test for the disease is 90% accurate. This means that if a person has the disease, the test will show positive 90% of the time:

$$P(\text{Positive Test}|\text{Disease}) = 0.90$$

   However, the test can also give false positives. Let's assume 5% of healthy patients test positive (false positive rate):

$$P(\text{Positive Test}|\text{No Disease}) = 0.05$$

3. **Calculate the Marginal Likelihood**
   We now need to compute the overall probability of a positive test result, $P(\text{Positive Test})$:

$$P(\text{Positive Test}) = P(\text{Positive Test}|\text{Disease}) \cdot P(\text{Disease}) + P(\text{Positive Test}|\text{No Disease}) \cdot P(\text{No Dise}$$

$$= (0.90 \cdot 0.01) + (0.05 \cdot 0.99) = 0.009 + 0.0495 = 0.0585$$

4. **Apply Bayes' Theorem**
   Now, we update the belief using Bayes' Theorem:

$$P(\text{Disease}|\text{Positive Test}) = \frac{P(\text{Positive Test}|\text{Disease}) \cdot P(\text{Disease})}{P(\text{Positive Test})}$$

$$= \frac{0.90 \cdot 0.01}{0.0585} = \frac{0.009}{0.0585} \approx 0.154$$

   Therefore, the updated probability that the patient has the disease after a positive test is approximately **15.4%**.

1.6.2 Example 2: Coin Tossing

Consider a scenario where you are unsure whether a coin is fair or biased toward heads. Initially, you believe the coin is fair, but after tossing the coin several times and observing the outcomes, you update your belief.

1. **Define the Prior**
   You start with the prior belief that the coin is fair:

$$P(\text{Fair}) = 0.50$$

   and that it is biased:

$$P(\text{Biased}) = 0.50$$

2. **Define the Likelihood**
   Suppose you toss the coin 10 times, and you observe 7 heads. If the coin is fair, the probability of observing 7 heads is:

$$P(7 \text{ heads}|\text{Fair}) = \binom{10}{7} \cdot (0.5)^7 \cdot (0.5)^3 = 0.117$$

   If the coin is biased, suppose it lands heads 70% of the time, so:

$$P(7 \text{ heads}|\text{Biased}) = \binom{10}{7} \cdot (0.7)^7 \cdot (0.3)^3 = 0.267$$

3. **Calculate the Marginal Likelihood**
   The overall probability of observing 7 heads is:

$$P(\text{7 heads}) = P(\text{7 heads}|\text{Fair}) \cdot P(Fair) + P(\text{7 heads}|\text{Biased}) \cdot P(Biased)$$

$$= (0.117 \cdot 0.50) + (0.267 \cdot 0.50) = 0.0585 + 0.1335 = 0.192$$

4. **Apply Bayes' Theorem**
   Now, update the belief:

$$P(\text{Fair}|\text{7 heads}) = \frac{P(\text{7 heads}|\text{Fair}) \cdot P(Fair)}{P(\text{7 heads})}$$

$$= \frac{0.117 \cdot 0.50}{0.192} = 0.305$$

   Similarly, update for the biased hypothesis:

$$P(\text{Biased}|\text{7 heads}) = \frac{P(\text{7 heads}|\text{Biased}) \cdot P(Biased)}{P(\text{7 heads})}$$

$$= \frac{0.267 \cdot 0.50}{0.192} = 0.695$$

   Thus, after observing 7 heads, the updated belief is that there's approximately a **30.5%** chance that the coin is fair and a **69.5%** chance that it is biased toward heads.

1.7 Applications of Bayes' Theorem

Bayes' Theorem finds application in a wide range of areas:

- **Medical Diagnosis**: Helps in updating probabilities of diseases after observing test results.
- **Spam Filtering**: Used by email services to filter spam based on word frequencies.
- **Forensic Science**: Applied in drug testing or DNA matching to calculate the likelihood of guilt or innocence.
- **Machine Learning and AI**: Forms the foundation for Bayesian classifiers, like the Naive Bayes classifier.
- **Predictive Analytics**: Used to update predictions in various domains, including finance, weather forecasting, and market analysis.

Practical Exercise

Question 2.1: In a clinical study, 5% of the population is known to carry a particular virus. A new test is developed with the following characteristics:

- If a person has the virus, the test is positive 92% of the time.

- If a person does not have the virus, the test is positive 8% of the time.

Suppose a random person from the population tests positive. Calculate:

(i) The probability that the person actually has the virus.

(ii) The probability of getting a positive test result for this population.

Question 2.2: Using the following study data of a population of 10,000 individuals where a disease test is conducted:

- Total population (N): 10,000 individuals
- Number of people with the disease (D^+): 500 individuals
- Number of people without the disease (D^-): 9,500 individuals

The test characteristics are as follows:

- True Positive (TP): 450 individuals
- False Negative (FN): 50 individuals
- True Negative (TN): 8,550 individuals
- False Positive (FP): 950 individuals

Calculate the following:

(a) The marginal probability of testing negative, $P(\neg D)$.

(b) The probability that a patient has the disease given that they tested negative, $P(\theta|\neg D)$.

Bayes' theorem is the foundation of Bayesian inference. It is expressed as:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

where:

- $P(\theta|D)$ is the posterior probability of the parameter θ given data D.
- $P(D|\theta)$ is the likelihood of the data given the parameter θ.
- $P(\theta)$ is the prior probability of the parameter θ.
- $P(D)$ is the marginal likelihood or evidence.

Bayes' theorem enables the updating of prior beliefs $P(\theta)$ using observed data D to obtain the posterior distribution $P(\theta|D)$. This process is iterative and allows for:

- Dynamic incorporation of new evidence.
- Improved decision-making as more data becomes available.

Practical Examples Comparing Bayesian and Frequentist Approaches

Example 1: Coin Tossing

Scenario: A coin is flipped 10 times, resulting in 7 heads. What is the probability of the coin landing heads in future flips?

- **Frequentist approach:** Estimate the probability using the sample proportion: $\hat{p} = 7/10 = 0.7$. This provides a point estimate but no direct measure of uncertainty.
- **Bayesian approach:** Assume a prior distribution (e.g., Beta(1, 1)) for the probability of heads. Using Bayes' theorem, update the prior with the observed data to obtain a posterior distribution (e.g., Beta(8, 4)), which allows for interval estimates and predictions.

Example 2: Drug Effectiveness

Scenario: A new drug is tested on a small sample, showing promising results. How confident can we be about its effectiveness?

- **Frequentist approach:** Perform a hypothesis test (e.g., t-test) and calculate a p-value to assess significance.
- **Bayesian approach:** Use prior knowledge about similar drugs to construct a prior distribution. Update this with the observed data to obtain a posterior probability distribution, providing a more intuitive measure of confidence.

Key Insights from Comparisons

- Bayesian methods provide a richer output (e.g., full posterior distributions) compared to frequentist point estimates or p-values.
- Bayesian approaches are more flexible in incorporating prior information and adapting to small sample sizes.
- Frequentist methods are simpler to implement in standard cases and do not rely on subjective priors.

2 Prior Distributions

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.

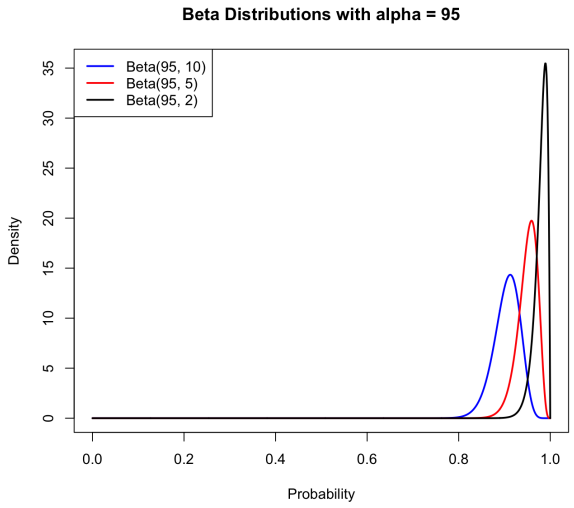


Figure 2.1: prior of Beta(95,β)

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.

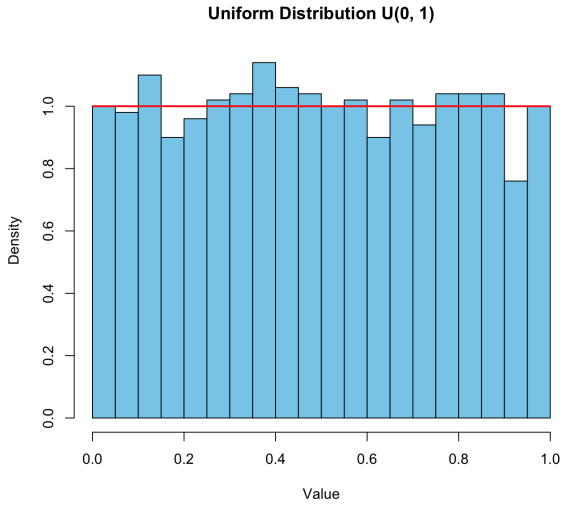


Figure 2.2: Uniform Prior

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.

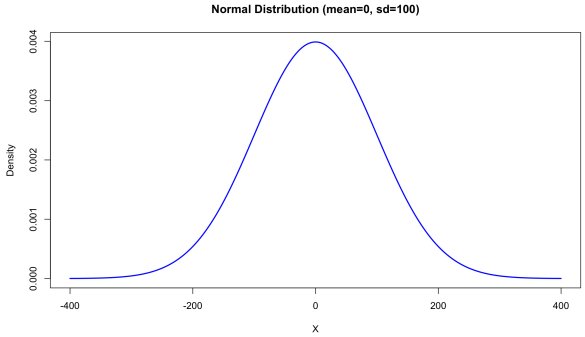


Figure 2.3: Normal prior with hi variance (week informative)

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.

2.2.7 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.

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: Consider a fair six-sided die. The probability of rolling a specific number (say 3) is:

$$P(\text{rolling a 3}) = \frac{1}{6}.$$

The sum of the probabilities for all possible outcomes is:

$$\sum_{i=1}^{6} P(\text{rolling } i) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = 1.$$

Now suppose we observe a data point $x = 5$ from a normal distribution with unknown mean μ and known standard deviation $\sigma = 2$. The likelihood function for μ is:

$$L(\mu|x) = \frac{1}{\sqrt{2\pi(2^2)}} \exp\left(-\frac{(5-\mu)^2}{2 \times 2^2}\right).$$

This function indicates the plausibility of different values of μ given the observed data. The likelihood does not sum to 1 over all possible values of μ.

Example: 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.

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.

2.2.8 Posterior Distribution

In Bayesian statistics, the **posterior distribution** represents the updated belief about a parameter θ after observing the data D. It is computed using **Bayes' theorem**, which combines prior beliefs with the likelihood of the observed data.

Posterior Distribution: Intuition

The posterior distribution $P(\theta|D)$ is the central object of interest in Bayesian inference. It combines two important components:

1. **Prior Distribution** $P(\theta)$: This reflects your initial belief or uncertainty about the parameter θ, before observing any data. For example, in the case of estimating the probability of a coin being heads, the prior distribution might reflect your belief that the coin is fair (e.g., $P(\theta) = 0.5$) or some other prior information.
2. **Likelihood** $P(D|\theta)$: This represents how likely the observed data D is for different values of θ. The likelihood reflects the compatibility of the data with different parameter values.

The posterior distribution represents a balance between these two components. It updates the prior belief based on how well different values of θ explain the observed data. The posterior tells us which values of θ are more likely after considering both prior knowledge and the observed data.

2.2.9 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.

2.2.10 Deriving posterior distribution with beta prior and binomial likelihood

We will derive the posterior distribution by combining the Beta prior with the Binomial likelihood using Bayes' Theorem.

1. Beta Prior Distribution

The Beta distribution is defined for a probability parameter $\theta \in [0, 1]$. Its probability density function (PDF) is:

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

Where:

- θ is the unknown probability parameter (e.g., the probability of heads),
- α and β are the shape parameters of the Beta distribution,
- $B(\alpha, \beta)$ is the Beta function:

$$B(\alpha, \beta) = \int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

2. Binomial Likelihood Function

The likelihood function for k successes out of n Bernoulli trials with success probability θ is given by the Binomial distribution:

$$P(D|\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$$

Where:

- k is the number of successes (e.g., heads),
- n is the total number of trials (e.g., tosses),
- θ is the probability of success.

This likelihood function gives the probability of observing k successes in n trials, given a specific value of θ.

3. Bayes' Theorem

Bayes' Theorem provides the posterior distribution by updating the prior distribution using the likelihood of the observed data:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

Where:

- $P(\theta|D)$ is the posterior distribution, representing the updated belief about θ after observing the data D,
- $P(D|\theta)$ is the likelihood, the probability of observing data D given θ,
- $P(\theta)$ is the prior distribution of θ,
- $P(D)$ is the marginal likelihood, a normalizing constant.

4. Deriving the Posterior Distribution

Now, we multiply the prior and likelihood to get the unnormalized posterior:

$$P(\theta|D) \propto P(D|\theta)P(\theta)$$

Substitute the Binomial likelihood and Beta prior:

$$P(\theta|D) \propto \left(\binom{n}{k}\theta^k(1-\theta)^{n-k}\right) \times \left(\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}\right)$$

Since $\binom{n}{k}$ and $B(\alpha,\beta)$ are constants with respect to θ, we can drop them for proportionality:

$$P(\theta|D) \propto \theta^k(1-\theta)^{n-k} \times \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Simplifying the exponents:

$$P(\theta|D) \propto \theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1}$$

5. Recognizing the Posterior as a Beta Distribution

The expression $\theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1}$ corresponds to the form of a Beta distribution. Therefore, the posterior distribution is a Beta distribution with updated parameters:

$$\theta|D \sim \text{Beta}(\alpha+k, \beta+n-k)$$

Where:

- $\alpha+k$ is the updated shape parameter for successes,
- $\beta+n-k$ is the updated shape parameter for failures.

Example: Beta-Binomial Model for Coin Tosses

As an example, consider estimating the probability θ of heads in a series of coin tosses. Suppose we observe 10 tosses, out of which 7 are heads. We can use Bayesian inference to estimate θ, the probability of heads.

- **Prior Distribution**: Assume we use a **Beta prior**, specifically $\theta \sim \text{Beta}(2, 2)$. The Beta distribution is a common prior for probabilities, as it is conjugate to the Binomial distribution.
- **Likelihood**: The data follows a **Binomial distribution**. For 7 heads out of 10 tosses, the likelihood is proportional to $\theta^7(1-\theta)^3$.
- **Posterior Distribution**: Using Bayes' Theorem, the posterior distribution is updated by combining the prior and the likelihood. Since the Beta distribution is conjugate to the Binomial, the posterior is also a Beta distribution. Specifically:

  $$\theta|D \sim \text{Beta}(2+7, 2+3) = \text{Beta}(9, 5)$$

  This posterior distribution reflects our updated belief about θ after observing the data.

The posterior Beta(9,5) distribution suggests that the probability of heads is now most likely around $\frac{9}{14} \approx 0.64$, which represents a more informed belief after observing 7 heads out of 10 tosses.

Deriving posterior distribution with uniform prior and normal likelihood

To derive the posterior distribution with a uniform prior and a normal likelihood, we can use Bayes' theorem. Let's denote the following:

- θ: the parameter we want to estimate. - x: the observed data. - $p(\theta)$: the prior distribution of θ. - $p(x \mid \theta)$: the likelihood of the data given the parameter θ. - $p(\theta \mid x)$: the posterior distribution of θ given the data x.

1. Set the Prior

Assume a uniform prior for θ:

$$p(\theta) = c$$

where c is a constant. The uniform prior does not provide information about the parameter, so it can be considered constant over the parameter's support.

2. Set the Likelihood

Assume the likelihood is normally distributed:

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

where σ^2 is the variance of the normal distribution.

3. Apply Bayes' Theorem

$$p(\theta \mid x) = \frac{p(x \mid \theta)p(\theta)}{p(x)}$$

Since $p(x)$ does not depend on θ, we can ignore it for the purpose of deriving the form of the posterior distribution.

4. Substitute the Prior and Likelihood

$$p(\theta \mid x) \propto p(x \mid \theta)p(\theta) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right)\right) \cdot c$$

$$p(\theta \mid x) \propto \exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right)$$

5. Identify the Form of the Posterior

The expression $\exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right)$ is the kernel of a normal distribution. Thus, the posterior distribution can be recognized as a normal distribution.

6. Normalize the Posterior

The posterior distribution is normally distributed:

$$p(\theta \mid x) \sim \mathcal{N}(x, \sigma^2)$$

This indicates that the posterior distribution of θ given the observed data x is a normal distribution centered at the observed value x with the same variance σ^2.

When using a uniform prior with a normal likelihood, the posterior distribution of the parameter θ is a normal distribution:

$$\theta \mid x \sim \mathcal{N}(x, \sigma^2)$$

This reflects the fact that observing x gives us the best estimate of θ while retaining the variance inherent in the likelihood.

Example

A public health researcher wants to estimate the average number of daily visits (θ) to a local health clinic. The researcher believes that the average could be anywhere between 0 and 100 visits (uniform prior), and they have data from 10 days showing the following number of visits:

$$x = \{20, 25, 30, 35, 40, 45, 50, 55, 60, 65\}$$

Steps to Derive the Posterior Distribution

1. **Calculate the Sample Mean and Variance:**
First, we compute the sample mean $\bar{x}$ and sample variance s^2:

$$\bar{x} = \frac{20 + 25 + 30 + 35 + 40 + 45 + 50 + 55 + 60 + 65}{10} = 42.5$$

To calculate the sample variance s^2:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Where $n = 10$. - Now calculate the variance:

$$s^2 = \frac{2500}{10 - 1} = \frac{2500}{9} \approx 277.78$$

Thus, $\sigma^2 = s^2 \approx 277.78$.

2. **Set the Prior:**
Assume a uniform prior for θ:

$$p(\theta) = c \quad \text{for } \theta \in [0, 100]$$

3. **Set the Likelihood:**
The likelihood is:

$$p(x \mid \theta) = \frac{1}{\sqrt{2\pi s^2}} \exp\left(-\frac{(x - \theta)^2}{2s^2}\right)$$

For our calculations, we'll use the sample mean and variance in the likelihood.

4. **Apply Bayes' Theorem:**
The posterior is proportional to the product of the likelihood and the prior:

$$p(\theta \mid x) \propto p(x \mid \theta) p(\theta)$$

Ignoring constants not dependent on θ:

$$p(\theta \mid x) \propto \exp\left(-\frac{(42.5 - \theta)^2}{2 \cdot 277.78}\right)$$

5. **Identify the Form of the Posterior:**
The posterior distribution is:

$$p(\theta \mid x) \sim \mathcal{N}\left(\bar{x}, s^2\right) \text{ where } \bar{x} = 42.5 \text{ and } s^2 \approx 277.78$$

Thus, the posterior distribution of θ is:

$$\theta \mid x \sim \mathcal{N}(42.5, 277.78)$$

Normal-Normal Model

1. Prior Distribution Assume that the prior distribution for the parameter μ is normally distributed:

$$\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$$

where μ_0 is the mean and σ_0^2 is the variance of the prior.

2. Likelihood Assume we have observed data x which follows a Normal distribution centered at μ:

$$x|\mu \sim \mathcal{N}(\mu, \sigma^2)$$

where σ^2 is the variance of the likelihood.

Deriving the Posterior Distribution

Using Bayes' theorem, the posterior distribution $P(\mu|x)$ is proportional to the product of the likelihood and the prior:

$$P(\mu|x) \propto P(x|\mu)P(\mu)$$

Substituting in the expressions for the likelihood and the prior:

1. **Likelihood**:

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

2. **Prior**:

$$P(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)$$

Combining the Terms

Now we can combine these to find the posterior:

$$P(\mu|x) \propto \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\right)\left(\frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)\right)$$

Simplifying the Expression

Combining the exponentials:

$$P(\mu|x) \propto \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} - \frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)$$

This can be rearranged by combining the terms in the exponent. We want to complete the square for the term:

$$-\frac{(x-\mu)^2}{2\sigma^2} - \frac{(\mu-\mu_0)^2}{2\sigma_0^2}$$

Completing the Square

Completing the square in the expression:

$$P(\mu|x) \propto \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}\right)\left(\mu - \frac{\frac{x}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}}{\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}}\right)^2 + \text{constant}\right)$$

Posterior Mean and Variance

From the completed square form, the posterior distribution is also normal, with mean and variance given by:

$$\mu_n = \frac{\frac{x}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}}{\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}}$$

$$\sigma_n^2 = \frac{1}{\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}}$$

Thus, the posterior distribution is:

$$\mu|x \sim \mathcal{N}\left(\mu_n, \sigma_n^2\right)$$

Example:

Suppose we have a sample of $n = 10$ measurements with sample mean $\bar{X} = 5$ and known variance $\sigma^2 = 4$. If we use a prior $\mu \sim \mathcal{N}(4, 1)$, the posterior mean and variance are computed as:

$$\mu_n = \frac{\frac{10}{4} \cdot 5 + \frac{1}{1} \cdot 4}{\frac{10}{4} + 1} = \frac{12.5 + 4}{3.5} = 4.71$$

and

$$\tau_n^2 = \left(\frac{10}{4} + 1\right)^{-1} = \frac{1}{3.5} = 0.286$$

Thus, the posterior distribution is $\mu|X \sim \mathcal{N}(4.71, 0.286)$.

Gamma-Poisson Model

For a Poisson likelihood, where the parameter λ represents the rate of occurrence, the conjugate prior is the Gamma distribution.

Model Setup

1. **Data Model:** Let Y be the number of events (counts), and assume that:

$$Y|\lambda \sim \text{Poisson}(\lambda)$$

where λ is the rate of the Poisson distribution.
2. **Prior Distribution**: Assume a Gamma prior for the rate parameter λ:

$$\lambda \sim \Gamma(\alpha, \beta)$$

Here, α is the shape parameter and β is the rate parameter of the Gamma distribution.

Derivation of the Posterior Distribution

Using Bayes' theorem, the posterior distribution of λ given the observed data $Y = y$ can be computed:

$$P(\lambda|Y) \propto P(Y|\lambda)P(\lambda)$$

Step 1: Likelihood Function

The likelihood of the data given λ is:

$$P(Y = y|\lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$$

Step 2: Prior Distribution

The prior distribution is given by:

$$P(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

Step 3: Posterior Distribution

Now, substituting the likelihood and prior into Bayes' theorem:

$$P(\lambda|Y) \propto \left(\frac{\lambda^y e^{-\lambda}}{y!}\right)\left(\frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}\right)$$

Ignoring constants that do not depend on λ, we have:

$$P(\lambda|Y) \propto \lambda^{y+\alpha-1}e^{-(1+\beta)\lambda}$$

Step 4: Identify the Posterior Distribution

The above expression is recognized as the kernel of a Gamma distribution:

$$P(\lambda|Y) \sim \Gamma(y+\alpha, 1+\beta)$$

2.2.11 Predictive and Marginal Distributions

The **predictive distribution** allows us to make predictions about future observations based on the current model and the posterior distribution. It represents the distribution of future data given the observed data and is obtained by averaging the likelihood over the posterior distribution of the parameter:

Derivation of the Predictive Distribution

We want to derive the predictive distribution:

$$P(x_{\text{new}}|D) = \int P(x_{\text{new}}|\theta)P(\theta|D)d\theta$$

where $P(x_{\text{new}}|D)$ is the probability of a new data point x_{new} given the observed data D.

Derivation

1. **Start with the definition of the predictive distribution** :

The predictive distribution of a new data point x_{new} given the observed data D can be expressed as the marginal probability over the parameter θ:

$$P(x_{\text{new}}|D) = \int P(x_{\text{new}}, \theta|D)d\theta$$

2. **Apply the chain rule of probability**:

The joint probability $P(x_{\text{new}}, \theta|D)$ can be decomposed using the chain rule:

$$P(x_{\text{new}}, \theta|D) = P(x_{\text{new}}|\theta, D)P(\theta|D)$$

3. **Simplify** $P(x_{\textbf{new}}|\theta, D)$:

Since x_{new} depends only on θ and not directly on the dataset D once θ is known, we can simplify:

$$P(x_{\text{new}}|\theta, D) = P(x_{\text{new}}|\theta)$$

4. **Substitute back into the integral**:

Substituting the decomposed expression into the integral, we get:

$$P(x_{\text{new}}|D) = \int P(x_{\text{new}}|\theta)P(\theta|D)d\theta$$

Explanation of Components

- $P(x_{\text{new}}|\theta)$: The likelihood of the new data point x_{new} given the parameter θ.
- $P(\theta|D)$: The posterior distribution of the parameter θ after observing the data D.
- $\int P(x_{\text{new}}|\theta)P(\theta|D)d\theta$: The integral sums over all possible values of θ, weighted by the posterior probability $P(\theta|D)$.

Coin Toss Example: Predictive Distribution

We are interested in computing the probability of getting heads on the next toss, given that we have already observed 9 heads and 5 tails. We'll use Bayesian inference to update our belief about the probability of heads, θ, and compute the predictive distribution.

Step 1: Prior Distribution

We start by specifying a prior distribution for θ, the probability of getting heads. We use a Beta distribution as the prior, $\text{Beta}(\alpha_0, \beta_0)$, which is conjugate to the binomial likelihood (coin tosses).

Assuming a uniform prior, $\text{Beta}(1, 1)$, which corresponds to no strong prior belief about θ, we have:

$$P(\theta) = \text{Beta}(1, 1)$$

This is equivalent to the uniform distribution over $[0, 1]$.

Step 2: Likelihood

The likelihood function represents the probability of observing the data D (9 heads and 5 tails) given θ. Since the data follows a binomial distribution, the likelihood is:

$$P(D|\theta) = \theta^9(1-\theta)^5$$

Step 3: Posterior Distribution

Using Bayes' theorem, we update the prior distribution based on the observed data to obtain the posterior distribution for θ:

$$P(\theta|D) \propto P(D|\theta)P(\theta)$$

Substituting the expressions for the likelihood and the prior, we get:

$$P(\theta|D) \propto \theta^9(1-\theta)^5 \cdot \theta^{1-1}(1-\theta)^{1-1}$$

Simplifying this expression:

$$P(\theta|D) \propto \theta^9(1-\theta)^5$$

This is the kernel of a Beta distribution. Specifically:

$$P(\theta|D) = \text{Beta}(9+1, 5+1) = \text{Beta}(10, 6)$$

Thus, the posterior distribution for θ after observing 9 heads and 5 tails is $\text{Beta}(10, 6)$.

Step 4: Predictive Distribution

The predictive distribution for the next toss being heads is:

$$P(\text{heads}|9,5) = \int_0^1 P(\text{heads}|\theta)P(\theta|9,5)d\theta$$

where:

$$P(\text{heads}|\theta) = \theta$$

is derived from the definition of θ in the context of a Bernoulli process (such as a coin toss), where θ represents the probability of observing a "head" in a single trial (coin toss)

and:

$$P(\theta|9,5) = \text{Beta}(10,6)$$

Step 5. Compute the Predictive Probability

The mean of the Beta distribution $\text{Beta}(\alpha,\beta)$ is:

$$\mathbb{E}[\theta] = \frac{\alpha}{\alpha+\beta}$$

For $\alpha = 10$ and $\beta = 6$, the mean is:

$$\mathbb{E}[\theta] = \frac{10}{10+6} = \frac{10}{16} = 0.625$$

3 Bayesian Estimation and Loss Functions

Bayesian estimation is a statistical approach that treats the parameter θ as a random variable. Unlike frequentist methods, which consider θ as a fixed but unknown constant, Bayesian methods use probability distributions to represent uncertainty about θ. By combining prior beliefs with observed data, Bayesian estimation updates knowledge about θ using Bayes' theorem:

$$\pi(\theta|x) = \frac{f_X(x|\theta)\pi(\theta)}{f_X(x)},$$

where:

- $\pi(\theta)$: The **prior distribution**, representing beliefs about θ before observing data.
- $f_X(x|\theta)$: The **likelihood function**, describing the probability of the observed data x given the parameter θ.
- $f_X(x)$: The **marginal likelihood**, ensuring that the posterior distribution integrates to 1. It is computed as:

$$f_X(x) = \int f_X(x|\theta)\pi(\theta)d\theta.$$

- $\pi(\theta|x)$: The **posterior distribution**, which updates the prior belief using the information provided by the observed data.

The posterior distribution combines the prior distribution and the likelihood function, balancing prior beliefs with evidence from the data to provide an updated, probabilistic summary of θ.

Example 1: Estimating a Proportion (*Beta-Binomial Model*)

Suppose we want to estimate the proportion p of voters who support a particular candidate. We begin with a prior belief about p and collect data from a random sample of voters.

- **Prior Distribution**: Assume that p follows a Beta distribution:

$$\pi(p) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}, \quad 0 \leq p \leq 1,$$

where $B(\alpha,\beta)$ is the Beta function, and $\alpha, \beta > 0$ are shape parameters reflecting prior beliefs about p.

- **Likelihood Function**: Suppose we observe X successes (votes for the candidate) in n trials. The likelihood function is:

$$f_X(x|p) = \binom{n}{x} p^x (1-p)^{n-x}.$$

- **Posterior Distribution**: Using Bayes' theorem, the posterior distribution is:

$$\pi(p|x) \propto f_X(x|p)\pi(p),$$

which simplifies to:

$$\pi(p|x) = \frac{p^{x+\alpha-1}(1-p)^{n-x+\beta-1}}{B(x+\alpha, n-x+\beta)},$$

showing that $p|x \sim Beta(x+\alpha, n-x+\beta)$.

Illustration: If we initially believe that p is likely close to 0.5, we could use a prior $Beta(2,2)$. After observing $x = 6$ successes in $n = 10$ trials, the posterior becomes $Beta(8,6)$, shifting our belief to reflect the observed data.

Example 2: Estimating a Mean (*Normal-Normal Model*)

Suppose we want to estimate the mean μ of a population based on sample data $x_1, x_2, \ldots, x_n$, assuming the population variance σ^2 is known.

- **Prior Distribution**: Assume $\mu \sim N(\mu_0, \sigma_0^2)$, reflecting prior knowledge about the mean.

$$P(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)$$

- **Likelihood Function**: If the data are normally distributed $X_i \sim N(\mu, \sigma^2)$, the likelihood is:

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

$$f_X(x|\mu) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right).$$

- **Posterior Distribution**: Using Bayes' theorem, the posterior is:

$$P(\mu|x) \propto \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\right)\left(\frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)\right)$$

Combining the exponentials:

$$P(\mu|x) \propto \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} - \frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)$$

This can be rearranged by combining the terms in the exponent. We want to complete the square for the term:

$$-\frac{(x-\mu)^2}{2\sigma^2} - \frac{(\mu-\mu_0)^2}{2\sigma_0^2}$$

Completing the square in the expression:

$$P(\mu|x) \propto \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\left(\mu-\frac{\frac{x}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}}{\frac{1}{\sigma^2}+\frac{1}{\sigma_0^2}}\right)^2+\text{constant}\right)$$

From the completed square form, the posterior distribution is also normal, with mean and variance given by:

$$\mu_n = \frac{\frac{x}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}}{\frac{1}{\sigma^2}+\frac{1}{\sigma_0^2}}$$

$$\sigma_n^2 = \frac{1}{\frac{1}{\sigma^2}+\frac{1}{\sigma_0^2}}$$

Thus, the posterior distribution is:

$$\mu|x \sim \mathcal{N}\left(\mu_n, \sigma_n^2\right)$$

$$\mu|x \sim N\left(\frac{\frac{\mu_0}{\sigma_0^2}+\frac{\bar{x}}{\sigma^2}}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}}, \left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)^{-1}\right),$$

where $\bar{x}$ is the sample mean.

Illustration: If $\mu_0 = 50$, $\sigma_0^2 = 25$, and we observe $n = 10$ samples with $\bar{x} = 55$ and $\sigma^2 = 16$, the posterior mean is:

$$\mu|x = \frac{\frac{50}{25}+\frac{10\cdot 55}{16}}{\frac{1}{25}+\frac{10}{16}} = 53.57,$$

indicating an updated belief closer to the observed data.

3.1 Loss Functions

In statistics, a *loss function*, denoted $L(\theta, a)$, quantifies the cost or penalty associated with estimating a parameter θ as a. It reflects how "bad" the estimate a is if the true value of the parameter is θ. Different choices of loss functions allow us to tailor the estimation process to specific objectives.

In the Bayesian framework, the goal is to minimize the expected posterior loss, $h(a)$, which is defined as:

$$h(a) = \int L(\theta, a)\pi(\theta|x)\, d\theta,$$

where:

- $\pi(\theta|x)$ is the posterior distribution of θ, given the observed data x.
- The integral represents averaging the loss over all possible values of θ, weighted by the posterior probability of each θ.

The value of a that minimizes $h(a)$ is called the *Bayes estimator*.

3.1.1 Common Loss Functions

Different loss functions lead to different Bayes estimators. Below are three widely used loss functions and their implications.

1. **Squared Error Loss**:
$$L(\theta, a) = (\theta - a)^2.$$
This loss function penalizes large errors more severely than small ones, making it ideal when we want to minimize the average size of squared deviations.

Bayes Estimator: The posterior mean minimizes the expected squared error loss:
$$\hat{\theta} = \mathbb{E}[\theta|x] = \int \theta\pi(\theta|x)\, d\theta.$$

Why Does the Posterior Mean Minimize the Expected Squared Error Loss?

1. Start with the Expected Posterior Loss The expected posterior loss under squared error loss is defined as:
$$h(a) = \int L(\theta, a)\pi(\theta|x)\, d\theta,$$
where:

- $L(\theta, a) = (\theta - a)^2$ is the squared error loss function.
- $\pi(\theta|x)$ is the posterior distribution of θ, given the data x.
- $h(a)$ represents the "average penalty" incurred when estimating θ as a.

2. Substitute the Loss Function Substituting $L(\theta, a) = (\theta - a)^2$ into $h(a)$, we get:
$$h(a) = \int (\theta - a)^2\pi(\theta|x)\, d\theta.$$

3. Expand the Squared Term Expanding $(\theta - a)^2$ using the formula for squares gives:
$$h(a) = \int \left(\theta^2 - 2\theta a + a^2\right)\pi(\theta|x)\, d\theta.$$
Using the linearity of integration, split the integral into three terms:
$$h(a) = \int \theta^2\pi(\theta|x)\, d\theta - 2a\int \theta\pi(\theta|x)\, d\theta + \int a^2\pi(\theta|x)\, d\theta.$$

4. Simplify the Terms - The first term, $\int \theta^2\pi(\theta|x)\, d\theta$, is the expected value of θ^2, denoted $\mathbb{E}[\theta^2|x]$. - The second term, $\int \theta\pi(\theta|x)\, d\theta$, is the expected value of θ, denoted $\mathbb{E}[\theta|x]$. - The third term, $\int a^2\pi(\theta|x)\, d\theta$, simplifies to a^2, as a is independent of θ.

Thus, $h(a)$ becomes:
$$h(a) = \mathbb{E}[\theta^2|x] - 2a\mathbb{E}[\theta|x] + a^2.$$

5. Minimize the Expected Posterior Loss To find the value of a that minimizes $h(a)$, take the derivative of $h(a)$ with respect to a and set it equal to zero:

$$\frac{d}{da}h(a) = \frac{d}{da}\left(\mathbb{E}[\theta^2|x] - 2a\mathbb{E}[\theta|x] + a^2\right).$$

The derivative is:

$$\frac{d}{da}h(a) = -2\mathbb{E}[\theta|x] + 2a.$$

Setting this to zero:

$$-2\mathbb{E}[\theta|x] + 2a = 0.$$

Solve for a:

$$a = \mathbb{E}[\theta|x].$$

6. Conclusion The posterior mean, $\hat{\theta} = \mathbb{E}[\theta|x]$, minimizes the expected posterior loss under squared error loss.

Intuitive Explanation

- Squared error loss penalizes larger errors more heavily than smaller ones. For example:
  - An error of 1 gives a loss of $1^2 = 1$,
  - An error of 2 gives a loss of $2^2 = 4$,
  - Larger errors grow much faster in their penalty.
- The posterior mean balances these penalties by finding the "center of gravity" of the posterior distribution.
- If you choose a point to the left or right of the mean, the penalties grow asymmetrically, increasing the overall loss.

Example 1: Posterior Mean for a Normal Distribution

Suppose the posterior distribution is $\pi(\theta|x) \sim N(10, 4)$, where the mean is 10 and the variance is 4.

- The posterior mean is $\mathbb{E}[\theta|x] = 10$.
- Estimating θ as $\hat{\theta} = 10$ minimizes the expected squared error loss.
- If you estimate θ as 9 or 11, the penalties increase asymmetrically, resulting in a higher average loss.

Example 2

Suppose the posterior distribution of θ is $\pi(\theta|x) \sim N(5, 2^2)$ (normal distribution with mean 5 and variance 4).

1. The posterior mean is:

$$\mathbb{E}[\theta|x] = 5.$$

2. To verify, consider another estimate $a = 6$. The expected squared error loss becomes:

$$h(6) = \int (\theta - 6)^2 \pi(\theta|x)\, d\theta.$$

but

$$\mathbb{E}[(\theta - 6)^2|x] = (\mathbb{E}[\theta|x] - 6)^2 + \text{Var}(\theta|x).$$

Using properties of the normal distribution:

$$h(6) = (6 - 5)^2 + \text{Variance} = 1 + 4 = 5.$$

3. If $a = 5$, the loss is:

$$h(5) = (5 - 5)^2 + \text{Variance} = 0 + 4 = 4.$$

Thus, the posterior mean minimizes the loss compared to any other estimate.

2. **Absolute Error Loss**:

$$L(\theta, a) = |\theta - a|.$$

This loss function treats all errors equally, regardless of their size. It is often used when outliers might skew results under squared error loss.

Bayes Estimator: The posterior median minimizes the expected absolute error loss:

$$\hat{\theta} = \text{Median}(\pi(\theta|x)).$$

3.1.2 Bayes Estimator: Posterior Median and Absolute Error Loss

The **posterior median** minimizes the *expected absolute error loss*. Let us demonstrate this statement step by step.

Absolute Error Loss Function

The *absolute error loss function* is defined as:

$$L(\theta, a) = |\theta - a|,$$

where:

- θ: True value of the parameter.
- a: Estimated value (in this case, the Bayes estimator $\hat{\theta}$).

The Bayesian approach minimizes the expected loss:

$$h(a) = \int |\theta - a| \pi(\theta|x)\, d\theta,$$

where $\pi(\theta|x)$ is the posterior distribution of θ, given the observed data x.

Why the Posterior Median Minimizes Absolute Error Loss

To find the a that minimizes $h(a)$, we split the integral into two parts:

$$h(a) = \int_{-\infty}^{a} (a-\theta)\pi(\theta|x)\,d\theta + \int_{a}^{\infty} (\theta-a)\pi(\theta|x)\,d\theta.$$

- For $\theta < a$, the loss is $a-\theta$, so this is represented by the first integral.
- For $\theta > a$, the loss is $\theta-a$, so this is represented by the second integral.

To minimize $h(a)$, differentiate with respect to a:

$$\frac{d}{da}h(a) = \int_{-\infty}^{a} \pi(\theta|x)\,d\theta - \int_{a}^{\infty} \pi(\theta|x)\,d\theta.$$

At the minimum, $\frac{d}{da}h(a) = 0$, which implies:

$$\int_{-\infty}^{a} \pi(\theta|x)\,d\theta = \int_{a}^{\infty} \pi(\theta|x)\,d\theta = 0.5.$$

This means a must split the posterior distribution into two equal parts, with 50% of the probability on each side. Therefore:

$$\hat{\theta} = \text{Median}(\pi(\theta|x)).$$

Example: Posterior Median for a Triangular Distribution

Suppose the posterior distribution of θ is symmetric and triangular, defined as:

$$\pi(\theta|x) = \begin{cases} 2(1-|\theta-5|), & 4 \le \theta \le 6, \\ 0, & \text{otherwise.} \end{cases}$$

Step 1: Visualize the Distribution The posterior peaks at $\theta = 5$ and tapers linearly to 0 at $\theta = 4$ and $\theta = 6$.

Step 2: Find the Median The posterior median $\hat{\theta}$ satisfies:

$$\int_{4}^{\hat{\theta}} 2(1-|\theta-5|)\,d\theta = 0.5.$$

For $\theta \le 5$, $1-|\theta-5| = \theta-4$. Integrate:

$$\int_{4}^{\hat{\theta}} 2(\theta-4)\,d\theta = 2\left[\frac{(\hat{\theta}-4)^2}{2}\right] = (\hat{\theta}-4)^2.$$

Set the integral equal to 0.5:

$$(\hat{\theta}-4)^2 = 0.25 \quad \Rightarrow \quad \hat{\theta}-4 = 0.5 \quad \Rightarrow \quad \hat{\theta} = 4.5.$$

Step 3: Verify The median $\hat{\theta} = 4.5$ splits the distribution such that 50% of the probability lies on either side, confirming it minimizes the absolute error loss.

3. **0-1 Loss**:

This loss function assigns no penalty if the estimate is exactly correct but imposes a fixed penalty otherwise. It is best suited when we want the most likely value of θ.

The **0-1 loss function** is defined as:

$$L(\hat{\theta}, \theta) = \begin{cases} 0, & \text{if } \hat{\theta} = \theta, \\ 1, & \text{if } \hat{\theta} \neq \theta. \end{cases}$$

This loss function assigns:

- Zero loss if the estimate $\hat{\theta}$ is equal to the true parameter θ,
- A loss of 1 if $\hat{\theta} \neq \theta$.

In Bayesian decision theory, the goal is to minimize the **expected loss**, which is given by:

$$\mathbb{E}[L(\hat{\theta}, \theta)|x] = \int L(\hat{\theta}, \theta)\pi(\theta|x)\, d\theta.$$

Using the definition of $L(\hat{\theta}, \theta)$, this integral splits into two cases:

$$\mathbb{E}[L(\hat{\theta}, \theta)|x] = \int_{\theta \neq \hat{\theta}} 1 \cdot \pi(\theta|x)\, d\theta + \int_{\theta = \hat{\theta}} 0 \cdot \pi(\theta|x)\, d\theta.$$

The second term vanishes because it is multiplied by 0, leaving:

$$\mathbb{E}[L(\hat{\theta}, \theta)|x] = \int_{\theta \neq \hat{\theta}} \pi(\theta|x)\, d\theta.$$

The posterior distribution $\pi(\theta|x)$ integrates to 1 over all possible values of θ, i.e.,

$$\int_{\theta \in \mathbb{R}} \pi(\theta|x)\, d\theta = 1.$$

This integral can be split into two disjoint regions:

$$\int_{\theta \in \mathbb{R}} \pi(\theta|x)\, d\theta = \int_{\theta = \hat{\theta}} \pi(\theta|x)\, d\theta + \int_{\theta \neq \hat{\theta}} \pi(\theta|x)\, d\theta.$$

Rearranging, we find:

$$\int_{\theta \neq \hat{\theta}} \pi(\theta|x)\, d\theta = 1 - \pi(\hat{\theta}|x).$$

Thus, the expected loss under the 0-1 loss function simplifies to:

$$\mathbb{E}[L(\hat{\theta}, \theta)|x] = 1 - \pi(\hat{\theta}|x).$$

For the 0-1 loss, this simplifies to:

$$\mathbb{E}[L(\hat{\theta}, \theta)|x] = 1 - \pi(\hat{\theta}|x),$$

where $\pi(\hat{\theta}|x)$ is the posterior probability density at $\hat{\theta}$.

MAP as the Minimizer of Expected 0-1 Loss

To minimize the risk, we want to minimize:

$$1 - \pi(\hat{\theta}|x).$$

Since 1 is constant, minimizing the loss is equivalent to maximizing the posterior probability $\pi(\theta|x)$. That is:

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} \pi(\theta|x).$$

Thus, the MAP estimate corresponds to the **posterior mode**—the value of θ that has the highest posterior probability.

MAP stands for **Maximum A Posteriori**. It refers to the estimate of a parameter θ that maximizes the posterior distribution $\pi(\theta|x)$, given observed data x.

This shows that the **posterior mode** minimizes the expected loss under the 0-1 loss function, making it the optimal choice when the loss is defined in this way.

Key Points

- The MAP estimate provides the parameter value that is most probable under the posterior distribution $\pi(\theta|x)$.
- Under the 0-1 loss, the expected loss is minimized by choosing $\hat{\theta}$ to maximize $\pi(\theta|x)$, which corresponds to the posterior mode.
- MAP incorporates both prior information ($\pi(\theta)$) and data ($\pi(x|\theta)$) to provide a balanced estimate.

Example: For a bimodal posterior distribution (e.g., $\pi(\theta|x)$ peaks at $\theta = 2$ and $\theta = 5$), the MAP estimate would be the mode with the highest probability density.

Note:

- Different loss functions serve different purposes. The choice of a loss function should align with the problem's objectives and the consequences of estimation errors.
- The posterior mean, median, and mode correspond to different optimality criteria:
  - **Mean**: Minimizes squared error loss, sensitive to outliers.
  - **Median**: Minimizes absolute error loss, robust to outliers.
  - **Mode**: Maximizes posterior density, ideal for identifying the most probable value.
- In practice, understanding the nature of the problem and the properties of the posterior distribution is crucial for choosing the appropriate loss function.

3.2 Bayesian Point and Interval Estimation

In Bayesian inference, estimation is based on the posterior distribution $\pi(\theta|x)$, which combines prior beliefs with observed data through Bayes' theorem. Bayesian estimation can be categorized into *point estimation* and *interval estimation*.

3.2.1 Point Estimation

A Bayesian point estimate is a single value that best represents the unknown parameter θ based on the posterior distribution. The choice of the optimal point estimate depends on the loss function, which quantifies the penalty for estimation errors. Different loss functions lead to different Bayesian estimators:

- **Posterior Mean**: Given by

$$\hat{\theta}_{\text{mean}} = \mathbb{E}[\theta|x] = \int \theta\pi(\theta|x)d\theta.$$

  This estimator minimizes the **squared error loss**, which penalizes large deviations quadratically. It provides a measure of central tendency and is useful when the posterior distribution is symmetric.

- **Posterior Median**: Defined as

$$P(\theta \leq \hat{\theta}_{\text{median}}|x) = P(\theta \geq \hat{\theta}_{\text{median}}|x) = 0.5.$$

  This estimator minimizes the **absolute error loss**, making it more robust to skewed distributions compared to the posterior mean.

- **Posterior Mode (MAP Estimate)**: Given by

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} \pi(\theta|x).$$

  This estimator minimizes the **0-1 loss function**, which considers only whether the estimate is correct or incorrect. The MAP estimate corresponds to the mode of the posterior distribution and is useful when seeking the most probable parameter value.

3.2.2 Interval Estimation

While point estimates provide a single best estimate of θ, they do not account for uncertainty. *Credible intervals* offer a Bayesian alternative to frequentist confidence intervals. Given a confidence level $1 - \alpha$, a credible interval satisfies:

$$P(\theta \in [a, b]|x) = 1 - \alpha.$$

Unlike confidence intervals, which rely on long-run frequency properties, credible intervals have a direct probabilistic interpretation: the true parameter θ lies within the interval with probability $1 - \alpha$, given the observed data.

The common credible intervals is Equal-Tailed Interval

- **Equal-Tailed Interval**: An equal-tailed credible interval ensures that the probability of θ falling below the lower bound a is the same as the probability of it exceeding the upper bound b. Mathematically, it satisfies:

$$P(\theta < a|x) = P(\theta > b|x) = \frac{\alpha}{2}.$$

This means that a total probability mass of α is excluded from the interval, with $\frac{\alpha}{2}$ in each tail. The remaining probability $1-\alpha$ is contained within the interval $[a, b]$, capturing the central $(1-\alpha)$ fraction of the posterior distribution.

Key Features of the Equal-Tailed Interval:

  - **Symmetric Tail Probabilities:** The interval cuts off an equal probability $\frac{\alpha}{2}$ from both tails of the posterior distribution. This ensures that extreme values are treated symmetrically.

  - **Application in Bayesian Inference:** It is useful when the posterior distribution is symmetric (e.g., a normal posterior), in which case the equal-tailed interval often coincides with the highest posterior density (HPD) interval.
  - **Limitations:** If the posterior distribution is skewed, the equal-tailed interval may exclude more probable values while including less likely ones. Additionally, in the case of a multi-modal posterior (having multiple peaks), it may not capture the most significant mode.

Example: Consider a Bayesian estimation problem where the posterior distribution of a parameter θ follows a normal distribution:

$$\theta|x \sim \mathcal{N}(\mu, \sigma^2).$$

Suppose we want to construct a 95% credible interval for θ, meaning that $1-\alpha = 0.95$, so $\alpha = 0.05$. The equal-tailed credible interval ensures that the probability mass in each tail is:

$$P(\theta < a|x) = P(\theta > b|x) = \frac{\alpha}{2} = 0.025.$$

For a normal posterior distribution $\mathcal{N}(\mu, \sigma^2)$, the 95% equal-tailed credible interval is found by solving:

$$P\left(\mu - z_{0.975}\sigma < \theta < \mu + z_{0.975}\sigma\right) = 0.95.$$

Using standard normal quantiles, $z_{0.975} \approx 1.96$, so the credible interval is:

$$[a, b] = \left[\mu - 1.96\sigma, \mu + 1.96\sigma\right].$$

Interpretation: Since the posterior is normal (symmetric), the equal-tailed interval places 2.5% probability in each tail, ensuring that extreme values are treated symmetrically. This means:

- There is a 2.5% probability that θ is less than $\mu - 1.96\sigma$. - There is a 2.5% probability that θ is greater than $\mu + 1.96\sigma$. - The middle 95% of the probability mass is contained within $[a, b]$.

4 Bayesian Hypothesis Testing and Model Averaging

4.1 Introduction to Bayesian Hypothesis Testing

Bayesian hypothesis testing is a framework for statistical inference that uses probabilities to quantify uncertainty about hypotheses. Unlike frequentist methods, which rely on p-values and rejection regions, Bayesian methods incorporate prior information and update beliefs through Bayes' theorem.

4.1.1 The Bayesian Paradigm

Bayesian hypothesis testing differs from the frequentist approach by making direct probability statements about hypotheses. It is built on three main principles:

- **Prior Probability ($P(H)$)**: Represents the belief in a hypothesis before observing any data.
- **Likelihood ($P(D|H)$)**: The probability of the observed data given a hypothesis.
- **Posterior Probability ($P(H|D)$)**: The updated belief in the hypothesis after incorporating the observed data using Bayes' theorem.

4.1.2 Bayes' Theorem in Hypothesis Testing

Bayes' theorem states:

Given data D, the posterior probability of a hypothesis H is given by:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

where:

- $P(H|D)$ is the posterior probability of the hypothesis given the data.
- $P(D|H)$ is the likelihood of the data given the hypothesis.
- $P(H)$ is the prior probability of the hypothesis.
- $P(D)$ is the marginal probability of the data, also known as the evidence:

$$P(D) = \sum_i P(D|H_i)P(H_i)$$

4.1.3 Advantages of Bayesian Hypothesis Testing

Bayesian hypothesis testing has several advantages:

- **Probability Interpretation**: Direct probability statements about hypotheses.
- **Incorporation of Prior Information**: Allows the use of existing knowledge.
- **Flexibility**: Suitable for small sample sizes and complex models.
- **Avoids Arbitrary Significance Levels**: Does not rely on a fixed significance threshold like 0.05.

4.1.4 Bayesian Decision Rule

The Bayesian approach to decision-making in hypothesis testing is grounded in *Bayes' theorem*, which allows us to update our beliefs about a hypothesis given observed data. This approach provides a probabilistic framework for making decisions by comparing the posterior probabilities of competing hypotheses.

1. Formulation of the Bayesian Decision Rule

Given two competing hypotheses:

- H_0 (null hypothesis)
- H_1 (alternative hypothesis)

We use *Bayes' theorem* to compute the posterior probabilities:

$$P(H_1|D) = \frac{P(D|H_1)P(H_1)}{P(D)}$$

$$P(H_0|D) = \frac{P(D|H_0)P(H_0)}{P(D)}$$

where:

- $P(H_1)$ and $P(H_0)$ are the *prior probabilities* of the hypotheses.
- $P(D|H_1)$ and $P(D|H_0)$ are the *likelihoods* of the observed data D given each hypothesis.
- $P(D)$ is the *marginal likelihood* or evidence, which ensures probabilities sum to 1:

$$P(D) = P(D|H_1)P(H_1) + P(D|H_0)P(H_0)$$

2. Bayesian Decision Rule in Hypothesis Testing

A decision in favor of H_1 over H_0 is made if:

$$P(H_1|D) > P(H_0|D)$$

or equivalently, using the ratio:

$$\frac{P(H_1|D)}{P(H_0|D)} > 1$$

which can be rewritten using Bayes' theorem as:

$$\frac{P(D|H_1)P(H_1)}{P(D|H_0)P(H_0)} > 1$$

This means that if the likelihood of observing the data under H_1 is sufficiently large relative to that under H_0, and considering prior beliefs, we favor H_1.

3. Decision-Making with Bayes Factors

Rather than relying solely on posterior probabilities, we can compare hypotheses using the *Bayes factor*, which is defined as:

$$BF = \frac{P(D|H_1)}{P(D|H_0)}$$

- If $BF > 1$, the data favor H_1 over H_0.
- If $BF < 1$, the data favor H_0 over H_1.
- If $BF = 1$, the data provide no preference between H_1 and H_0.

4. Interpretation of Bayes Factors

Jeffreys (1961) proposed a guideline for interpreting Bayes factors:

Bayes Factor BF	Interpretation
$BF < 1$	Evidence for H_0
$1 < BF < 3$	Weak evidence for H_1
$3 < BF < 10$	Moderate evidence for H_1
$10 < BF < 30$	Strong evidence for H_1
$30 < BF < 100$	Very strong evidence for H_1
$BF > 100$	Decisive evidence for H_1

Bayes factors provide a continuous measure of evidence and do not rely on arbitrary significance levels like p-values in frequentist hypothesis testing.

5. Advantages of the Bayesian Decision Rule

- *Incorporates prior knowledge*: Unlike frequentist methods, Bayesian inference allows the incorporation of prior beliefs.
- *Provides a probabilistic interpretation*: The Bayesian approach directly quantifies the probability of hypotheses given the observed data.
- *Avoids issues of p-values*: Traditional null hypothesis significance testing (NHST) relies on p-values, which do not quantify the strength of evidence but rather the probability of observing extreme data under H_0.
- *Flexible decision-making*: Bayes factors allow for more nuanced decision-making, rather than relying on a strict cutoff.

The Bayesian Decision Rule provides a principled way to compare hypotheses by updating beliefs using observed data. By leveraging *posterior probabilities* and *Bayes factors*, it allows for *rational, evidence-based decision-making* without relying on arbitrary thresholds like the 0.05 significance level.

4.1.5 Example1: Comparing Two Models

Suppose we have two hypotheses:

- H_0: A coin is fair ($P(H) = 0.5$).
- H_1: The coin is biased ($P(H) = 0.7$).

Observed data: 8 heads in 10 flips.

Using a binomial likelihood:

$$P(D|H_0) = \binom{10}{8}(0.5)^8(0.5)^2 = 0.044$$

$$P(D|H_1) = \binom{10}{8}(0.7)^8(0.3)^2 = 0.057$$

Bayes factor:

$$BF = \frac{0.057}{0.044} = 1.295$$

This suggests weak evidence in favor of H_1.

4.1.6 Example 3: Medical Diagnosis using a Normal Distribution

Suppose we are testing the effectiveness of a new drug for lowering cholesterol. We have two hypotheses:

- H_0: The drug has no effect on cholesterol levels ($\mu = 200$, where μ is the mean cholesterol level).
- H_1: The drug lowers cholesterol levels ($\mu = 180$).

We observe the following data: 20 patients, sample mean cholesterol level = 185, and sample standard deviation = 15.

We can model the data using a normal distribution with the likelihood function for each hypothesis:

For H_0 (no effect):

$$P(D|H_0) = \frac{1}{\sqrt{2\pi 15^2}} \exp\left(-\frac{(185-200)^2}{2 \cdot 15^2}\right)$$

For H_1 (drug lowers cholesterol):

$$P(D|H_1) = \frac{1}{\sqrt{2\pi 15^2}} \exp\left(-\frac{(185-180)^2}{2 \cdot 15^2}\right)$$

Calculating the likelihoods:

$$P(D|H_0) = 0.026 \quad \text{and} \quad P(D|H_1) = 0.028$$

The Bayes factor is:

$$BF = \frac{0.028}{0.026} = 1.077$$

This suggests weak evidence in favor of H_1 (the drug lowers cholesterol).

4.1.7 Example 2: Marketing Campaign Analysis using a Poisson Distribution

In a marketing campaign, we are testing whether an online advertisement results in more customer visits to a website. We have two hypotheses:

- H_0: The advertisement has no effect on the number of visits ($\lambda = 5$, where λ is the average number of visits per day).
- H_1: The advertisement increases the number of visits ($\lambda = 7$).

We observe the following data: 7 visits in one day.

We model the number of visits using a Poisson distribution. The likelihood for each hypothesis is given by the Poisson probability mass function:

For H_0 (no effect):

$$P(D|H_0) = \frac{5^7 e^{-5}}{7!} = 0.127$$

For H_1 (advertisement effect):

$$P(D|H_1) = \frac{7^7 e^{-7}}{7!} = 0.139$$

The Bayes factor is:

$$BF = \frac{0.139}{0.127} = 1.094$$

This suggests weak evidence in favor of H_1 (the advertisement increased visits).

4.2 Bayesian Model Averaging (BMA)

Bayesian Model Averaging (BMA) is a statistical technique used to address model uncertainty in the context of parameter estimation. In many cases, we may have several plausible models that could explain the data, but we don't know which one is correct. BMA provides a way to combine the predictions from multiple models, weighted by how likely each model is, given the data.

Instead of relying on a single model, BMA takes the idea that all models should be weighted according to their posterior probability, which reflects how well each model explains the observed data.

The main idea is to average over all possible models M_i, where each model is weighted by its posterior probability $P(M_i|D)$. This allows BMA to incorporate model uncertainty into the final inference about the parameter of interest θ.

The formula for BMA is as follows:

$$P(\theta|D) = \sum_i P(\theta|M_i, D)P(M_i|D)$$

Where:

- θ is the parameter of interest (e.g., the regression coefficients, means, or other model parameters).
- M_i represents different candidate models.
- $P(\theta|M_i, D)$ is the posterior distribution of θ under model M_i, given the data D.
- $P(M_i|D)$ is the posterior probability of model M_i, given the data D. This represents the weight assigned to each model based on how well it fits the data.

Key Concepts in Bayesian Model Averaging

- **Model Uncertainty**: In many scientific fields, it's common to face uncertainty about which model best describes the data. In such cases, it is beneficial to consider a set of models instead of relying on a single model. BMA accounts for this uncertainty by averaging over all models, weighted by their posterior probabilities. This provides a more robust and well-calibrated inference.
- **Posterior Model Probability**: The term $P(M_i|D)$ represents the posterior probability of model M_i, which is calculated using Bayes' theorem:

$$P(M_i|D) = \frac{P(D|M_i)P(M_i)}{P(D)}$$

  Where:

  - $P(D|M_i)$ is the likelihood of the data under model M_i.
  - $P(M_i)$ is the prior probability of model M_i, reflecting any prior beliefs or information about the plausibility of the model.
  - $P(D)$ is the marginal likelihood of the data, which ensures that all model probabilities sum to 1.

- **Weighted Averaging**: The final posterior distribution $P(\theta|D)$ is obtained by averaging over the parameter θ across all models M_i, weighted by the posterior probability of each model $P(M_i|D)$. This allows the models with stronger support from the data to have a larger influence on the final estimate.

$$P(\theta|D) = \sum_i P(\theta|M_i, D)P(M_i|D)$$

  The idea is that models that better explain the data will be assigned a higher weight, and thus their parameter estimates will influence the final result more.

- **Computational Considerations**: One of the main challenges with BMA is calculating the posterior probabilities for each model, as it requires computing the marginal likelihood $P(D|M_i)$ for all candidate models. This can be computationally intensive, especially for complex models or large datasets. However, there are several techniques for approximating the marginal likelihood, such as **Laplace approximation**, **importance sampling**, and **Markov Chain Monte Carlo (MCMC)** methods.

- **Interpretation**: The advantage of Bayesian Model Averaging is that it incorporates the uncertainty across multiple models, which can help prevent overfitting or reliance on a potentially incorrect model. The final distribution $P(\theta|D)$ is not biased toward a single model but reflects the full range of possibilities that the models account for.

Example: Linear Regression with Two Models

We consider two competing models for predicting a response variable y using a predictor x:

Model 1: Simple Linear Regression

$$y = \beta_0 + \beta_1 x + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2) \tag{4.1}$$

Model 2: Quadratic Regression

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2) \tag{4.2}$$

where $\beta_0, \beta_1, \beta_2$ are unknown parameters.

Data

The observed dataset is given below:

Prior Beliefs

We assume equal prior probabilities for both models:

$$P(M_1) = P(M_2) = 0.5 \tag{4.3}$$

x	y
1	2
2	3
3	5
4	7
5	11

Table 4.1: Observed Data

For the regression parameters and noise variance, we assume weakly informative priors:

$$\beta_j \sim N(0, 10^2),$$
$$\sigma^2 \sim \text{Inverse-Gamma}(1, 1)$$

Illustration

We are going to:

1. Fit both models to the data.
2. Calculate the posterior distributions for the parameters of each model.
3. Compute the posterior model probabilities $P(M_1|D)$ and $P(M_2|D)$.
4. Perform Bayesian Model Averaging to compute the weighted posterior distribution for the parameters.

Step 1: Fit the Models

Assume we have the data given above.

We will fit both the simple linear regression model and the quadratic regression model to this data.

Model 1 (Simple Linear Regression)

The formula for the slope $\hat{\beta}_1$ and the intercept $\hat{\beta}_0$ in simple linear regression is:

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

1.1 Compute the Means

First, we compute the means of x and y:

$$\bar{x} = \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$
$$\bar{y} = \frac{2+3+5+7+11}{5} = \frac{28}{5} = 5.6$$

1.2 Compute $\hat{\beta}_1$
Now, we compute the terms for $\hat{\beta}_1$:

$$\sum(x_i-\bar{x})(y_i-\bar{y}) = (1-3)(2-5.6)+(2-3)(3-5.6)+(3-3)(5-5.6)+(4-3)(7-5.6)+(5-3)(11-5.6)$$

$$= (-2)(-3.6) + (-1)(-2.6) + (0)(-0.6) + (1)(1.4) + (2)(5.4)$$

$$= 7.2 + 2.6 + 0 + 1.4 + 10.8 = 22$$

Next, compute $\sum(x_i - \bar{x})^2$:

$$\sum(x_i - \bar{x})^2 = (1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2$$

$$= 4 + 1 + 0 + 1 + 4 = 10$$

Thus,

$$\hat{\beta}_1 = \frac{22}{10} = 2.2$$

1.3 Compute $\hat{\beta}_0$
Now, using the formula for $\hat{\beta}_0$:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

$$\hat{\beta}_0 = 5.6 - (2.2)(3) = 5.6 - 6.6 = -1$$

Thus, the estimated coefficients for Model 1 are:

$$\hat{\beta}_0 = -1, \quad \hat{\beta}_1 = 2.2$$

—

Model 2 (Quadratic Regression)

For Model 2, we use the quadratic regression equation:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

We will use the OLS formula for a multiple linear regression. We need to compute the following system of normal equations to solve for the coefficients $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$:

$$\mathbf{X}^T\mathbf{X}\beta = \mathbf{X}^T\mathbf{y}$$

Where:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \\ 1 & x_4 & x_4^2 \\ 1 & x_5 & x_5^2 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix}$$

The entries of matrix $\mathbf{X}$ are given by 1, x_i, and x_i^2, and the response vector $\mathbf{y}$ is the given y values.

2.1 Construct the Design Matrix and Response Vector

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 1^2 \\ 1 & 2 & 2^2 \\ 1 & 3 & 3^2 \\ 1 & 4 & 4^2 \\ 1 & 5 & 5^2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \\ 1 & 5 & 25 \end{bmatrix}$$

$$\mathbf{y} = \begin{bmatrix} 2 \\ 3 \\ 5 \\ 7 \\ 11 \end{bmatrix}$$

2.2 Solve for the Parameters

We solve the system of equations $\mathbf{X}^T\mathbf{X}\beta = \mathbf{X}^T\mathbf{y}$. We compute the following:

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 5 & 15 & 55 \\ 15 & 55 & 225 \\ 55 & 225 & 979 \end{bmatrix}$$

$$\mathbf{X}^T\mathbf{y} = \begin{bmatrix} 28 \\ 104 \\ 446 \end{bmatrix}$$

We then solve for $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$ using matrix inversion or other numerical methods. The final solution gives the estimated parameters:

$$\hat{\beta}_0 = -0.4, \quad \hat{\beta}_1 = 1.2, \quad \hat{\beta}_2 = 0.8$$

We now have the OLS estimates for both models:
- Model 1 (Simple Linear Regression):

$$\hat{\beta}_0 = -1, \quad \hat{\beta}_1 = 2.2$$

- Model 2 (Quadratic Regression):

$$\hat{\beta}_0 = -0.4, \quad \hat{\beta}_1 = 1.2, \quad \hat{\beta}_2 = 0.8$$

These are the estimated parameters for both regression models using the OLS method.

Step 2: Calculate Posterior Distributions for Parameters

For each model, the parameters β_0, β_1, and β_2 have a posterior distribution, which depends on the likelihood of the data given the model and the prior distribution for the parameters.

Prior Distribution:

We assume non-informative priors for all parameters, i.e., we assume they have a uniform distribution or a normal distribution with large variance. For simplicity, we use the following prior for each parameter:

$$P(\beta_0) = P(\beta_1) = P(\beta_2) \sim \mathcal{N}(0, \text{large variance})$$

This indicates that the parameters are equally likely to take any value within a large range, representing our lack of prior knowledge about them.

Likelihood:

We assume that the likelihood of the data, given the model, is a normal distribution:

$$P(y|\beta_0, \beta_1, \beta_2, D) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - f(x_i, \beta_0, \beta_1, \beta_2))^2}{2\sigma^2}\right)$$

where:

- y_i are the observed values of the response variable,
- $f(x_i, \beta_0, \beta_1, \beta_2)$ is the model prediction for the response at the i-th observation,
- σ^2 is the variance of the error term, and
- n is the total number of data points.

Posterior Distribution:

Using Bayes' Theorem, we can calculate the posterior distribution of the parameters β_0, β_1, and β_2 as the product of the likelihood and the prior:

$$P(\beta_0, \beta_1, \beta_2|y, X) \propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - f(x_i, \beta_0, \beta_1, \beta_2))^2}{2\sigma^2}\right) \cdot P(\beta_0) \cdot P(\beta_1) \cdot P(\beta_2)$$

This distribution represents our updated belief about the parameters after observing the data. However, calculating the posterior distribution directly is often complex, so we use numerical methods to approximate it.

Markov Chain Monte Carlo (MCMC):

To sample from the posterior distribution, we use **Markov Chain Monte Carlo (MCMC)** methods. MCMC generates a sequence of samples from the posterior by iteratively updating the parameters β_0, β_1, and β_2 based on the likelihood and prior. Two popular MCMC methods are:

- **Gibbs Sampling**: A special case of MCMC where the parameters are updated one at a time, conditioned on the current values of the other parameters.
- Metropolis-Hastings: A more general MCMC method that generates a proposal for the parameters and accepts or rejects it based on a probabilistic criterion.

After running the MCMC algorithm, we obtain a set of samples for β_0, β_1, and β_2 that approximate their posterior distributions.

Estimating the Posterior:

From the MCMC samples, we can compute the following estimates of the posterior:

- **Posterior mean**: The expected value of the parameters given the data:

$$\hat{\beta}_0 = \frac{1}{N}\sum_{i=1}^{N}\beta_0^{(i)}, \quad \hat{\beta}_1 = \frac{1}{N}\sum_{i=1}^{N}\beta_1^{(i)}, \quad \hat{\beta}_2 = \frac{1}{N}\sum_{i=1}^{N}\beta_2^{(i)}$$

  where $\beta_0^{(i)}$, $\beta_1^{(i)}$, and $\beta_2^{(i)}$ are the individual samples from the posterior, and N is the total number of samples.

- **Posterior variance**: The uncertainty in the parameter estimates can be calculated as:

$$\text{Var}(\beta_0) = \frac{1}{N}\sum_{i=1}^{N}(\beta_0^{(i)} - \hat{\beta}_0)^2$$

  and similarly for β_1 and β_2.

Finally, we can plot the posterior distributions of the parameters to visualize the uncertainty in the estimates. These plots show the distribution of the parameters given the data and reflect both the prior and the likelihood.

Python Implementation :

Click HERE to view the implementation in Python

Step 3: Compute Posterior Model Probabilities

The posterior probability of each model is given by:

$$P(M_j|D) = \frac{P(D|M_j)P(M_j)}{P(D)} \tag{4.4}$$

where $P(D|M_j)$ is the marginal likelihood, and $P(D)$ is a normalizing constant. The Bayes Factor (BF) is:

$$BF = \frac{P(D|M_2)}{P(D|M_1)} \tag{4.5}$$

Thus, the posterior model probabilities are:

$$P(M_1|D) = \frac{1}{1+BF}, \tag{4.6}$$

$$P(M_2|D) = \frac{BF}{1+BF} \tag{4.7}$$

Using numerical computation, suppose we obtain: This can be done using python see HERE

$$P(D|M_1) = 0.02,$$
$$P(D|M_2) = 0.05$$

Then,

$$BF = \frac{0.05}{0.02} = 2.5 \tag{4.8}$$

The posterior probabilities are:

$$P(M_1|D) = \frac{1}{1+2.5} = 0.286,$$
$$P(M_2|D) = \frac{2.5}{1+2.5} = 0.714$$

Bayesian Model Averaging (BMA)

Instead of selecting a single model, we use BMA to compute a weighted posterior distribution for the parameters:

$$p(\boldsymbol{\beta}|D) = P(M_1|D)p(\boldsymbol{\beta}|D, M_1) + P(M_2|D)p(\boldsymbol{\beta}|D, M_2) \tag{4.9}$$

Suppose the posterior estimates for each model are: (*These one are gotten from the MCMC python simulation*)

- **Model 1 (Linear Regression)**: $\hat{\beta}_0 = 0.5, \quad \hat{\beta}_1 = 2.0$
- **Model 2 (Quadratic Regression)**: $\hat{\beta}_0 = 0.2, \quad \hat{\beta}_1 = 1.5, \quad \hat{\beta}_2 = 0.3$

The Bayesian Model Averaged (BMA) estimates are:

$$\hat{\beta}_0^{BMA} = (0.286)(0.5) + (0.714)(0.2) = 0.314,$$
$$\hat{\beta}_1^{BMA} = (0.286)(2.0) + (0.714)(1.5) = 1.643,$$
$$\hat{\beta}_2^{BMA} = (0.286)(0) + (0.714)(0.3) = 0.214$$

Thus, the final BMA regression model is:

$$y = 0.314 + 1.643x + 0.214x^2 \tag{4.10}$$

- Model 2 (Quadratic) has a higher posterior probability (0.714), indicating stronger support from the data. - Instead of selecting a single model, Bayesian Model Averaging (BMA) combines both models probabilistically. - The final BMA regression equation incorporates information from both models.

Step 4: Perform Bayesian Model Averaging

Now that we have the posterior probabilities of the models $P(M_1|D)$ and $P(M_2|D)$, we can compute the weighted posterior distribution for the parameters. We will average over the parameters $\theta = (\beta_0, \beta_1, \beta_2)$ across both models:

$$P(\theta|D) = P(\theta|M_1, D)P(M_1|D) + P(\theta|M_2, D)P(M_2|D)$$

Assume the following values for the posterior distributions of the parameters:

For Model 1:

$$P(\theta|M_1, D) = N(\hat{\beta}_0 = 1, \hat{\beta}_1 = 2, \sigma = 0.5)$$

For Model 2:

$$P(\theta|M_2, D) = N(\hat{\beta}_0 = 1.5, \hat{\beta}_1 = 1.5, \hat{\beta}_2 = 0.5, \sigma = 0.4)$$

Suppose we have two models for predicting a response variable y based on predictors x:

- Model 1 (M_1): Simple linear regression model $y = \beta_0 + \beta_1 x$.
- Model 2 (M_2): A quadratic regression model $y = \beta_0 + \beta_1 x + \beta_2 x^2$.

For each model, we calculate the posterior distribution of the parameters ($\theta = \beta_0, \beta_1, \beta_2$) given the data D. Let's assume that we have calculated the posterior probabilities for each model:

- $P(M_1|D) = 0.7$ (Model 1 is more probable given the data).
- $P(M_2|D) = 0.3$ (Model 2 is less probable).

To combine these models using BMA, we compute the weighted posterior distribution for θ:

$$P(\theta|D) = P(\theta|M_1, D)P(M_1|D) + P(\theta|M_2, D)P(M_2|D)$$

This means we combine the posterior distributions of the parameters from both models, weighted by their posterior probabilities.

Advantages of Bayesian Model Averaging

- **Improved Prediction**: BMA often leads to better predictive performance compared to using a single model, especially when there is model uncertainty.
- **Quantification of Uncertainty**: It allows for the quantification of model uncertainty, providing a more robust estimate.
- **Flexibility**: BMA can be applied to a wide range of problems, such as regression, classification, and time-series forecasting.

Disadvantages and Challenges

- **Computational Cost**: Calculating the marginal likelihoods for each model can be computationally expensive, especially for complex models.
- **Model Selection**: BMA requires the user to specify a set of candidate models, which may be challenging if the correct model is not included in the set.
- **Prior Sensitivity**: The results of BMA can be sensitive to the choice of priors for the models, so careful selection of priors is essential.

Summary

Bayesian Model Averaging is a powerful technique that allows for a more nuanced approach to parameter estimation when there is uncertainty about the best model. It combines the predictions from multiple models, weighting them by their posterior probabilities, and provides a more robust final estimate. By incorporating model uncertainty, BMA offers a more comprehensive and reliable approach to statistical inference, particularly in complex or uncertain settings.

4.3 Sensitivity Analysis in Bayesian Modelling

Sensitivity analysis examines how posterior results depend on prior assumptions and model choices. It helps:

- Assess robustness of conclusions.
- Compare different prior choices.
- Identify influential data points.

A common method involves testing different priors and observing the posterior changes. Graphical methods such as prior-posterior plots help visualize sensitivity.

4.3.1 Example: Impact of Prior Choices

Consider a Bayesian model estimating a population proportion θ using a binomial likelihood with a beta prior:

$$\theta \sim \text{Beta}(\alpha, \beta) \tag{4.11}$$

$$X|\theta \sim \text{Binomial}(n, \theta) \tag{4.12}$$

If we choose different priors, the posterior distribution changes significantly. We demonstrate this using Python.

4.3.2 Python Demonstration

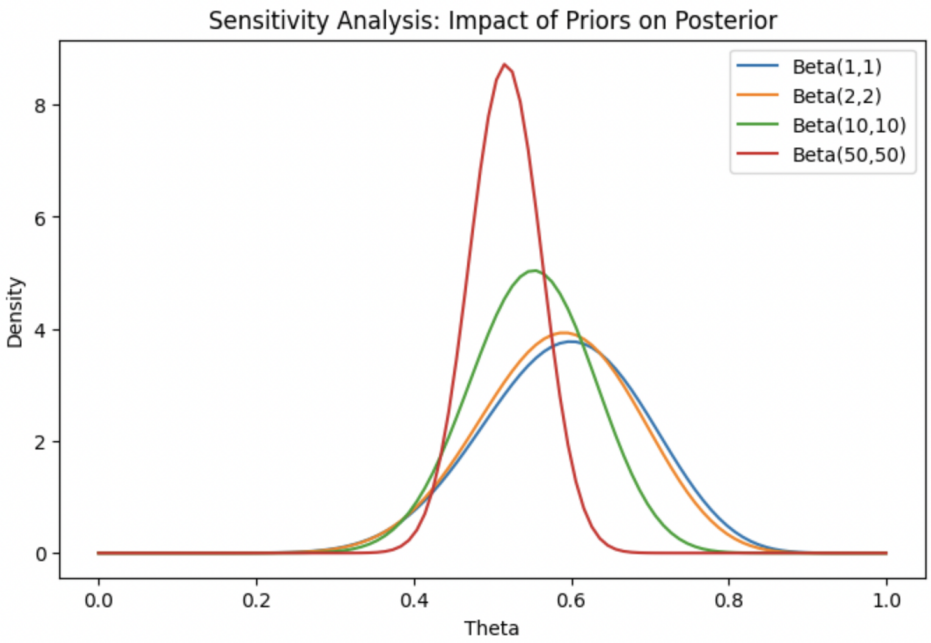
Listing 4.1: Sensitivity Analysis with Different Priors

```
import numpy as np
import scipy.stats as stats
import matplotlib.pyplot as plt

# Define priors
priors = [(1, 1), (2, 2), (10, 10), (50, 50)]  # Different Beta priors
n, x = 20, 12  # Observed data: 12 successes in 20 trials

x_vals = np.linspace(0, 1, 100)
plt.figure(figsize=(8, 5))

for a, b in priors:
    prior = stats.beta(a, b)
```




```
    posterior = stats.beta(a + x, b + (n - x))
    plt.plot(x_vals, posterior.pdf(x_vals), label=f'Beta({a},{b})')

plt.title('Sensitivity Analysis: Impact of Priors on Posterior')
plt.xlabel('Theta')
plt.ylabel('Density')
plt.legend()
plt.show()
```

4.3.3 Interpretation

From the graph, we observe that:

- A weak prior (e.g., Beta(1,1)) results in a posterior dominated by the likelihood. *The Beta(1,1) prior is uniform, meaning it provides no strong preference for any value of θ. The posterior, in this case, is primarily shaped by the likelihood, which is based on observed data. The resulting posterior distribution closely follows the likelihood function.*
- A strong prior (e.g., Beta(50,50)) maintains its influence even with observed data.
- Priors with different shapes can lead to different posterior distributions.

This highlights the importance of sensitivity analysis in Bayesian modelling.

5 Bayesian Inference in Regression Models

Bayesian inference provides a probabilistic approach to estimating regression parameters by incorporating prior knowledge. The posterior distribution is obtained using Bayes' theorem:

$$P(\theta|D) \propto P(D|\theta)P(\theta)$$

where:

- $P(\theta|D)$ is the posterior distribution of parameters given data D,
- $P(D|\theta)$ is the likelihood function,
- $P(\theta)$ is the prior distribution of parameters.

5.1 Bayesian Inference for Binomial Regression

In binomial regression, the response variable follows a binomial distribution:

$$Y_i \sim \text{Binomial}(n_i, p_i)$$

where Y_i represents the number of successes out of n_i trials, and p_i is the probability of success.

A common link function used in binomial regression is the logit function:

$$\log\left(\frac{p_i}{1-p_i}\right) = X_i\beta$$

which ensures that p_i remains between 0 and 1.

A Bayesian approach requires specifying:

- **Likelihood**: Based on the binomial distribution:

$$P(Y_i|\beta) = \binom{n_i}{Y_i} p_i^{Y_i}(1-p_i)^{n_i-Y_i}$$

- **Priors**: Common choices include normal priors for β:

$$\beta \sim \mathcal{N}(0, \sigma^2)$$

  where σ^2 is a variance hyperparameter that can be set based on prior knowledge.

- **Posterior Distribution**: Using Bayes' theorem, we update our beliefs about β given the data:

$$P(\beta|Y) \propto P(Y|\beta)P(\beta)$$

This posterior is typically estimated using Markov Chain Monte Carlo (MCMC) methods.

5.1.1 Worked Example: Bayesian Logistic Regression

Suppose we are modeling the probability of a student passing an exam ($Y_i = 1$) based on the number of study hours (X_i). We assume a logistic regression model:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_i$$

Given the following dataset:

Hours Studied (X_i)	Passed Exam (Y_i)
1	0
2	0
3	1
4	1
5	1

We assume the following priors:

$$\beta_0 \sim \mathcal{N}(0, 10)$$

$$\beta_1 \sim \mathcal{N}(0, 10)$$

The likelihood function is:

$$P(Y|\beta) = \prod_{i=1}^{n} p_i^{Y_i}(1-p_i)^{1-Y_i}$$

Using MCMC methods (such as Gibbs sampling or Hamiltonian Monte Carlo), we estimate the posterior distributions of β_0 and β_1. The resulting posterior means provide an estimate of the effect of study hours on passing the exam. See the results HERE

After running MCMC, we obtain posterior samples for β_0 and β_1. Suppose the estimated values are:

$$\beta_0 = -11, \quad \beta_1 = 4.9$$

This means that for each additional hour of study, the log-odds of passing increases by 0.8. Converting back to probability:

$$p = \frac{1}{1 + e^{-(-1.5+0.8X_i)}}$$

For a student studying 3 hours:

$$p = \frac{1}{1 + e^{-(-1.5+0.8\times 3)}} \approx 0.73$$

Thus, the probability of passing after studying for 3 hours is approximately 73%.

5.2 Bayesian Inference for Ordinal Regression

Ordinal regression is used when the dependent variable is ordinal—that is, it has a natural order but unknown intervals between categories. A common model for ordinal regression is the **proportional odds model**, which assumes that the relationship between each pair of outcome groups is the same. This is defined as:

$$P(Y_i \leq k) = \frac{\exp(\alpha_k - X_i\beta)}{1 + \exp(\alpha_k - X_i\beta)}$$

where:

- Y_i is the ordinal response variable.
- α_k are cutpoints (thresholds) that separate the categories.
- X_i is a vector of predictors.
- β is a vector of regression coefficients.

5.2.1 Bayesian Inference Components

Bayesian inference in ordinal regression involves specifying prior distributions, defining the likelihood, and performing posterior inference. The key components include:

- **Priors on Cutpoints (α_k)**: Typically, weakly informative priors such as normal or uniform distributions are used to ensure proper identification of threshold parameters.
- **Priors on Regression Coefficients (β)**: A common choice is a normal prior with mean zero and large variance (e.g., $\beta \sim N(0, 10^2)$) to represent prior uncertainty.
- **Likelihood**: The likelihood is defined using cumulative probabilities derived from the proportional odds model.
- **Posterior Inference**: Computed using Markov Chain Monte Carlo (MCMC) methods such as Gibbs sampling or Hamiltonian Monte Carlo (HMC) to obtain posterior distributions for model parameters.

5.2.2 Model Implementation

To implement Bayesian ordinal regression, the following steps are typically followed:

1. Define the prior distributions for α_k and β.
2. Specify the likelihood function based on the proportional odds model.
3. Use MCMC methods to sample from the posterior distribution.
4. Assess model convergence using trace plots and effective sample size (ESS).
5. Interpret posterior estimates, including credible intervals for coefficients.

5.2.3 Applications

Bayesian ordinal regression is widely used in various fields, including:

- **Medicine**: Modeling patient outcomes on an ordinal scale (e.g., mild, moderate, severe).
- **Social Sciences**: Analyzing survey responses with ordered categories (e.g., strongly disagree to strongly agree).
- **Economics**: Studying consumer satisfaction levels.

This Bayesian approach allows for a flexible and principled framework for ordinal regression analysis, incorporating prior knowledge and quantifying uncertainty effectively.

5.2.4 Manually Worked Example

Consider a dataset where students rate a course as 'Poor', 'Average', or 'Good' (encoded as 1, 2, 3). Assume:

- Cutpoints: $\alpha_1 = -1$, $\alpha_2 = 1$.
- Coefficient: $\beta = 0.5$.

For a student who studies 4 hours per week ($X = 4$):

$$\begin{aligned}
\eta &= X\beta = 4 \times 0.5 = 2 \\
P(Y \leq 1) &= \frac{\exp(-1-2)}{1+\exp(-1-2)} \approx 0.047 \\
P(Y \leq 2) &= \frac{\exp(1-2)}{1+\exp(1-2)} \approx 0.269 \\
P(Y = 1) &= P(Y \leq 1) = 0.047 \\
P(Y = 2) &= P(Y \leq 2) - P(Y \leq 1) = 0.222 \\
P(Y = 3) &= 1 - P(Y \leq 2) = 0.731
\end{aligned}$$

5.2.5 Applications

Bayesian ordinal regression is widely used in various fields, including:

- **Medicine**: Modeling patient outcomes on an ordinal scale (e.g., mild, moderate, severe).
- **Social Sciences**: Analyzing survey responses with ordered categories (e.g., strongly disagree to strongly agree).
- **Economics**: Studying consumer satisfaction levels.

This Bayesian approach allows for a flexible and principled framework for ordinal regression analysis, incorporating prior knowledge and quantifying uncertainty effectively.

6 Bayesian Hierarchical Models

6.1 Understanding Bayesian Hierarchical Models

Bayesian hierarchical models (BHMs) are statistical models that involve multiple levels of parameters, allowing for structured dependency among observations. These models are particularly useful when dealing with grouped data, where each group has its own parameters that are linked through a common higher-level distribution.

6.1.1 Definition and Structure

- **Likelihood Function** $P(y|\theta)$
  - Defines the probability distribution of the observed data y given the model parameters θ.
  - Example: If we assume normally distributed data:

$$y_i \sim \mathcal{N}(\mu, \sigma^2)$$

- **Prior Distributions** $P(\theta)$
  - Expresses prior beliefs about the parameters θ before observing data.
  - Example: A normal prior on μ:

$$\mu \sim \mathcal{N}(\mu_0, \tau^2)$$

- **Hyperpriors** $P(\lambda)$
  - Represents higher-level uncertainty by placing distributions on prior parameters.
  - Example: An inverse-gamma prior for σ^2:

$$\sigma^2 \sim \text{Inverse-Gamma}(\alpha, \beta)$$

Example: Suppose we have test scores from different schools, where each school has its own mean but shares a common variance structure:

$$\begin{aligned} Y_{ij}|\mu_j, \sigma^2 &\sim \mathcal{N}(\mu_j, \sigma^2), \quad j = 1, \ldots, J; \quad i = 1, \ldots, n_j \\ \mu_j|\mu_0, \tau^2 &\sim \mathcal{N}(\mu_0, \tau^2) \\ \mu_0 &\sim \mathcal{N}(0, 10^2), \quad \tau^2 \sim \text{Inverse-Gamma}(2, 1) \\ \sigma^2 &\sim \text{Inverse-Gamma}(2, 1) \end{aligned}$$

6.1.2 Hierarchical Structure Representation

Bayesian hierarchical models are structured in three levels:

1. **Data Level (Observations)**:
   - The first level of the hierarchy represents the observed data, denoted as y, which follows a likelihood model governed by unknown parameters θ.
   - This level captures the variability in the data, typically assuming an underlying probability distribution such as Normal, Poisson, or Binomial.
   - Example: If we are modeling student test scores y_{ij} in school j, the likelihood function might be:
     $$y_{ij} \sim \mathcal{N}(\mu_j, \sigma^2)$$
     where μ_j represents the mean test score for school j, and σ^2 represents measurement variability.
2. **Parameter Level (Prior Distributions)**:
   - The second level introduces prior distributions on the parameters θ to encode domain knowledge and uncertainty before observing data.
   - These priors help stabilize estimates, especially in cases with small sample sizes or limited data.
   - Example: We may assume that the school-specific mean test scores μ_j follow a normal distribution around an overall mean:
     $$\mu_j \sim \mathcal{N}(\mu_0, \tau^2)$$
     where μ_0 is the overall population mean, and τ^2 is the variance capturing differences across schools.
3. **Hyperparameter Level (Hyperpriors)**:
   - The third level further models the uncertainty in prior distributions using hyperpriors on hyperparameters τ^2.
   - Hyperpriors introduce additional flexibility and allow the model to adapt based on the data structure.
   - Example: We can model the prior variance τ^2 with an inverse-gamma hyperprior to allow for automatic regularization:
     $$\tau^2 \sim \text{Inverse-Gamma}(\alpha, \beta)$$
     where α and β are chosen based on prior beliefs about the variability in means across different groups.

This hierarchical approach enables **borrowing of strength**, where information from all schools informs the estimation of each school's mean.

6.1.3 Posterior Predictive Checks

Posterior predictive distributions are used to check if simulated data resemble real data:

$$Y_{rep}|Y \sim p(Y_{rep}|\theta)$$

If replicated data deviate significantly from observed data, the model might be misspecified.

6.1.4 Information Criteria

Common Bayesian model selection metrics include:

- **Deviance Information Criterion (DIC)**:

$$DIC = D(\hat{\theta}) + 2p_D,$$

  where p_D is the effective number of parameters.

- **Widely Applicable Information Criterion (WAIC)**: A fully Bayesian alternative to AIC.

6.2 Bayesian Model Checking and Model Selection

Evaluating the adequacy of a Bayesian model is crucial to ensure that it properly represents the observed data. Two commonly used techniques for model checking and selection are **Posterior Predictive Checks** and **Information Criteria**.

6.2.1 Posterior Predictive Checks

Posterior predictive checks assess the validity of a model by comparing replicated data generated from the posterior distribution to the observed data. The idea is to simulate new datasets under the fitted model and check whether these resemble the real data. If significant discrepancies are found, the model may be misspecified.

Posterior Predictive Distribution: Given observed data Y and a posterior distribution $p(\theta|Y)$, the posterior predictive distribution is defined as:

$$Y_{rep}|Y \sim p(Y_{rep}|Y) = \int p(Y_{rep}|\theta)p(\theta|Y)d\theta.$$

Here, - Y_{rep} represents new (replicated) data generated from the posterior. - θ are the model parameters. - $p(Y_{rep}|\theta)$ is the likelihood of new data given parameters. - $p(\theta|Y)$ is the posterior distribution of the parameters.

Checking for Model Fit: The replicated datasets Y_{rep} should be statistically similar to the observed dataset Y. Discrepancies may indicate model misspecification. One common way to quantify such discrepancies is through test statistics:

$$T(Y) \text{ vs. } T(Y_{rep})$$

where $T(\cdot)$ is a function measuring some characteristic of the data (e.g., mean, variance, quantiles). The posterior predictive p-value is computed as:

$$p_{ppc} = P(T(Y_{rep}) \geq T(Y)|Y).$$

A good fit is indicated when the posterior predictive p-value p_{ppc} is close to 0.5.

- If $p_{ppc} \approx 0.5$: The replicated data Y_{rep} behaves similarly to the observed data Y, meaning the model is likely well-specified.
- If p_{ppc} is close to 0 or 1: The replicated data differs significantly from the observed data, suggesting model misspecification.

A posterior predictive p-value too extreme (very small or very large) indicates that the model is not capturing key characteristics of the data well, requiring further model refinement.

6.2.2 Information Criteria

Information criteria provide a quantitative way to compare Bayesian models by balancing goodness-of-fit and model complexity. Two commonly used criteria are the **Deviance Information Criterion (DIC)** and the **Widely Applicable Information Criterion (WAIC).**

Deviance Information Criterion (DIC): DIC is a generalization of AIC for Bayesian models and is defined as:

$$DIC = D(\hat{\theta}) + 2p_D,$$

where:
- $D(\hat{\theta}) = -2 \log p(Y|\hat{\theta})$ is the deviance evaluated at the posterior mean $\hat{\theta} = E[\theta|Y]$.
- $p_D = E[D(\theta)] - D(\hat{\theta})$ is the effective number of parameters, representing the complexity of the model.

A lower DIC value indicates a better trade-off between fit and complexity. However, DIC is reliable primarily for models with relatively large sample sizes and asymptotic properties.

Widely Applicable Information Criterion (WAIC): WAIC is a fully Bayesian alternative to AIC and DIC, derived using the log pointwise predictive density:

$$WAIC = -2 \sum_{i=1}^{n} \left(\log E_{\theta|Y}[p(Y_i|\theta)] - V_{\theta|Y}(\log p(Y_i|\theta)) \right),$$

where:
- $E_{\theta|Y}[p(Y_i|\theta)]$ is the posterior mean of the likelihood for each observation.
- $V_{\theta|Y}(\log p(Y_i|\theta))$ is the variance of the log-likelihood over the posterior distribution.

WAIC provides a more accurate estimate of predictive performance than DIC and is particularly useful for complex hierarchical models.

- **Posterior Predictive Checks** help diagnose model misspecification by comparing simulated data with observed data.
- **DIC** is a Bayesian analogue to AIC but relies on asymptotic assumptions.
- **WAIC** is a fully Bayesian criterion that accounts for the full posterior distribution and is more reliable for complex models.

- Lower values of DIC and WAIC indicate better model fit, but they should be used in conjunction with other diagnostic tools.

These methods together provide a robust framework for evaluating Bayesian models and ensuring their predictive reliability.

6.2.3 Case Study: Bayesian Hierarchical Modeling with Real Data

We analyze real educational test score data using Python. its implementation can be found HERE

This Bayesian hierarchical approach provides more robust estimates by accounting for data structure.

Practicals

Bayesian Linear regression model

Watch this Video.
you can practice the same using the python code HERE

Bayesian Multiple Regression

Go through this Bayesian Multiple Regression
data can be found HERE

Results

Key Metrics

- $R^2 = 0.819$: The model explains 81.9% of the variance in the "Chance of Admit" variable.
- Adj. $R^2 = 0.815$: Adjusted for the number of predictors, still a strong fit.
- F-statistic = 220.9 with $p < 0.001$: The overall model is statistically significant.

Parameter Estimates

Variable	Coefficient (β)	Std. Error	t-value	p-value
Intercept	-1.2936	0.120	-10.775	<0.001
GRE Score	0.0018	0.001	3.129	0.002
TOEFL Score	0.0037	0.001	3.487	0.001
University Rating	0.0088	0.005	1.903	0.058
SOP	0.0001	0.005	0.018	0.985
LOR	0.0215	0.005	4.041	<0.001
CGPA	0.1053	0.012	8.786	<0.001
Research	0.0244	0.008	3.185	0.002

Table 6.1: OLS Regression Results

Interpretation of OLS Results

- GRE Score, TOEFL Score, LOR, CGPA, and Research experience have a significant positive effect on admission probability.
- SOP is not significant, meaning it has little predictive power.
- University Rating is marginally significant (p = 0.058).
- The residuals show some signs of non-normality (skewness = -0.955, kurtosis = 4.737), which may indicate some violations of OLS assumptions.

Variable	Mean	Std. Dev.	95% Credible Interval
Intercept	0.615	5.819	(-9.969, 10.491)
GRE Score	0.002	0.001	(0.001, 0.003)
TOEFL Score	0.004	0.001	(0.002, 0.006)
University Rating	0.007	0.010	(-0.012, 0.027)
SOP	0.001	0.005	(-0.010, 0.010)
LOR	0.022	0.005	(0.013, 0.031)
CGPA	0.104	0.012	(0.083, 0.126)
Research	0.024	0.008	(0.009, 0.039)

Table 6.2: Bayesian Regression Results

Key Bayesian Findings

- The Bayesian estimates are similar to the OLS estimates, confirming the findings.
- Bayesian credible intervals are wider than OLS confidence intervals, showing higher uncertainty quantification.
- The mean estimates align well with OLS coefficients, but uncertainty in the Intercept and University Rating is high.
- The effective sample size (ess_bulk and ess_tail) is low for some parameters, suggesting some convergence issues.

Bayesian Hierarchical Linear Regression

Go through this Bayesian Hierarchical Linear Regression
data can be found HERE

Bayesian Time series Model

Go through this Bayesian Time Series model