STA 250: Statistics

Notes 5. Choosing Test Statistic: the Maximum Likelihood Approach

Book chapters: 7.5,7.6

1 Choice of Test Statistic: Neyman-Pearson Lemma

Last week we discussed about choosing the cut-off c for a testing rule "reject H_0 if $T(x) > c$" when a test statistic $T(X)$ has been decided upon. This week we will focus on how to choose $T(X)$. We start with a definitive result, albeit within a very simple statistical model.

Lemma 1 (Neyman-Pearson Lemma)**.** *Consider a statistical model $X \sim f(x|\theta)$, $\theta \in \Theta$ where the parameter set contains only two points: $\Theta = \{\theta_0, \theta_1\}$ and suppose we want to test $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$. Consider the "likelihood ratio" testing rule "reject H_0 if $f(x|\theta_1)/f(x|\theta_0) > k$" with size $\alpha(k) = P_{[X|\theta_0]}(f(X|\theta_1)/f(X|\theta_0) > k)$. Then, among all testing rules for H_0 (based on any test statistic and any cut-off) with size less than or equal to $\alpha(k)$, the likelihood ratio test has maximum power at $\theta = \theta_1$.*

For example, recall the drug effectiveness study and consider the model: $X_1, \cdots, X_n \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma^2)$, $\sigma = 3$ is known, μ is unknown but is assumed to equal either 0 or 2000, i.e., $\Theta = \{0, 2000\}$. Suppose we want to test $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$. For this model, for any $k > 0$ there is a $c > 0$ such that $f(x|\mu = 2000)/f(x|\mu = 0) > k \iff \sqrt{n}\bar{X}/\sigma > c$, and hence the size α likelihood ratio test is precisely: "reject H_0 if $\sqrt{n}\bar{X}/\sigma > z(2\alpha)$". By Neyman-Pearson lemma, among all size α tests, this likelihood ratio test has maximum power at $\mu = 2000$.

2 The likelihood function

Neyman-Pearson's lemma describes how to construct an "optimal" testing rule of a given size. Although the result is not directly applicable to most models (none of which will be a two point set), it leads to a pretty useful and unified approach of choosing testing rules for a large number of models. In most cases these testing rules enjoy a similar optimality property. The unified approach is known as the maximum likelihood (ML) testing. To introduce this notion, we first need to define the "likelihood function".

Suppose a statistical model $\{f(x|\theta) : \theta \in \Theta\}$ has been constructed for data X, with each θ representing a different theory. When we observe data $X = x$, we can compare two parameter values (i.e., two theories) $\theta = \theta_1$ and $\theta = \theta_2$ by looking at the ratio $f(x|\theta_1)/f(x|\theta_2)$. If this ratio equals 2, then the data $X = x$ is twice as likely to be observed under $\theta = \theta_1$ than it is under $\theta = \theta_2$. Such comparisons can be done based on the **likelihood function**

$$L_x(\theta) := f(x|\theta), \theta \in \Theta.$$

Note that $L_x(\theta)$ is a function over the variable θ taking values in the set Θ.

For all technical purposes, one can work with $L_x(\theta)$ in the log-scale. That is, define the log-likelihood function

$$\ell_x(\theta) = \log L_x(\theta) = \log f(x \mid \theta).$$

Log-scale comparisons between theories are then done by differences $\ell_x(\theta_1) - \ell_x(\theta_2)$. The likelihood function $L_x(\theta)$ (or the log-likelihood function $\ell_x(\theta)$) gives scores to parameter values $\theta \in \Theta$ as to how well they explain the observed data $X = x$.

3 Maximum likelihood testing

Maximum likelihood (ML) testing takes as a test statistic the (maximum) likelihood ratio statistic $LR(X)$ defined by

$$LR(x) = \frac{\max_{\theta\in\Theta} L_x(\theta)}{\max_{\theta\in\Theta_0} L_x(\theta)}$$

which equals the ratio of the highest likelihood score among all parameter values to the highest score among all parameter values matching H_0. The larger the ratio, the stronger the evidence against H_0, because some parameter value outside of Θ_0 must have explained the data way better than any value inside Θ_0.

So the ML testing rule is "reject H_0 if $LR(x) > k$" for some cut-off k. We deliberately use "k" to denote the cut-off here, because in all applications of ML testing, we will simplify $LR(x)$ to a more familiar test statistic $T(x)$ and use "c" for cut-offs to be applied to $T(X)$.

4 Calculating the numerator of $LR(x)$: MLE

To calculate the numerator of $LR(x)$, we need to maximize $L_x(\theta)$ over $\theta \in \Theta$ (recall x is fixed at the actual recorded data). Any point $\theta \in \Theta$ where $L_x(\theta)$ attains its maximum is called a maximum likelihood estimate (MLE), and is denoted $\hat{\theta}_{\text{MLE}}(x)$. For many models there is a single point where this happens, so the MLE is unique, and we can talk about *the* MLE. Note that since log is a monotone transform, we also have $\ell_x(\hat{\theta}_{\text{MLE}}(x)) = \max_{\theta\in\Theta} \ell_x(\theta)$, i.e., the MLE maximizes the log-likelihood function over Θ.

The MLE has several nice properties. Among them is the probability result that for many models of the form $X_i \overset{\text{IID}}{\sim} g(x_i|\theta)$, if we fixed a value θ_0 for θ and simulated our data $X = (X_1, \cdots, X_n)$ with $X_i \overset{\text{IID}}{\sim} g(x_i|\theta_0)$ and evaluated the corresponding MLE value $\hat{\theta}_{\text{MLE}}(X)$, it would be fairly close θ_0 on average, with the closeness improving with larger n. In fact we will see the result that $\hat{\theta}_{\text{MLE}}(X)$ will be approximately distributed as a normal variable with mean θ_0 and a variance that is proportional to $1/n$. Because of this proximity of $\hat{\theta}_{\text{MLE}}(X)$ to the true value of θ, the MLE is taken to be a good estimate of the model parameter.

5 Finding the MLE

A standard technique to find the MLE relies on the following observation. If $L_x(\theta)$, or equivalently, $\ell_x(\theta)$ is a differentiable function over Θ with a unique maxima inside Θ, then

its first derivative vanishes at the maximum. Thus, if θ is a p-dimensional vector $\theta = (\theta_1, \theta_2, \cdots, \theta_p)$ then the MLE $\hat{\theta}_{\text{MLE}}(x)$ can be found by solving the simultaneous equations

$$\frac{\partial}{\partial \theta_j} \ell_x(\theta) = 0, \quad j = 1, 2, \cdots, p,$$

in θ. In many cases these equations can be solved analytically, and we'd see some examples shortly. In many other cases, these equations can be solved by running a suitable computer algorithm.

Example (Opinion poll)**.** In our opinion poll example data X is modeled by $\mathsf{Binomial}(n, p)$, $p \in [0, 1]$ and the likelihood function based on observation $X = x$ is given by

$$L_x(p) = \binom{n}{p} p^x (1-p)^{n-x}$$

and so the log-likelihood function is given by

$$\ell_x(p) = \text{const} + x \log p + (n - x) \log(1 - p),$$

with $p \in [0, 1]$. To find the MLE we set up the equation

$$0 = \frac{\partial}{\partial p} \ell_x(p) = \frac{x}{p} - \frac{n-x}{1-p}$$

which is solved at $p = x/n$. Hence $\hat{p}_{\text{MLE}}(x) = x/n$. For $n = 500$ and observed data $X = 200$, the MLE is 0.40. This is the researcher's "estimate", based on the ML approach, of the unknown proportion of supporters in the entire college.

Example (Drug effectiveness)**.** In this study we modeled increase in sleep hours $X_1, \cdots, X_n$ of $n = 10$ patients by $X_i \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma^2)$. For the moment assume σ is fixed (say $\sigma = 3$) and the only model parameter is $\mu \in (-\infty, \infty)$. So the joint pdf of the data X equals

$$f(x|\mu, \sigma^2) = \prod_{i=1}^{n} g(x_i|\mu, \sigma^2), \quad x = (x_1, \cdots, x_n) \in \mathbb{R}^n.$$

Let's start by writing the log-likelihood function

$$\begin{aligned}
\ell_x(\mu) = \log f(x|\mu, \sigma^2) &= \sum_{i=1}^{n} \log g(x_i|\mu, \sigma^2) \\
&= \sum_{i=1}^{n} \left[-\frac{1}{2} \log(2\pi) - \frac{1}{2} \log \sigma^2 - \frac{(x_i - \mu)^2}{2\sigma^2} \right] \\
&= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{2\sigma^2}
\end{aligned}$$

which is a quadratic function in μ (here $\sigma = 3$ is known, but we retain the symbol σ to keep the calculations general and adaptable to other values of σ). At this stage we use the identity that for any n numbers $x_1, \cdots, x_n$ and another number a,

$$\sum_{i=1}^{n} (x_i - a)^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - a)^2$$

where $\bar{x} = \sum_{i=1}^n x_i/n$ is the average of $x_1, \cdots, x_n$. Using this above we see

$$\begin{aligned}\ell_x(\mu) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{2\sigma^2} - \frac{n(\bar{x}-\mu)^2}{2\sigma^2} \\ &= \text{const} - \frac{n(\bar{x}-\mu)^2}{2\sigma^2}\end{aligned}$$

where "const" absorbs all additive terms that do not involve the argument μ of the log-likelihood function.

To find the MLE we now set up the equation

$$0 = \frac{\partial}{\partial\mu}\ell_x(\mu) = \frac{n(\bar{x}-\mu)}{\sigma^2}$$

which is solved at $\mu = \bar{x}$ and hence $\hat{\mu}_{\text{MLE}}(x) = \bar{x}$. For our actual data $\bar{x} = 2.33$, hence the MLE of μ is 2.33.

Example (Drug effectiveness, Cond.)**.** Now consider the case where both μ and σ are unknown model parameters. Working as before we get

$$\ell_x(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{2\sigma^2} - \frac{n(\bar{x}-\mu)^2}{2\sigma^2}.$$

To find the MLE we set up the equations:

$$\begin{aligned}0 &= \frac{\partial}{\partial\mu}\ell_x(\mu, \sigma^2) = \frac{n(\bar{x}-\mu)}{\sigma^2} \\ 0 &= \frac{\partial}{\partial\sigma^2}\ell_x(\mu, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{2(\sigma^2)^2} + \frac{n(\bar{x}-\mu)^2}{2(\sigma^2)^2}\end{aligned}$$

which are solved at $\mu = \bar{x}$, $\sigma^2 = \sum_{i=1}^n (x_i - \bar{x})^2/n$. Hence $\hat{\mu}_{\text{MLE}}(x) = \bar{x}$ and $\hat{\sigma}^2_{\text{MLE}}(x) = \sum_{i=1}^n (x_i - \bar{x})^2/n = \frac{n-1}{n}s_x^2$. For our actual recorded data, $\hat{\mu}_{\text{MLE}}(x) = \bar{x} = 2.33$ and $\hat{\sigma}_{\text{MLE}}(x) = \sqrt{(n-1)/n} \times s_x = \sqrt{0.9} \times 2 = 1.898$.

6 Computing $LR(x)$

Although $LR(x)$ could be computed by computing its numerator and denominator (which would require maximizing $L_x(\theta)$ or $\ell_x(\theta)$ over $\theta \in \Theta_0$), there is an easier way out. For observed data x, and any $b \in [0, 1]$, define the ML interval $A_b(x)$ as:

$$A_b(x) = \{\theta : L_x(\theta) > bL_x(\hat{\theta}_{\text{MLE}}(x))\},$$

i.e., $A_b(x)$ collects all θ at which the likelihood score is within a fraction b of the maximum score. Then the testing rule "reject H_0 if $LR(x) > k$" is equivalent to the rule "reject H_0 if $A_{1/k}(x)$ and Θ_0 have no overlap". We will see in the next lecture that characterizing an ML interval $A_b(x)$ is relatively easy.