ST202 Lent Term – Handout 3

Yudong Chen
y.chen276@lse.ac.uk

March, 2023

1 A brief discussion on the regularity conditions

Throughout this handout, we assume $Y_1, \ldots, Y_n$ to be a random sample of size n. We use $I_{\underline{\mathbf{Y}}}(\cdot)$ to denote the Fisher information based on the entire sample (of size n), and $I_Y(\cdot)$ to denote that based on a single observation Y_1. Since $Y_1, \ldots, Y_n$ are i.i.d., we have $I_{\underline{\mathbf{Y}}}(\cdot) = nI_Y(\cdot)$.

We say a parametric model $\{f(\cdot;\theta), \theta \in \Theta\}$ is regular (or satisfies the regularity conditions) if

1. $\frac{\partial}{\partial\theta} f(y;\theta)$ exists almost everywhere;
2. The integral of $f(y;\theta)$ can be differentiated under the integral sign with respect to θ;
3. The support of $f(y;\theta)$ does not depend on θ.

You are not expected to know the above conditions for the exam, though the following discussions can be very helpful for you to know. Most of the models in this course are regular. However, speical attention needs to be paid to the 3rd condition above. For example, Uniform$[0,\theta]$ is not regular, as its support depends on the parameter of interest.

Many properties of the score function $s(\theta;\underline{\mathbf{Y}})$ and the Fisher information $I_{\underline{\mathbf{Y}}}(\theta) := \mathbb{E}[(s(\theta;\underline{\mathbf{Y}}))^2]$ rely on the regularity conditions. These results include $\mathbb{E}[s(\theta;\underline{\mathbf{Y}})] = 0$, $I_{\underline{\mathbf{Y}}}(\theta) = \mathrm{Var}(s(\theta;\underline{\mathbf{Y}})) = -\mathbb{E}\big[\frac{\partial}{\partial\theta} s(\theta;\underline{\mathbf{Y}})\big]$, Cramer–Rao lower bound, the consistency and the asymptotic normality of the MLE (this actually requires more conditions than the regularity alone). When the regularity conditions do not hold, even though the Fisher information can be computed from the definition, it will not have these good properties, so it is not that interesting to study it.

2 A worked example of an irregular model

We study $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \text{Uniform}[0,\theta]$. The parameter of interest is θ. The density is $f(y;\theta) = \frac{1}{\theta}$ for $0 \le y \le \theta$. In a regular model, we are usually not concerned about the support. But here, the support of the density is $[0,\theta]$, the right endpoint of which is our parameter of interest. In this case, extra care needs to be taken. We need to write the density as $f(y;\theta) = \frac{1}{\theta}\mathbb{1}_{[0,\theta]}(y)$, or $f(y;\theta) = \frac{1}{\theta}\mathbb{1}\{0 \le y \le \theta\}$. Both expressions mean that the density is $\frac{1}{\theta}$ when $0 \le y \le \theta$ and the density is 0 otherwise. Now we first calculate the likelihood function:

$$\begin{aligned} L(\theta;\underline{\mathbf{Y}}) &= \prod_{i=1}^n f(Y_i;\theta) = \frac{1}{\theta^n}\prod_{i=1}^n \mathbb{1}\{0 \le Y_i \le \theta\} = \frac{1}{\theta^n}\mathbb{1}\{0 \le Y_1, Y_2, \ldots, Y_n \le \theta\} \\ &= \frac{1}{\theta^n}\mathbb{1}\big\{0 \le \min_{1\le i\le n} Y_i \le \max_{1\le i\le n} Y_i \le \theta\big\} = \frac{1}{\theta^n}\mathbb{1}\big\{\min_{1\le i\le n} Y_i \ge 0\big\}\mathbb{1}\big\{\max_{1\le i\le n} Y_i \le \theta\big\}. \end{aligned}$$

Let $c(\underline{\mathbf{y}}) = \mathbb{1}\{\min_{1\le i\le n} y_i \ge 0\}$ and $b(\theta, h(\underline{\mathbf{y}})) = \frac{1}{\theta^n}\mathbb{1}\{\max_{1\le i\le n} y_i \le \theta\}$, with $h(\underline{\mathbf{y}}) = \max_{1\le i\le n} y_i$. We see that

$$h(\underline{\mathbf{Y}}) = \max_{1\le i\le n} Y_i$$

is a sufficient statistic for θ by the factorisation criterion $L(\theta; \underline{\mathbf{y}}) = b(\theta, h(\underline{\mathbf{y}}))c(\underline{\mathbf{y}})$. It is also straightforward to check that it is minimal sufficient.

Recall that MLE $\hat{\theta}$ maximises the likelihood $L(\theta; \underline{\mathbf{Y}})$. Equivalently, it maximises $b(\theta, h(\underline{\mathbf{Y}})) = \frac{1}{\theta^n}\mathbb{1}\{\max_{1\le i\le n} Y_i \le \theta\}$. We are thinking of this as a function of θ. When $\theta < \max_{1\le i\le n} Y_i$, this is 0; when $\theta \ge \max_{1\le i\le n} Y_i$, this is $\frac{1}{\theta^n}$, which is decreasing in θ. Plot it! Hence, the MLE is also

$$\hat{\theta} = \max_{1\le i\le n} Y_i.$$

Is $\hat{\theta}$ consistent for θ here? Since the regularity conditions do not hold, we cannot directly quote the general result about the MLE. However, we can check this from first principles. Recall that in order to show whether $\hat{\theta}$ is consistent, we only need to work out whether $\hat{\theta}$ converges to θ in probability. For any small $\varepsilon > 0$, we have

$$\begin{aligned}\mathbb{P}\big(|\hat{\theta} - \theta| < \varepsilon\big) &= \mathbb{P}\big(|\max_{1\le i\le n} Y_i - \theta| < \varepsilon\big) = \mathbb{P}\big(\theta - \varepsilon < \max_{1\le i\le n} Y_i \le \theta\big) = 1 - \mathbb{P}\big(\max_{1\le i\le n} Y_i \le \theta - \varepsilon\big) \\ &= 1 - \left(\frac{\theta - \varepsilon}{\theta}\right)^n = 1 - \left(1 - \frac{\varepsilon}{\theta}\right)^n \to 1 - 0 = 1,\end{aligned}$$

as $n \to \infty$. Thus, $\hat{\theta}$ is indeed consistent.

The asymptotic normality, however, does not hold. Notice that $\sqrt{n}(\hat{\theta} - \theta)$ can only take non-positive values as $\hat{\theta} = \max Y_i \le \theta$, hence it cannot be getting closer to a mean 0 normal distribution.

Related Problem set questions: PS5Q3, PS5Q4, PS6Q1(d)(e).

3 Using MLE to construct (approximate) confidence intervals

In this section, we use the asymptotic normality property of the MLE to derive an approximate pivotal, and proceed to construct an approximate/asymptotic confidence interval for the parameter of interest θ in a regular model. Before going into the details below, make sure you are familiar with Section 1 of Handout 2 (Workflow for constructing an exact confidence interval).

Under regularity conditions (plus some other), we have the asymptotic normality of the MLE:

$$\sqrt{n}\big(\hat{\theta} - \theta\big) \xrightarrow{d} N\big(0, I_Y^{-1}(\theta)\big).$$

Here, $I_Y(\theta)$ is the Fisher information based on one single observation. This result actually holds for θ in any dimension. When the parameter of interest θ is multivariate, the right hand side limiting distribution is a multivariate normal distribution, with covariance matrix $I_Y^{-1}(\theta)$. Our main focus, though, will be the univariate case.

Recall that a pivotal quantity is a function of the data and the parameter of interest, and its distribution does not depend on the parameter. The left hand side quantity above $\sqrt{n}\big(\hat{\theta} - \theta\big)$ is not yet an approximate pivotal, as its asymptotic distribution still depend on θ. To resolve this issue, we introduce the *observed Fisher information*:

$$i_Y\big(\hat{\theta}\big) = I_Y(\theta^*)\Big|_{\theta^*=\hat{\theta}}. \qquad \text{and} \qquad i_{\underline{\mathbf{Y}}}\big(\hat{\theta}\big) = I_{\underline{\mathbf{Y}}}(\theta^*)\Big|_{\theta^*=\hat{\theta}}.$$

We simply replace the parameter θ in the expression of the Fisher information by its esimator $\hat{\theta}$. Since the MLE is close to the parameter, we reasonably have

$$\sqrt{n}\big(\hat{\theta}-\theta\big) \approx N\big(0, i_Y^{-1}(\hat{\theta})\big).$$

Now we can see that $\sqrt{n}\big(\hat{\theta}-\theta\big)$ is an approximate pivotal quantity. We use the usual CI construction procedure from Section 1 of Handout 2 to proceed:

$$\mathbb{P}\bigg(-z_{1-\alpha/2} \le \sqrt{n \cdot i_Y\big(\hat{\theta}\big)}\big(\hat{\theta}-\theta\big) \le z_{1-\alpha/2}\bigg) \approx 1-\alpha.$$

Note that we still have $i_{\underline{\mathbf{Y}}}\big(\hat{\theta}\big) = n \cdot i_Y\big(\hat{\theta}\big)$. An approximate $(1-\alpha)$-level confidence interval for θ is

$$\bigg[\hat{\theta} - \frac{z_{1-\alpha/2}}{\sqrt{i_{\underline{\mathbf{Y}}}(\hat{\theta})}}, \hat{\theta} + \frac{z_{1-\alpha/2}}{\sqrt{i_{\underline{\mathbf{Y}}}(\hat{\theta})}}\bigg].$$

Related Problem set question: PS6Q2.

3.1 A worked example

Let $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$. We want to construct an approximate $(1-\alpha)$-level confidence interval for p.

- Likelihood: $L(p; \underline{\mathbf{Y}}) = p^{\sum Y_i}(1-p)^{n-\sum Y_i}$;
- Log-likelihood: $\ell(p; \underline{\mathbf{Y}}) = \sum Y_i \log(p) + (n - \sum Y_i)\log(1-p)$;
- Score function: $s(p; \underline{\mathbf{Y}}) = \frac{\sum Y_i}{p} - \frac{n-\sum Y_i}{1-p}$;
- MLE: $\hat{p} = \bar{Y}$;
- Fisher information (based on all sample): $I_{\underline{\mathbf{Y}}}(p) = -\mathbb{E}\big[\frac{\partial}{\partial p} s(p; \underline{\mathbf{Y}})\big] = \frac{n}{p(1-p)}$.
- Observed Fisher information (simply replace p by $\hat{p}$ in the above Fisher information expression): $i_{\underline{\mathbf{Y}}}\big(\hat{p}\big) = \frac{n}{\hat{p}(1-\hat{p})} = \frac{n}{\bar{Y}(1-\bar{Y})}$.
- An approximate $(1-\alpha)$-level confidence interval for p is

$$\bigg[\bar{Y} - z_{1-\alpha/2}\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}}, \bar{Y} + z_{1-\alpha/2}\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}}\bigg].$$

4 Wilks' Theorem

Wilks' theorem offers an asymptotic distribution of the log-likelihood ratio statistic under the null. Suppose we want to test $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$. The likelihood ratio statistic is:

$$r(\underline{\mathbf{Y}}) := \frac{\sup_{\theta\in\Theta_1} L(\theta; \underline{\mathbf{Y}})}{\sup_{\theta\in\Theta_0} L(\theta; \underline{\mathbf{Y}})}.$$

In other words, we need to derive the MLE $\hat{\theta}_0 \in \Theta_0$ in the null parameter space and the MLE $\hat{\theta}_1 \in \Theta_1$ in the alternative parameter space. In this course, we usually do not differentiate the alternative space Θ_1 from the full parameter space Θ. You can view them as the same thing. The statistic can be then written as $r(\underline{\mathbf{Y}}) = \frac{L(\hat{\theta}_1;\underline{\mathbf{Y}})}{L(\hat{\theta}_0;\underline{\mathbf{Y}})}$. Wilks' Theorem states that, under regularity (and some other) conditions, we have

$$2\log r(\underline{\mathbf{Y}}) \xrightarrow{d} \chi^2_q,$$

as $n \to \infty$, under the null, where $q = \dim(\Theta_1) - \dim(\Theta_0)$, the difference between the dimension of the alternative/full parameter space and the dimension of the null parameter space.

Recall that the likelihood ratio $r(\underline{\mathbf{Y}})$ is an indication of the departure from the null. We thus reject the null hypothesis when $r(\underline{\mathbf{Y}})$, or equivalently $2\log r(\underline{\mathbf{Y}})$, is big (i.e. one-sided test). Using Wilks' Theorem, we can construct an asymptotically $100\alpha\%$ significance level test: we reject the null hypothesis when $r(\underline{\mathbf{Y}}) > \chi^2_{q,1-\alpha}$.

4.1 Degree of freedom

We briefly discuss the degree of freedom $q = \dim(\Theta_1) - \dim(\Theta_0)$ by giving some examples:

- $Y_1, \ldots, Y_n \overset{\text{i.i.d}}{\sim} N(\mu, 1)$. We test $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$.
  $\dim(\Theta_0) = 0$: the null parameter space is a single point 0, i.e. μ fixed;
  $\dim(\Theta_1) = 1$: in the alternative, μ can be any real number, and the real line has dimension 1.
  $q = \dim(\Theta_1) - \dim(\Theta_0) = 1$.

- $Y_1, \ldots, Y_n \overset{\text{i.i.d}}{\sim} N(\mu, \sigma^2)$. We test $H_0 : \mu = 0, \sigma^2 = 1$ against $H_1 : \mu \neq 0, \sigma^2 \neq 1$.
  $\dim(\Theta_0) = 0$: the null parameter space is a single point $(0, 1)$, i.e. both μ and σ^2 are fixed;
  $\dim(\Theta_1) = 2$: in the alternative, μ can be any real number, and σ^2 can be any positive number;
  Θ_1 thus has dimension 2;
  $q = \dim(\Theta_1) - \dim(\Theta_0) = 2$.

- $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} N(\mu_1, 1)$, $Y_1, \ldots, Y_n \overset{\text{i.i.d}}{\sim} N(\mu_2, 1)$ We test $H_0 : \mu_1 = \mu_2$, against $H_1 : \mu_1 \neq \mu_2$.
  $\dim(\Theta_0) = 1$: in the null space, $\mu_1 = \mu_2$; they are equal but can be any real number.
  $\dim(\Theta_1) = 2$: in the alternative, both μ_1 and μ_2 can be any real number; thus dimension 2;
  $q = \dim(\Theta_1) - \dim(\Theta_0) = 1$.

Related Problem set question: PS6Q4, PS6Q5.

4.2 A worked example

We provide a solution for PS6Q5 below.
Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Exp}(\theta_1)$ and an independent random sample $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \text{Exp}(\theta_2)$. We want to test $H_0 : \theta_1 = \theta_2$ against $H_1 : \theta_1 \neq \theta_2$.
The likelihood ratio statistic is

$$r(\underline{\mathbf{X}}, \underline{\mathbf{Y}}) := \frac{\sup_{\theta_1 \in \mathbb{R}^+, \theta_2 \in \mathbb{R}^+} L(\theta_1, \theta_2; \underline{\mathbf{X}}, \underline{\mathbf{Y}})}{\sup_{\theta_1 = \theta_2 \in \mathbb{R}^+} L(\theta_1, \theta_2; \underline{\mathbf{X}}, \underline{\mathbf{Y}})},$$

where the likelihood is

$$L(\theta_1, \theta_2; \underline{\mathbf{X}}, \underline{\mathbf{Y}}) = \left(\theta_1^n e^{-\theta_1 \sum_{i=1}^n X_i}\right)\left(\theta_2^n e^{-\theta_2 \sum_{i=1}^n Y_i}\right).$$

Under the null constraint ($\theta_1 = \theta_2$), we would like to maximise $L(\theta, \theta; \underline{\mathbf{X}}, \underline{\mathbf{Y}})$ over $\theta \in \mathbb{R}$.

$$L(\theta, \theta; \underline{\mathbf{X}}, \underline{\mathbf{Y}}) = \theta^{2n} e^{-\theta\left(\sum X_i + \sum Y_i\right)}.$$

The MLE under the null is $\hat{\theta}_1 = \hat{\theta}_2 = \hat{\theta} = \frac{2n}{\sum X_i + \sum Y_i}$. The other way to think about this is that under H_0, all $2n$ observations come from the same distribution, so we can directly deduce the above MLE. The maximum likelihood under the null is

$$\sup_{\theta_1 = \theta_2 \in \mathbb{R}^+} L(\theta_1, \theta_2; \underline{\mathbf{X}}, \underline{\mathbf{Y}}) = L(\hat{\theta}, \hat{\theta}; \underline{\mathbf{X}}, \underline{\mathbf{Y}}) = \hat{\theta}^{2n} e^{-\hat{\theta}\left(\sum X_i + \sum Y_i\right)} = \left(\frac{2n}{\sum X_i + \sum Y_i}\right)^{2n} e^{-2n}.$$

Under the alternative, the MLE is easy to derive: $\hat{\theta}_1 = 1/\bar{X}$ and $\hat{\theta}_2 = 1/\bar{Y}$, with the maximum likelihood under the alternative being

$$\sup_{\theta_1 \in \mathbb{R}^+, \theta_2 \in \mathbb{R}^+} L(\theta_1, \theta_2; \underline{\mathbf{X}}, \underline{\mathbf{Y}}) = L(\hat{\theta}_1, \hat{\theta}_2; \underline{\mathbf{X}}, \underline{\mathbf{Y}}) = \hat{\theta}_1^n e^{-\hat{\theta}_1 \sum X_i} \hat{\theta}_2^n e^{-\hat{\theta}_2 \sum Y_i} = \left(\frac{1}{\bar{X}\bar{Y}}\right)^n e^{-2n}.$$

Thus the likelihood ratio statistic is

$$r(\underline{\mathbf{X}}, \underline{\mathbf{Y}}) = \frac{\left(\frac{1}{\bar{X}\bar{Y}}\right)^n e^{-2n}}{\left(\frac{2n}{\sum X_i + \sum Y_i}\right)^{2n} e^{-2n}} = \left(\frac{\left(\bar{X} + \bar{Y}\right)^2}{4\bar{X}\bar{Y}}\right)^n.$$

The asymptotic distribution of $2 \log r(\underline{\mathbf{X}}, \underline{\mathbf{Y}}) = 2n \log\left(\frac{\left(\bar{X}+\bar{Y}\right)^2}{4\bar{X}\bar{Y}}\right)$ is χ^2_q, where $q = \dim(\Theta_1) - \dim(\Theta_0)$. The alternative/full parameter space has dimension 2, as both θ_1 and θ_2 can take any positive real number. The null parameter space has dimension 1, as the equal rate can be any positive real number: $\theta_1 = \theta_2 \in \mathbb{R}^+$. Thus $q = 1$.

For an $100\alpha\%$ level test, we reject the null hypothesis when

$$2 \log r(\underline{\mathbf{X}}, \underline{\mathbf{Y}}) = 2n \log\left(\frac{\left(\bar{X} + \bar{Y}\right)^2}{4\bar{X}\bar{Y}}\right) > \chi^2_{1,1-\alpha},$$

where $\chi^2_{1,1-\alpha}$ is the $(1-\alpha)$-quantile, or equivalently the top α-quantile, of a χ^2_1 distribution.

5 Extra exercises

1. Let $Y_1, \ldots, Y_n$ be i.i.d. with density $f(y; \lambda, \theta) = \lambda e^{-\lambda(y-\theta)}$ for $y > \theta$, where $(\lambda, \theta)^\top$ is our parameter of interest in $\mathbb{R}^+ \times \mathbb{R}$.
(a) Find a sufficient statistic for $(\lambda, \theta)^\top$. (Hint: a 2d statistic)
(b) First find a statistic $\hat{\theta}$ that satisfies $L(\lambda, \hat{\theta}; \underline{\mathbf{Y}}) \geq L(\lambda, \theta; \underline{\mathbf{Y}})$ for all $\lambda > 0$. Then derive a statistic $\hat{\lambda}$ such that $L(\hat{\lambda}, \hat{\theta}; \underline{\mathbf{Y}}) \geq L(\lambda, \hat{\theta}; \underline{\mathbf{Y}})$. Briefly explain why $(\hat{\lambda}, \hat{\theta})^\top$ is an MLE for $(\lambda, \theta)^\top$.
(c) Show that $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{p} 0$. Does this imply the consistency of $\hat{\theta}$?
(d) Prove that $\hat{\lambda}$ is consistent.
(e)* Does $\hat{\lambda}$ satisfy the asymptotic normality? Justify your answer.

2. Let $Y_1, \ldots, Y_n$ be i.i.d exponential random variables with rate parameter $\lambda > 0$.
(a) Find the MLE $\hat{\lambda}$. Is $\hat{\lambda}$ unbiased?

(b) Show that $\hat{\lambda}$ is consistent.
(c) Calculate the Fisher information $I_{\underline{\mathbf{Y}}}(\lambda)$ (based on n observations).
(d) Use the asymptotic normality of $\hat{\lambda}$ to construct an approx. 95%-level confidence interval for λ.

3. Let $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$. The variance σ^2 is unknown and our parameter of interest is $(\mu, \sigma^2)^\top$. We want to test $H_0 : \mu = \mu_0$ (a given value) against $H_1 : \mu \neq \mu_0$.
(a) Work out the MLE under the null and the alternative respectively. Note again that our unknown parameter is $(\mu, \sigma^2)^\top$.
(b) Use Wilks' Theorem to construct an asymptotically $100\alpha\%$ significance level test. What is the difference in dimension between full parameter space and the null space?
(c)* Can you derive an *exact* $100\alpha\%$ significance level likelihood ratio test? (Hint: a t-distribution)