4 Estimation

ENSY SILVER[1]

Monday 10$^{\text{th}}$ August, 2020

[1]Thanks to my family, my friend and freedom.

1 Introduction

In fact, those distribution models only works with given parameters, but we do not know the parameters naturally. So we need to estimate those parameters from the experimental data.

2 Maximum likelihood

A natural thought to estimate the parameters is to create a function which evaluate the correspondence of the parameters and experimental data. Mathematicians named the function 'likelihood'. Now we introduce the formal definition of **likelihood**.

Theorem 2.1. *Let $k_1, \cdots, k_n$ be a random sample of size n from the discrete pdf $p_X(k;\theta)$, where θ is an unknown parameter. The likelihood function, $L(\theta)$, is the product of the pdf evaluated at k_j, where $1 \le j \le n$. That is,*

$$L(\theta) = \prod_{j=1}^{n} p_X(k_j;\theta)$$

If $y_1, \cdots, y_n$ is a random sample of size n from a continuous pdf, $f_Y(y;\theta)$, where θ is an unknown parameter, the likelihood function is written

$$L(\theta) = \prod_{j=1}^{n} f_Y(y_j;\theta)$$

Theorem 2.2. *Let $L(\theta) = \prod_{j=1}^{n} p_X(k_j;\theta)$ and $L(\theta) = \prod_{j=1}^{n} f_Y(y_j;\theta)$ be the likelihood functions corresponding to random samples $k_1, \cdots, k_n$ and $y_1, \cdots, y_n$ drawn from the discrete pdf $p_X(k;\theta)$ and continuous pdf $f_Y(y;\theta)$. In each case, let θ_e be a value of the parameter such that $L(\theta_e) \ge L(\theta)$ for all possible values of θ. Then θ_e is called a **maximum likelihood estimate** for θ.*

The general method to get the θ_e is to calculate the derivatives, when $dL(\theta)/d\theta = 0$ or $d\log L(\theta)/d\theta = 0$, we have the θ_e.

However, there exists situations that $dL(\theta)/d\theta = 0$ or $d\log L(\theta)/d\theta = 0$ are not meaningful anymore, for example, $f_Y(y;\theta) = 1/\theta$, $0 \le y \le \theta$. These occur when the range of the pdf from which the data are drawn is a function of the parameter being estimated. The maximum likelihood estimates in these cases will be an order statistic, typically either $y_{\min}$ or $y_{\max}$.

When more than one parameters are not determined, for example, there are

n unknown parameters $\theta_1, \cdots, \theta_n$, we need to solve equations

$$\begin{aligned} \frac{\partial \log L(\theta_1, \cdots, \theta_n)}{\partial \theta_1} &= 0 \\ \frac{\partial \log L(\theta_1, \cdots, \theta_n)}{\partial \theta_2} &= 0 \\ \vdots \quad \vdots \quad \vdots & \\ \frac{\partial \log L(\theta_1, \cdots, \theta_n)}{\partial \theta_n} &= 0 \end{aligned}$$

3 Moments

Another common way of estimation is the *method of moments*. The intuition comes from the equation of moments.

$$E(X^k) = \int_{-\infty}^{\infty} x^k f_X(x; \theta)\, dy$$

If there are n parameters $\theta_1, \cdots, \theta_n$, then we can list the moment of order 1 to moment of order n, and get n equations. Sovling the equations, we finally have the value of $\theta_1, \cdots, \theta_n$.

Theorem 3.1. *Let $y_1, \cdots, y_n$ be a random sample from the continuous pdf $f_Y(y; \theta_1, \cdots, \theta_s)$. The method of moments estimates, $\theta_{1e}, \cdots, \theta_{se}$, for the model' s unknown parameters are the solutions of the s simultaneous equations*

$$\int_{-\infty}^{\infty} y^k f_Y(y; \theta_1, \cdots, \theta_s)\, dy = \frac{1}{n} \sum_{j=1}^{n} y_j^k \quad k = 1, \cdots, s$$

If the underlying random variable is discrete with pdf $p_X(x; \theta_1, \cdots, \theta_s)$ the method of moments estimates are the solutions of the system of equations,

$$\sum_{k=-\infty}^{\infty} x^k p_X(x; \theta_1, \cdots, \theta_s) = \frac{1}{n} \sum_{j=1}^{n} x_j^k \quad k = 1, \cdots, s$$

4 Interval estimation

Those method introduced above is to estimate the parameters of our models. Unfortunately, those values we calculated are not 100% accurate, so we need to find a method to measure the uncertainty.

The usual way to quantify the amount of uncertainty in an estimator is to construct a **confidence interval**. In principle, confidence intervals are ranges of numbers that have a high probability of "containing" the unknown parameter as an interior point. By looking at the width of a confidence interval, we can get a good sense of the estimator' s precision.

It is hard to find a general description for all sorts of confidence intervals. However, we can still introduce the intuition.

Suppose we have chosen the model with a undetermined parameter to fit the data, than we estimate the parameter from the experimental data. We take those data into our models, and express the estimator Z, with pdf $f_Z(z)$. The estimator satisfies standard distribution.(In fact, the *central limit theorem* allows us to do so). Suppose $z_{\alpha/2}$ is defined to be the value for which $P(\geq z_{\alpha/2}) = \alpha/2$. It is $100(1-\alpha)\%$ that

$$-z_{\alpha/2} \leq Z \leq z_{\alpha/2}$$

For a special parameter μ, a $100(1-\alpha)\%$ *confidence interval* for μ is the range of numbers

$$(\mu_{\text{estimated}} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \mu_{\text{estimated}} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) \tag{1}$$

For a specific case, confidence intervals for binomial parameter p, we take 1, and have the theorem.

Theorem 4.1. *Let k be a number of successes in n independent trials, where n is large and $p = P(success)$ is unknown. An appropriate $100(1-\alpha)\%$ confidence interval for p is the set of numbers*

$$\left[\frac{k}{n} - z_{\alpha/2}\sqrt{\frac{(k/n)(1-k/n)}{n}}, \frac{k}{n} + z_{\alpha/2}\sqrt{\frac{(k/n)(1-k/n)}{n}}\right]$$

4.1 Median test

Suppose $y_1, y_2, \cdots, y_n$ denote measurements presumed to have come from a continuous pdf $f_Y(y)$. Let k denote the number of y_j' s that are less than the median of $f_Y(y)$. If the sample is random, we would expect the difference between k/n and $1/2$ to be small. More specifically, a 95% confidence interval based on k should contain the value 0.5.

4.2 Margin of error

Generally, at a confidence level γ , a sample sized n of a population having expected standard deviation σ,

$$MOE_\gamma = z_\gamma\sqrt{\frac{\sigma^2}{n}}$$

4.3 Chossing sample sizes

It is obvious that we need a large sample size to get a higher accuracy. We now parametrize the relationship among sample size n, $z_{\alpha/2}$ and distance d between estimator and unknown parameter.

The key point is, when we transform the random variable to a standard random variable Z, Z can be expressed by a fraction with a denominator $\sqrt{n}$, where n is the sample size. Now we give a theorem.

Theorem 4.2. *Let X be a continuous random variable, parameter θ is unknown, and the estimator is denoted by Y. In order for Y to have at least a $100(1-\alpha)\%$ probability of being within a distance d of θ, the sample size n should be no smaller than*

$$n = \frac{z_{\alpha/2}^2}{4d^2}$$

The proof is a simple application of central limit theorem.

5 Properties of estimators

The method of maximum likelihood and the method of moments both use very reasonable criteria to identify estimators for unknown parameters, yet the two do not always yield the same answer.

More generally, the fact that parameters have multiple estimators (actually, an infinite number of θ's can be found for any given θ) requires that we investigate the statistical properties associated with the estimation process. What qualities should a "good" estimator have? Is it possible to find a "best" θ? These and other questions relating to the theory of estimation will be addressed in the next several sections.

we must first keep in mind that every estimator is a function of a set of random variables—that is, $\theta = h(Y_1, Y_2, \cdots, Y_n)$. As such, any $\hat{\theta}$, itself, is a random variable: It has a pdf, an expected value, and a variance, all three of which play key roles in evaluating its capabilities.

We will denote the pdf of an estimator (at some point u) with the symbol $f_{\hat{\theta}}(u)$ or $p_{\hat{\theta}}(u)$, depending on whether θ is a continuous or a discrete random variable.

5.1 Unbiasedness

Theorem 5.1. *Suppose that $Y_1, Y_2, \cdots, Y_n$ is a random sample from the continuous pdf $f_Y(y;\theta)$, where θ is an unknown parameter. An estimator $\theta = h(Y_1, Y_2, \cdots, Y_n)$ is said to be unbiased (for θ) if $E(\hat{\theta}) = \theta$ for all θ. [The same concept and terminology apply if the data consist of a random sample $X_1, X_2, \cdots, X_n$ drawn from a discrete pdf $p_X(k;\theta)$].*

The key point is, we have already known the relationship between expected θ and random sample Y, then we examine whether the expected value $E(\hat{\theta})$ of estimator $\hat{\theta}$ equals to θ. In the process of estimating unbiasedness, we do not know the exact value of θ.

5.2 Efficiency

The unbiasedness is not the only property for estimators, we also use the efficiency to measure the unbiasedness.

Theorem 5.2. *Let θ_1 and θ_2 be two unbiased estimators for a parameter θ. If*

$$\mathrm{Var}(\hat{\theta}_1) < \mathrm{Var}(\hat{\theta}_2)$$

we say that $\hat{\theta}_1$ is more efficient than $\hat{\theta}_2$. Also, the relative efficiency of $\hat{\theta}_1$ with respect to $\hat{\theta}_2$ is the ratio $\mathrm{Var}(\hat{\theta}_1)/\mathrm{Var}(\hat{\theta}_2)$.

6 Minimum-variance estimators: the Cramér-Rao lower bound

Theorem 6.1 (**Cramér-Rao Inequality**). *Let $f_Y(y;\theta)$ be a continuous pdf with continuous first-order and second-order derivatives. Also, suppose that the set of y values, where $f_Y(y;\theta) \neq 0$, does not depend on θ.*

Let $Y_1, \cdots, Y_n$ be a random sample from $f_Y(y;\theta)$, and let $\hat{\theta} = h(Y_1, Y_2, \cdots, Y_n)$ be any unbiased estimator of θ. Then

$$\mathrm{Var}(\hat{\theta}) \geq \left\{ nE\left[\left(\frac{\partial \ln f_Y(Y;\theta)}{\partial \theta} \right)^2 \right] \right\} = \left\{ nE\left[\frac{\partial^2 \ln f_Y(Y;\theta)}{\partial \theta^2}^2 \right] \right\}$$

The proof is in [1].

Theorem 6.2. *Let Θ denote the set of all estimators $\hat{\theta} = h(Y_1, \cdots, Y_n)$ that are unbiased for the parameter θ in the continuous pdf $f_Y(y;\theta)$. We say that $\hat{\theta}^*$ is a best (or minimum-variance) estimator if $\hat{\theta}^* \in \Theta$ and*

$$\mathrm{Var}(\hat{\theta}^*) \leq \mathrm{Var}(\hat{\theta}) \quad \text{for all} \quad \hat{\theta} \in \Theta$$

Theorem 6.3. *Let $Y_1, \cdots, Y_n$ be a random sample of size n drawn from the continuous pdf $f_Y(y;\theta)$. Let $\hat{\theta} = h(Y_1, \cdots, Y_n)$ be an unbiased estimator for θ.*

1. *The unbiased estimator $\hat{\theta}$ is said to be efficient if the variance of $\hat{\theta}$ equals the Cramér-Rao lower bound associated with $f_Y(y;\theta)$.*
2. *The efficiency of an unbiased estimator θ is the ratio of the Cramér-Rao lower bound for $f_Y(y;\theta)$ to the variance of $\hat{\theta}$.*

7 Sufficient estimators

Theorem 7.1. *Let $X_1 = k_1, \cdots, X_n = k_n$ be a random sample of size n from $p_X(k;\theta)$. The statistic $\hat{\theta} = h(X_1, \cdots, X_n)$ is sufficient for θ if the likelihood function, $L(\theta)$, factors into the product of the pdf for $\hat{\theta}$ and a constant that does not involve θ, that is, if*

$$L(\theta) = \prod_{j=1}^{n} p_X(k_j;\theta) = p_{\hat{\theta}}(\theta_e;\theta) b(k_1, \cdots, k_n)$$

A similar statement holds if the data consist of a random sample $Y_1 = y_1, \cdots, Y_n = y_n$ drawn from a continuous pdf $f_Y(y;\theta)$.

Using theorem 7.1 to verify that a statistic is sufficient requires that the pdf $p_{\hat{\theta}}[h(k_1, \cdots, k_n); \theta]$ or $f_{\hat{\theta}}[h(y_1, \cdots, y_n); \theta]$ be explicitly identified as one of the two factors whose product equals the likelihood function. If $\hat{\theta}$ is complicated, though, finding its pdf may be prohibitively difficult. The next theorem gives an alternative factorization criterion for establishing that a statistic is sufficient. It does not require that the pdf for $\hat{\theta}$ be known.

Theorem 7.2. *Let $X_1 = k_1, \cdots, X_n = k_n$ be a random sample of size n from the discrete pdf $p_X(k; \theta)$. The statistic $\hat{\theta} = h(X_1, \cdots, X_n)$ is sufficient for θ if and only if there are functions $g[h(k_1, \cdots, k_n); \theta]$ and $b(k_1, \cdots, k_n)$ such that*

$$L(\theta) = g[h(k_1, \cdots, k_n)] \cdot b(k_1, \cdots, k_n)$$

where the function $b(k_1, \cdots, k_n)$ does not involve the parameter θ . A similar statement holds in the continuous case.

8 Consistency

Theorem 8.1. *An estimator $\hat{\theta}_n = h(W_1, W_2, \cdots, W_n)$ is said to be consistent for θ if it converges in probability to θ, that is, if for all $\epsilon > 0$,*

$$\lim_{n \to \infty} P(|\hat{\theta}_n - \theta| < \epsilon) = 1$$

9 Bayesian estimation

Theorem 9.1. *Let W be a statistic dependent on a parameter θ. Call its pdf $f_W(w|\theta)$. Assume that θ is the value of a random variable , whose **prior distribution** is denoted by $p_\Theta(\theta)$, if Θ is discrete, and $f_\Theta(\theta)$, if Θ is continuous. The **posterior distribution** of Θ, given that $W = w$, is the quotient*

$$g_\theta(\theta|W = w) = \frac{p_W(w|\theta) f_\Theta(\theta)}{\int_{-\infty}^{\infty} p_W(w|\theta) f_\Theta(\theta)\, d\theta}$$

If W is continuous, replace p_W with f_W. If θ is discrete, replace the integrations and f_Θ with summation and p_Θ.

Theorem 9.2. *Let $\hat{\theta}$ be an estimator for θ based on a statistic W. The loss function associated with $\hat{\theta}$ is denoted $L(\hat{\theta}, \theta)$, where $L(\hat{\theta}, \theta) \geq 0$ and $L(\theta, \theta) = 0$.*

Theorem 9.3. *Let $L(\hat{\theta}, \theta)$ be the loss function associated with an estimate of the parameter θ. Let $g_\theta(\theta|W = w)$ be the posterior distribution of the random variable. Then the risk associated with $\hat{\theta}$ is the expected value of the loss function with respect to the posterior distribution of θ.*

$$risk = \begin{cases} \int_\theta L(\hat{\theta}, \theta) g_\Theta(\theta|W = w)\, d\theta, & \text{if } \Theta \text{ is continuous} \\ \sum_\theta L(\hat{\theta}, \theta) g_\Theta(\theta|W = w), & \text{if } \Theta \text{ is discrete} \end{cases}$$

Given that the risk function represents the expected loss associated with the estimator $\hat{\theta}$, it makes sense to look for the $\hat{\theta}$ that minimizes the risk. Any $\hat{\theta}$ that achieves that objective is said to be a Bayes estimate. In general, finding the **Bayes estimate** requires solving the equation $d(\text{risk})/d\hat{\theta} = 0$. For two of the most frequently used loss functions, $L(\hat{\theta}, \theta) = |\hat{\theta} - \theta|$ and $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$, though, there is a much easier way to calculate $\hat{\theta}$.

Theorem 9.4. *Let $g_\Theta(\theta|W = w)$ be the posterior distribution for the unknown parameter θ.*

1. *If the loss function associated with $\hat{\theta}$ is $L(\hat{\theta}, \theta) = |\hat{\theta} - \theta|$, then the Bayes estimate for θ is the median of $g_\Theta(\theta|W = w)$.*
2. *If the loss function associated with $\hat{\theta}$ is $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$, then the Bayes estimate for θ is the mean of $g_\Theta(\theta|W = w)$.*

References

[1] Adam Merberg and Steven J. Miller. *Course Notes for Math 162: Mathematical StatisticsThe Cramér-Rao Inequality*. `https://web.williams.edu/Mathematics/sjmiller/public_html/BrownClasses/162/Handouts/CramerRaoHandout08.pdf`. May 2008.