LIKELIHOOD RATIO TEST

DATA MINING

UCT633

Submitted to: Submitted by:

Dr. Jatin Bedi Himanshi(102018009)

Divija (102018056)

Aniket Sahran(102018065)

**ABSTRACT**

This report provides an overview of the likelihood ratio test in data mining. The likelihood ratio test is a statistical method used to compare two models and determine which one better fits a given dataset. The report explains the concept of the likelihood ratio and how it is used to calculate the statistical significance of the difference between the two models. Additionally, the report describes the practical applications of the likelihood ratio test in the industry, including its use in feature selection, variable reduction, and model selection. The report also discusses the limitations of the likelihood ratio test and its alternatives that are used in data mining. Overall, this report serves as a useful resource for individuals interested in understanding the likelihood ratio test and its business applications in real life.

**INTRODUCTION**

The likelihood ratio (LR) test is a statistical tool used to compare two models, a null model and an alternative model, and determine which one better fits a given set of data. The test is based on the principle that a model with a higher likelihood of generating the observed data is a better fit. The LR test is widely used in various fields, including business, finance, healthcare, and marketing, to assess the significance of various factors and variables. It can help businesses make informed decisions, thereby improving their efficiency.

Suppose X₁, X₂, X₃, …, Xₙ is a random sample from a distribution with Probability Density Function f(x | θ). Then given X = x is observed, the **likelihood function** is the function of θ defined by:A

**Maximum Likelihood Function:** The maximum likelihood estimator is the set of parameter values that maximize the likelihood function, i.e., the values that make the observed data most probable given the model.

The Likelihood-Ratio Test (LRT) is a statistical test used to compare the goodness of fit of two competing statistical models based on the ratio of their likelihoods. It assesses if the constraint (i.e. the null hypothesis) is supported by the observed data and can be used to find out which model is the best fit.

**HOW IS LRT PERFORMED?**

The LRT is performed in the following steps:

1. **Develop the null and alternative hypotheses:** The null hypothesis states that the simpler model is sufficient to explain the data, while the alternative hypothesis states that the more complex model provides a better fit to the data.
2. **Calculate the likelihood ratio test statistic:** The likelihood ratio test statistic is calculated as ratio of maximum likelihood estimator over the null hypothesis and the estimator over the entire space. The likelihood ratio test is a measure of how well a model fits the data.
3. **Determine the threshold value:** Consider the level of significance and determine an appropriate value for the threshold, keeping in mind the trade-off between type I and type II error. If the likelihood ratio test statistic lies below a threshold value, the null hypothesis is rejected.
4. **Interpret the results:** If the null hypothesis is rejected, it indicates that the more complex model provides a better fit to the data than the simpler model.

**NEED AND SIGNIFICANCE**

The likelihood ratio test (LRT) is a significant statistical tool in data analysis with several important implications, including:

* **Model selection:** The LRT is commonly used to compare the fit of two nested models and select the one that provides a better fit to the data. This is critical in data analysis as it helps researchers to identify the most appropriate model for a given dataset and reduce the risk of over-fitting or under-fitting.
* **Hypothesis testing:** The LRT can be used to test the significance of a parameter in a statistical model. For example, in linear regression, the LRT can be used to test whether a variable significantly contributes to the model fit.
* **Variable selection:** The LRT can be used to select the most significant variables in a model by comparing the fit of a model with and without the variable of interest. This can help reduce the number of variables in the model and improve its predictive accuracy.
* **Goodness-of-fit testing:** The LRT can be used to evaluate the fit of a model to the data by comparing the fit of a proposed model to the fit of a null model.
* **Confidence interval estimation:** The LRT can be used to construct confidence intervals for model parameters by comparing the fit of the model with the parameter of interest to the fit of a model without that parameter.

**APPLICATIONS**

* **Credit Risk Assessment in Finance:**

Credit risk assessment is crucial for financial institutions to assess and mitigate the potential risks of loan defaults, bankruptcies, and other credit-related losses. Effective credit risk management helps institutions to maintain a healthy portfolio, reduce credit losses, and ensure long-term profitability. Without proper credit risk management, financial institutions may face significant financial losses and reputation damage.

**How LRT is useful in Credit Risk Management**

One such financial institution that uses the LRT in credit risk management is American Express. It has a sophisticated credit risk management system that uses various statistical models, including the LRT, to evaluate the creditworthiness of customers and make lending decisions. The company uses data analytics to analyze customers' payment behavior, credit history, income, and other factors to assess the likelihood of default. Based on this analysis, the company sets credit limits and interest rates for its customers.

In this study, the financial institution was looking to improve its credit risk management practices. It wants to determine if there is a significant difference in default rates between two groups of borrowers: those with good credit scores and those with poor credit scores. To perform this analysis, the financial institution decides to use a likelihood ratio test.

: no significant difference in default rates between the two groups,

: there is a significant difference.

It calculates the likelihood of the observed data given the null hypothesis and the alternative hypothesis. Using the likelihood ratio test, the bank compares the ratio of the two likelihoods to a threshold value to determine if the difference in default rates is statistically significant or not.

It found that the likelihood ratio is above the threshold, indicating that the difference in default rates between the two groups is statistically significant and not just due to chance. Based on this analysis, the bank decided to adjust its credit risk management practices, such as increasing the interest rates or tightening the lending criteria for borrowers with poor credit scores to reduce the risk of default.

**Model Selection**: Financial institutions can use the likelihood ratio test to evaluate the power of two different models for assessing credit risk.

Financial institutions can define two different models for predicting credit risk or worthiness, a simple model with fewer parameters and a complex model with more parameters:

*Model A:* Parameters: credit utilization, payment history, and length of credit history (traditional credit scoring factors)

*Model B:* Parameters: traditional credit scoring factors + income, employment history, and education level

To determine which model is a better fit for assessing credit risk, it used the likelihood ratio test. The test involved comparing the likelihood of observing the credit data given Model A and the likelihood of observing the same data given Model B. It was found that Model B provides a significantly better fit to the data, then concluded that Model B is a more reliable and accurate model for assessing credit risk.

**Conclusion**

The final inference shows that the likelihood ratio test is a useful tool for comparing the relative merits of different models and can provide valuable insights into the accuracy and reliability of credit risk assessments.  The LRT helps in identifying significant differences in default rates between different groups of borrowers and in determining the most accurate model for assessing credit risk. By adopting effective credit risk management practices based on the LRT, financial institutions can reduce potential losses and maintain a healthy loan portfolio.

* **Quality Control in Manufacturing:**

With the growing market competition in the modern days, it has become essential for companies to keep a track on their product quality, and control it throughout. This helps firms and organizations to maintain a good reputation in the headlines and in the consciousness of their consumers. All industries, be it pharmaceuticals, food production, or manufacturing use Likelihood Ratio Test as a statistical tool for product quality control to compare two or more manufacturing processes and determine whether they produce products of similar quality.

The LRT involves estimating the likelihood of data under two hypotheses: the null hypothesis that the mean quality of the two products is the same, and alternative hypothesis that the mean quality of the new product is greater than that of the original product. To perform the said test, companies might randomly sample their products to test them for quality to calculate a likelihood ratio statistic and test it against a predetermined threshold.

One company that has a well-established quality control system is Toyota. Toyota's production system, known as the Toyota Production System (TPS), is based on the principles of just-in-time manufacturing and continuous improvement. TPS includes several quality control tools and techniques, including the use of statistical process control (SPC) and the LR test.

To illustrate the use of the LR test in quality control, consider a company which produces switches, and they want to know if a change in their manufacturing process brings about a major difference in the quality of the product. The company will start by choosing a parameter to compare the two batches of the product, let’s say tensile strength. The next step would be forming the hypothesis. The null hypothesis will be that both the processes produce switches with same tensile strength, while the alternative will be that the tensile strength of products in the two batches is different.

Let’s say this company takes a sample 20 units of switches from original and new batch each. On conducting studies, it is found that average tensile strength of the new product batch is 55 N/mm2 with a variance of 25 N/mm2. On the other hand, the original product batch has an average tensile strength of 50 N/mm2, with a variance of 25 N/mm2. Let’s consider the tensile strength of the switches in both batches to be normally distributed.

To calculate the likelihood of the null hypothesis, we assume that the two batches have the same mean and estimate a common mean by using the concept of weighted mean which comes out to be 52.5 N/mm2. The likelihood of null hypothesis comes out to be 2.57 \* 10-19. On the other hand, the likelihood of alternate hypothesis comes out to be 1.38 \* 10-19. When put into the formula for likelihood ratio, these values give us 2.04.

Now, to have a threshold value, let’s consider the level of significance to be 0.05. The critical value of chi-square distribution with one degree of freedom is 3.84. Since, our LR value is 2.04, which is less than the critical value, we fail to reject the null hypothesis. This means that the company can safely conclude that mean tensile strength of the new product is more or less the same with the previous product.

In conclusion, quality control is a critical process in manufacturing that ensures products meet certain standards and specifications. Companies use statistical process control and the LR test to monitor and control the production process and identify any issues that may arise. The use of these tools helps companies maintain consistent product quality and reduce the likelihood of product recalls or other quality-related issues.

* **Market Analysis:**

Marketing analysis is an important application of the likelihood ratio (LR) test in business. The LR test can be used to compare different marketing campaigns and assess which one generates the most sales or conversions. This allows companies to make data-driven decisions about their marketing strategies and optimize their marketing budget.

For example, Procter & Gamble (P&G) is a multinational consumer goods company that has used the LR test to evaluate the effectiveness of its marketing campaigns. P&G conducted a study to assess whether its online video ads generated more sales than traditional TV ads. The company used the LR test to compare the likelihood of generating sales from the two types of ads. The results showed that the online video ads generated significantly more sales than TV ads. As a result, P&G increased its investment in online video ads and reduced its investment in TV ads.

Another example is Amazon, which uses the LR test to evaluate the effectiveness of its email marketing campaigns. The company conducted a study to compare two versions of an email campaign promoting a new product. One version of the email had a call-to-action button, while the other did not. Amazon used the LR test to compare the likelihood of generating clicks and sales from the two versions of the email. The results showed that the version with the call-to-action button generated significantly more clicks and sales. As a result, Amazon implemented the call-to-action button in all of its email marketing campaigns.

**Conclusion:**

In conclusion, the LRT is a valuable tool in marketing analysis and can be used to evaluate the effectiveness of various marketing campaigns. P&G's example illustrates how LRT can be applied to test the significance of differences in sales between two different ad variations. However, it is important to ensure that the sample size is large enough to yield reliable results and that the null and alternative hypotheses are carefully defined before conducting the test. Overall, LRT can provide valuable insights for companies looking to optimize their marketing strategies to increase profitability.

**ADVANTAGES**

The likelihood ratio test (LRT) is a powerful statistical tool with several advantages, including:

* **Flexibility:** The LRT can be applied to a wide range of statistical models, including linear regression, logistic regression, mixed-effects models, and generalized linear models. This makes it a versatile tool that can be used in various fields of study.
* **Statistical power:** The LRT is known for its high statistical power, which means that it is effective in detecting differences in fit between two models when they exist. This makes it an ideal tool for hypothesis testing and model selection.
* **Provides a measure of fit:** The LRT provides a measure of the difference in fit between two models, which can be used to evaluate the performance of the models and select the one that provides a better fit to the data.
* **Controls for overfitting:** The LRT can be used to compare the fit of a more complex model to a simpler model, which helps to control for overfitting. This is important in data analysis, as overfitting can lead to poor generalization performance and model instability.
* **Simplicity:** The LRT is a simple statistical test that can be easily implemented in most statistical software packages. This makes it accessible to researchers with varying levels of statistical expertise.

**LIMITATIONS**

While the likelihood ratio test (LRT) is a powerful statistical tool with broad applications, it has several limitations, including:

* **Assumption of nested model structure:** The LRT assumes that the simpler model is nested within the more complex model, which means that the simpler model can be obtained by setting some of the parameters in the more complex model to zero. This assumption may not always hold in practice, particularly in non-parametric models.
* **Sensitivity to sample size:** The LRT can be sensitive to sample size, particularly when comparing two nested models that are very similar. In such cases, small differences in fit between the two models can be statistically significant, leading to spurious results.
* **Dependence on the null model:** The LRT is dependent on the choice of the null model, which is often the simpler model. If the null model is misspecified, the LRT may fail to detect significant differences in fit between the two models..

**ALTERNATIVES**

There are several alternatives to the LRT for comparing the fit of statistical models, including the Wald test, score test, AIC, BIC, and cross-validation. The choice of method depends on the specific research question, the type of data, and the complexity of the model.

* **Wald test:** The Wald test is a statistical test that compares the fit of two models by comparing the difference in their parameter estimates to the standard error of the estimates. The Wald test is less powerful than the LRT but is simpler to implement and is less sensitive to sample size.
* **Score test:** The score test is a statistical test that compares the fit of two models by comparing the difference in their log-likelihoods to the derivative of the log-likelihood with respect to the parameter of interest. The score test is less powerful than the LRT but is more robust to model misspecification and can be used in non-parametric models.

**SUMMARY**

The likelihood ratio test (LRT) is a statistical method used to compare the fit of two nested models, where one model is a simplified version of the other. It is a valuable tool in data mining, with applications in a wide range of fields including business, economics, and finance.

The LRT provides a measure of the difference in fit between the two models and tests whether the more complex model provides a better fit to the data than the simpler model. It is flexible and can be applied to a variety of statistical models, including linear regression, logistic regression, and mixed models.

The LRT is a powerful tool with high statistical power, providing a measure of the difference in fit between the two models. It also controls overfitting, a common issue in data analysis that can lead to poor generalization performance and model instability. However, the LRT has limitations, including the assumption of independent observations and its sensitivity to sample size.

Overall, the LRT is a valuable tool for business analysis, providing insights into model selection, hypothesis testing, and statistical inference. It can be used to identify the best-fitting model for sales forecasting, pricing models, market research, and credit scoring models. However, researchers must be aware of the assumptions and limitations of the LRT and consider alternative methods depending on the research question and type of data.

**REFERENCES**

* <https://jnu.ac.bd/journal/assets/pdf/7_1_211.pdf>
* <https://www.probabilitycourse.com/chapter8/8_4_5_likelihood_ratio_tests.php>
* Introduction to Probability by Hossein Pishro-Nik
* Probability and Statistical Inference by Robert V. Hogg, Elliot A. Tanis and Jagan Mohan Rao

**CONTRIBUTION**

Marketing - Himanshi

Credit Risk Management in Finance - Divija.

Quality control in manufacturing – Aniket