1 Introduction

Have you ever wondered how statisticians extract meaningful insights from data? Or how models are calibrated to make predictions about the future? At the heart of these questions lies a powerful concept: likelihood. This concept serves as a cornerstone of statistical inference, bridging data and theory to uncover hidden truths.

In this article series, I will examine statistical models from a frequentist perspective. To start this series off, we first have to look at a basic element, i.e., likelihood. While I will introduce key ideas and formulas, I will not delve into the detailed process of calculating maximum likelihood estimations; for that, I recommend consulting more specialized resources.

2 Likelihood

Likelihood is a foundational concept in statistical inference. It allows us to measure how well a particular set of parameters explains the observed data. Formally, given a statistical model with parameters θ and observed data $X = (x_1, x_2, \ldots, x_n)$, the likelihood function is defined as:

$$L(\theta; x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i; \theta),$$

where $f(x_i; \theta)$ is the probability density function (PDF) for continuous variables or probability mass function (PMF) for discrete variables, and x_i is the i-th observation. The likelihood function treats x_i as fixed and θ as the variable, indicating how plausible different parameter values are given the data.

2.1 Maximum Likelihood Estimation

One of the most frequent use cases of likelihood is Maximum Likelihood Estimation (MLE). With MLE, we aim to estimate the parameter θ that maximizes the likelihood of observing the given data. Mathematically, the maximum likelihood estimator $\hat{\theta}$ is given by:

$$\hat{\theta} = \arg\max_{\theta} L(\theta; x_1, x_2, \ldots, x_n).$$

Because working with products of many probabilities can be cumbersome, we often take the natural logarithm of the likelihood function to transform the product into a sum. This also improves numerical stability. The log-likelihood function is:

$$\ell(\theta; x_1, x_2, \dots, x_n) = \log L(\theta; x_1, x_2, \dots, x_n) = \sum_{i=1}^{n} \log f(x_i; \theta).$$

Maximizing $\ell(\theta; x)$ instead of $L(\theta; x)$ is equivalent since the logarithm is a strictly increasing function. This transformation simplifies differentiation and other analytical tasks, especially for complex models or large datasets.

3 Parametric Models

Parametric models are statistical models that are characterized by a finite set of parameters. This assumption implies that the data-generating process can be fully described by a specific probability distribution with a finite number of parameters. Formally, a parametric model is usually defined as the family of distributions:

$$\{f(x; \theta) : \theta \in \Theta\},$$

where:

- $f(x; \theta)$ is the probability density function (PDF) or probability mass function (PMF),
- θ is the parameter vector defining the distribution,
- Θ is the parameter space, a subset of $\mathbb{R}^k$ for some finite k.

Any distribution can serve as a parametric model. For example:

- If we assume a normal distribution, the model is parameterized by μ (mean) and σ^2 (variance), and the parameter space Θ is $\{(\mu, \sigma^2) \mid \mu \in \mathbb{R}, \sigma^2 > 0\}$.
- For an exponential distribution, the model is defined by a single parameter $\lambda > 0$ with probability density function $f(x; \lambda) = \lambda \exp(-\lambda x)$ for $x \geq 0$. Here the parameter space is $\Theta = \{\lambda > 0\}$.

The choice of parametric model has profound implications for inference, model fit, and predictive performance.

4 Score Function

The score function is a tool used in the context of maximum likelihood estimation. It measures the sensitivity of the log-likelihood function to changes in the parameter θ. Specifically, the score function $U(\theta)$ is the gradient (or derivative) of the log-likelihood with respect to θ:

$$U(\theta) = \frac{\partial}{\partial \theta} \ell(\theta; x_1, x_2, \ldots, x_n).$$

The score function points in the direction in which the log-likelihood increases most rapidly. At the maximum likelihood estimate $\hat{\theta}$, the score function is zero (under regularity conditions), because the log-likelihood is at a local maximum. By iteratively updating θ in the direction of the score (using techniques like Newton-Raphson or other optimization algorithms), we can find $\hat{\theta}$.

Beyond optimization, the score function provides insight into how strongly the data supports changes in parameter values. Large values of $U(\theta)$ suggest that small changes in θ lead to significant increases in likelihood, indicating that the parameter is sensitive and potentially very informative about the data.

5 Fisher Information

The Fisher Information quantifies the amount of information that an observable random variable X carries about an unknown parameter θ. It is defined as the variance of the score function or equivalently as the expected value of the observed information. For a single observation X, the Fisher Information $I(\theta)$ is given by:

$$I(\theta) = \mathrm{Var}\left(\frac{\partial}{\partial \theta} \log f(X; \theta) \right) = -\mathbb{E}\left[\frac{\partial^2}{\partial \theta^2} \log f(X; \theta) \right],$$

where the expectation is taken with respect to the distribution of X given θ. For a sample of n independent observations, the total Fisher Information is usually the sum of the individual Fisher Informations, assuming independence.

Fisher Information plays a crucial role in the asymptotic theory of MLE. Under regularity conditions, the MLE is approximately normally distributed around the true parameter value for large n, with variance equal to the inverse of the Fisher Information. This relationship underpins many results in statistical estimation and hypothesis testing.

6 Moments of the Score Function

The moments of the score function provide further insights into the distribution of the parameter estimates:

- **First Moment:** The expected value of the score function is zero under regularity conditions:
$$\mathbb{E}[U(\theta)] = 0.$$
This property reflects that, on average, the log-likelihood is at a stationary point with respect to θ.

- **Second Moment:** The variance of the score function is the Fisher Information:
$$\mathrm{Var}[U(\theta)] = I(\theta).$$
This moment explains the dispersion of the parameter estimates around the true value.

- **Third Moment:** The third moment of the score function relates to the skewness of the distribution of the estimates:
$$\mathbb{E}\Big[\Big(U(\theta) - \mathbb{E}[U(\theta)]\Big)^3\Big].$$
If this value is zero, the distribution of the score (and approximately the distribution of the estimator) is symmetric. A positive value indicates right skewness, while a negative value suggests left skewness.

- **Higher Moments:** Higher-order moments (fourth and beyond) describe additional properties such as kurtosis (tailedness) and other shape characteristics of the distribution of the estimator. Exploring these can provide a deeper understanding but often requires more complex calculations and is beyond the scope of this introduction.

7 Relative Likelihood

The relative likelihood quantifies the plausibility of a candidate parameter value relative to the maximum likelihood estimate (MLE). While the MLE gives a single best estimate $\hat{\theta}$, it does not capture uncertainty or the plausibility of alternative parameter values. Relative likelihood addresses this by comparing the likelihood at any θ to the maximum likelihood.

Formally, the relative likelihood $R(\theta)$ is defined as:

$$R(\theta) = \frac{L(\theta; \mathbf{x})}{L(\hat{\theta}; \mathbf{x})},$$

where:

- $L(\theta; \mathbf{x})$ is the likelihood function for a given parameter value θ,
- $L(\hat{\theta}; \mathbf{x})$ is the maximum likelihood, attained at $\hat{\theta}$.

The value $R(\theta)$ lies between 0 and 1, with $R(\hat{\theta}) = 1$. By examining how quickly $R(\theta)$ drops off from 1, we can assess the certainty of our estimate $\hat{\theta}$. A sharp drop indicates high certainty, while a gradual decline suggests a range of plausible values.

Relative likelihood can also be used to construct confidence intervals. For a given confidence level, we consider all parameter values with relative likelihood above a specified threshold. Specifically, parameter values satisfying

$$2 \log R(\theta) \geq -\chi^2_{1,\,1-\alpha},$$

where $\chi^2_{1,\,1-\alpha}$ is the $(1-\alpha)$-quantile of the chi-squared distribution with one degree of freedom, form a confidence interval for θ.

8 Likelihood Ratio Statistics

The likelihood ratio statistic is used to compare how well two nested models fit the observed data. Typically, we compare a more complex, unrestricted model to a simpler, restricted model. The null hypothesis H_0 states that the simpler model fits the data as well as the more complex one.

The likelihood ratio λ is defined as:

$$\lambda = \frac{\sup_{\theta \in \Theta_0} L(\theta; \mathbf{x})}{\sup_{\theta \in \Theta} L(\theta; \mathbf{x})},$$

where Θ_0 is the parameter space under the null hypothesis, and Θ is the unrestricted parameter space.

In practice, we often work with the log-likelihood ratio statistic:

$$-2 \log \lambda = -2 \Big[\log L(\hat{\theta}_0; \mathbf{x}) - \log L(\hat{\theta}; \mathbf{x}) \Big],$$

where $\hat{\theta}_0$ is the MLE under the null hypothesis, and $\hat{\theta}$ is the unrestricted MLE. Under certain regularity conditions, as the sample size increases, the distribution of $-2 \log \lambda$ approaches a chi-squared distribution with degrees of freedom equal to the difference in the number of parameters between the unrestricted and restricted models. This result allows us to perform hypothesis tests: if the observed value of $-2 \log \lambda$ exceeds the critical value from the chi-squared distribution for a chosen significance level, we reject the null hypothesis in favor of the alternative.