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:

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

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

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

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

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

   Similarly, update for the biased hypothesis:

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

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

   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.

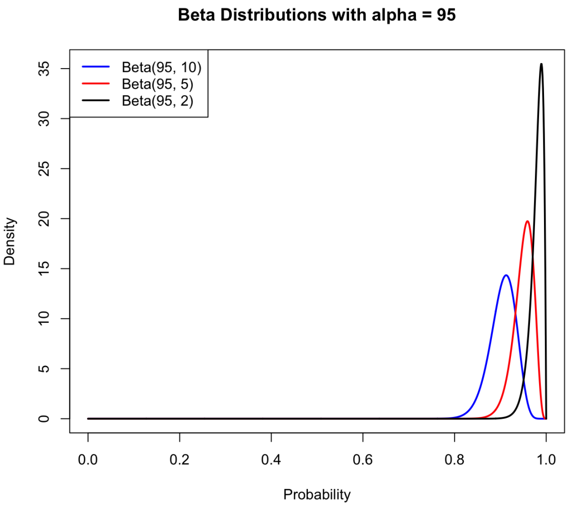

Figure 2.1: prior of Beta(95,β)

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.

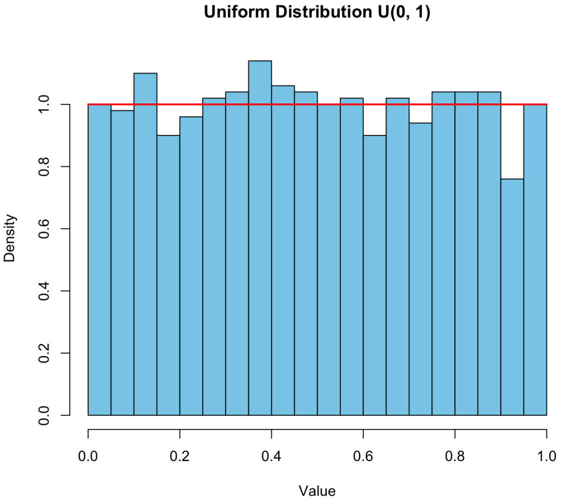

Figure 2.2: Uniform Prior

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.

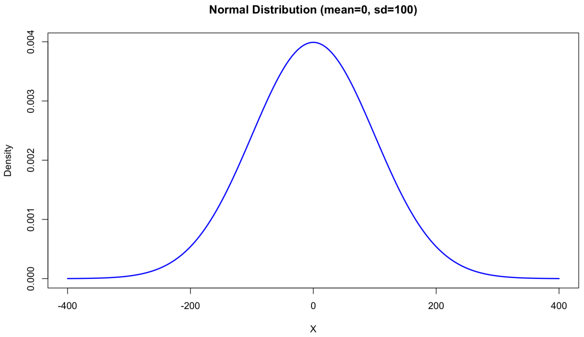

Figure 2.3: Normal prior with hi variance (week informative)

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 Beta(α,β) 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

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

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 expected 0-1 loss, we select $\hat{\theta}$ that maximizes $\pi(\theta|x)$. Thus, the Bayes estimator under the 0-1 loss is given by:

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

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.

3.1.3 Key Takeaways

- 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.

3.1.4 Bayesian Risk

The *Bayesian risk* is defined as:

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

and the optimal Bayesian estimator minimizes this risk.

3.2 Bayesian Point and Interval Estimation

3.2.1 Point Estimation

The type of loss function determines the optimal Bayesian point estimate:

- **Posterior Mean**: Minimizes squared error loss.
- **Posterior Median**: Minimizes absolute error loss.
- **Posterior Mode (MAP)**: Minimizes 0-1 loss.

3.2.2 Interval Estimation

Credible intervals provide Bayesian alternatives to confidence intervals. For a given $1-\alpha$, a credible interval satisfies:

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

- **Equal-Tailed Interval**: Ensures equal probability in the tails.
- **Highest Posterior Density (HPD) Interval**: The shortest interval containing $1-\alpha$ posterior probability.

3.3 Case Studies and Applications

3.3.1 Beta-Binomial Model

Scenario: Estimating a population proportion p.

- **Prior**: $p \sim Beta(\alpha, \beta)$.
- **Likelihood**: $X \sim Binomial(n,p)$.
- **Posterior**:
$$p|X \sim Beta(\alpha + X, \beta + n - X).$$
- **Posterior Mean**:
$$\hat{p} = \frac{\alpha + X}{\alpha + \beta + n}.$$

3.3.2 Mortality Rate Estimation (Beta Prior)

Example8203;:contentReference[oaicite:0]index=0: Estimating hospital mortality rates:

- Prior: $\theta \sim Beta(3, 27)$, reflecting prior knowledge.
- Data: No deaths in the first 10 operations.
- Posterior:
$$\pi(\theta|x) \sim Beta(3+0, 37-0) = Beta(3, 37).$$
- Posterior Mean:
$$\hat{\theta} = \frac{3}{40} = 0.075.$$

3.3.3 Coin Toss (Discrete Parameter)

Setup: Three coins with biases $\theta = \{0.25, 0.5, 0.75\}$ and equal prior probabilities.

- Observed one head ($X = 1$).
- Posterior probabilities:

$$\pi(\theta = 0.75|X = 1) = 0.5, \quad \pi(\theta = 0.5|X = 1) = 0.333.$$

3.4 Comparison with Classical Estimation

3.4.1 Key Differences

- Bayesian estimators incorporate prior knowledge, whereas classical estimators rely solely on data.
- Bayesian credible intervals provide a probabilistic interpretation of parameter uncertainty.

3.4.2 Connections

Bayesian estimators often converge to frequentist estimators (e.g., MLE) as sample size increases.

3.5 Conclusion

Bayesian estimation provides a flexible framework for parameter estimation by integrating prior beliefs and observed data. Loss functions allow for tailored decision-making, with applications ranging from estimating population proportions to analyzing hospital mortality rates.