677_Final_Project

Chenghao Xia

2024-05-03

Evolution of Statistical Techniques

Background: Before electronic computers were common after World War II, the ways people could do statistics were limited by the mechanical tools they had. These limits made statistics very math-focused but only useful for certain things. When electronic computers became popular, these limits were greatly reduced. This change allowed statistics to be used for a wider range of practical things.

Impact on Statistical Theory: The introduction of electronic computation made it possible to use methods that needed a lot of computing power, which were not practical before. This change let statisticians work with more complex models and greatly expanded what they could do with statistical analysis. It also set the stage for using techniques that handle a lot of data, which are common in statistics today.

Introduction to Empirical Bayes

Definition and Importance: Empirical Bayes is a method in statistics that mixes ideas from Bayesian statistics with actual data to estimate prior distributions. This method is especially useful when working with big datasets because it lets you update what you previously thought based on new data. This improves the accuracy of the conclusions drawn from the data.

Basic Concept: In the Empirical Bayes method, the parameters of the prior distribution are estimated from the data itself, instead of being set in advance. This flexibility makes it a strong tool in modern statistics, especially when dealing with large datasets where traditional methods might not work as well.

Robbins' Formula

Scenario Description: In one part of the story, there's a car insurance company in Europe. They want to figure out how many times people will ask for money because of accidents in the future. They're using information from the year before to make guesses about this.

Mathematical Expression: Robbins' formula is introduced as a method to estimate future claims based on past data:

$$E[\mu|x] = \frac{(x+1)f(x+1)}{f(x)}$$

where x represents the number of claims made by an individual in the previous year, $f(x)$ is the proportion of policyholders who made x claims, and $E[\mu|x]$ is the expected number of claims for the next year.

Application: This equation uses past information to guess how likely it is for there to be claims in the future. When the insurance company uses Robbins' formula, it helps them handle risks better. They can adjust their insurance plans based on how likely it is for people to make claims.

The Missing-Species Problem

Introduction to the Problem: This question, looked at by R.A. Fisher, is about guessing how many different types of things there are in a place that we haven't seen yet. Fisher's method uses a way of guessing called empirical Bayes to figure out these hidden types based on what we already know.

Application of Empirical Bayes: Fisher used Empirical Bayes to guess how many types of butterflies we haven't seen yet, using information collected during wartime by naturalist Alexander Corbet. The method mixes what we already know about the types we've seen with some math to guess the total number of types of butterflies, even the ones we haven't found yet.

Example: The information had numbers showing how many times different types of butterflies were seen. Fisher used Empirical Bayes to guess how many more types of butterflies we might find if we kept looking. This helped us understand more about the variety of butterflies, even though we couldn't see them all directly in the data we had.

Computational Methods in Empirical Bayes

Empirical Bayes methods use the information we already have to guess what the distribution was like before. This affects how we make guesses about what will happen next. This part of the book talks about how to do this on a computer, and it gives examples using Python. It also talks about using the Poisson and Gamma distributions in these guesses.

Poisson Distribution Model

Empirical Bayes methods often use the Poisson distribution to model how often things happen. This works well when we're counting stuff, like how many times something happens, and each event happens on its own without being influenced by the others. We use this when we already know the total number of events.

Example with R:

In a dataset, if insurance claims per policyholder each year follow a Poisson pattern, we can predict how many future claims there might be using Robbins' formula.

```
claims <- c(0, 1, 2, 3, 4, 5, 6, 7)
policyholders <- c(8000, 1500, 300, 50, 20, 10, 5, 2)

total_policyholders <- sum(policyholders)
f_x <- policyholders / total_policyholders

f_x_plus_1 <- c(f_x[-1], 0)

expected_claims <- (claims + 1) * f_x_plus_1 / f_x
expected_claims[is.na(expected_claims) | is.infinite(expected_claims)] <- 0
print(data.frame(Claims = claims, ExpectedClaims = expected_claims))
```

```
##   Claims ExpectedClaims
## 1      0         0.1875
## 2      1         0.4000
## 3      2         0.5000
## 4      3         1.6000
```

```
## 5      4         2.5000
## 6      5         3.0000
## 7      6         2.8000
## 8      7         0.0000
```

When we're dealing with rates or durations, we often use the Gamma distribution as a starting guess in Bayesian analysis. This R script creates a Gamma guess and shows its shape on a graph.

```
library(ggplot2)
```

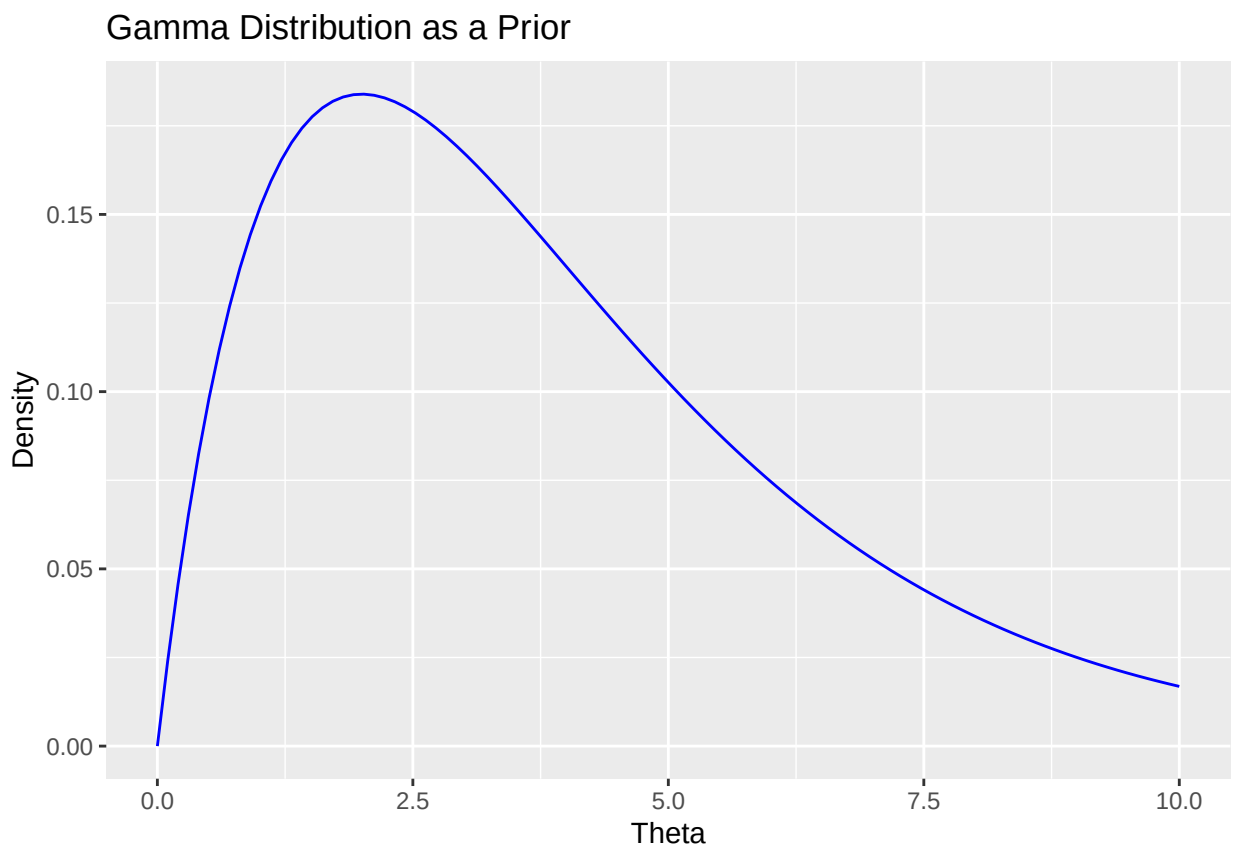
```
## Warning:   'ggplot2' R 4.3.2
```

```
shape <- 2
scale <- 2

theta_values <- seq(0, 10, length.out = 100)

gamma_density <- dgamma(theta_values, shape = shape, scale = scale)

ggplot(data.frame(Theta = theta_values, Density = gamma_density), aes(x = Theta, y = Density)) +
  geom_line(color = "blue") +
  ggtitle("Gamma Distribution as a Prior") +
  xlab("Theta") +
  ylab("Density")
```



These scripts will do the math and make graphs to help with your chapter notes, showing how we can use Empirical Bayes methods in R.

Mathematical Underpinnings of Empirical Bayes

Empirical Bayes methods are about using what we see in the data to change what we thought before about the shape of a distribution. This part of the book will explain the main math ideas behind Empirical Bayes methods, which are talked about more in Chapter 6.

Bayesian Inference

Bayesian Framework: Bayesian inference is based on the idea that as we get more data, we can change what we think about a parameter we don't know yet. This is formalized through Bayes' theorem:

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

where:

- $P(\theta|X)$ is the posterior probability of the parameter θ given the data X.
- $P(X|\theta)$ is the likelihood of the data given the parameter values.
- $P(\theta)$ is the prior probability of the parameter.
- $P(X)$ is the probability of the data, which acts as a normalizing constant.

Empirical Bayes Adaptation: In Empirical Bayes, the prior $P(\theta)$ is estimated from the data itself, allowing for a more data-driven approach to updating our beliefs about θ.

The Poisson-Gamma Model

Empirical Bayes methods are often used to analyze count data. Here, we use the Poisson distribution to represent the data, and the Gamma distribution to guess before we see the data.

Poisson Distribution: Used to model the number of events in a fixed interval, space, or time, the probability mass function is given by:

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

where
lambda is the rate at which events occur.

Gamma Distribution as Prior: We pick the Gamma distribution as the guess for λ in Poisson processes because it makes the math easier when we update our guess using new data.

$$\lambda|X \sim \text{Gamma}(\alpha + x, \beta + 1)$$

where α and β are the shape and rate parameters of the prior Gamma distribution.

Robbins' Formula

Robbins' formula, a fundamental aspect of this chapter, provides an empirical way to estimate future occurrences from past data, without needing a known prior. The formula derives from rearranging the terms of the expected posterior estimate under a Poisson model.

$$E[\lambda|x] = \frac{(x+1)f(x+1)}{f(x)}$$

This formula uses the data to make a guess about what the guess should be, which is handy when we don't have regular guess information available.

The Missing-Species Problem

Fisher's solution to the Missing-Species Problem is another excellent application of Empirical Bayes methods. The problem involves estimating the number of unseen species in a biological context, and Fisher's approach used empirical data to estimate this unseen diversity.

Mathematical Expression: Fisher developed an estimator for the unseen species by extrapolating from the observed data, using a mixture of Poisson processes and empirical estimates:

$$E[S_{unseen}] = \sum_{x=0}^{max} \frac{x(n_x+1)f(n_x+1)}{f(n_x)}$$

where n_x is the number of species seen x times.

Historical Context of Empirical Bayes

Learning about how Empirical Bayes methods came to be helps us understand why they're important and how they've changed over time in statistics. This part of the book talks about where these methods came from, who played important roles, and how they made people think differently about statistics.

Emergence of Empirical Bayes Methods

Post-War Technological Advances: The growth of Empirical Bayes methods was greatly shaped by the rise of electronic computers after World War II. With better computers, we could work with bigger sets of data and more complicated statistical models. This made it possible for new methods like Empirical Bayes to come into being.

Early Adaptations and Usage: In the 1950s and 1960s, statisticians started to pay more attention to Empirical Bayes methods. They were looking for ways to use data better, especially when they weren't sure about the initial guesses. Empirical Bayes offered a way to guess parameters even when we didn't know the exact starting distributions, by making educated guesses from the data itself.

Key Contributions and Figures

Herbert Robbins: Herbert Robbins played a big role in the history of Empirical Bayes. In the middle of the 20th century, he talked about "Empirical Bayes" in some really important papers. Robbins helped make this approach more official, blending the analysis of real data with Bayesian ideas.

Bradley Efron: Bradley Efron also made a big impact on Empirical Bayes. He built on Robbins' ideas and made Empirical Bayes methods even better. Efron focused on using these techniques to handle big sets of data, which are often seen in fields like genomics and bioinformatics.

Theoretical and Practical Impacts

Shift in Statistical Paradigms: When Empirical Bayes methods came around, they changed how we did statistics. Instead of just relying on theory, we started to use more data to guide our decisions. This shift mirrored what was happening across statistics, as we got more data and better computers, making it possible and important to use more empirical and computational methods.

Influence on Modern Statistical Practices: Empirical Bayes methods have had a big impact on how we do statistics today, especially in fields where we have to work with lots of data and make educated guesses. They're crucial in fields like insurance and environmental science, where we often have to predict things even when we don't have all the information we need.

Integration into Broader Statistical Methodology

Expansion and Refinement: Empirical Bayes methods have evolved and grown stronger over time, thanks to improvements in how we use computers and our understanding of statistics. This ongoing progress has kept these methods useful and effective for making guesses based on data.

Educational and Applied Usage: Nowadays, Empirical Bayes methods are taught in advanced statistics classes all over the globe and are built into different statistical software programs. This shows how valuable and useful they still are for solving real-world problems.

Statistical Practice Implications of Empirical Bayes Methods

Empirical Bayes methods are used in many different areas, showing how flexible and essential they are in statistics. This part of the book talks about how these methods are used in real-life situations and how they help us make decisions based on statistics.

Applications in Various Fields

Insurance Industry: As seen with Robbins' formula, Empirical Bayes is widely used in the insurance industry to guess how many claims there might be in the future using past data. This helps insurance companies figure out risk better and decide on premiums that match the real chances of claims happening.

Medical Research: In medical statistics, Empirical Bayes methods are used to study data from clinical trials, especially when there aren't many patients involved. These methods help researchers make stronger conclusions even when they have limited data, which improves the process of developing new treatments and therapies.

Environmental Science: In environmental studies, Empirical Bayes methods are used to look at things like how many different species there are or what might happen with climate change. By looking at past data, researchers can make smart guesses about what might happen in the environment, which helps with efforts to protect it.

Sports Analytics: In sports, Empirical Bayes methods can help forecast how well players might do in the future by looking at how they've done before. They consider things like how tough the other teams are. This helps with picking who plays, planning the game, and deciding if players should switch teams.

Enhancing Decision-Making

Data-Driven Decisions: Empirical Bayes methods help organizations make smarter choices by using solid statistical analysis. This means companies and institutions can lower risks and get better results in different parts of their operations.

Handling Uncertainty: Empirical Bayes methods are really useful when we don't have all the data we need or when our initial guesses aren't very sure. They give us a clear way to change our beliefs and make guesses based on real data.

Challenges and Limitations

Data Dependency: Although Empirical Bayes methods are strong, they rely a lot on how good and how much data we have. If the data isn't very good or there isn't enough of it, the estimates we get might be wrong or not very trustworthy.

Complexity of Interpretation: The interpretation of results obtained from Empirical Bayes methods can be complex, especially when communicating findings to stakeholders who may not have a statistical background.

Future Trends

Integration with Machine Learning: As statistical inference and machine learning come together, Empirical Bayes methods are being mixed with machine learning techniques more often. This lets us use fancier models and make better predictions, especially when we're dealing with lots of data.

Advancements in Computational Techniques: As computers get better and our algorithms improve, Empirical Bayes methods are likely to become even more useful and important for statistics in lots of different fields.

Additional Resources and References

This section lists key texts, articles, and online resources that can provide further insights into Empirical Bayes methods, as well as the primary sources used in preparing this assignment.

Key Textbooks and Articles

- *Computer Age Statistical Inference* by Bradley Efron and Trevor Hastie: This book serves as the primary source for the chapter on Empirical Bayes and offers a comprehensive overview of modern statistical techniques.
- *Bayesian Data Analysis* by Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin: This textbook provides a deeper dive into Bayesian methods, including Empirical Bayes, with extensive examples and applications.
- *The Theory of the Design of Experiments* by Herbert Robbins: Explore the foundational work by Herbert Robbins that introduces Empirical Bayes in the context of experimental design.

Online Resources

- Stanford Online: This free course offers lectures and materials that cover various statistical methods, including Empirical Bayes.
- Cross Validated (Stack Exchange): A question and answer site for statistics, machine learning, data analysis, data mining, and data visualization, where topics on Empirical Bayes are frequently discussed.

Journals and Research Papers

- "Empirical Bayes Estimation of the Multinomial Probabilities" by Herbert Robbins (1956): This seminal paper by Robbins introduces Empirical Bayes estimation methods in the context of multinomial probabilities.
- "Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction" by Bradley Efron (2010): In this influential work, Efron further explores Empirical Bayes methods within the framework of large-scale inference, making it crucial for understanding contemporary applications.