Bayesian inference

Saurabh Kumar*, Ankit Kumar Singh†

January 2024

1 Introduction

Bayesian inference is a statistical method based on Bayes' theorem, which provides a framework for updating probabilities of hypotheses based on new evidence. In the context of biology, Bayesian inference is commonly used to make predictions, estimate parameters, and analyze data, which allows us to update our knowledge about the parameters of a random process using experimental data.

This approach assumes that the data is generated by a probabilistic model, and the parameters of this model are random variables. It uses Bayes' theorem to update the beliefs about these parameters based on the observed data. In Bayesian inference, we combine our prior knowledge about a parameter, expressed through a prior distribution, with data to update our knowledge and obtain the posterior distribution. The posterior distribution represents our updated beliefs about the parameter after seeing the data.

2 Bayes Theorem

Before understanding Bayes' theorem, it's essential to grasp the concept of joint probability. Joint probability deals with the probability of two or more events happening simultaneously. It calculates the likelihood of two or more events occurring together and at the same point in time. There are various notations for joint probability. Consider two events, A and B. The joint probability of A and B, denoted as $p(A \cap B)$ or $p(A, B)$ or $p(A\&B)$, represents the probability that both A and B occur together. We will use the notation $p(A, B)$ for the joint probability. $p(B, A)$ is the same as $p(A, B)$ because the intersection of events is commutative.

$$p(A, B) = p(A) \times p(B \mid A)$$

$$p(B, A) = p(B) \times P(A \mid B)$$

Bayes' Theorem forms the backbone of Bayesian inference, generating the posterior distribution that updates our understanding of parameters with each new data point. Bayes' Theorem calculates the posterior using the likelihood, prior, and evidence.

$$Posterior = \frac{Likelihood * Prior}{Evidence}$$

Let us consider the data y as random and the also treat the parameters θ as random variables. Then according to Bayes' theorem:

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

where $p(\theta \mid y)$ is the posterior probability, $p(y \mid \theta)$ is the likelihood, $p(\theta)$ is the prior probability and $p(y)$ is the marginal likelihood or evidence, $p(y) = \int_{-\infty}^{\infty} p(y \mid \theta) f(\theta)\, d\theta$

*Email: saurabh20541@iiitd.ac.in
†Email: ankit21310@iiitd.ac.in

Proof of Bayes' Theorem:

$$p(y, \theta) = p(y/\theta)p(\theta)$$
$$p(\theta, y) = p(\theta/y)p(y)$$

Since both joint probabilities are are equal, therefore:

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

3 Discrete distributions

3.1 Binomial Distribution

1. Binomial Distribution: The binomial distribution is a discrete probability distribution that models the number of successes in a fixed number of independent and identically distributed Bernoulli trials. Each trial has only two possible outcomes: success or failure. The binomial distribution is named after the fact that it involves two ("bi") outcomes. Here are the key components and properties of the binomial distribution:

(i) **Bernoulli Trials:** Each experiment is called a Bernoulli trial. There are only two possible outcomes: success (usually denoted as 1) or failure (usually denoted as 0).

(ii) **Parameters:** n: The number of trials or experiments. p: The probability of success on a single trial.

(iii **Probability Mass Function (PMF):** The probability mass function describes the probability of getting exactly k successes in n trials.

$$P(X = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$

where $\binom{n}{k}$ is the binomial coefficient, representing the number of ways to choose k successes from n trials.

(iv) **Mean and Variance:** The mean (expected value) of a binomial distribution is given by $\mu = np$. The variance is given by $\sigma^2 = np(1-p)$.

(v) **Cumulative Distribution Function (CDF):** The cumulative distribution function gives the probability that a random variable X is less than or equal to a certain value k. The CDF is often expressed as a sum of individual PMF values.

(vi) **Shape of the Distribution:** The binomial distribution is bell-shaped, symmetric (for $p = 0.5$), and skewed towards the side with fewer successes for extreme values of p.

(vii) **Applications:** The binomial distribution is used in various fields, such as statistics, genetics, quality control, and finance, to model situations involving a fixed number of independent trials with two possible outcomes.

(viii) **Normal Approximation:** For large values of n and moderate values of p, the binomial distribution can be approximated by a normal distribution using the central limit theorem.

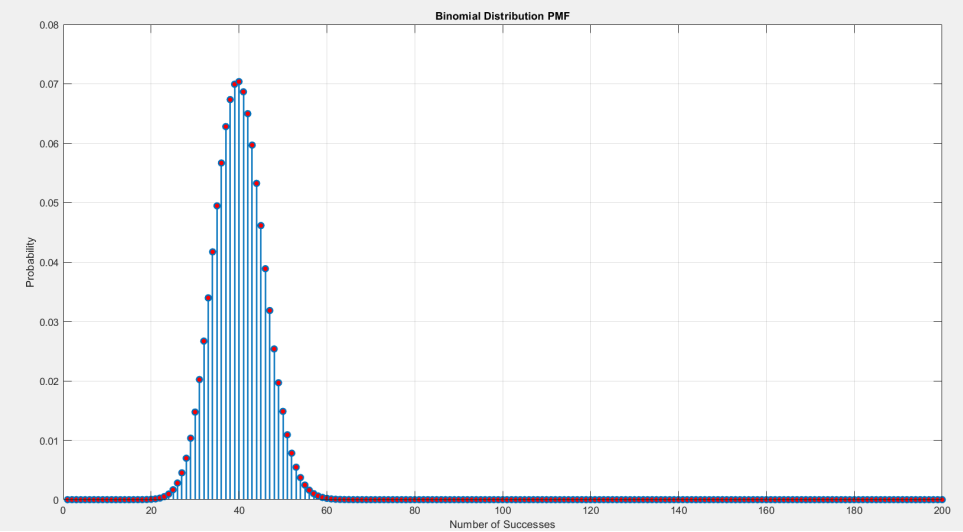


Figure 1: Binomial Distribution with $n = 200$ and $p = 0.2$

Code link: Click here to download the code

Let's derive the mean and variance of a binomial distribution.

Mean (Expected Value) of a Binomial Distribution:

$\rightarrow$ Mean (Expected value) of a Binomial Distribution $\Rightarrow$

$$\mu = np$$

Proof:

The expected value of random variable x is given by:

$$E(x) = \sum_{k=0}^{n} k \cdot P(x = k)$$

For a binomial distribution, the probability mass function (PMF) is given by:

$$P(x = k) = \binom{n}{k} \cdot p^k \cdot (I - p)^{n-k}$$

So, the expected value becomes:

$$E(x) = \sum_{k=0}^{n} n \cdot \binom{n-1}{k-1} \cdot p^k \cdot (1 - p)^{n-k}$$

Let's factor out n and write the summation:

$$E(x) = n \cdot \sum_{k=0}^{n} \binom{n-1}{k-1} \cdot p^k \cdot (1 - p)^{n-k}$$

This summation is equivalent to the probability mass function of a binomial distribution with $n - 1$ trials:

$$E(x) = n \cdot \sum_{k=0}^{n-1} \binom{n-1}{k} \cdot p^k \cdot (2 - p)^{n-1-k}$$

* Now, recognizing this is the PMF of a binomial distribution with $h - 1$ trials, the sum is equal to 1.

$$E(x) = n - 1 = n$$

therefore the mean (μ) of a binomial distribution is np

→ **Variance of a Binomial Distribution: The variance of a binomial distribution is given by:**

$$\sigma^2 = np(1-p)$$

Proof: The variance $\left(\sigma^2\right)$ of a random variable x is given by:

$$\sigma^2 = E\left(x^2\right) - [E(x)]^2$$

We have already established that $E(x) = np$. Now, let's find $E(x)^2$.

$$E\left(x^2\right) = \sum_{k=0}^{n} k^2 \cdot p(x=k)$$

Substitution the binomial PMF, we get:

$$E\left(x^2\right) = \sum_{k=0}^{n} k^2 \cdot \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$

This Expression can be quite complex, but by using some algebric manipulation and identities, we can simplify it to:

$$E\left(x^2\right) = np(1-p) + n(n-1)p^2$$

now, substitue this into the variance formula

$$\sigma^2 = np(1-p) + n(n-1)p^2 - (np)^2$$

Simplify the expression:

$$\sigma^2 = np(1-p) + np^2(n-1) - np^2$$

combine like terms:

$$\sigma^2 = np(1-p)$$

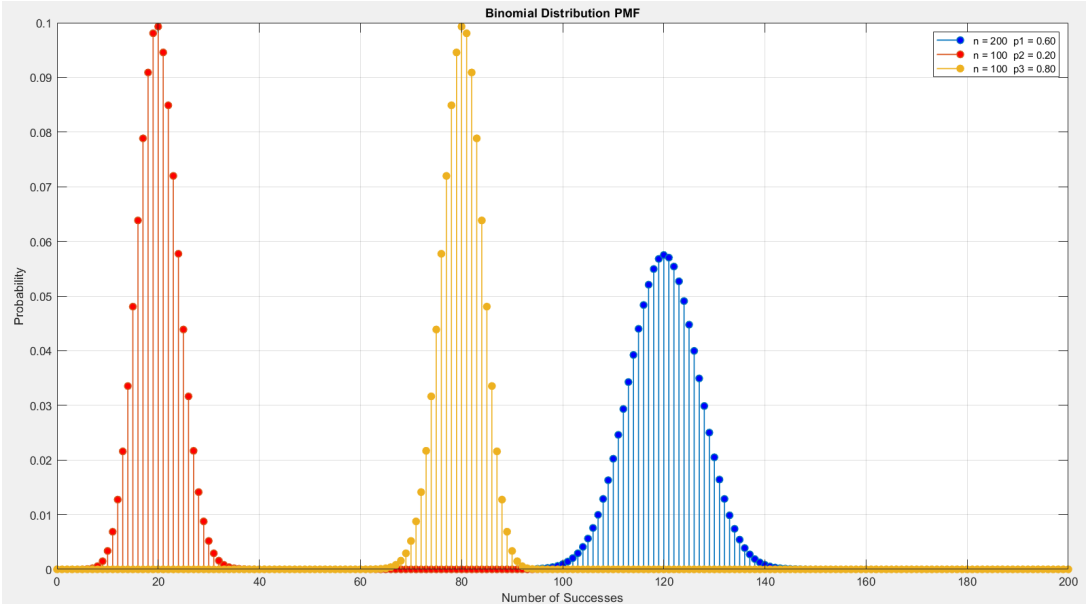


Figure 2: Binomial Distribution with different p and n values

Click here to download the code

proof of the both mean and variance in the complex version Lets, proof in comply version:-

Proof using Generating Functions: (PGF): The Probability generating Function (PGF) for a random variable x with a probability mass Function $P(x = k)$ is given by:

$$G_x(t) = \sum_{k=0}^{n} P(x = k) \cdot t^k$$

For PMF in binomial distribution, the PMF is $P(x = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$. we want to find the generating function $G_x(t)$ for this distribution

$$G_x(t) = \sum_{k=0}^{n} \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} \cdot t^k$$

Using the binomial theorem, which states that $(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} \cdot a^k \cdot b^{n-k}$, we can rewrite the expression as:

$$G_x(t) = \sum_{k=0}^{n} \binom{n}{k} \cdot (p \cdot t)^k \cdot (1-p)^{n-k}$$

Now, recognize that this is the binomial expression of $(p \cdot t + (1-p))^n$. Therefore:

$$G_x(t) = (p \cdot t + (1-p))^n$$

The coefficient of t^k in the expansion of $(p \cdot t + (1-p))^n$ is exactly $P(x = k)$, which proves that $G_x(t)$ is the PGF for a binomial distribution.

Now, we can find the mean by evaluating $\mathrm{G}'_x(t)$ the first derivative of $\mathrm{G}_x(t)$ with respect to t evaluate at t=1):

$\mathrm{G}'x(t) = n \cdot p \cdot n(p \cdot t + (1-p))^{n-1}$

$G'x(1) = n \cdot p$

This is consistent with the mean of the binomial distribution, which $\mu = n.p$.

For the variance, we need to Find $\mathrm{cr}''(x)(1)$ (the second derivative of $\mathrm{G}_x(t)$ with respect to t evaluate at $t = I$):

$$G''_x(t) = n \cdot p \cdot (n \cdot p \cdot t + n \cdot (1-p)) \cdot p(p \cdot t + (1-p))^{n-2}$$
$$G''x(1) = n \cdot p \cdot (n \cdot p + n(1-p)))$$
$$G''x(1) = n \cdot p \cdot (n \cdot p + n - n \cdot p))$$
$$G''x(1) = n \cdot p \cdot n$$

This is consistent with the variance of the binomial distribution, which is $\sigma^2 = np(1-p)$. Using PGF, we can derive the mean and variance of the binomial distribution.

The proof of the binomial distribution using the Gamma function involves the utilization of the gamma function, which is often employed in more advanced mathematical contexts. The gamma function allows for a generalization of factorials to non-integer values. $\Gamma(x)$ **is an extension of the factorial function to real and complex numbers, and it is defined as** :

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t}\, dt$$

The gamma function has applications in various areas of mathematics, physics, and statistics. In statistics, it is often used to define the gamma distribution, a continuous probability distribution that generalizes the exponential distribution.

Gamma Distribution and Bayesian Inference:

- **Gamma Distribution:** The gamma distribution is frequently used in Bayesian statistics as a conjugate prior to certain likelihood functions. If a prior distribution and likelihood function belong to the same family, the resulting posterior distribution will also belong to that family. The gamma distribution is often chosen as a conjugate prior to the exponential likelihood, which is common in Bayesian modeling.

- **Exponential Distribution:** The exponential distribution is often used in Bayesian inference to model the time until an event occurs. The gamma distribution serves as a prior distribution for the rate parameter of the exponential distribution.

Bayesian Inference: In Bayesian inference, the gamma distribution can be used as a prior when modeling unknown parameters, particularly those related to rates or scales.

Connections with Bio-Bayesian Inference: In the context of bio-Bayesian inference, where Bayesian methods are applied to biological data, the gamma distribution might be chosen as a prior for parameters related to rates of biological processes or events.

Link to Poisson Process: The gamma distribution is also connected to the Poisson process, which models the number of events occurring in a fixed interval of time or space. The gamma distribution can be used to model the waiting time until a certain number of events occur in a Poisson process.

While the gamma function itself may not be directly involved in bio-Bayesian inference, its associated gamma distribution plays a role in specifying prior distributions for certain types of parameters commonly encountered in statistical modeling, including those in the field of biology. The choice of prior distribution depends on the characteristics of the data and the underlying assumptions of the model.

This expression using the gamma function is an alternative representation of the binomial PMF. While it introduces the gamma function, it doesn't necessarily simplify the computation of probabilities in practice. The gamma function is often used in more advanced mathematical analysis and is not typically required for introductory statistics. It's worth noting that the gamma function reduces to factorials when its argument is an integer, so in most introductory statistics courses, you'd see factorials directly in the binomial distribution derivation.

PROOF:

$\rightarrow$ proof of the Binomial Distribution using Gamma function. Proof Using Gamma function: The gamma function is defined as $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$ for $n > 0$. Now, consider the binomial cofficient $\binom{n}{k}$ in the binomial PMF:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Let's express the factorials using the gamma function:

$$n! = \Gamma(n+1) = \int_0^\infty t^n e^{-1}dt$$
$$k! = \Gamma(k+1) = \int_0^\infty \mu^k e^{-\mu}du$$
$$(n-k)! = \Gamma(n-k+1) = \int_0^\infty v^{n-k}e^{-v}dv$$

$$\binom{n}{k} = \frac{\Gamma(n+1)}{\Gamma(k+1)\Gamma(n-k+1)}$$

Now, Let's incorporate this into the binomial PMF:

$$P(x=k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$
$$P(x=k) = \frac{\Gamma(n+1)}{\Gamma(k+1)\Gamma(n-k+1)} \cdot p^k \cdot (1-p)^{n-k}$$

Now, Substitue these expression back into the binomial cofficient:

The Cumulative Distribution Function (CDF) for a binomial distribution gives the probability that a random variable X is less than or equal to a certain value k. The CDF is often denoted as $F(X \leq k)$ and is expressed as a sum of individual Probability Mass Function (PMF) values.

The CDF for a binomial distribution with parameters n (number of trials) and p (probability of success) is given by:

$$F(X \leq k) = \sum_{i=0}^{k} \binom{\lfloor}{n} i]p^i(1-p)^{n-i}$$

Using the binomial PMF, which is

$$P(X = k) = \binom{[}{n} k] p^k (1-p)^{n-k}$$

This expression represents the cumulative probability of observing k or fewer successes in n trials. It's important to note that the CDF accumulates the probabilities as you move from 0 to k. The CDF will reach 1 when k is equal to or greater than n because the sum of all probabilities in the distribution is 1. The CDF is a useful tool for understanding the probability distribution of a random variable and is particularly helpful in calculating probabilities associated with specific ranges of values.

- **When Binomial Turns into Poisson Distribution:**

The Poisson distribution is often used as an approximation of the binomial distribution under certain conditions. This approximation occurs when the number of trials n is large and the probability of success p is small, while the product np remains moderate or large. This condition is expressed mathematically as $np \to \lambda$, where λ is the mean of the Poisson distribution.

Proof:

Consider a binomial distribution with parameters n and p, where n is the number of trials and p is the probability of success.

Probability Mass Function (PMF) of Binomial Distribution: $\to$ The PMF of the Binomial distribution is given by:

$$P(x = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$

Stirling's Approximation : Stirling's approximation states that for large n, n ! can be approximated as:

$$n! \approx \sqrt{2\pi n} \left(\frac{n}{e}\right)^n$$

we will use this approximation for a large value of n. Binomial coefficient Approximation Apply to simplify, many terms will cancel out, leaving:

$$\binom{n}{k} \approx \frac{n^k}{k!} \cdot \frac{1}{\sqrt{2\pi n}} \cdot \frac{1}{\sqrt{2\pi(n-k)}}$$

The simplification is valid for large n.

After simplifying, many terms cancel we take combining terms. Substitute the approximated binomial coefficient into the PMF of the binomial distribution.

Taking the limit As n approach infinity, p approaches zero in such a way that np remains constant, denoted as λ Rewrite $P = \frac{\lambda}{h}$ and let n approach infinity. As n goes to infinity, terms involving n will domains leave:

$$P(x = k) \approx \lim_{n \to \infty} \left(\frac{n^k}{k!} \cdot \frac{1}{\sqrt{2\pi n}} \cdot \frac{1}{\sqrt{2\pi(n-k}} \right) \cdot \left(\frac{\lambda^k}{n^k} \right) \cdot \left(\frac{(1-\lambda/n)^{n-k}}{\sqrt{1-\lambda/n}} \right)$$

8. Simplify further. $\to$ After simplifying and taking the limit as n: approaches infinity, we obtain the PMF of the Poisson distribution.

$$P(x = k) = \frac{e^{-\lambda} \cdot \lambda^k}{k!}$$

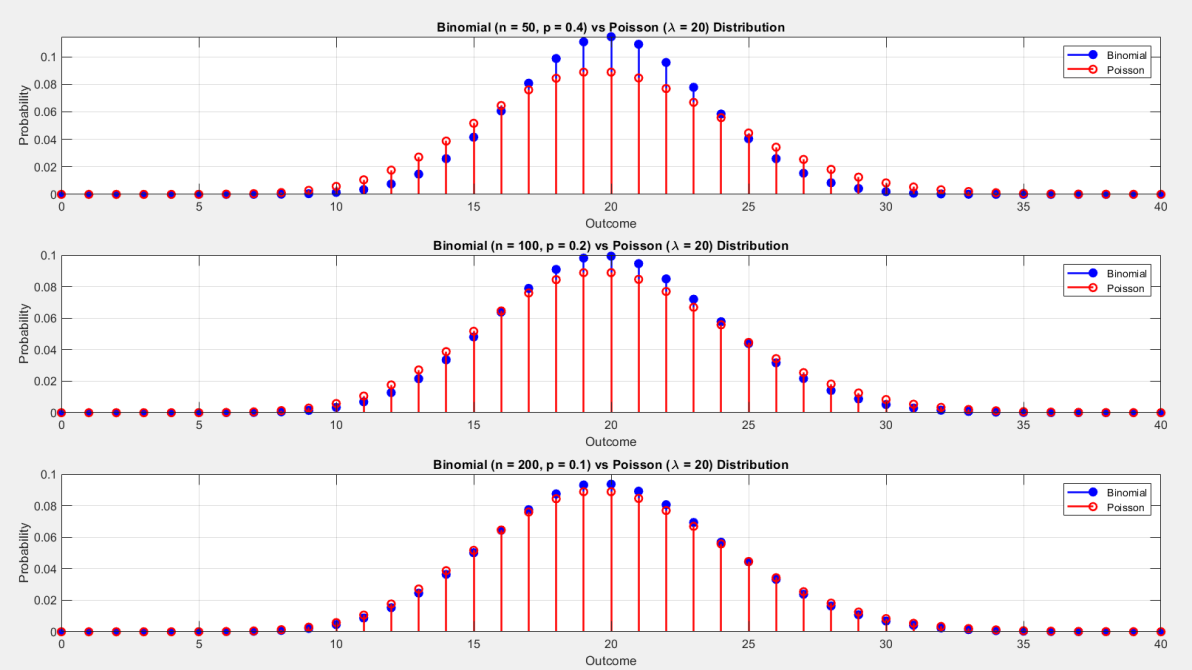


Figure 3: Binomial vs Poisson (with different values of n and p)

Click here to download the code

- **When Binomial Turns into a Geometric Distribution:** The geometric distribution is often seen as a special case of the binomial distribution. Specifically, when the binomial distribution is considered over a sequence of independent Bernoulli trials, and we're interested in the probability of the first success occurring on the k-th trial, where k can range from 1 to infinity, then the resulting distribution is the geometric distribution.

Proof:

Binomial Distribution → considers a sequence of Independent Bernoulli trials with probability of success p and probability of failure $1-p$.

Let X be the random variable representing the number of trials until the first success.

Probability of first success on the k-th trials. → In the binomial distribution, the probability of observing the first success on the $k-th$ trials is given by the Binomial Pmf:

$$P(x=k) = \binom{k-1}{0} \cdot p^1 \cdot (1-p)^{k-1}$$

This simplifies to:

$$p(x=k) = p.(1-p)^{k-1}$$

Simplify the Expression: When $r=1$ (i.e. we are interested in the first success), the binomial pmf further simplifies to:

$$p(x=k) = (1-p)^{k-1} \cdot p$$

3.2 Bernoulli Distribution

The Bernoulli distribution is a fundamental concept in probability theory and statistics, named after Swiss mathematician Jacob Bernoulli. It models a random experiment with only two possible outcomes, conventionally labeled as success (usually denoted as 1) and failure (usually denoted as 0). It's often used to represent simple yes/no or true/false situations.

Definition:

The Bernoulli distribution is characterized by a single parameter p, which represents the probability of success in a single trial of the experiment. The probability of failure is $1-p$.

Probability Mass Function (PMF):
The probability mass function of a Bernoulli random variable X is given by:
$P(X = x) = p^x(1-p)^{1-x}$

Where:
x is the outcome of the random experiment (either 0 or 1). p is the probability of success (when $x = 1$). $1 - p$ is the probability of failure (when $x = 0$). $x \in \{0, 1\}$ and p is the probability of success.

Mean and Variance:
The mean (μ) and variance (σ^2) of a Bernoulli random variable are given by: $\mu = p$, $\sigma^2 = p(1-p)$.

Expectation and Standard Deviation:
The expectation (or expected value) of a Bernoulli random variable is simply the probability of success p, denoted as $\mathrm{E}(X) = p$. The standard deviation (σ) is the square root of the variance.

Relation to Other Distributions:
The Bernoulli distribution is a special case of the Binomial distribution, which represents the number of successes in a fixed number of independent Bernoulli trials. As the number of trials in a Binomial distribution approaches infinity with the success probability held constant, it converges to a Normal distribution. Understanding the Bernoulli distribution is crucial as it serves as a building block for more complex distributions and models.

3.3 Geometric Distribution

The Geometric distribution is a discrete probability distribution that models the number of trials needed to achieve the first success in a sequence of independent Bernoulli trials, where each trial has a constant probability of success p. It's often used to model situations such as the number of coin flips needed to get the first heads or the number of attempts needed to make the first successful sale in sales calls.

Definition:

The Geometric distribution is characterized by a single parameter p, which represents the probability of success in each individual trial. The random variable X follows a Geometric distribution if it represents the number of trials needed to achieve the first success.

Probability Mass Function (PMF):
The probability mass function of a Geometric random variable X is given by: $P(X = k) = (1-p)^{k-1}p$.

Mean and Variance:
The mean (μ) of a Geometric distribution is given by $\mu = 1/p$. The variance (σ^2) of a Geometric distribution is given by $\sigma^2 = (1-p)/p^2$.

Expectation and Standard Deviation:
The expectation (or expected value) of a Geometric random variable X is $\mathrm{E}(X) = 1/p$. The standard deviation (σ) of X is $\sigma = \sqrt{(1-p)/p^2}$.

Applications:
Modeling the number of trials until the first success in repeated experiments with a constant probability of success. Used in reliability engineering to model time to failure. Analyzing the number of attempts needed for success in various scenarios, such as games, sales, or quality control processes.

Relation to Other Distributions:
The Geometric distribution is a special case of the Negative Binomial distribution, where the number of successes is fixed at 1. As the number of trials in a Negative Binomial distribution approaches infinity with the success probability held constant, it converges to a Poisson distribution.

3.4 Negative Binomial Distribution

The Negative Binomial distribution is a discrete probability distribution that generalizes the Geometric distribution. It models the number of independent and identically distributed Bernoulli trials needed to achieve a specified number of successes r, where each trial has a constant probability p of success.

Definition:

The Negative Binomial distribution is characterized by two parameters: r, the number of successes required, and p, the probability of success in each individual trial. The random variable X follows a Negative Binomial distribution if it represents the number of trials needed to achieve the r-th success.

Probability Mass Function (PMF):

The probability mass function of a Negative Binomial random variable X is given by:

$$P(X = k) = \binom{k-1}{r-1} \cdot p^r \cdot (1-p)^{k-r}$$

Where:

- k is the total number of trials needed to achieve the r-th success.
- r is the number of successes required.
- p is the probability of success in each individual trial.
- $\binom{r-1}{k-1}$ is the binomial coefficient, representing the number of combinations of $k-1$ trials with $r-1$ successes.

Mean and Variance: Mean and Variance:

The mean (μ) of a Negative Binomial distribution is given by $\mu = \dfrac{r}{p}$.

The variance (σ^2) of a Negative Binomial distribution is given by $\sigma^2 = \dfrac{r(1-p)}{p^2}$.

Expectation and Standard Deviation:

The standard deviation (σ) of X is $\sigma = \left(\frac{r(1-p)}{p^2}\right)^{1/2}$.

Applications:

- Modeling the number of trials until a specified number of successes occur in repeated experiments with a constant probability of success.
- Used in areas such as quality control, reliability engineering, and biology to analyze the number of trials needed to achieve a certain goal or outcome.

Relation to Other Distributions:

- The Negative Binomial distribution is related to the Poisson distribution. As the number of successes r approaches infinity with the success probability p decreasing proportionally, the Negative Binomial distribution converges to a Poisson distribution.

3.5 When Discrete Distributions Approach Normality

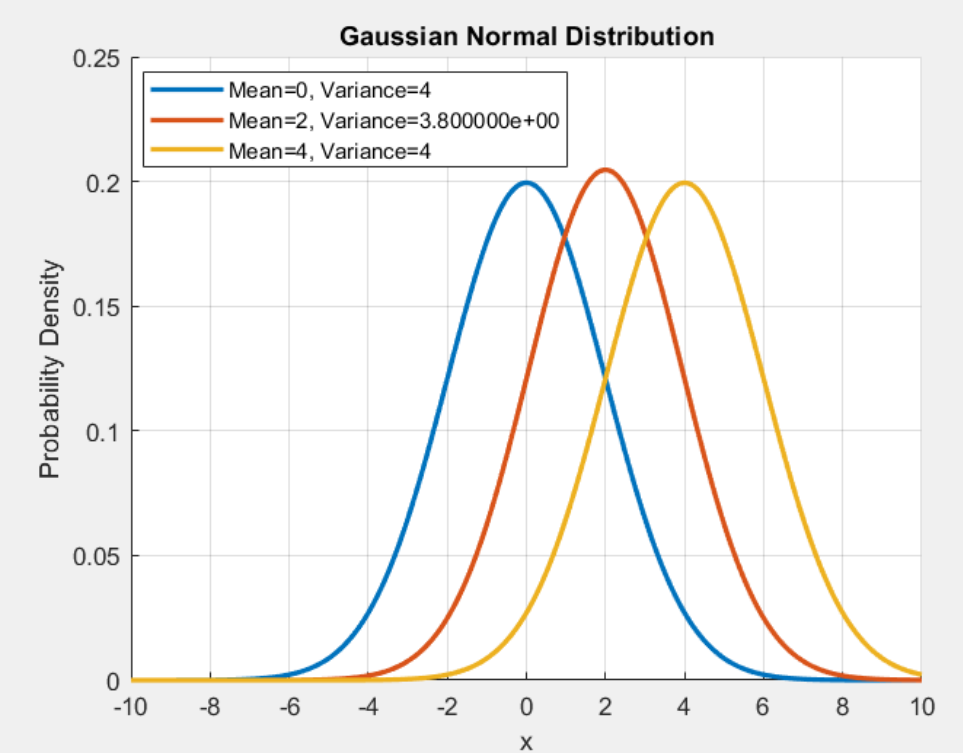


Figure 4: Normal Distribution

Click here to download the code

1. **Binomial Distribution:**

   The Central Limit Theorem (CLT) provides the mathematical basis for the convergence of the binomial distribution to the normal distribution under certain conditions. The CLT states that the sum (or average) of a large number of independent and identically distributed random variables, regardless of their original distribution, will be approximately normally distributed, provided that certain conditions are met. Conditions for Convergence:

   - Independence: The trials in the binomial distribution must be independent of each other. This means that the outcome of one trial does not affect the outcome of another.
   - Identically Distributed: Each trial must follow the same probability distribution. In the case of the binomial distribution, each trial follows a Bernoulli distribution with the same probability of success p.
   - Large Sample Size: The sample size n must be sufficiently large. While there isn't a fixed rule for what constitutes "sufficiently large," a common guideline is that np and $n(1p)$ should both be greater than or equal to 5. This ensures that the binomial distribution isn't too skewed.

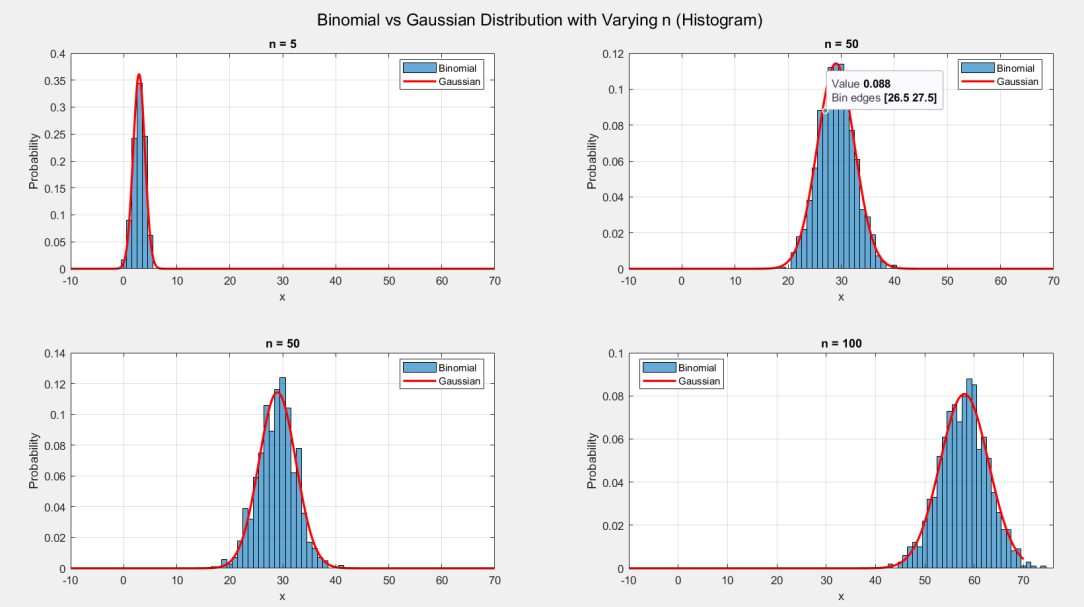


Figure 5: Binomial vs Normal distribution with p = 0.58 and n = [5, 50, 50, 100]

Click here to download the code

2. **Bernoulli Distribution:**

   The Bernoulli distribution itself does not directly converge to a Normal distribution. It represents a single trial with only two possible outcomes: success (usually denoted as 1) and failure (usually denoted as 0). As such, it's a discrete distribution. However, when you have a sequence of independent Bernoulli trials (i.e., a Binomial distribution), and the number of trials becomes sufficiently large, the Binomial distribution can approximate a Normal distribution under certain conditions. This is due to the Central Limit Theorem (CLT), which states that the sum (or average) of a large number of independent and identically distributed random variables tends toward a Normal distribution, regardless of the original distribution of the variables. So, in the case of the Bernoulli distribution, when you have a large number of independent trials with a fixed probability of success p, the resulting Binomial distribution approaches a Normal distribution as the number of trials increases. This approximation becomes better as n, the number of trials, increases.

3. **Geometric Distribution:**

   The Geometric distribution does not directly converge to a Normal distribution. However, under certain conditions, when the number of trials becomes sufficiently large, it can exhibit behavior resembling a Normal distribution due to the Central Limit Theorem (CLT). The Central Limit Theorem states that the sum (or average) of a large number of independent and identically distributed random variables, regardless of their original distribution, tends towards a Normal distribution as the sample size increases. In the case of the Geometric distribution, if p (the probability of success) is not too close to 0 or 1, then the distribution of the sum of n independent Geometric random variables, where n is sufficiently large, may approximate a Normal distribution. For example, consider a scenario where you repeatedly conduct Geometric trials (e.g., flipping a biased coin) to observe the number of trials needed until the first success. If you conduct a large number of such trials and record the average number of trials needed, then according to the Central Limit Theorem, the distribution of these averages could be approximately Normal, provided that p is not extremely close to 0 or 1 and the sample size is sufficiently large. However, it's important to note that the convergence to a Normal distribution may not be perfect, especially if p is close to 0 or 1 or if the conditions of the Central Limit Theorem are not fully met. Additionally, the Geometric distribution is inherently discrete, while the Normal distribution is continuous, so the approximation is not exact.

4. **Negative Binomial Distribution:**

   The Negative Binomial distribution does not directly converge to a Normal distribution. However, in certain conditions, when the parameters of the Negative Binomial distribution are such that the

number of trials becomes sufficiently large, it can exhibit behavior resembling a Normal distribution due to the Central Limit Theorem (CLT). The Central Limit Theorem states that the sum (or average) of a large number of independent and identically distributed random variables, regardless of their original distribution, tends towards a Normal distribution as the sample size increases. In the case of the Negative Binomial distribution, if r (the number of successes required) is sufficiently large and p (the probability of success) is not too close to 0 or 1, then the distribution of the sum of r independent Negative Binomial random variables may approximate a Normal distribution. However, it's important to note that the convergence to a Normal distribution may not be perfect, especially if the parameters are not within certain ranges or if the conditions of the Central Limit Theorem are not met. Additionally, the Negative Binomial distribution is inherently a discrete distribution, while the Normal distribution is continuous, so the approximation is not exact.

3.5.1 Parameter Estimation

1. **Case 1:**

   Mean and Variance known then are not needed for estimation since parameters are already known.

2. **Case 2:**

   Mean known and Variance unknown.

   Then Maximum likelihood estimation (MLE) for the variance is the sample variance.

$$s^2 = \left(\sum (x_i - \mu)^2\right)/n$$

   where x_i is the observed data point and n is the sample size.

   PROOF

   Case II

   Random Sample size $\rightarrow N$ known μ, unknown σ^2 Likelihood function

$$L\left(\sigma^2 \mid x_i, x_2, x_3 \ldots x_N\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left(\frac{n_1-\mu}{2\sigma^2}\right)^2}$$

   MLE for σ^2, we maximize the likelihood function with respect to $\sigma 2$ taking the derivative of the Log-likelihood function.

$$\frac{d \log L}{d\sigma^2}\left(\sigma^2/n_1, n_2 \ldots x_n\right) = 0$$
$$\frac{d}{d\sigma^2}\left(\frac{-n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(n_i - \mu\right)^2\right) = 0$$
$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(n_i - \mu\right)^2$$

   without Bessel's correction since μ is known.

3. **Case 3:**

   Both Mean and Variance Unknown

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

   PROOF

Case -III MLE for both Sample $(\bar{u}) = \frac{1}{n}\sum x_i$ Sample $\left(s^2\right) = \frac{1}{n-1}\varepsilon\left(x_i - \mu\right)^2$ where x_i are the observed data point and n is the sample size.

$$L\left(\mu, \sigma^2 \mid n_i, n_2, n_3 \ldots n_n\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{\left(n_i - \mu\right)^2}{2\sigma^2}}$$

To find MLE, we maximize the likelihood function with respect to μ and σ^2.
Estimate mean (μ)

$$\frac{\partial}{\partial \mu} \log L\left(\mu, \sigma^2 / x_j, x_2, n_3 \ldots, n_n\right) = 0$$
$$\frac{\partial}{\partial \mu}\left(\frac{-n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i - \mu\right)^2\right) = 0$$
$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Estimate variance $\left(\sigma^2\right)$

$$\frac{\partial}{\partial \sigma^2} \log L\left(\mu, \sigma^2 \mid x_1, x_2, x_3 \ldots x_n\right) = 0$$
$$\frac{\partial}{\partial \sigma^2}\left(-\frac{n}{2}\log\left(2\pi\sigma^2\right) = \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i - \mu\right)^2\right) = 0$$
$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \hat{\mu}\right)^2$$

However, it is well known that results in statistics that the MLE, for the variance σ^2, is braised When estimated with $\frac{1}{n}$. so, to correct For this bias, the unbiased, estimation of the variance.

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(n_i - \mu\right)^2$$

known as Bessel's correction

These methods are commonly used for parameter estimation when dealing with different combinations of known and unknown mean and variance in a dataset. The choice of method depends on the specific characteristics of the data and the objectives of the analysis.

4 Conjugate Priors

A conjugate prior is a prior probability distribution that, when combined with a specific likelihood function, results in a posterior probability distribution that belongs to the same parametric family as the prior distribution. Meaning that, a prior distribution $P(\theta)$ is said to be conjugate to the likelihood function $L(\theta \mid x)$ if the resulting posterior distribution $P(\theta \mid x)$ belongs to the same parametric family as the prior distribution $P(\theta)$. Mathematically, this can be expressed as:

$$P(\theta \mid x) \in \Phi \quad \Longleftrightarrow \quad P(\theta) \in \Phi$$

Where Φ represents a parametric family of distributions.

Table 1: A collection of Conjugate Priors for Commonly Encountered Likelihoods [**?**]

Form of the Likelihood Function as a Function of the Parameter of Interest	**Parameter**	**Conjugate prior**	**Posterior distribution of the parameter** $f_{\theta\mid\mathbf{x}}(\theta \mid \underline{\mathbf{x}})$
1. Bernoulli likelihood: $f_{\mathbf{x}\mid\theta}(\mathbf{x} \mid p) \propto p^{n\bar{x}}(1-p)^{n(1-\bar{x})}$ where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	$\theta = p$	$\theta \sim \text{beta}(\alpha,\beta)$ with $\alpha > 0$ and $\beta > 0$, $g(\theta) = \begin{cases} \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} \text{ if } 0 < \theta < 1 \\ 0 \text{ otherwise} \end{cases}$	$\theta \mid \mathbf{x} \sim \text{beta}(\alpha',\beta')$, where $\alpha' = \alpha + n\bar{x}$, $\beta' = \beta + n(1-\bar{x})$
2. Poisson likelihood: $f_{\mathbf{x}\mid\theta}(\mathbf{x} \mid \mu) \propto \begin{cases} \mu^{n\bar{x}}e^{-n\mu} \text{ if } x_i = 0,1,\ldots, \text{ for } i = 1,2,\ldots,n \\ 0 \text{ otherwise,} \end{cases}$ where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	$\theta = \mu$	$\theta \sim G(\alpha,\beta)$ with $\alpha > 0$ and $\beta > 0$, $g(\theta) = \begin{cases} \frac{\theta^{\alpha-1}e^{-\theta/\beta}}{\Gamma(\alpha)\beta^{\alpha}} \text{ if } \theta > 0 \\ 0 \text{ otherwise} \end{cases}$	$\theta \mid \mathbf{x} \sim G(\alpha',\beta')$, where $\alpha' = \alpha + n\bar{x}$, $\beta' = \frac{\beta}{1+n}$
3. Negative binomial likelihood: $f_{\mathbf{x}\mid\theta}(\mathbf{x} \mid p) \propto \begin{cases} p^{nr\bar{x}}(1-p)^{n(\bar{x}-r)} \text{ if } x_i \geq r \text{ for } i = 1,2,\ldots,n \\ 0 \text{ otherwise} \end{cases}$ where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	$\theta = p$	$\theta \sim \text{beta}(\alpha,\beta)$ with $\alpha > 0$ and $\beta > 0$, $g(\theta) = \begin{cases} \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} \text{ if } 0 < \theta < 1 \\ 0 \text{ otherwise} \end{cases}$	$\theta \mid \mathbf{x} \sim \text{beta}(\alpha',\beta')$, where $\alpha' = \alpha + nr$, $\beta' = \beta + n(\bar{x}-r)$
4. Uniform likelihood: $f_{\mathbf{x}\mid\theta}(\mathbf{x} \mid L) \propto \begin{cases} \frac{1}{L^n} \text{ if } L > \max\{x_i\} \\ 0 \text{ otherwise} \end{cases}$	$\theta = L$	$\theta \sim \text{Pareto}(\alpha,\beta)$ with $\alpha > 0$ and $\beta > 0$, $g(\theta) = \begin{cases} \frac{\alpha\beta^{\alpha}}{\theta^{\alpha+1}} \text{ if } 0 > \beta \\ 0 \text{ otherwise} \end{cases}$	$\theta \mid \mathbf{x} \sim \text{Pareto}(\alpha',\beta')$, where $\alpha' = \alpha + n$, $\beta' = \max\{x_1,\ldots,x_n,\beta\}$
5. Pareto likelihood: $\mathbf{x} \mid b \sim \text{Pareto}(a,b)$, $f_{\mathbf{x}\mid\theta}(\mathbf{x} \mid b) \propto \begin{cases} b^{na} \text{ if } \min\{x_1,\ldots,x_n\} > b \\ 0 \text{ otherwise} \end{cases}$	$\theta = b$	$\theta \sim \text{Pareto}(\alpha,\beta)$ with $\alpha > 0$ and $\beta > 0$, $g(\theta) = \begin{cases} \frac{\alpha\beta^{\alpha}}{\theta^{\alpha+1}} \text{ if } 0 > \beta \\ 0 \text{ otherwise} \end{cases}$	$\theta \mid \mathbf{x} \sim \text{Pareto}(\alpha',\beta')$, where $\alpha' = \alpha - \text{an}$, $\beta' = \beta$ with $\alpha > na$

4.1 Binomial distribution is Conjugate prior for a Binomial likelihood

Binomial distribution

$$f(y)\tau) = \binom{n}{y}\tau^{y}(1-\tau)^{n-y}$$

Shape depends on $\tau^{y}(1-\tau)^{n-y}$

$$\tau \in [0,1]$$

$$f(y) = \exp\left\{\frac{\theta y - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

Beta (a,b) distribution $\rightarrow$ valid on $[0,1]$

$$\text{beta } (a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}m^{a-1}(1-m)^{b-1}$$

$$\text{prior for } \tau \quad g(\tau) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\tau^{a-1}(1-\tau)^{b-1}$$

posterior $\propto$ prior * likelihood

$$\begin{aligned} &\propto \tau^{a-1}(1-\tau)^{b-1} \times \tau^{y}(1-\tau)^{n-y} \\ &\propto \tau^{(a+y-1}(1-\tau)^{(b+n-y)-1} \\ &= \text{Beta}(a+y, b+n-y) \\ &= \frac{\Gamma(a+b+n)}{\Gamma(a+y)\Gamma(b+n-y)}\tau^{a+y-1}(1-\tau)^{b+n-y-1} \end{aligned}$$

4.1.1 Example of beta-binomial model

In the beta-binomial model, the problem involves estimating the parameter θ of a binomial distribution given observed data y_N. The data are assumed to be drawn from a binomial distribution with a fixed

number of trials n and an unknown success probability θ. The goal is to estimate the probability θ based on the observed counts of successful trials y_i.

The likelihood function for the binomial distribution is given by:

$$p(y_i|\theta) = \binom{n}{y_i} \theta^{y_i} (1-\theta)^{n-y_i}$$

where y_i represents the number of successful trials out of n total trials.

For Bayesian inference, a prior distribution is specified for the parameter θ. In this example, a beta distribution is chosen as the prior distribution for θ, denoted as $\text{Be}(\theta; a, b)$. The beta distribution is parameterized by two shape parameters a and b, which can be adjusted to reflect prior beliefs about the value of θ. A flat prior, where $a = b = 1$, is chosen in this case, representing a lack of prior knowledge about θ.

The posterior distribution of θ given the observed data y_N can be obtained using Bayes' theorem:

$$p(\theta|y_N) \propto \left(\prod_{i=1}^{N} p(y_i|\theta) \right) p(\theta)$$

Due to the conjugacy between the beta prior and the binomial likelihood, the posterior distribution $p(\theta|y_N)$ also follows a beta distribution. The parameters of the posterior beta distribution, denoted as A and B, can be calculated based on the observed data and the prior parameters a and b:

$$p(\theta|y_N) = \text{Be}(\theta; A, B)$$

where:

$$A = a + \sum_{i=1}^{N} y_i$$

$$B = b + nN - \sum_{i=1}^{N} y_i$$

This posterior distribution provides a complete characterization of the uncertainty in the parameter θ based on the observed data y_N and the prior information.

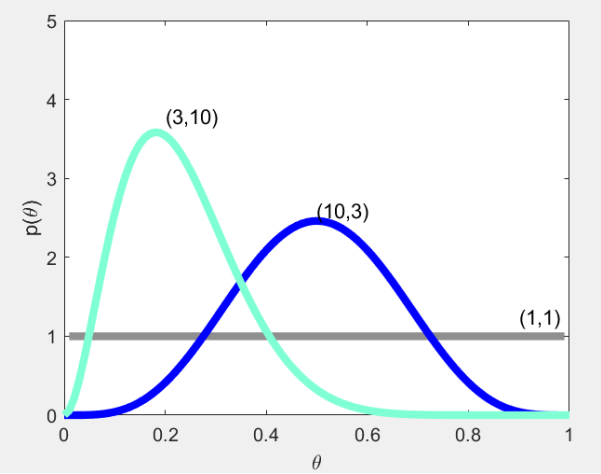


Figure 6: The beta distribution is shown with three different parameters

Click here to download the code

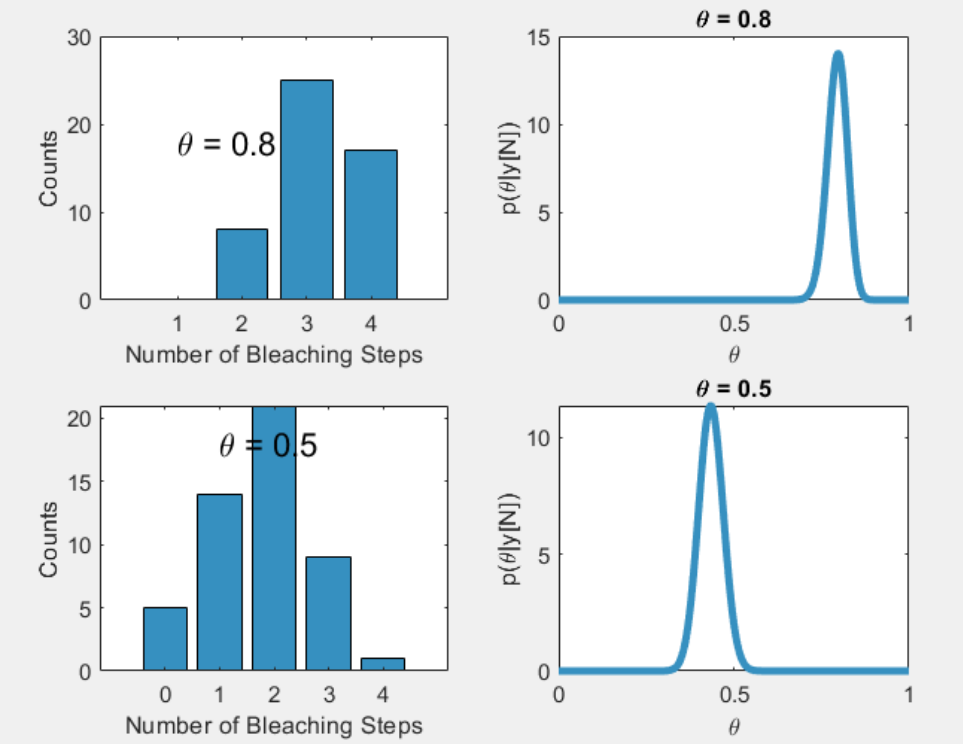


Figure 7: Posterior estimation in the beta-binomial model. (Left) Samples drawn from a binomial distribution with (Right) The resulting posterior distributions of θ.

Click here to download the code

4.2 Gamma distribution is Conjugate prior for Poisson Likelihood

Conjugate Prior is Gamma distribution

$$g(\lambda; r, v) = \frac{v\lambda^{r-1}e^{-v\lambda}}{\Gamma(r)} \propto \lambda^{r-1}e^{-v\lambda}$$

Prior X Likelihood

$$= \frac{V^r}{\Gamma(r)}\lambda^{r-1}e^{-v^{\lambda}} \times \frac{1}{y!}\lambda^y e^{-\lambda}$$
$$\propto \lambda^{r-1}e^{-v\lambda}\lambda^y e^{-\lambda}$$
$$= \lambda^{r+y-1}e^{-(v+1)\lambda}$$
$$\text{Gamma}\,(\lambda_;(r+y),(v+1))$$
$$= \frac{(v+1)^{r+y}\lambda^{r+y-1}e^{-(v+1)\lambda}}{\Gamma(r+y)}$$

4.3 Normal prior Normal likelihood Normal posterior distribution

Likelihood, n iid samples from a Normal; $N\left(\mu, \sigma^2\right)$

$$N\left(\bar{y}, \frac{\sigma^2}{n}\right)$$

Prior : $N\left(m, s^2\right)$

Prior $\times Likelihood will be \mathrm{N}\left(\mu_n, \sigma_n^2\right)$

As we know s^2 and σ, we can ignore the constant term

$$\left[\frac{1}{\sqrt{2\pi\pi^2}}\right]^n \times \left[\frac{1}{\sqrt{2\pi\pi^2}}\right]$$

$$
\begin{aligned}
&g\left(\mu \mid y_1, \cdots, y_n\right) \propto g(\mu) \times f\left(y_1, \ldots, y_n \mid \mu\right) \\
&\quad \alpha \exp\left\{-\frac{1}{2s^2}(\mu - m)^2\right\} \times \exp\left\{-\frac{1}{2\sigma^2}(\mu - \bar{y})^2\right\} \\
&\quad = \exp\left\{-\frac{1}{2s^2}\left(\mu^2 + m^2 - 2\mu m\right) - \frac{n}{2\sigma^2}\left(\mu^2 + \bar{y}^2 - 2\mu\bar{y}\right)\right\} \\
&\quad = \exp\left\{-\frac{1}{2}\left[\mu^2\left(\frac{1}{s^2} + \frac{n}{\sigma^2}\right) - 2\mu\left(\frac{m}{s^2} + \frac{n\bar{y}}{\sigma^2}\right) + \frac{m^2}{s^2} + \frac{\bar{y}^2}{\sigma^2}\right]\right\} \\
&\text{Expect } \exp\left\{-\frac{1}{2\pi^2}\left[\mu^2 + \mu^2 - 2\mu m\right]\right.
\end{aligned}
$$

Expect:

$\exp\left\{-\frac{1}{2\sigma_n^3}\left[\mu^2 + \mu_n^2 - 2\mu\mu_n\right]\right\}$

$$
+\frac{2\mu\mu_n}{2\sigma_n^2} = \frac{\mu\mu_n}{\sigma_n^2} = \mu\left(\frac{m}{s^2} + \frac{n\bar{y}}{\sigma}\right)
$$

$$
\frac{\mu_n}{\sigma_n^2} = \frac{m\sigma^2 + n\bar{y}s^2}{\sigma_{s^2}^2}
$$

$$
\begin{aligned}
&\frac{\mu_n}{\sigma_n^2} = \frac{m\sigma^2 + n\bar{y}s^2}{\sigma^2 s^2} \\
&\mu_n = \sigma_n^2\left(\frac{m\sigma^2 + n\bar{y}s^2}{\sigma^2 s^2}\right) = \frac{\sigma^2 s^2}{\sigma^2 + s^2 n} + \left(\frac{m\sigma^2 + n\bar{y}s^2}{\sigma^2 s}\right) \\
&\quad = \frac{m\sigma^2 + n\bar{y}s^2}{\sigma^2 + s^2 n} = \frac{\sigma^2}{\sigma^2 + s^2 n}m + \frac{s^2}{\sigma^2 + s^2 n}n\bar{y}
\end{aligned}
$$

4.4 Conjugate Prior for Variance of Normal Distribution with known mean

Normal distribution with a known mean μ, but unknown variance σ^2 Likelihood $f\left(x_1, x_n \mid \mu, \sigma^2\right) = f\left(x_n \mid \mu, \sigma^2\right) = \prod \frac{2}{\sqrt{2\pi\sigma^{-2}}} \exp\left\{-\frac{(x_i - \mu)^2}{2\sigma^2}\right\}$

$$
f\left(x_n \mid \mu, \sigma^2\right) = (2\pi)^{-\frac{n}{2}}\left(\sigma^2\right)^{-\frac{1}{2}} e^{-\frac{1}{\sigma}\cdot\sum_{i=1}^{n}\frac{(x_1 - \mu)^2}{2}}
$$

Inverse Gamma: $f(y; \alpha, \beta) = \frac{\beta^-}{\Gamma(\alpha)} y^{-(\alpha+1)} e^{-\beta/y}$ $g\left(\sigma^2\right) = \frac{\beta^\alpha}{\Gamma(\alpha)}\left(\sigma^2\right)^{-(\alpha+1)} e^{-\beta\sigma^2}$ be the prior.

prior $\times$*likelihood* $\propto$ *posterior*

Ignore constants [prior: $\frac{\beta^\Gamma}{\Gamma(\alpha)}$) likelihood $(2\pi)^{-\frac{n}{2}}$]

prior $\times$*likelihood* $\propto \left(\sigma^2\right)^{-(\alpha+1)} e^{-\frac{\beta}{\sigma^2}} * \left(\sigma^2\right)^{-\frac{n}{2}} e^{-\frac{1}{\sigma^\sigma}\sum_{i=1}^{n}\frac{(x_i - \mu)}{2}}$

$$
\equiv \left(\sigma^2\right)^{-\left(\alpha + \frac{n}{2} + 1\right)} e^{-\frac{1}{\sigma^2}\left(\beta + \sum_{i=1}^{n}\left(\frac{1}{2} - \mu\right)^2\right)}
$$

$\propto$ *posterior*

Inverse Gamma with parameters $\left(\alpha + \frac{n}{2}, \beta + \sum_{i=1}^{n}\left(x_i - \mu\right)^2\right)$

$$
g\left(\sigma^2 \mid x_n, \mu\right) = \frac{\left[\beta + \frac{\sum(x_i - \mu)^2}{2}\right]^{\alpha + \frac{n}{2}}}{\Gamma\left(\alpha + \frac{n}{2}\right)}\left(\sigma^2\right)^{-\left(\alpha + \frac{n}{2} +\right)} e^{-\frac{1}{\sigma^2}\left(\beta + \sum_{i=1}^{n}\frac{(x_i - \mu)}{2}\right)}
$$

4.5 Conjugate Prior for Precision of Normal Distribution with known mean

Normal with known mean μ, but unknown precision τ (Recall that $\tau = 1/\sigma^2$)

Likelihood: $f\left(x_1, \ldots, x_n \mid \mu_1\tau\right) = \prod_{i=1}^{n}\sqrt{\frac{\tau}{2\pi}} \exp\left\{-\frac{\tau}{2}\left(x_i - \mu\right)^2\right\}$

$$
f\left(x_1, \ldots, x_n \mid \mu, \tau\right) = \tau^{\frac{n}{2}}(2\pi)^{-\frac{n}{2}} e^{-\tau\sum_{i=1}^{n}\frac{(x_i - \mu))^2}{2}}
$$

Gamma distribution

$$f(y;\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y}$$

Prior $\quad g(\tau) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \tau^{\alpha-1} \beta^{-\beta\tau}$
Prior: $g(\tau) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \tau^{\alpha-1} \beta^{-\beta\tau}$

Posterior $\propto$ *Prior* $\times$ *Likelihood*

$$\begin{aligned}
&\propto \tau^{\alpha-1} e^{-\beta\tau} * \tau^{n/2} e^{-\tau \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2}} \\
&\equiv \tau^{(\alpha+\hat{\beta}_2)-1} e^{-\tau\left(\beta+\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2}\right)} \\
&\text{Gamma}\left(\alpha + \frac{n}{2}, \beta + \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2}\right) \\
&= \frac{\left[\beta + \sum_{i=1}^{n} \frac{(x_1-\mu)^2}{2}\right]^{(\alpha+n/2)}}{\Gamma\left(\alpha+\frac{n}{2}\right)} \tau^{\left(\alpha+\frac{n}{2}\right)-1} e^{-\tau\left(\beta+\sum_{i=1}^{\left(\frac{x_i}{}-\mu^2\right)^2}\right)}
\end{aligned}$$

5 MCMC: Markov Chain Monte Carlo

5.1 Gibbs Sampling

Gibbs sampling is a Markov chain Monte Carlo (MCMC) algorithm used for sampling from complex probability distributions, especially when the joint distribution of variables is difficult to compute directly. The key insight behind Gibbs sampling is that even if the joint distribution is complex, the conditional distributions of each variable given the others may be simpler and more tractable. Gibbs sampling leverages this by iteratively sampling each variable from its conditional distribution while holding all other variables fixed at their current values.

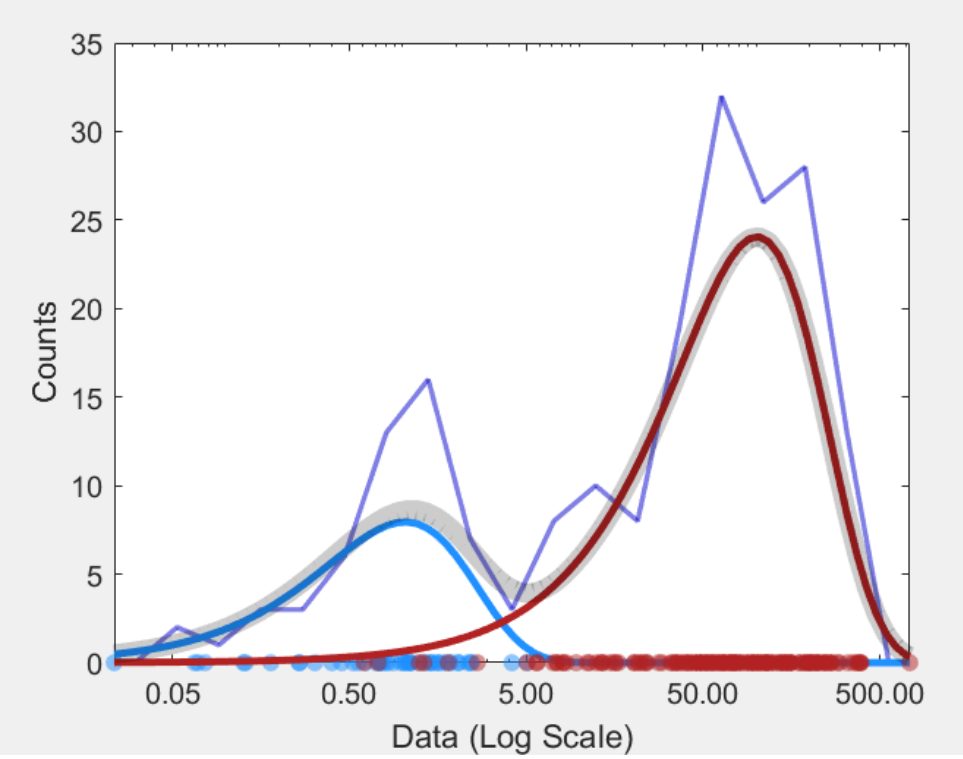


Figure 8: An exponential mixture model. Simulated data drawn from a mixture of two exponential distributions. Data are plotted logarithmically for visualization. The result of using a Gibbs sampler to infer the parameters of a two-component mixture model are also shown.

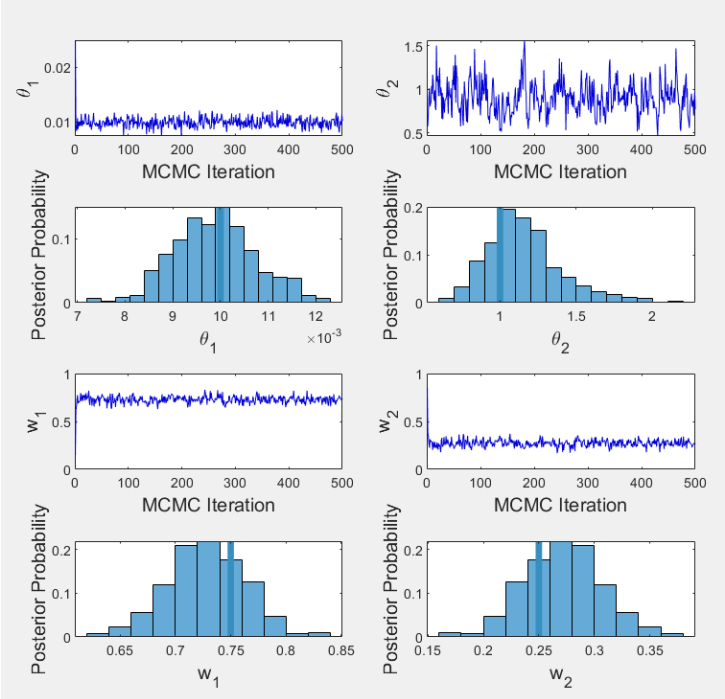


Figure 9: MCMC trajectories and marginal posterior distribution

Click here to download the code for Gibbs sampler

5.2 Metropolis-Hastings algorithm

Metropolis-Hastings is a specific MCMC algorithm designed for sampling from a target probability distribution by generating a Markov chain. In the Metropolis-Hastings algorithm, a proposal distribution is used to generate candidate samples, and then a decision is made whether to accept or reject each proposed sample based on an acceptance probability. The acceptance probability is determined by comparing the probability of the proposed sample under the target distribution to the probability of the current sample

The Metropolis-Hastings algorithm is based on the principle of detailed balance, which is a fundamental concept in statistical mechanics. The *Principle of Detailed Balance* states that for a system in equilibrium, the rate of transition from one state to another should be equal to the rate of transition from the second state to the first. Mathematically, for any two states i and j, the product of the probability of being in state i and the transition probability from i to j should be equal to the product of the probability of being in state j and the transition probability from j to i. In terms of probabilities, it can be written as $\pi(i) \cdot P(i \to j) = \pi(j) \cdot P(j \to i)$, where:

- $\pi(i)$ is the equilibrium probability of being in state i.
- $P(i \to j)$ is the transition probability from state i to state j.

The Metropolis-Hastings algorithm leverages this principle to sample from a target probability distribution $\pi(x)$ by constructing a Markov chain with the desired equilibrium distribution.

Initialization: Start with an initial state x_0 and set $t = 0$.

Proposal Distribution: Propose a candidate state x' from a proposal distribution $q(x'|x_t)$. This distribution should be symmetric, meaning $q(x'|x_t) = q(x_t|x')$.

Acceptance Probability: Compute the acceptance probability $\alpha(x_t, x')$ as the ratio of the target distribution probabilities:

$$\alpha(x_t, x') = \min\left(1, \frac{\pi(x')}{\pi(x_t)} \cdot \frac{q(x_t|x')}{q(x'|x_t)}\right)$$

Accept or Reject: Generate a uniform random number u from $[0, 1]$. If $u \leq \alpha(x_t, x')$, accept the proposed state x' and set $x_{t+1} = x'$. Otherwise, reject the proposed state and set $x_{t+1} = x_t$.

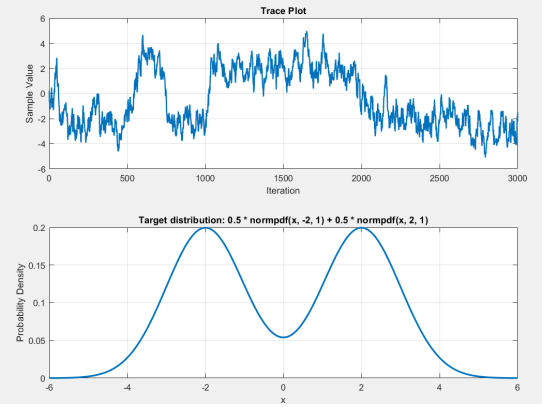
Iteration: Increment t and repeat steps 2-4 until a sufficient number of samples are obtained.

The key insight here is that the acceptance probability ensures that the detailed balance principle is satisfied. If $\pi(x') > \pi(x_t)$, the proposed state is more probable than the current state, so it's always accepted. If $\pi(x') < \pi(x_t)$, it might still be accepted with a probability proportional to the ratio of their probabilities, which ensures detailed balance.

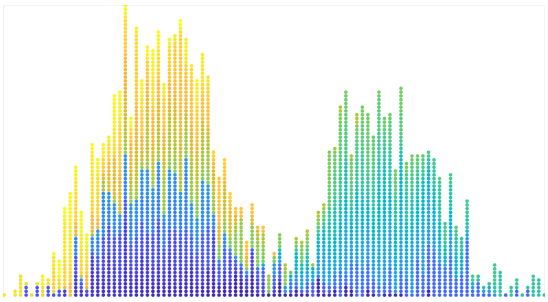
By iterating this process, the Markov chain eventually reaches equilibrium, and the samples drawn from the chain converge to the desired distribution $\pi(x)$.

5.2.1 Example 1: Sampling from Gaussian mixture

We implement a Metropolis-Hastings algorithm to sample from a mixture of two Gaussian distributions and visualize the sampling process using static and animated plots. The mixture of Gaussians is defined by two standard normal distributions, one centered at -2 and the other at 2, with equal weights. The code generates a sequence of samples from this distribution using the Metropolis-Hastings algorithm and visualizes the sampling process using a trace plot and an animated plot.



(a) MCMC trajectry and target distribution



(b) Generated samples

Figure 10: Metropolis-Hastings algorithm for sampling from a mixture of Gaussian distributions

Click here to download the code

5.2.2 Example 2: Voltage-gated Ion-channel

$$P(\epsilon) = \frac{\exp(-\varepsilon/kT)}{2} - (i)$$

T-) Temperature

$$\varepsilon \to p(v)$$

$k \to$ BoLtzman constant. $Z \to$ partition function, which normalizes, the distribution So that probability sum to 1

$$E = -(v - a) - (ii)$$

$\varepsilon \propto v$

$a \to$ midpoint of voltage

(ii) in (i)

$$P(v) = \exp((v - a)/k\tau)/2$$

To simplify this function we can define a new variable, the slope factor (h) as $b = kT$ now.

$$P(v) = \exp((v - a)/b)/z$$

now to normalize the distribution we need to find partition function F(2) which is given by

$$z = \int \exp(v - a)/b$$
$$z = b \int \exp(\mu)du$$
$$z = be^{\mu}$$
$$z = be\text{^}((v - a)(b)).$$

Substituting this back to express for $p(v)$ we get $p(v)\exp((v - a)/b)/be\text{^}((v - a)/b)$ Simplify this express we get.

$$P(v) = \frac{1}{1 + \exp(-(v - a)(b)).}$$

Multiply a and b by $\exp(v - a)/b)$

$$p(v) = \exp((v - a)/b)/\exp((v - a)/b) + 1)$$

by changing New variable

$$F(a,b,v) = P(v)$$
$$F(a,b,v) = \frac{1}{1+\exp(-(v-a)b)).}$$

$y_i \sim f\left(a,b,V_i\right) + N\left(0,\sigma^2\right), where \mathrm{N}(\mu,\sigma)$ denotes a normal distribution with mean μ and standard deviation σ. Given this, our likelihood function is simply a normal distribution centered at f and with variance σ^2,

$$p\left(y_i \mid \ldots\right) = N\left(f\left(a,b,V_i\right),\sigma^2\right).$$

We assume that each data point arises from f and some independent and identically distributed noise, so the posterior distribution is

$$p\left(a,b,\sigma^2 \mid y_N\right) \propto \left(\prod_{i=1}^{N} N\left(f\left(a,b,V_i\right),\sigma^2\right)\right) p(a)p(b)p\left(\sigma^2\right).$$

$$y_i \sim F\left(a,b,v_i\right) + N\left(0,\sigma^2\right)$$

$F($ centred at mean $)$

$$.P\left(y_i \mid \ldots\right) = N\left(F\left(a,b,v_i\right),\sigma^2\right)$$
$$y_i = y_1, y_2, y_3 \ldots y_N$$

$P\left(y_N \mid \ldots\right)$ when it is cores to Normal distribution $\rightarrow$ Let this be

$$P(A \mid B) = N\left(F,\left(a,b,v_i\right),\sigma^2\right)$$

So,

$$P(B/A) = \frac{P(A/B)\cdot P(B)}{\text{evidence } P(A)}$$
$$P\left(a,b,\sigma^2/y_N\right) = N\left(F\left(a,b,v,\cdot,\sigma^2\right)\cdot P(a)\cdot P(b)\cdot P\left(\sigma^2\right)\right.$$

Here,
Posterior $\propto$ *likelihood* $\times$ *prior*.

$$P\left(a,b,\sigma^2\right) y_N\right) = \pi_{i=1}^{N}\left(N\left(F\left(a,b,v_i\right),\sigma^2\right) P(a)P(b)\cdot\sigma^2\right)$$

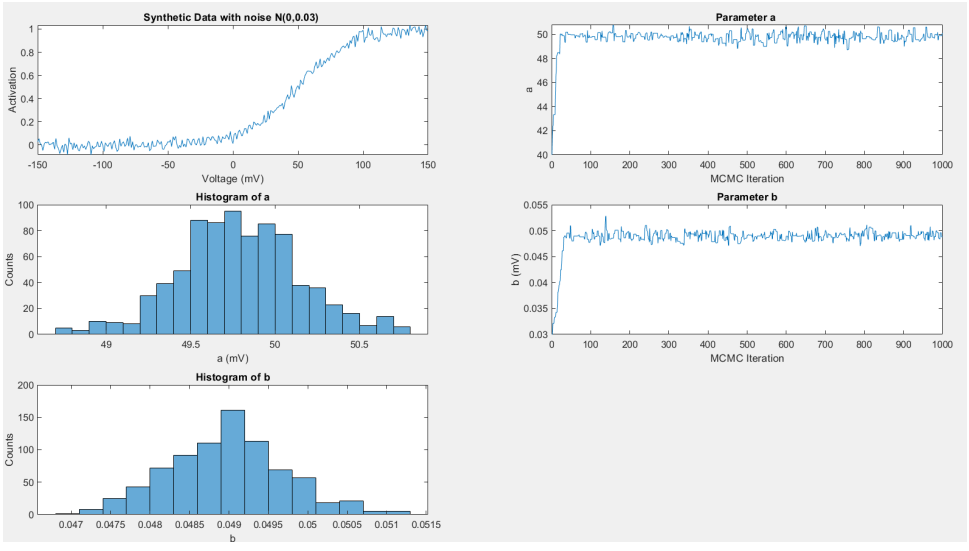


Figure 11: Demonstration of Metropolis-Hastings algorithm to analyze ion channel activation data

Click here to download the code

6 Approximate Bayesian Computation

Approximate Bayesian Computation(ABC) methods provide a robust framework for Bayesian inference when dealing with complex models where likelihoods are difficult to compute. ABC methods allow us to estimate posterior distributions without needing to calculate the likelihoods. Many dynamical systems in various fields, particularly biology, are modeled using differential equations, but reliable parameter information is often missing, and multiple models may exist for the same system. Therefore it is very difficult to compute the likelihood for such systems.

6.1 Methods

6.1.1 ABC Rejection Algorithm

The simplest ABC algorithm is the *ABC* rejection sampler (Pritchard et al. 1999), which is as follows.
$R1$ Sample θ^* from $\pi(\theta)$.
R2 Simulate a dataset x^* from $f\left(x \mid \theta^*\right)$.
$R3$ If $d\left(x_0, x^*\right) \leq \epsilon$, accept θ^*, otherwise reject.
$R4$ Return to $R1$.
Now we have tried to implement the ABC Rejection sampler for normal distribution with variance 1 and the true mean parameter as 5. The goal is to estimate the mean (parameter) of a normal distribution based on observed data. The observed data is a sequence of 100 random numbers generated from a normal distribution with the true mean and a standard deviation of 1. The prior distribution for the parameter is taken to be uniform, $\mu \sim \mathrm{U}(-10, 10)$. The distance function calculates the absolute difference between the means of the observed and simulated data sets. The tolerance level (epsilon) is set to 0.1, and the number of accepted particles (N) is set to 1000. The code takes 388.104 seconds to run. The code for the implementation can be found at the following link: https://drive.google.com/file/d/1hVYyKl8WuFF6Gt5kDSAaswPlSz57MO5y/view?usp=sharing.

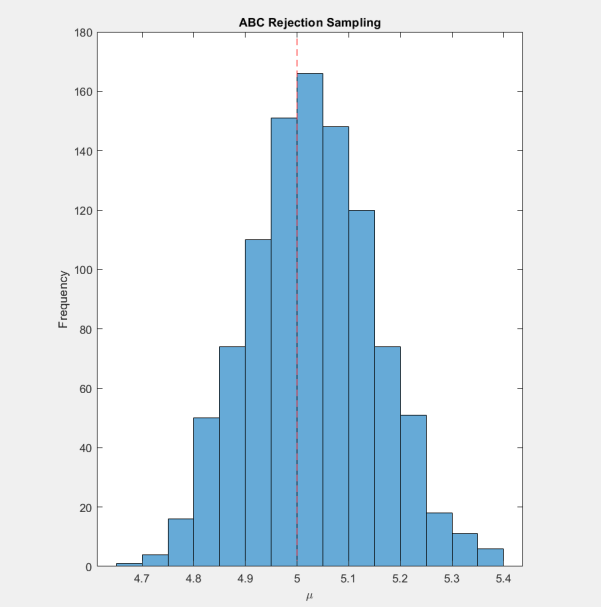


Figure 12: Histogram of the parameter μ, the red dotted line is the true mean

The main disadvantage of ABC rejection sampling is its low acceptance rate, especially when the prior distribution significantly differs from the posterior distribution. This occurs because rejection sampling relies on randomly sampling parameters from the prior distribution, which may not efficiently explore regions of high posterior probability. As a result, many proposed parameter values are rejected, leading to slow convergence and inefficient sampling of the posterior distribution.

6.1.2 ABC SMC Algorithm

In ABC SMC, a number of sampled parameter values (called particles), $\left\{\theta^{(1)}, \ldots, \theta^{(N)}\right\}$, sampled from the prior distribution $\pi(\theta)$, are propagated through a sequence of intermediate distributions, $\pi\left(\theta \mid d\left(x_0, x^*\right) \leq \boldsymbol{\epsilon}_i\right), i = 1, \ldots, T-1$, until it represents a sample from the target distribution $\pi\left(\theta \mid d\left(x_0, x^*\right) \leq \epsilon_T\right)$. The tolerances ϵ_i are chosen such that $\epsilon_1 > \cdots > \epsilon_T \geq 0$, thus the distributions gradually evolve towards the target posterior. Particles taken from the previous distribution are labeled with one asterisk. After being perturbed, these particles are labeled with two asterisks.

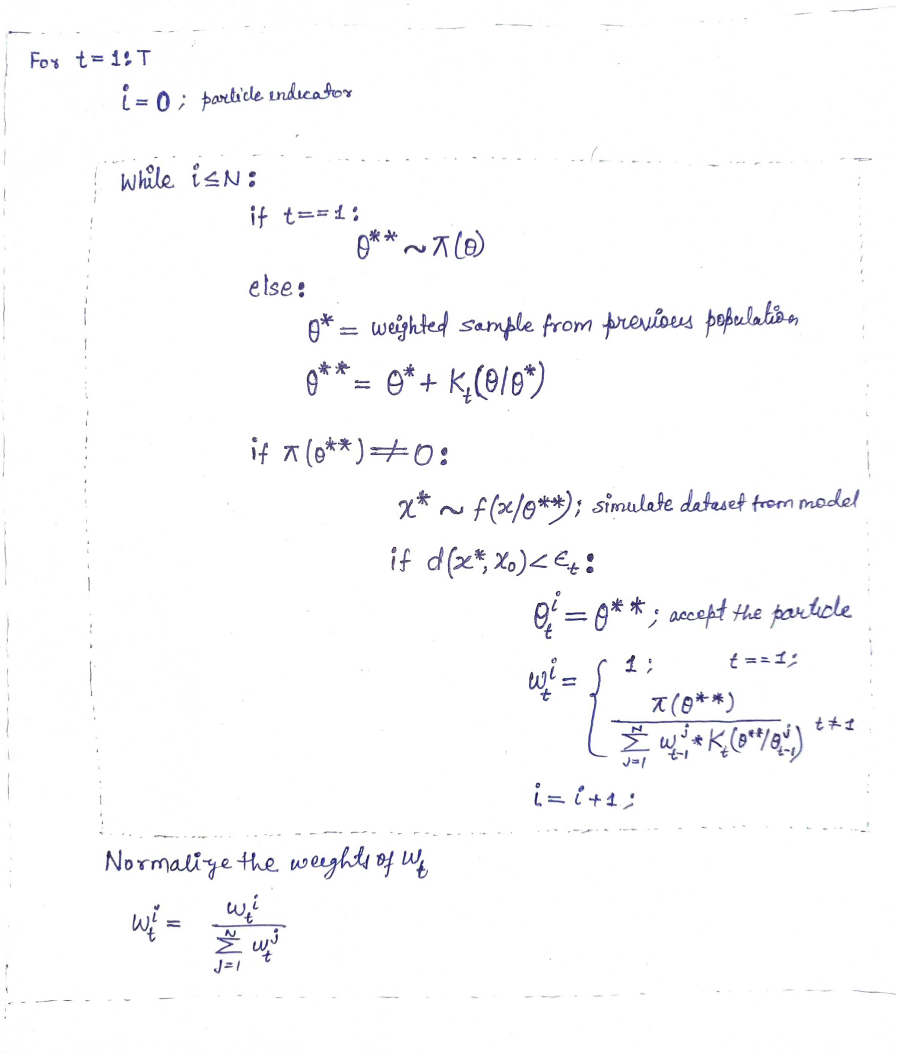


Figure 13: Pseudo code for ABC-SMC

6.2 Deterministic Lotka–Volterra system

Lotka-Volterra (LV) model describes the interaction between prey species, x, and predator species, y, with parameter vector $\theta = (a, b)$

$$\frac{\mathrm{d}x}{\mathrm{d}t} = ax - xy$$
$$\frac{\mathrm{d}y}{\mathrm{d}t} = bxy - y.$$

We generate noisy data points by simulating the system using parameter values of (a, b) = (1, 1) and adding Gaussian noise. We then compare these data points to the simulated data using a distance function, which calculates the sum of squared errors between the two sets of data.

6.2.1 ABC Rejection sampler on deterministic LV system

we implemented the Approximate Bayesian Computation (ABC) rejection sampler method with an epsilon value of 4.3. The prior distributions for parameters a and b were set as uniform distributions, a and b are taken to be uniform, $a, b \sim U(-10, 10)$. Our goal was to obtain 1000 accepted particles using this approach.

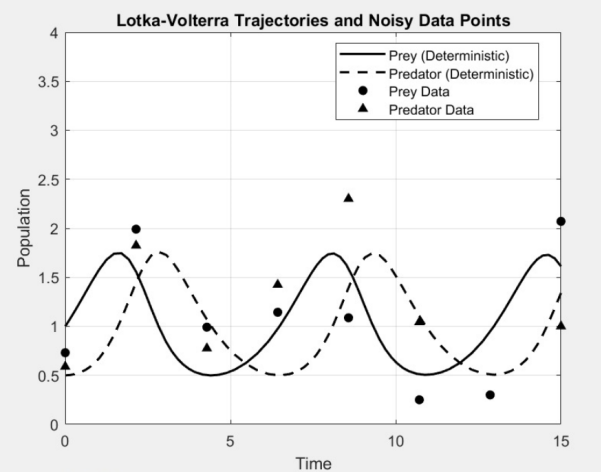


Figure 14: Trajectories of prey (solid curve) and predator (dashed curve) populations of the deterministic LV system and the data points (circles, prey data; triangles, predator data).

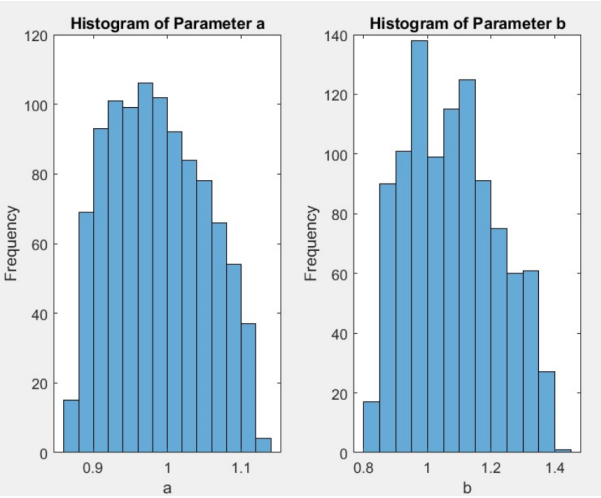


Figure 15: Parameters inferred by the ABC rejection sampler.

The above code with file name `LV_deterministic_ABC_Rejection.m` can be found through the following link: https://drive.google.com/file/d/1N7VwaFLnaIq5WTeetb8hSujH9xoT0vyz/view?usp=sharing.

- referencing material including proof and codes can be found through the following link:

https://drive.google.com/drive/folders/1qmoFwt8IKArDFR1RFiZi8TIl5Tg$_K kYJ$.