Miranda House
UNIVERSITY OF DELHI

STATISTICAL METHOD FOR DATA ANALYSIS IN COSMOLOGY

A Project Report Submitted for Summer Internship

At

DS KOTHARI RESEARCH CENTRE,
MIRANDA HOUSE COLLEGE

by

Ankit Sharma
Aryan Kumar Mandal
Keshav Sharma
Shruti Gera

Supervisors:	Dr. AbhaDev Habib
	Dr. Nisha Rani
Submission Date:	2024-07-04

Declaration

This report encapsulates the comprehensive study and application of various statistical methods essential for data analysis in cosmology. The primary techniques discussed and utilized in this project include Monte Carlo simulations, Bayes' theorem, chi-square fitting, likelihood analysis, and Markov Chain Monte Carlo (MCMC) methods using Python.

These methodologies have been explored and implemented to analyze cosmological data, thereby contributing to the understanding and interpretation of the universe. The project has been conducted with utmost sincerity and diligence, adhering to academic and research standards.

We declare that the work presented in this report is original and has not been submitted elsewhere for any degree or diploma.

Abstract

The field of cosmology relies heavily on statistical methods to analyze and interpret vast amounts of data gathered from observations of the universe. This project report, titled **"Statistical Method for Data Analysis in Cosmology,"** presents a detailed study and application of several key statistical techniques essential for cosmological data analysis. Conducted as part of a Summer Internship at **DS Kothari Research Centre, Miranda House College**, this project encompasses the following methodologies:

1. Monte Carlo Simulations: Used for generating random samples and understanding the probabilistic nature of cosmological phenomena.

2. Bayes' Theorem: Applied to update the probability estimates of hypotheses based on observational data.

3. Chi-Square Fitting: Utilized for assessing the goodness of fit between theoretical models and observed data.

4. Likelihood Analysis: Implemented to estimate the parameters of cosmological models by maximizing the likelihood function.

5. Markov Chain Monte Carlo (MCMC): Employed to sample from complex probability distributions and perform parameter estimation.

These statistical methods have been implemented using Python, allowing for efficient and effective data analysis. The application of these techniques to cosmological data has provided insights into the underlying physical processes and contributed to the broader understanding of the universe. This report not only demonstrates the practical utility of these methods in cosmology but also underscores their importance in the broader context of scientific research and data analysis.

The findings and methodologies detailed in this report are expected to aid future research in cosmology, offering robust tools and techniques for analyzing complex data sets.

Acknowledgements

We would like to express our heartfelt gratitude to our supervisors at the DS Kothari Research Centre, Miranda House College, for their unwavering support and guidance throughout the duration of this project. Their insights and feedback have been invaluable in shaping this report.

We extend our special thanks to Dr. Akshay Rana for his expert advice and continuous encouragement, which greatly facilitated our research. His mentorship has been a cornerstone of our learning experience during this internship.

Additionally, we would like to acknowledge the profound impact of Dr. Barbara Ryden's lectures, which have significantly deepened our understanding of cosmology. Her teachings have been a vital resource and inspiration for this project.

Finally, we are grateful to our peers and the entire staff at Miranda House College for creating a conducive and stimulating environment for research and learning.

Contents

Chapter 1

Introduction

1.1 Purpose of the Project

The primary purpose of this project is to explore, understand, and apply various statistical methods that are crucial for the analysis of cosmological data. Cosmology, the scientific study of the large scale properties of the universe as a whole, relies on vast and complex data sets obtained from observations such as the Cosmic Microwave Background (CMB), galaxy surveys, and other astronomical phenomena. Proper analysis of this data is essential for deriving meaningful scientific insights and advancing our understanding of the universe.

1.2 Key Objectives

1.2.1 Application of Statistical Technique:

- The project aims to delve into specific statistical methods including Monte Carlo simulations, Bayes' theorem, chi-square fitting, likelihood analysis, and Markov Chain Monte Carlo (MCMC). Each of these methods offers unique advantages in handling and interpreting cosmological data.

1.2.2 Improving Data Analysis Skills:

- Through the practical application of these statistical methods, the project seeks to enhance the data analysis skills of the participants. This includes learning how to preprocess data, apply statistical models, and interpret results within the context of cosmology.

1.2.3 Enhancing Understanding of Cosmological Phenomena:

- By analyzing real cosmological data, the project aims to provide insights into the underlying physical processes governing the universe. This includes studying the distribution of galaxies, understanding cosmic background radiation, and investigating dark matter and dark energy.

1.2.4 Implementation Using Python:

- By leveraging Python, a powerful programming language widely used in scientific computing, the project intends to implement these statistical techniques efficiently. Python's extensive libraries and tools facilitate sophisticated data analysis and visualization, making it an ideal choice for this study.

1.2.5 Contributing to Scientific Research:

- The methodologies and findings of this project are expected to contribute to the broader field of cosmological research. By providing robust tools and techniques for data analysis, the project supports ongoing and future research endeavors in cosmology.

1.3 Significance:

The significance of this project lies in its potential to bridge the gap between theoretical statistical methods and practical data analysis in cosmology. By doing so, it not only aids in the accurate interpretation of cosmological data but also fosters a deeper understanding of the universe's structure, origin, and evolution. The skills and knowledge gained through this project are valuable for students and researchers, preparing them for advanced studies and professional work in astrophysics and related fields.

Chapter 2

Bayesian Inference

2.1 Introduction

Bayes' Theorem is a fundamental principle in probability theory and statistics that describes how to update the probability of a hypothesis as more evidence or information becomes available. It is named after the Reverend Thomas Bayes, who first provided an equation that allows new evidence to update beliefs.

In the context of cosmology, Bayes' Theorem is used to update the probability of cosmological models or parameters given new observational data. This theorem plays a crucial role in Bayesian inference, a statistical method that incorporates prior knowledge along with new evidence to make probabilistic statements about parameters.

2.2 Formula

Bayes' Theorem is mathematically expressed as:

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

where:

- $P(H|E)$ is the posterior probability: the probability of the hypothesis H given the evidence E.
- $P(H|E)$ is the posterior probability: the probability of the hypothesis H given the evidence E.
- P(H) is the prior probability: the initial probability of the hypothesis H before observing the evidence.
- P(E) is the marginal likelihood or evidence: the total probability of the evidence E.

2.3 Application in Cosmology

In cosmology, Bayes' Theorem can be applied to update the probability of different cosmological models or parameters based on new observational data, such as

measurements of the Cosmic Microwave Background (CMB), supernovae data, or galaxy surveys.

2.4 Example

Suppose we have two competing cosmological models, $M_1 and M_2$,and we obtain new observational data D. We can use Bayes' Theorem to update our beliefs about the models based on this data.

1. Prior Probability $P(M_1) and P(M_2)$: These are our initial beliefs about the likelihood of each model before seeing the new data.

2. Likelihood $P(D|M_1) and P(D|M_2)$:These represent how probable the new data is under each model.

3. Marginal Likelihood $P(D)$:This is the total probability of observing the data, summed over all possible models.

Using Bayes' Theorem, we update our beliefs to find the posterior probabilities $P(M_1|D) and P(M_2|D)$, which tell us how likely each model is given the new data.

2.5 Practical Implementations

In practice, Bayesian inference in cosmology often involves computational techniques to estimate the posterior distributions, especially when dealing with complex models and large data sets. Markov Chain Monte Carlo (MCMC) methods are frequently used to sample from the posterior distribution.

2.6 Examples

Uniform and Normal PDF

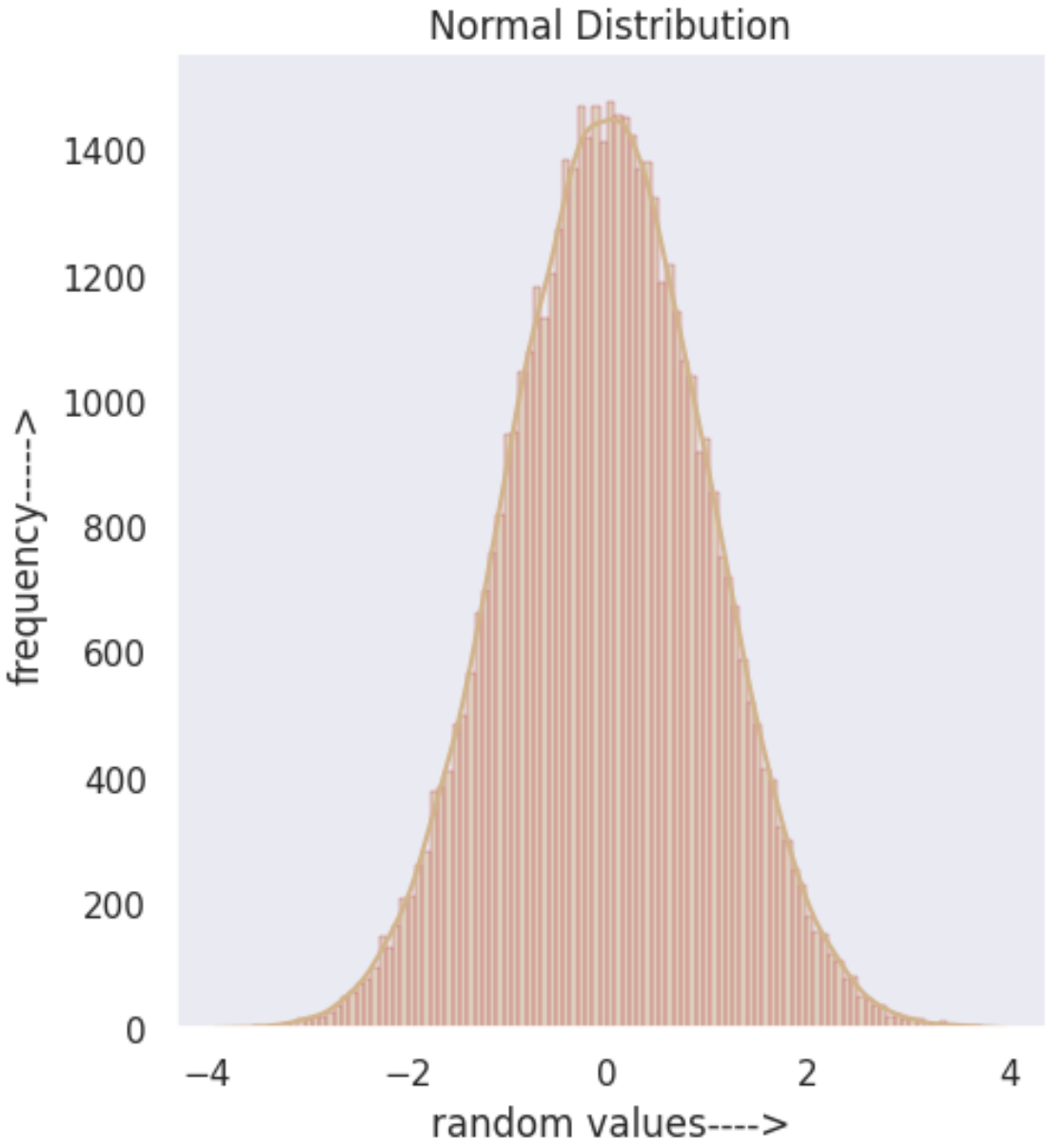


Figure 2.1: Normal Random Distribution

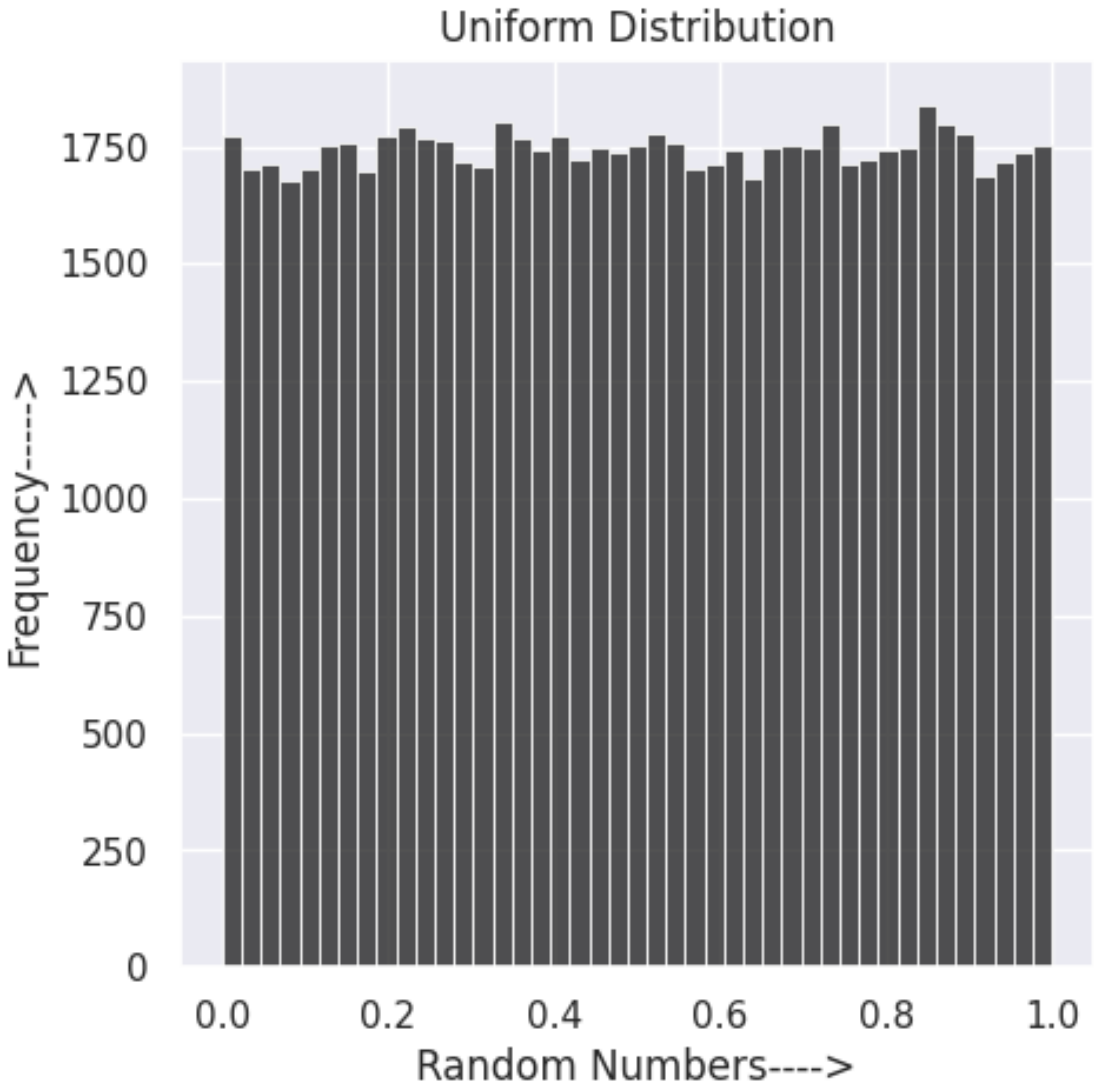


Figure 2.2: Uniform Random Distribution

2.7 Conclusion

Bayes' Theorem is a powerful tool in cosmology for updating our understanding of the universe as new data becomes available. By incorporating prior knowledge and new evidence, it allows for a more flexible and robust analysis of cosmological models and parameters.

Chapter 3

Monte Carlo Simulation

Monte Carlo Simulation Technique

Monte Carlo simulation is a computational technique that uses random sampling to obtain numerical results. It is widely used in various fields, including physics, finance, engineering, and cosmology, to model and analyze complex systems that are analytically intractable.

Principle

The basic principle of Monte Carlo simulation is to use random sampling to explore the behavior of a system. By generating a large number of random samples and calculating the results for each sample, one can approximate the desired quantity with high accuracy.

Steps in Monte Carlo Simulation

1. **Define the Problem**: Identify the system or process to be simulated and the quantity to be estimated.

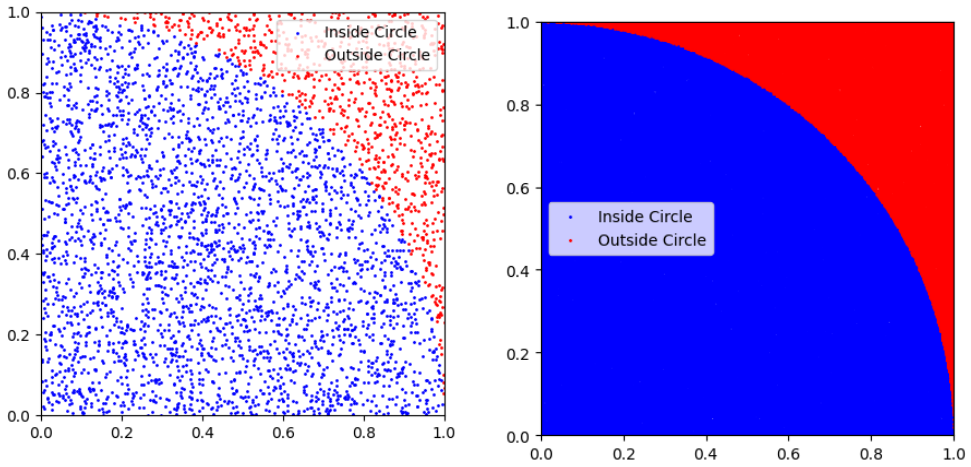
2. **Generate Random Samples**: Use random number generators to produce samples from the probability distributions of the input variables.

3. **Compute the Quantity of Interest**: For each random sample, compute the output based on the defined model or equations.

4. **Analyze the Results**: Aggregate the results from all samples to estimate the desired quantity and assess its uncertainty.

Application in Cosmology

In cosmology, Monte Carlo simulations are used to model and analyze various phenomena, such as the formation of large-scale structures, the distribution of galaxies, and the behavior of dark matter and dark energy. These simulations help researchers understand the probabilistic nature of cosmological processes and make predictions based on observational data.



(a) Initial (b) Final

Figure 3.1: A figure with two subfigures

Example1: Estimating the Value of π

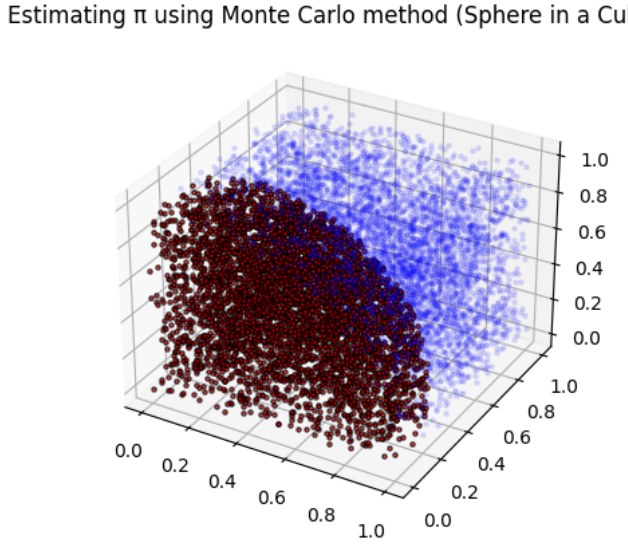
A classic example of Monte Carlo simulation is estimating the value of π. This can be done by randomly generating points in a unit square and counting how many fall inside a quarter circle.

Output:

$$Valueofpifrom MCsimulation = 3.092$$

Example2: SPHERE IN A CUBE

OUTPUT: The estimated value of pi is: 3.1482



Steps in Cosmological Monte Carlo Simulations

1. **Define the Cosmological Model**: Specify the cosmological parameters and the theoretical model to be studied (e.g., ΛCDM model).
2. **Generate Synthetic Data**: Use random sampling to create synthetic data sets that mimic real observational data (e.g., galaxy distributions, CMB maps).
3. **Compare with Observations**: Compute the likelihood of the observed data given the synthetic data generated by the model.
4. **Parameter Estimation**: Use the generated data to estimate the cosmological parameters and their uncertainties.

Link to Colab Notebook

For more detailed implementation, please refer to the Colab notebook available at: Colab Notebook Link

Conclusion

Monte Carlo simulation is a versatile and powerful technique for modeling complex systems in cosmology. By using random sampling to explore the behavior of these systems, researchers can gain insights into the probabilistic nature of cosmological processes and make predictions based on observational data. The flexibility and robustness of Monte Carlo methods make them essential tools in modern cosmological research.

Chapter 4

Chi Square Fitting

4.1 Chi Square Fitting

Chi-square fitting is a statistical method used to determine how well a theoretical model fits observed data. It is commonly used in various fields, including physics, biology, and economics. In cosmology, chi-square fitting is often employed to compare theoretical models with observational data, such as galaxy distributions or cosmic microwave background (CMB) measurements.

4.2 Principle

The chi-square statistic quantifies the difference between observed data and the expected values predicted by a model. For a set of N data points, the chi-square statistic is given by:

$$\chi^2 = \sum_{i=1}^{N} \frac{(O_i - E_i)^2}{E_i}$$

where O_i and E_i denote the observed and expected values, respectively. A lower chi-square value indicates a better fit of the model to the data.

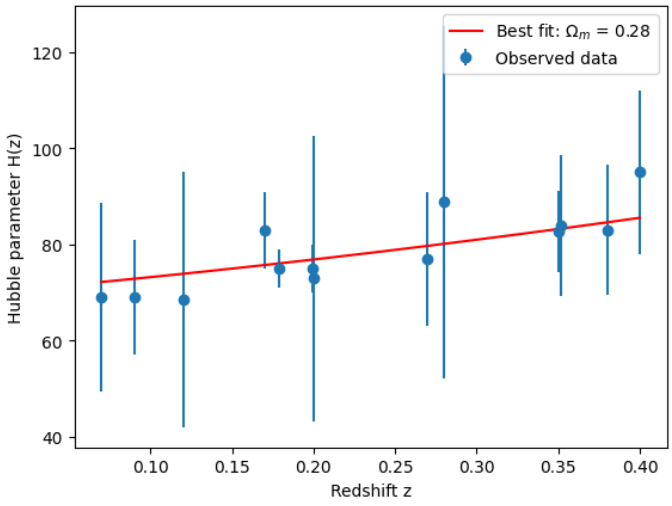
4.3 Steps in Chi-Square Fitting

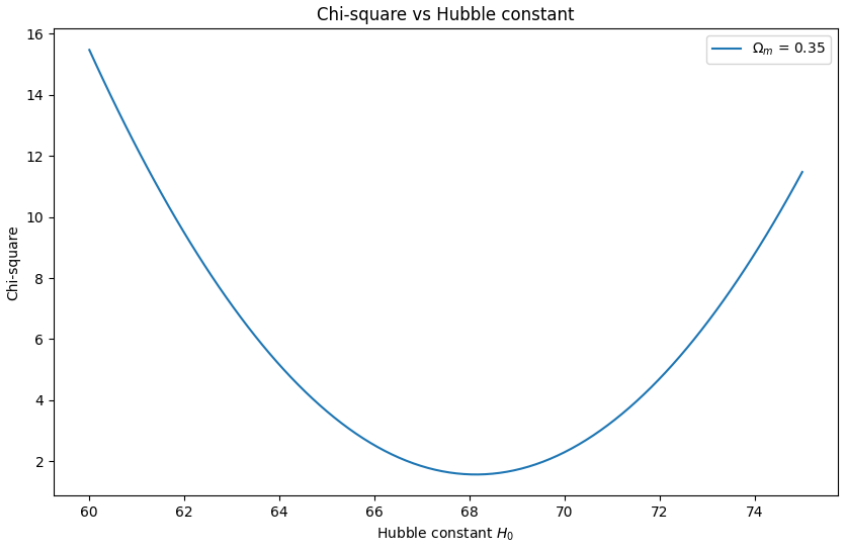
1. **Define the Model**: Specify the theoretical model and its parameters.
2. **Collect Data**: Gather observed data to be compared with the model predictions.
3. **Compute Expected Values**: Calculate the expected values using the theoretical model.
4. **Calculate Chi-Square Statistic**: Compute the chi-square statistic using the formula.
5. **Minimize Chi-Square**: Adjust model parameters to minimize the chi-square value, optimizing the fit.
6. **Assess Fit**: Evaluate the goodness of fit using the minimized chi-square value.

4.4 Application in Cosmology

In cosmology, chi-square fitting plays a crucial role in testing theoretical models against observational data. Examples include fitting the cosmic microwave background (CMB) power spectrum, galaxy distributions, and other large-scale structure formations. This method helps in estimating cosmological parameters such as the density of dark matter and dark energy, and the Hubble constant.

4.5 Example: Fitting a Linear Model

- 

- 

Link to Colab Notebook

For more detailed implementation, please refer to the Colab notebook available at:

1. Colab Notebook Link
2. Colab Notebook link 2

Chapter 5

Likelihood Analysis

5.1 Likelihood Analysis

Likelihood analysis is a fundamental statistical method used to estimate the parameters of a statistical model based on observed data. It involves maximizing the likelihood function, which measures how likely the observed data are under different parameter values.

5.1.1 Key Concepts

- **Likelihood Function**: The likelihood function $\mathcal{L}(\theta \mid \mathbf{x})$ expresses the probability of observing data $\mathbf{x}$ given the model parameters θ. For independent and identically distributed (i.i.d.) data points x_i, the likelihood function is typically the product of probability density functions or mass functions evaluated at each observation.

- **Log-Likelihood Function**: The log-likelihood function $\ell(\theta \mid \mathbf{x})$ is the natural logarithm of the likelihood function. It simplifies calculations and transforms products into sums, making it easier to work with in practice.

- **Maximum Likelihood Estimation (MLE)**: MLE is a method of estimating the parameters $\hat{\theta}$ that maximize the likelihood function $\mathcal{L}(\theta \mid \mathbf{x})$. It provides point estimates of the parameters and is often used for hypothesis testing and constructing confidence intervals.

5.1.2 Application in Cosmology

In cosmology, likelihood analysis plays a crucial role in fitting theoretical models to observational data. For example, it is used to estimate parameters such as the density of dark matter, dark energy equation of state, and cosmological parameters like the Hubble constant based on observations from experiments like the cosmic microwave background (CMB) measurements or galaxy surveys.

5.1.3 Likelihood Analysis And Chi square

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{5.1}$$

where:

- O_i represents the observed value.
- E_i represents the expected value.
- n is the number of observations.

Likelihood analysis involves estimating parameters of a model by maximizing the likelihood function. For a set of observations $\{x_1, x_2, \ldots, x_n\}$ and a parameter θ, the likelihood function $L(\theta)$ is given by:

$$L(\theta) = \prod_{i=1}^{n} f(x_i \mid \theta) \tag{5.2}$$

where $f(x_i \mid \theta)$ is the probability density function of x_i given the parameter θ. In practice, it is often more convenient to work with the log-likelihood function:

$$\log L(\theta) = \sum_{i=1}^{n} \log f(x_i \mid \theta) \tag{5.3}$$

The maximum likelihood estimate (MLE) of θ is the value that maximizes the log-likelihood function.

Link to Colab Notebook

For more detailed implementation, please refer to the Colab notebook available at:
Colab Notebook Link
The likelihood analysis estimates the parameters m and c that maximize the likelihood of observing the data (x_i, y_i).

Chapter 6

Marcov Chain Monte Carlo

6.1 Markov Chain Monte Carlo (MCMC)

Markov Chain Monte Carlo (MCMC) is a powerful statistical method used for sampling from complex probability distributions. It is particularly useful in situations where direct sampling is difficult or impractical. MCMC methods generate a Markov chain that asymptotically converges to the desired distribution, allowing for the estimation of properties such as means, variances, and quantiles.

6.1.1 Key Concepts

- **Markov Chain**: A sequence of random variables where the probability distribution of each variable depends only on the state of the previous one.

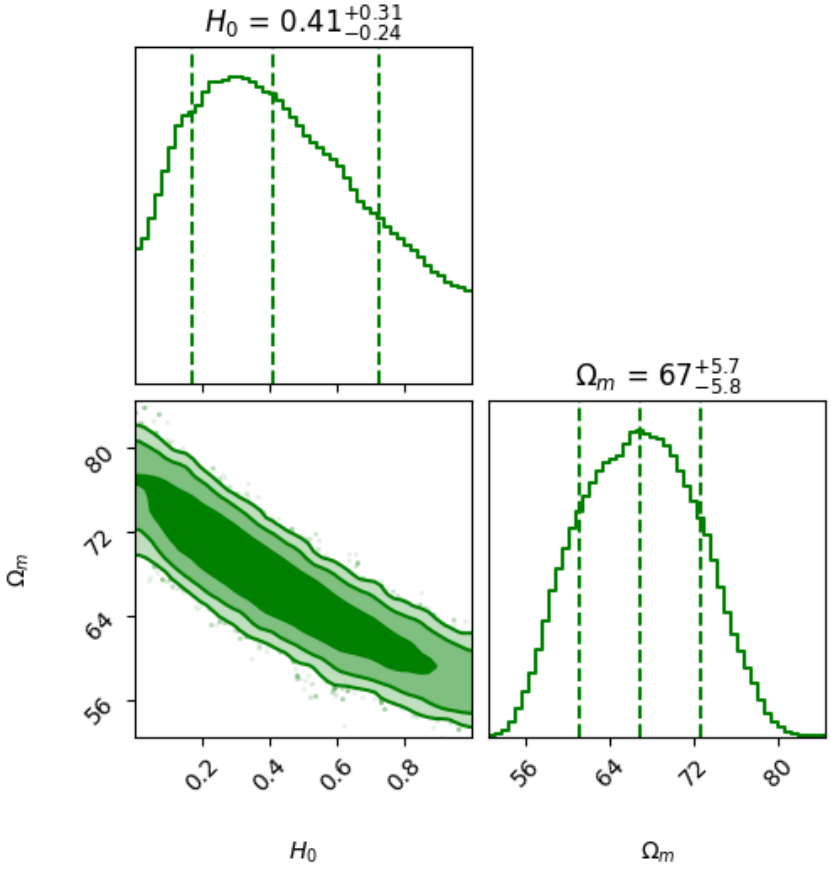
- **Monte Carlo Method**: A broad class of computational algorithms that rely on random sampling to obtain numerical results.

- **Metropolis-Hastings Algorithm**: One of the most common MCMC algorithms, it generates proposals for new states based on a proposal distribution and accepts or rejects them based on a acceptance probability derived from the target distribution.

6.1.2 Application in Cosmology

In cosmology, MCMC methods are used to explore parameter spaces and perform Bayesian inference. They are particularly valuable for estimating parameters of complex models from observational data, such as those derived from cosmic microwave background (CMB) experiments, galaxy surveys, or supernova observations.

6.1.3 Example: MCMC in Python



Link to Colab Notebook

For more detailed implementation, please refer to the Colab notebook available at: Colab Notebook Link

Chapter 7

Conclusion

7.1 Conclusion: Statistical Tools in Cosmology

Statistical methods play a crucial role in advancing our understanding of the universe through the analysis and interpretation of vast observational data in cosmology. In this report, we have explored several key statistical tools essential for cosmological research, including Monte Carlo simulations, Bayesian inference, chi-square fitting, likelihood analysis, and Markov Chain Monte Carlo (MCMC) methods.

7.1.1 Key Contributions

Each statistical method discussed offers unique capabilities in handling different aspects of cosmological data:

- **Monte Carlo Simulations**: These simulations provide a powerful means to generate random samples and understand the probabilistic nature of cosmological phenomena. They are instrumental in exploring parameter spaces and assessing model predictions against observational data.

- **Bayesian Inference**: Bayesian methods allow for updating probabilities of hypotheses based on observational data, providing a framework for incorporating prior knowledge and uncertainty quantification into cosmological models.

- **Chi-Square Fitting**: Used to evaluate the goodness of fit between theoretical models and observed data, chi-square fitting helps validate hypotheses and refine model parameters in cosmological studies.

- **Likelihood Analysis**: This method estimates the parameters of cosmological models by maximizing the likelihood function, offering robust parameter estimation and model comparison tools.

- **Markov Chain Monte Carlo (MCMC)**: MCMC methods enable sampling from complex probability distributions, facilitating Bayesian parameter estimation and uncertainty propagation in cosmological parameter space exploration.

7.1.2 Impact and Future Directions

The application of these statistical tools has significantly enhanced our ability to extract meaningful insights from cosmological observations. They have contributed to refining cosmological models, validating theoretical predictions, and uncovering new phenomena in the universe. As observational datasets continue to grow in complexity and size, the role of advanced statistical methods will only become more critical.

Looking ahead, future research in cosmology will benefit from further advancements in statistical techniques, such as machine learning algorithms for pattern recognition in large-scale surveys, hierarchical Bayesian models for complex data structures, and improved computational methods for handling big data challenges.

7.1.3 Conclusion

In conclusion, statistical tools are indispensable for advancing cosmological research, offering powerful means to analyze data, validate models, and infer parameters with quantified uncertainties. By harnessing these methods, cosmologists are poised to deepen our understanding of fundamental cosmic questions and uncover the mysteries of the universe.