STA 114: Statistics

Notes 4. ML Intervals

Well supported theories

Reporting the MLE, or any other single number summary of the likelihood function, undermines the inherent uncertainty associated with a statistical model. Instead we could report a subset of well supported theories as $A_k(x) = \{\theta \in \Theta : L_x(\theta) \geq kL_x(\hat{\theta}_{\text{MLE}}(x))\}$ for a fraction of our choice $k \in [0, 1]$. In the log-scale, the set $A_k(x)$ equals $B_c(x) = \{\theta \in \Theta : \ell_x(\theta) \geq \ell_x(\hat{\theta}_{\text{MLE}}(x)) - c^2/2\}$ where $c = \sqrt{2\log(1/k)} \geq 0$. When θ is a scalar parameter and $\ell_x(\theta)$ is a nice unimodal function with a unique maxima at $\hat{\theta}_{\text{MLE}}(x)$, the set $B_c(x)$ forms an interval around the MLE, and is called an ML interval. We now look at how to characterize and compute such ML intervals.

Characterization for normal model with known variance

Example (Lactic acid concentration, Contd.)**.** Consider again modeling n concentration measurements $X_1, \cdots, X_n$ by $X_i \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma^2)$, $\mu \in (-\infty, \infty)$, $\sigma = 1/3$. We previously derived that the log-likelihood function is given by:

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

with $\hat{\mu}_{\text{MLE}}(x) = \bar{x}$. Therefore,

$$\ell_x(\mu) - \ell_x(\hat{\mu}_{\text{MLE}}(x)) = -\frac{n(\bar{x} - \mu)^2}{2\sigma^2}$$

and so for any $c \geq 0$, the set $B_c(x) = \{\mu \in (-\infty, \infty) : \ell_x(\mu) \geq \ell_x(\hat{\mu}_{\text{MLE}}(x)) - c^2/2\}$ equals

$$B_c(x) = \left\{\mu \in (-\infty, \infty) : \frac{n(\bar{x} - \mu)^2}{2\sigma^2} \leq \frac{c^2}{2}\right\} = [\bar{x} - c\sigma/\sqrt{n}, \bar{x} + c\sigma/\sqrt{n}].$$

Therefore the set $B_c(x)$ of well supported theories forms an interval centered at the MLE $\bar{x}$ with half-width $c\sigma/\sqrt{n}$. For our data from the cheese manufacturer with $n = 10$ and $\bar{x} = 1.379$, this interval equals [1.17, 1.59] for a choice of $c = 1.96 = \sqrt{2\log(1/0.146)}$. $\square$

That the width of the ML interval $\bar{x} \mp c\sigma/\sqrt{n}$ should depend on σ and n is intuitive. With larger σ, there is less separation between the normal pdfs $\mathsf{Normal}(\mu_1, \sigma^2)$ and $\mathsf{Normal}(\mu_2, \sigma^2)$ and hence a less sharp comparison between theories is possible. Indeed, the interval gets wider with larger σ. However, sharp comparison should be eventually possible with more and more specimens being measured, i.e., with large n, which indeed shortens the width of the interval.

Characterization for normal model with unknown variance

Example (Lactic acid concentration, Contd.). Now consider the case where the variability component is not assumed known and our model for data is: $X_i \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma^2)$, $(\mu, \sigma) \in (-\infty, \infty) \times (0, \infty)$. We are still interested in reporting a set of valued of μ that are well supported by data. One way of constructing such a set is the following:

$$B_c(x) = \left\{ \mu \in (-\infty, \infty) : \max_{\sigma^2 \in (0,\infty)} \ell_x(\mu, \sigma^2) \geq \ell_x(\hat{\mu}_{\text{MLE}}(x), \hat{\sigma}^2_{\text{MLE}}(x)) - c^2/2 \right\}$$

that is, we report any value of μ which, for some σ, explains the data within a log-factor c of the best explanation offered by the MLE. The quantity $\ell^*_x(\mu) = \max_{\sigma^2} \ell_x(\mu, \sigma^2)$ is said to give the profile log-likelihood at μ. Equivalently, one can define the profile likelihood $L^*_x(\mu) = \max_{\sigma^2} L_x(\mu, \sigma^2)$. Evidently, $\ell^*_x(\mu) = \log L^*_x(\mu)$.

Note that $\max_\mu \ell^*_x(\mu) = \max_{\mu,\sigma^2} \ell_x(\mu)$ and hence the profile likelihood is maximized at the same $\hat{\mu}_{\text{MLE}}(x)$ which, coupled with $\hat{\sigma}^2_{\text{MLE}}(x)$ maximizes the original likelihood. So the MLE of μ based on the profile likelihood is the same as the original MLE. So the set $B_c(x)$ above then is same as what we would do for the scalar parameter μ but with its profile likelihood rather than the original likelihood: $B_c(x) = \{\mu \in (-\infty, \infty) : \ell^*_x(\mu) \geq \ell^*_x(\hat{\mu}_{\text{MLE}}(x)) - c^2/2\}$.

We previously derived

$$\ell_x(\mu, \sigma^2) = \text{const} - \frac{n}{2} \log \sigma^2 - \frac{n\{v_x + (\bar{x} - \mu)^2\}}{2\sigma^2}$$

and so to maximize this in σ^2, for a given μ, we set:

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

which is solved at $\sigma^2 = v_x + (\bar{x} - \mu)^2$. Plugging this into the log-likelihood we get the profile log-likelihood

$$\ell^*_x(\mu) = \max_{\sigma^2 \in (0,\infty)} \ell_x(\mu, \sigma^2) = \text{const} - \frac{n}{2} \log\{v_x + (\bar{x} - \mu)^2\} - \frac{n}{2}.$$

Plugging in the MLE $\hat{\mu}_{\text{MLE}}(x) = \bar{x}$, we get

$$\ell^*_x(\hat{\mu}_{\text{MLE}}(x)) = \text{const} - \frac{n}{2} \log v_x - \frac{n}{2}$$

and consequently, $\ell^*_x(\mu) - \ell^*_x(\hat{\mu}_{\text{MLE}}(x)) = -\frac{n}{2} \log\{1 + (\bar{x} - \mu)^2/v_x\}$. Therefore,

$$\begin{aligned} B_c(x) &= \left\{ \mu : \frac{n}{2} \log \left\{ 1 + \frac{(\bar{x} - \mu)^2}{v_x} \right\} \leq c^2/2 \right\} \\ &= \left\{ \mu : \frac{(\bar{x} - \mu)^2}{v_x} \leq e^{c^2/n} - 1 \right\} \\ &= \left[\bar{x} - v_x^{1/2} \sqrt{e^{c^2/n} - 1}, \bar{x} + v_x^{1/2} \sqrt{e^{c^2/n} - 1} \right] \end{aligned}$$

which is an interval centered at $\bar{x}$ with half-width $v_x^{1/2}\sqrt{e^{c^2/n}-1}$. For n moderately large, $e^{c^2/n}-1 \approx c^2/n$ (see Figure 1), and hence $B_c(x) \approx \bar{x} \mp c v_x^{1/2}/\sqrt{n}$. This interval looks just like the one we had for the known σ case, with the estimate $v_x^{1/2}$ in place of σ.

The sample variance of n numbers $x_1, \cdots, x_n$ is usually defined as $s_x^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$, which relates to v_x as $v_x = \frac{n-1}{n} s_x^2$. For large n, $v_x \approx s_x^2$ and therefore we could also write $B_c(x) \approx \bar{x} \mp c s_x/\sqrt{n}$.

For our observed data $\bar{x} = 1.379$ and $v_x = 0.31^2$. Therefore, for $c = 1.96$, $B_{1.96}(x) \approx [1.19, 1.57]$.

□

Characterization in general

The characterization of $B_c(x)$ in the above example used the fact that the log-likelihood function is quadratic, or log-quadratic in the parameter of interest. This is a special property of the normal models. For other models, the the log-likelihood may not be quadratic, making exact characterization of $B_c(x)$ difficult. Nonetheless, we can use a quadratic approximation to obtain a good approximate characterization of $B_c(x)$.

Suppose $X \sim f(x|\theta), \theta \in \Theta$ is our statistical model for data $X \in S$. We will assume θ is a scalar parameter, i.e., Θ is a subset of the real line. We have observed $X = x \in S$ and have constructed the log-likelihood function $\ell_x(\theta, \theta \in \Theta$ and suppose it is uniquely maximized at $\hat{\theta}_{\text{MLE}}(x)$ inside Θ. Fix a $c \in [0, \infty]$ and consider $B_c(x) = \{\theta \in \Theta : \ell_x(\theta) \geq \ell_x(\hat{\theta}_{\text{MLE}}(x)) - c^2/2\}$.

Use the notations $\dot{\ell}_x(\theta)$ and $\ddot{\ell}_x(\theta)$ to denote the first and second order derivatives $\frac{\partial}{\partial\theta}\ell_x(\theta)$ and $\frac{\partial^2}{\partial\theta^2}\ell_x(\theta)$ of the log-likelihood function. Because $\ell_x(\theta)$ is maximized at $\hat{\theta}_{\text{MLE}}(x)$ inside Θ, we must have $\dot{\ell}_x(\hat{\theta}_{\text{MLE}}(x)) = 0$ and $\ddot{\ell}_x(\hat{\theta}_{\text{MLE}}(x)) < 0$. Let $I_x = -\ddot{\ell}_x(\hat{\theta}_{\text{MLE}}(x))$, which is a positive number.

the derivative of $\ell_x(\theta)$ must vanish at the maximal points which is $\hat{\theta}_{\text{MLE}}(x)$. Now, use second order Taylor approximation of $\ell_x(\theta)$ around $\hat{\theta}_{\text{MLE}}(x)$ to write

$$\begin{aligned}
\ell_x(\theta) &\approx \ell_x(\hat{\theta}_{\text{MLE}}(x)) + (\theta - \hat{\theta}_{\text{MLE}}(x))\dot{\ell}_x(\hat{\theta}_{\text{MLE}}(x)) + \frac{1}{2}(\theta - \hat{\theta}_{\text{MLE}}(x))^2\ddot{\ell}_x(\hat{\theta}_{\text{MLE}}(x)) \\
&= \ell_x(\hat{\theta}_{\text{MLE}}(x)) - \frac{1}{2}(\theta - \hat{\theta}_{\text{MLE}}(x))^2 I_x
\end{aligned}$$

and consequently,

$$B_c(x) \approx \left\{\theta : \frac{I_x}{2}(\theta - \hat{\theta}_{\text{MLE}}(x))^2 \leq \frac{c^2}{2}\right\} = \left[\hat{\theta}_{\text{MLE}}(x) - \frac{c}{\sqrt{I_x}}, \hat{\theta}_{\text{MLE}}(x) + \frac{c}{\sqrt{I_x}}\right].$$

The quantity I_x is called the "observed information". It usually increases when more information are available from data. In particular if data $X = (X_1, \cdots, X_n)$ with X_i modeled as $X_i \overset{\text{IID}}{\sim} g(x|\theta)$, then I_x is roughly proportional to n.

Example (Opinion poll, Contd.)**.** For the opinion poll example, with the model $X \sim \mathsf{Binomial}(n, p)$, $p \in [0, 1]$, the log-likelihood function equals

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

with $\hat{p}_{\text{MLE}}(x) = x/n$. Differentiating twice we get, $\ddot{\ell}_x(p) = -x/p^2 - (n-x)/(1-p)^2$ and so

$$I_x = -\ddot{\ell}_x(\hat{p}_{\text{MLE}}(x)) = \frac{n^2}{x} + \frac{n^2}{n-x} = \frac{n}{\frac{x}{n}(1-\frac{x}{n})}.$$

For our data with $n = 500$ and $x = 200$, $\hat{p}_{\text{MLE}}(x) = 0.4$ and $I_x = 2083$. Therefore, for $c = 1.96 = \sqrt{2\log(1/0.146)}$, $B_{1.96}(x) \approx 0.4 \mp 0.043 = [0.357, 0.443]$.

Choice of the cutoff

So we now how to construct $B_c(x)$ for a choice of $c \geq 0$ (at least for some statistical models). But how do we decide upon c? Consider two choices of this cutoff $c = 1.96$ and $c = 3$. Qualitatively we understand that $B_3(x)$ includes more theories than $B_{1.96}(x)$, i.e., $c = 3$ has a lower standard than $c = 1.96$ of accepting a theory as a "good explanation" of the data. But is there a quantitative interpretation of the choice c?

The classical theory of statistics provides such a quantification. It involves a "what if" type thought experiment that we shall see in detail in the next two lectures.

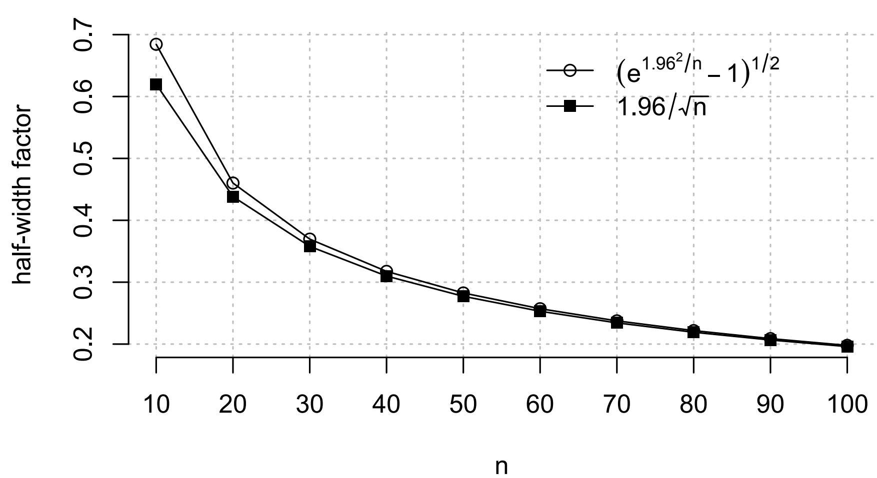

Figure 1: Comparison of $\sqrt{e^{c^2/n} - 1}$ and $c/\sqrt{n}$ for $c = 1.96$ over a range of n values.