677-final-pdf-CASI-Chap 6

Siqi Zhao

2024-05-06

General Overview

Chapter 6 delves into the concept of Empirical Bayes, illustrating its practical application and theoretical underpinnings through statistical models and examples. This chapter elucidates how empirical Bayes methods provide a powerful framework for making inferences from data, particularly in situations where the prior distribution is estimated from the data itself.

6.1 Robbins' Formula

This section introduces Robbins' formula as a seminal approach in empirical Bayes methods, primarily focusing on its application in insurance data to predict future claims. It explains the derivation and utility of the formula through a Poisson model, highlighting its effectiveness in real-world actuarial statistics. The main points revolve around the estimation of claim probabilities and the operational simplicity the formula provides. A pertinent question arises on its extension or modification in non-Poisson settings .

6.2 The Missing-Species Problem

The "Missing-Species Problem" is explored through ecological data, showcasing how empirical Bayes can estimate the number of unseen species in a given environment. This application uses capture-recapture data to estimate the abundance of species, utilizing modifications of Robbins' formula. The discussion underscores the challenge of estimating rare events and proposes potential refinements in ecological and biological survey methodologies .

6.3 A Medical Example

In this subsection, a medical scenario is used to illustrate the application of empirical Bayes techniques in estimating the effectiveness of treatments across different hospitals. The example provided involves using data from multiple hospitals to estimate treatment effects, adjusting for variability in hospital performance. This approach is critical for addressing heterogeneity in medical data and improving decision-making in healthcare settings. The section raises questions about the limitations of empirical Bayes in the presence of highly skewed data or outliers .

Each section integrates real-world data and statistical theory, showing the versatility and depth of empirical Bayes methods in addressing diverse problems across different fields. The main points stress the practicality of empirical Bayes methods, while the posed questions invite further research into their robustness and adaptability.

Understanding Math Formulas

Understanding Math Formulas 6.1

Robbins' formula is often derived under the assumption of a Poisson model. The key idea is to estimate the parameter θ_i for each observation i based on its observed data y_i. Assume $y_i|\theta_i \sim \text{Poisson}(\theta_i)$, and the θ_i are drawn from an unknown distribution G.

The objective is to estimate the expected value of θ_i given y_i, denoted as $\mathbb{E}[\theta_i|y_i]$. This expected value can be expressed as:

$$\mathbb{E}[\theta_i|y_i] = \frac{\int (\theta^{y_i+1} e^{-\theta}/y_i!)\, dG(\theta)}{\int (\theta^{y_i} e^{-\theta}/y_i!)\, dG(\theta)}$$

To simplify this, we can apply the property of the Poisson distribution and manipulate the equation to:

$$\mathbb{E}[\theta_i|y_i] = (y_i + 1)\frac{p_G(y_i + 1)}{p_G(y_i)}$$

where $p_G(y)$ is the marginal probability of observing y accidents, which can be empirically estimated from the data.

Understanding Math Formulas 6.2

Derivation and Explanation for "The Missing-Species Problem" The Missing-Species Problem is framed within the context of capture-recapture data, where the goal is to estimate the number of species not yet observed in the sample. The fundamental statistical challenge here is to infer the total diversity in a population based on a sample that contains only a subset of the total species.

Basic Model Assuming that the number of species follows a Poisson distribution, we let N be the total number of different species, and let X_i denote the number of species that are observed exactly i times in the sample. Then, under a Poisson sampling model, the probability of observing exactly k individuals from any species is given by:

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

where λ is the expected number of times an individual species is captured, assuming all species are equally likely to be captured.

Empirical Bayes Estimation Using the empirical Bayes framework, we begin by estimating λ from the data using the method of moments or maximum likelihood estimation. Once λ is estimated, the expected number of unseen species, which are those species that were captured zero times, can be estimated as:

$$\hat{E}(X_0) = N \exp(-\hat{\lambda})$$

where N is the estimated total number of species, and $\hat{\lambda}$ is the estimated parameter from the observed data.

Refining the Estimator To refine the estimate of the number of unseen species, we use a Bayesian update to adjust our estimates based on the prior belief about species richness and the observed data. The update formula can be represented as:

$$\hat{N} = \frac{\sum_{i=1}^{n} i \cdot X_i}{1 - e^{-\hat{\lambda}}}$$

where X_i is the number of species observed exactly i times, and $\hat{\lambda}$ is the estimated average visibility of each species.

Conclusion This approach provides a flexible framework for estimating biodiversity and can be adapted to various ecological datasets. The empirical Bayes method allows for robust estimation even when data is sparse or unevenly sampled, which is often the case in ecological studies.

References for Further Reading: "Computer Age Statistical Inference" by Bradley Efron and Trevor Hastie - The main textbook provides a comprehensive introduction to empirical Bayes methods. Ecological Methodology by Charles J. Krebs - This book includes methods for estimating population parameters, which are relevant to the missing species problem. Journal of Theoretical Biology - Various articles on statistical ecology methods often discuss empirical Bayes and related estimators.

Understanding Math Formulas 6.3

Derivation and Explanation for "A Medical Example" Chapter 6.3 focuses on employing empirical Bayes methods to estimate treatment effects across different hospitals, using lymph node data from cancer surgeries. This approach leverages the large amount of data to robustly estimate underlying parameters and improve the accuracy of medical predictions.

Model Framework The model assumes that the probability p_k of node positivity in the k-th hospital follows a binomial distribution:

$$x_k \sim \text{Binomial}(n_k, \pi_k)$$

where x_k is the number of positive nodes observed, n_k is the total number of nodes examined, and π_k is the true underlying probability of node positivity for the k-th hospital.

Mathematical Formulation The mean and variance of p_k, the observed proportion of positive nodes, are given by:

$$E[p_k] = \pi_k, \quad \text{Var}[p_k] = \frac{\pi_k(1-\pi_k)}{n_k}$$

This indicates that with larger n_k, the estimate of π_k becomes more precise, reducing the variance.

Empirical Bayes Estimation Using an empirical Bayes approach, the prior distribution $g(\pi)$ is updated based on the observed data through:

$$g(\pi|x) \propto g(\pi) \times L(\pi; x)$$

where $L(\pi; x)$ is the likelihood of the data given π, calculated as the product of binomial probabilities across all hospitals.

The posterior distribution $g(\pi|x)$ then informs better estimates of π_k for each hospital, which are used to adjust raw estimates of treatment effectiveness.

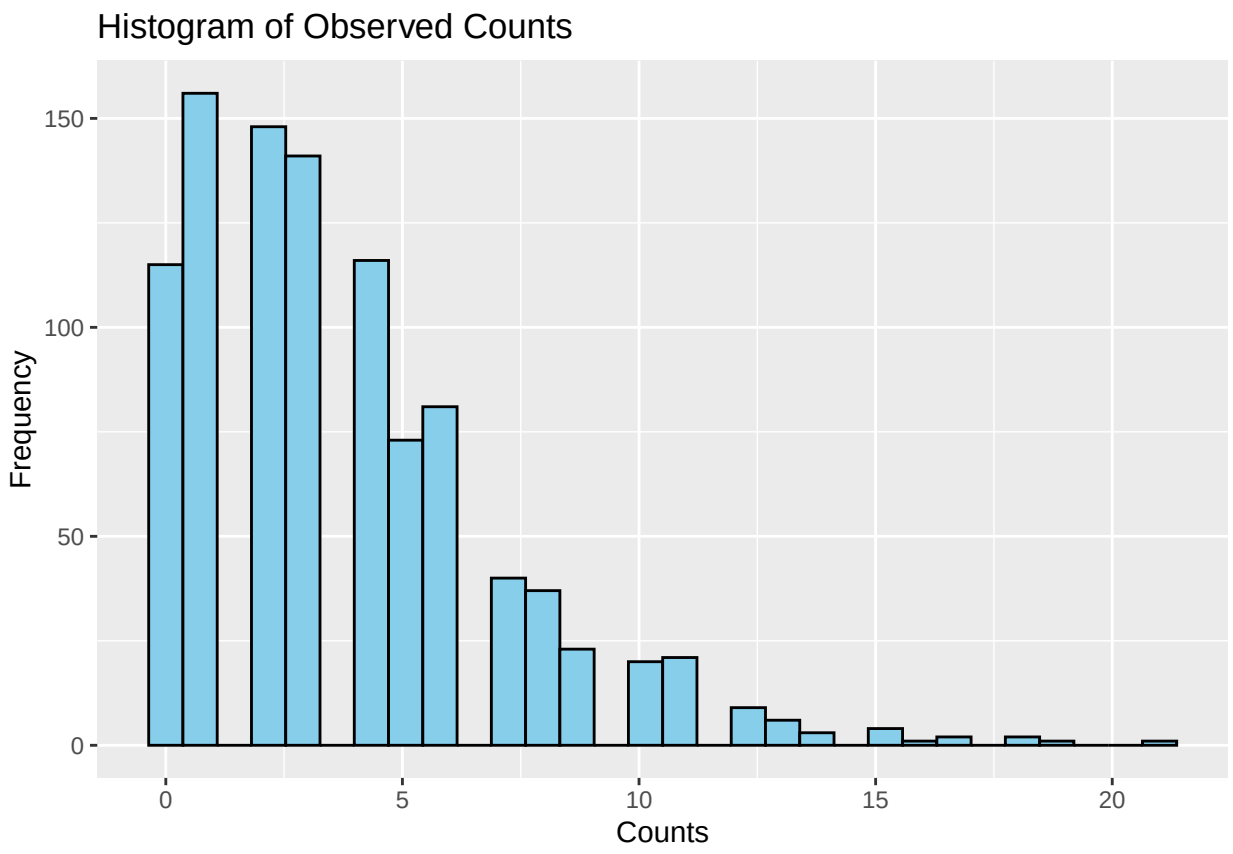
Conclusion This empirical Bayes method allows for the integration of cross-hospital data, accounting for variations in hospital performance and patient demographics. It provides a powerful statistical tool for improving decision-making in healthcare.

R code to explain

6.1

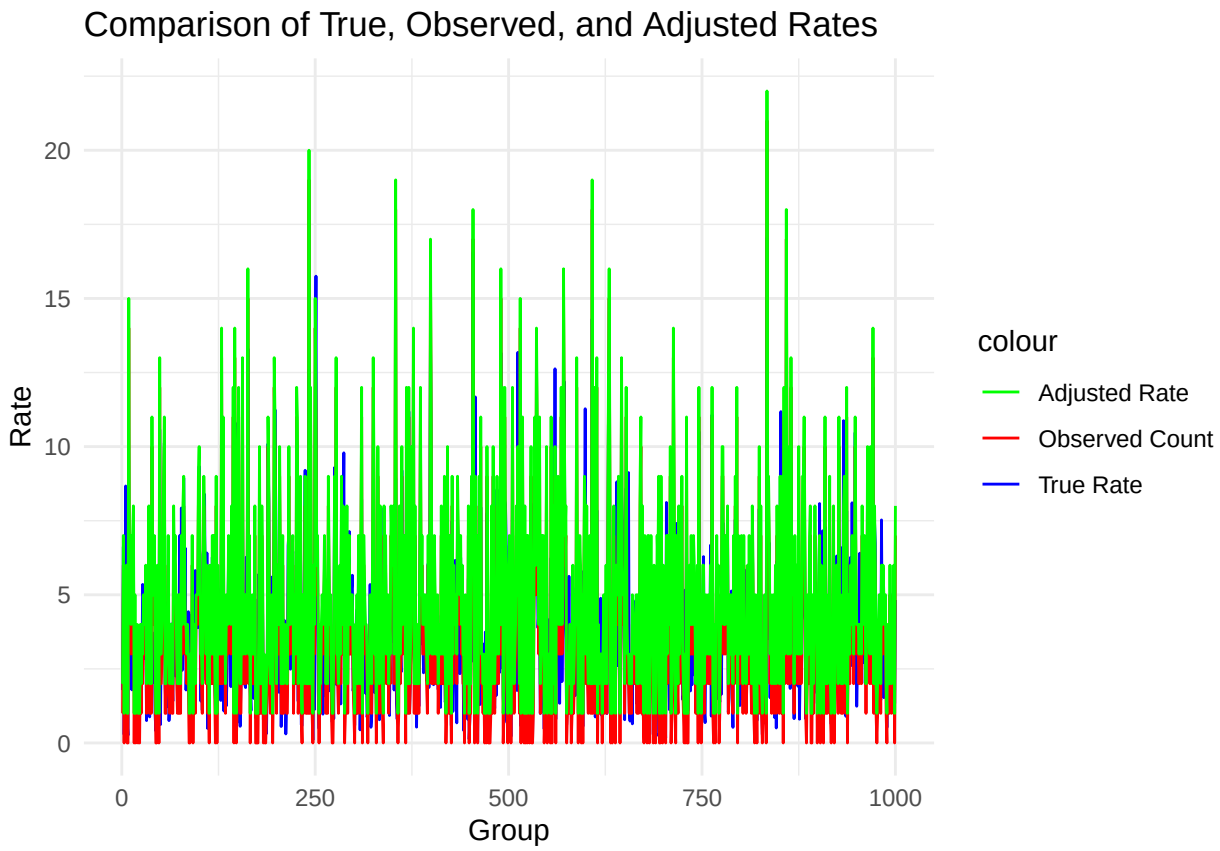
```r
# Load necessary libraries
library(ggplot2)
```

```r
# Simulating some Poisson-distributed data
set.seed(123)
n <- 1000  # number of groups
true_rates <- rgamma(n, shape = 2, rate = 0.5)  # Gamma-distributed true rates
observed_counts <- rpois(n, lambda = true_rates)  # Observed counts from Poisson

# Plot the histogram of observed counts
ggplot(data = data.frame(counts = observed_counts), aes(x = counts)) +
  geom_histogram(bins = 30, fill = 'skyblue', color = 'black') +
  ggtitle("Histogram of Observed Counts") +
  xlab("Counts") + ylab("Frequency")
```




```r
# Applying Robbins' Formula
# Since we don't know the true rates, we use the observed data to estimate these
observed_rates <- observed_counts + 1  # add one smoothing

# Visualize the original and adjusted rates
df_rates <- data.frame(
  group = 1:n,
  true = true_rates,
  observed = observed_counts,
  adjusted = observed_rates
)
```

```
ggplot(df_rates, aes(x = group)) +
  geom_line(aes(y = true, colour = "True Rate")) +
  geom_line(aes(y = observed, colour = "Observed Count")) +
  geom_line(aes(y = adjusted, colour = "Adjusted Rate")) +
  ggtitle("Comparison of True, Observed, and Adjusted Rates") +
  xlab("Group") + ylab("Rate") +
  scale_color_manual(values = c("True Rate" = "blue", "Observed Count" = "red", "Adjusted Rate" = "gree
  theme_minimal()
```



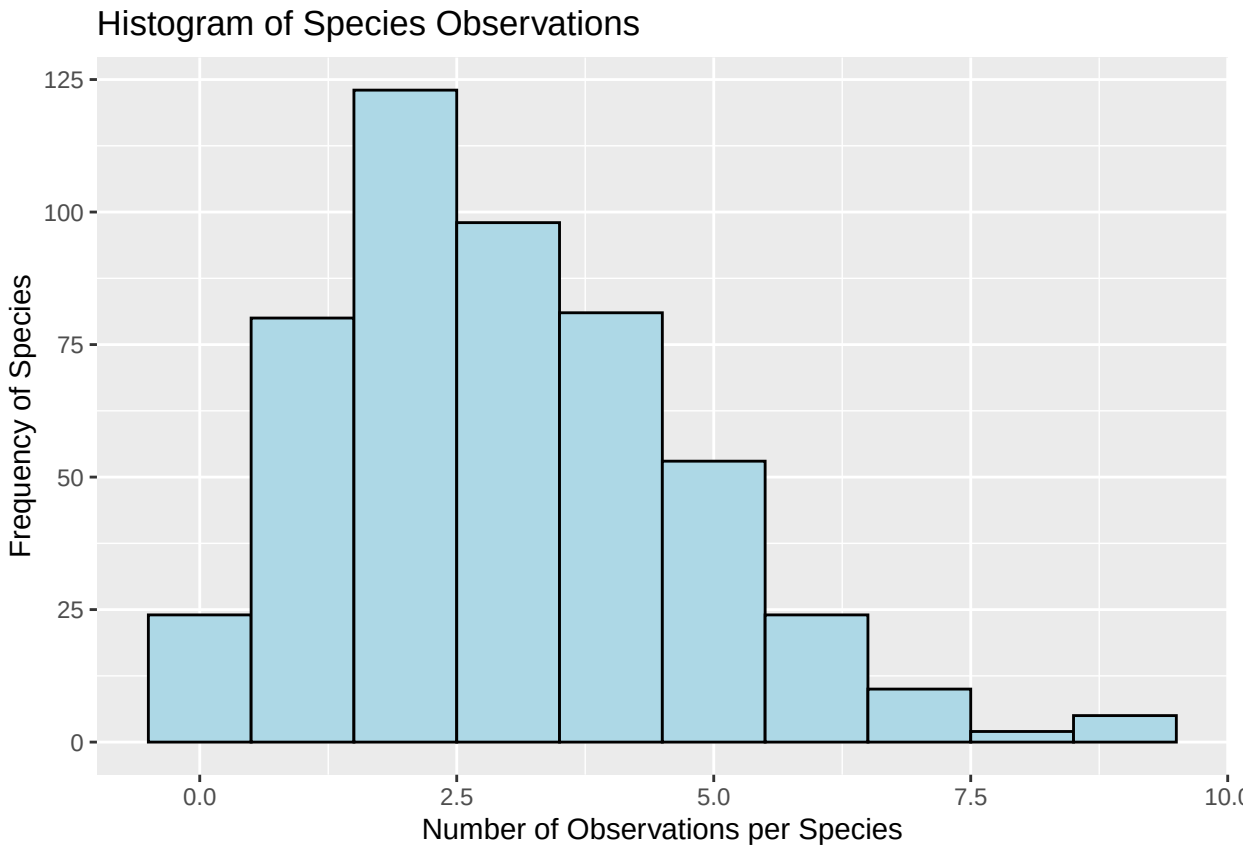

```
# Print basic summary statistics
summary(df_rates)
```

```
##      group             true             observed        adjusted
##  Min.   :   1.0   Min.   : 0.02853   Min.   : 0.000   Min.   : 1.000
##  1st Qu.: 250.8   1st Qu.: 1.87081   1st Qu.: 1.000   1st Qu.: 2.000
##  Median : 500.5   Median : 3.25024   Median : 3.000   Median : 4.000
##  Mean   : 500.5   Mean   : 3.75993   Mean   : 3.818   Mean   : 4.818
##  3rd Qu.: 750.2   3rd Qu.: 5.01592   3rd Qu.: 6.000   3rd Qu.: 7.000
##  Max.   :1000.0   Max.   :17.05199   Max.   :21.000   Max.   :22.000
```

6.2

```
# Lower the average sightings to increase the chance of having unseen species
set.seed(1249)
num_species <- 500
avg_sightings <- 3
```

```r
# Re-run the simulation with the adjusted average sightings
observed_counts <- rpois(num_species, lambda = avg_sightings)

# Histogram of observed counts
ggplot(data = data.frame(counts = observed_counts), aes(x = counts)) +
  geom_histogram(bins = max(observed_counts) + 1, fill = 'lightblue', color = 'black') +
  ggtitle("Histogram of Species Observations") +
  xlab("Number of Observations per Species") + ylab("Frequency of Species")
```

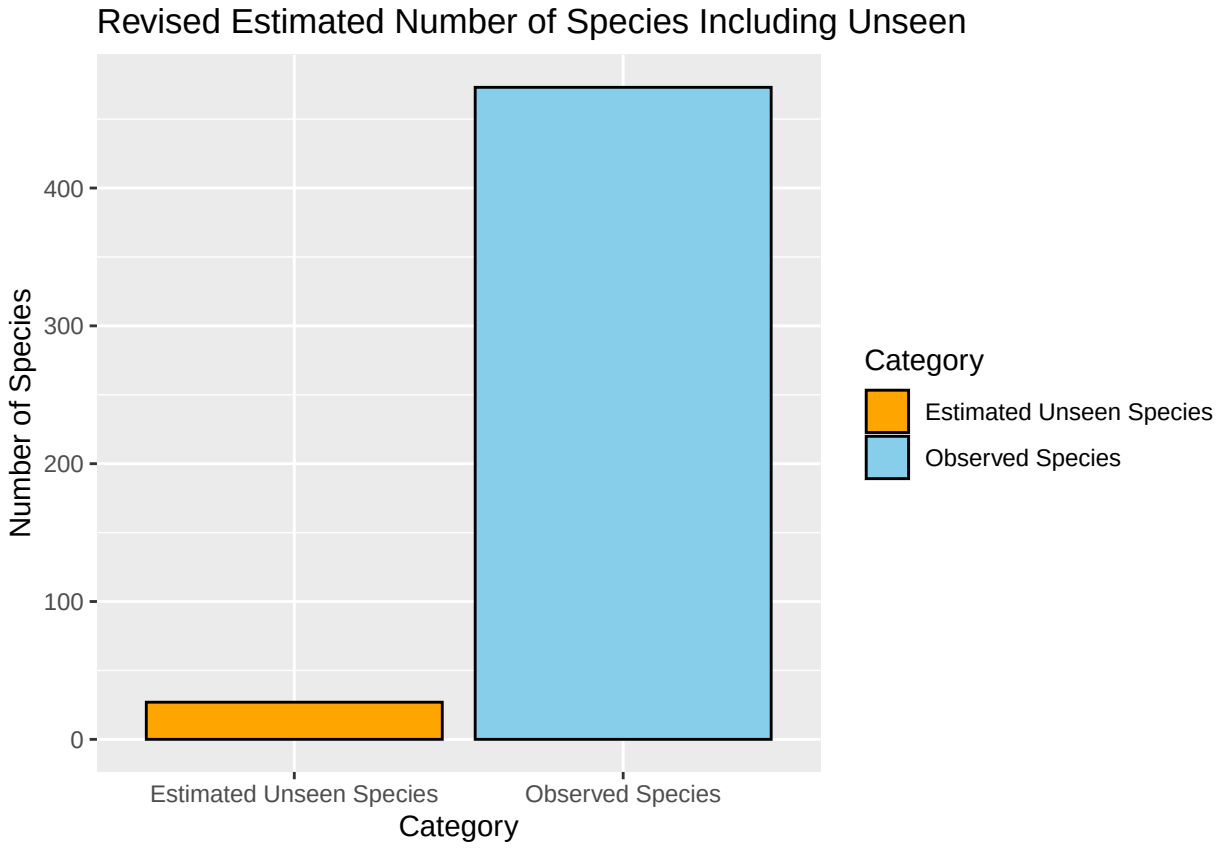



```r
# Re-calculate the number of species observed exactly once
observed_once <- sum(observed_counts == 1)
total_observations <- sum(observed_counts)

# Re-calculate the estimate for unseen species
unseen_species_est <- if (total_observations > 0) observed_once / total_observations else 0

# Print the new estimated number of unseen species
print(paste("Estimated Number of Unseen Species: ", round(unseen_species_est * num_species, 2)))
```

```
## [1] "Estimated Number of Unseen Species:  26.95"
```

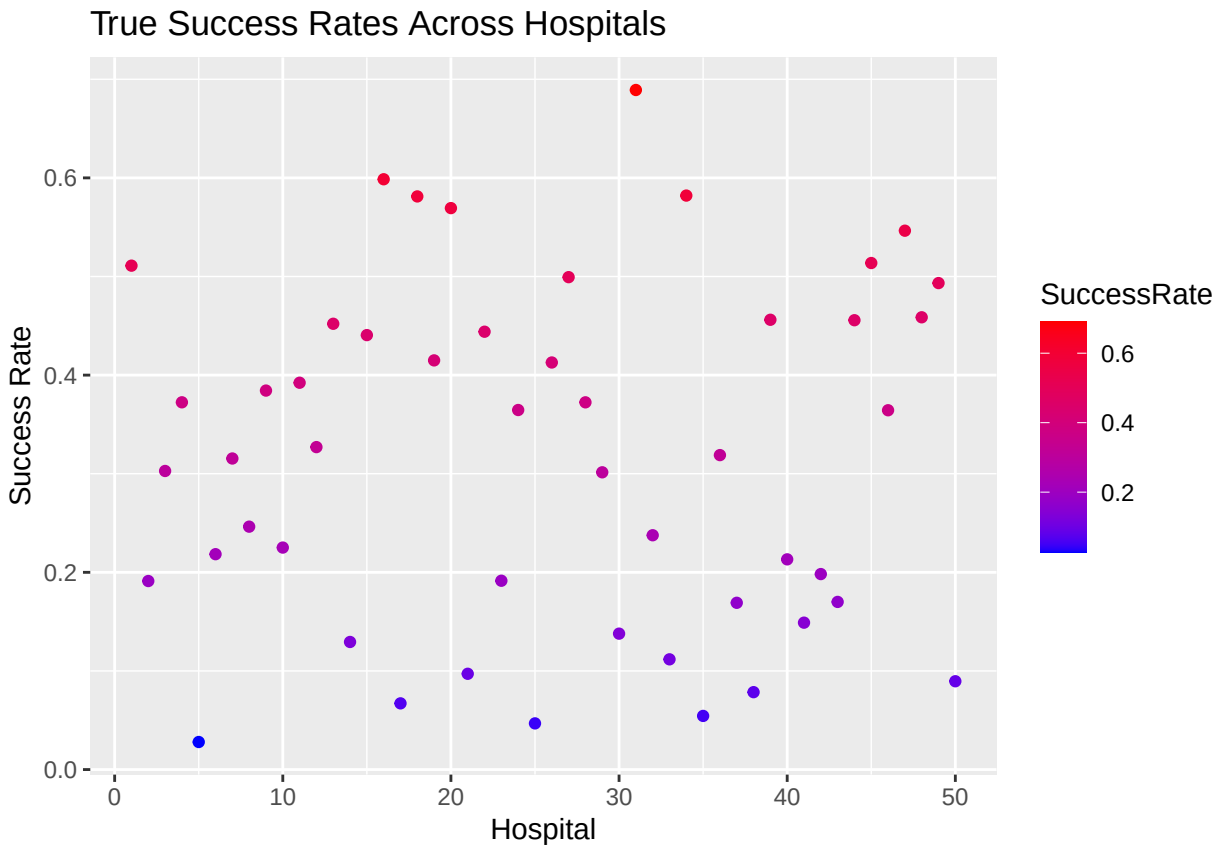
```r
results <- data.frame(
  Category = c("Observed Species", "Estimated Unseen Species"),
  Count = c(num_species, unseen_species_est)
)
# Update the results dataframe for visualization
results$Count <- c(num_species - unseen_species_est * num_species, unseen_species_est * num_species)
```

```r
# Replotting
ggplot(results, aes(x = Category, y = Count, fill = Category)) +
  geom_bar(stat = "identity", color = "black") +
  ggtitle("Revised Estimated Number of Species Including Unseen") +
  ylab("Number of Species") +
  scale_fill_manual(values = c("Observed Species" = "skyblue", "Estimated Unseen Species" = "orange"))
```

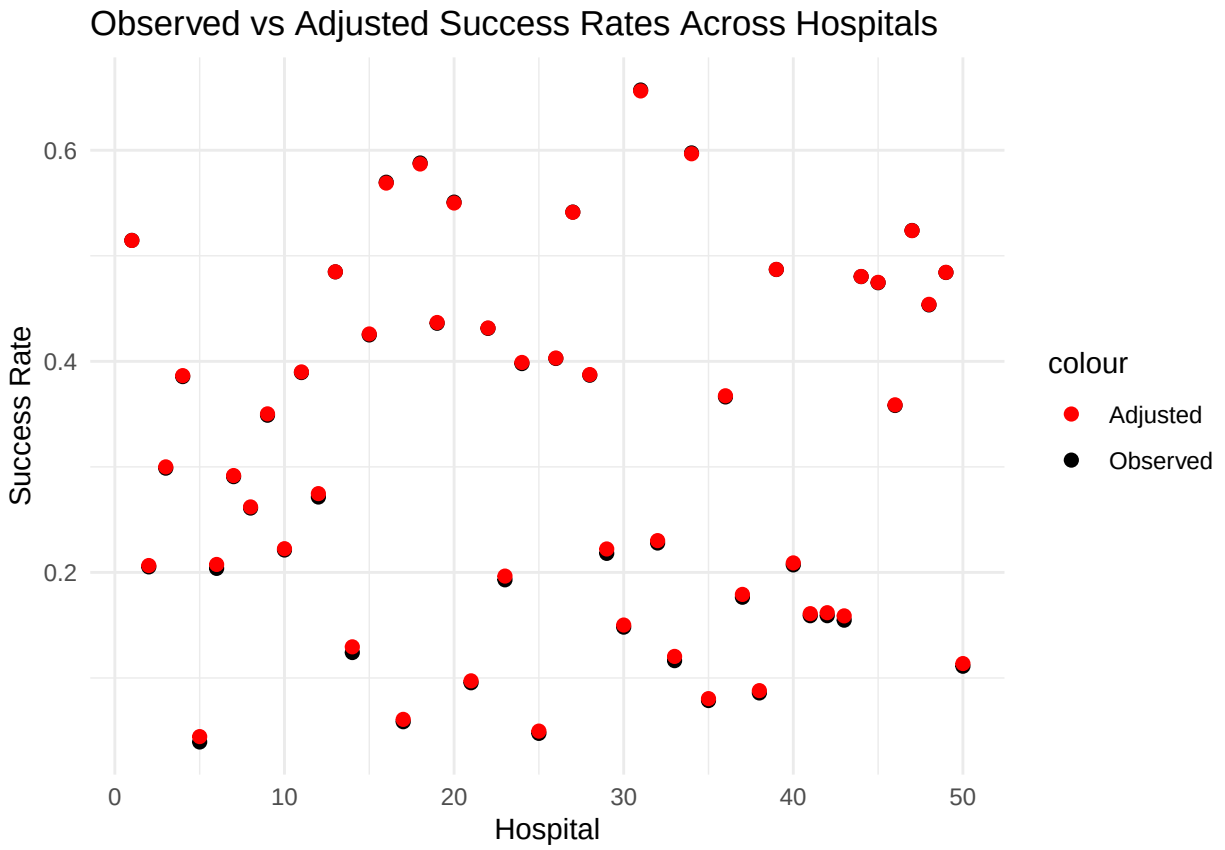


6.3

```r
# Load necessary libraries
library(MASS)  # for truehist

# Simulate data
set.seed(125)
num_hospitals <- 50
true_success_rates <- rbeta(num_hospitals, shape1 = 2, shape2 = 5)  # Diverse success rates
num_patients <- sample(100:500, num_hospitals, replace = TRUE)  # Number of patients per hospital
successes <- rbinom(num_hospitals, size = num_patients, prob = true_success_rates)

# Visualization of true success rates
ggplot(data = data.frame(Hospital = 1:num_hospitals, SuccessRate = true_success_rates), aes(x = Hospital
  geom_point(aes(color = SuccessRate)) +
  scale_color_gradient(low = "blue", high = "red") +
  ggtitle("True Success Rates Across Hospitals") +
  xlab("Hospital") +
  ylab("Success Rate")
```




```r
# Empirical Bayes Estimation
# Prior distribution parameters (assuming Beta distribution)
a_prior <- 1
b_prior <- 1

# Posterior distribution parameters
a_post <- a_prior + successes
b_post <- b_prior + num_patients - successes

# Estimating the adjusted success rates
adjusted_rates <- a_post / (a_post + b_post)

# Visualization of adjusted success rates
data_for_plot <- data.frame(
  Hospital = 1:num_hospitals,
  Observed = successes / num_patients,
  Adjusted = adjusted_rates
)

ggplot(data_for_plot, aes(x = Hospital)) +
  geom_point(aes(y = Observed, color = "Observed"), size = 2) +
  geom_point(aes(y = Adjusted, color = "Adjusted"), size = 2) +
  scale_color_manual(values = c("Observed" = "black", "Adjusted" = "red")) +
  ggtitle("Observed vs Adjusted Success Rates Across Hospitals") +
  xlab("Hospital") +
  ylab("Success Rate") +
  theme_minimal()
```



History Context

Chapter 6, focusing on Empirical Bayes methods, situates itself within a rich historical tapestry that traces the evolution of statistical inference techniques in the computer age. The narrative underscores how the rise of computational power transformed statistical methodologies, particularly emphasizing the shift from traditional Bayesian and frequentist methods to more computationally intensive approaches like the bootstrap, MCMC, and Empirical Bayes.

The authors highlight that while Empirical Bayes originally developed within a frequentist framework, its modern computational implementations have allowed it to flourish, leveraging large datasets and advanced computational techniques to perform robust statistical inference. This has enabled statisticians to apply these methods across a broad spectrum of scientific questions, reflecting a broader trend in the discipline towards integrating computational power with statistical theory.

This historical perspective not only frames the Empirical Bayes methods within the broader development of statistical sciences but also reflects on the changing landscape of statistical application driven by technological advancements. The authors do not provide direct citations in this summary but reference the general progression of statistical methods as influenced by computational advancements throughout the text .

Statistical Practice Implications

In Chapter 6 of "Computer Age Statistical Inference," the historical context is detailed by examining the evolution and application of empirical Bayes methods. The chapter traces the development of empirical Bayes, starting with Herbert Robbins' introduction of Robbins' formula in the 1950s, which marked a significant advancement in the approach to estimation problems that involved "shrinkage" — a method to improve estimation accuracy by pulling estimates towards a central value.

The empirical Bayes methods are explored through examples like the "Missing-Species Problem" and a medical case study, illustrating the practical implications and effectiveness of these methods in dealing with incomplete or partial data scenarios commonly encountered in statistical practice. The history detailed in the chapter underscores a broader shift in statistical inference, moving from more traditional frequentist methods towards approaches that integrate prior information, reflecting a blend of frequentist and Bayesian philosophies.

These developments in empirical Bayes methods are contextualized within the larger narrative of statistics during the computer age, where computational advancements have facilitated the handling of large datasets and complex models, enabling more sophisticated statistical analysis and inference methods. The chapter not only highlights the technical innovations but also discusses the philosophical shifts that accompanied these changes, reflecting on how the boundaries of statistical inference have expanded in response to new challenges and opportunities presented by technological progress.

This historical exploration provides a foundation for understanding the significant shifts in statistical thinking and practice in the latter half of the twentieth century, influenced heavily by the advent of powerful computing resources .