1 A review of results in statistical inference

This section, we review some results that you came across in STAT611 (or equivalent). We will review the Cramer-Rao bound and some properties of the likelihood. In later sections, we will use the likelihood as a means of parameter estimation (ie. the maximimum likelihood estimator which you would have done in previous courses) and heuristically argue why the Fisher information (which gives the Cramer-Rao bound) is extremely important.

1.1 The likelihood function

Let $\{X_i\}$ be iid random variables with probability function (or probability density function) $f(x;\theta)$, where f is known but the parameter θ is unknown.

The likelihood function is defined as

$$L(\underline{X};\theta) = \prod_{i=1}^{T} f(X_i;\theta) \tag{1}$$

and the log-likelihood is

$$\log L(\underline{X};\theta) = \mathcal{L}(\underline{X};\theta) = \sum_{i=1}^{T} \log f(X_i;\theta). \tag{2}$$

Example 1.1 *(i) Suppose that $\{X_t\}$ are iid normal random variables with mean μ and variance σ^2 the log likelihood is*

$$\mathcal{L}_T(\underline{X};\mu,\sigma^2) \propto T\sigma^2 + \sum_{t=1}^{T} \frac{(X_t-\mu)^2}{\sigma^2}$$

(ii) Suppose that $\{X_t\}$ are iid binomial random variables $X_t \sim Bin(n,\pi)$. Then the log likelihood is

$$\mathcal{L}_T(\underline{X};\pi) \propto \sum_{t=1}^{T} \log \binom{n}{X_t} + \sum_{t=1}^{T} \left(X_t \log\big(\frac{\pi}{1-\pi}\big) + n\log(1-\pi) \right).$$

(iii) Suppose that $\{X_t\}$ are independent binomial random variables such that $X_t \sim Bin(n_t,\pi_t)$, where the regressors z_t influence the mean of X_t, such that $\pi_t = g(\beta' x_t)$. Then the log likelihood is

$$\mathcal{L}_T(\underline{X};\pi) \propto \sum_{t=1}^{T} \log \binom{n_t}{X_t} + \sum_{t=1}^{T} \left(X_t \log\big(\frac{g(\beta' x_t)}{1-g(\beta' x_t)}\big) + n_t \log(1-g(\beta' x_t)) \right).$$

(iv) *Suppose that $\{X_t\}$ are independent exponential random variables which have the density $\theta^{-1}\exp(-x/\theta)$. The log-likelihood is*

$$\mathcal{L}_T(\underline{X};\theta) = \sum_{t=1}^{T}\Big(-\alpha\log\theta + \frac{Y_t}{\theta}\Big).$$

(v) *A generalisation of the exponential distribution which gives more freedom in terms of shape of the distribution is the Weibull. Suppose that $\{X_t\}$ are independent Weibull random variables which have the density $\frac{\alpha y^{\alpha-1}}{\theta^\alpha}\exp(-(y/\theta)^\alpha)$ where $\theta,\alpha>0$ (in the case that $\alpha=0$ we have the regular exponential) and y is defined over the positive real line. The log-likelihood is*

$$\mathcal{L}_T(\underline{X};\alpha,\theta) = \sum_{t=1}^{T}\Big(\log\alpha + (\alpha-1)\log Y_t - \alpha\log\theta - (\frac{Y_t}{\theta})^\alpha\Big).$$

In the case, that α is known, but θ is unknown the likelihood is proportional to

$$\mathcal{L}_T(\underline{X};\theta) = \sum_{t=1}^{T}\Big(-\alpha\log\theta - (\frac{Y_t}{\theta})^\alpha.\Big).$$

1.2 Bounds for the variance of an unbiased estimator

We require the following assumptions, often called the regularity assumptions. We state the assumptions and result scalar θ, but they can easily be extended to the case that θ is a vector.

Assumption 1.1 (Regularity Conditions 1) *Let us suppose that $L_T(\cdot)$ is the likelihood with true parameter θ, and*

(i) $\int \frac{\partial\log L_T(\underline{x};\theta)}{\partial\theta} L_T(\underline{x};\theta)d\underline{x} = 0$ *(for iid this is equivalent to* $\int \frac{\partial\log f(x;\theta)}{\partial\theta} f(x;\theta)dx = 0$*).*

(ii) $\frac{\partial}{\partial\theta}\int L_T(\underline{x};\theta)d\underline{x} = \int \frac{\partial L_T(\underline{x};\theta)}{\partial\theta}d\underline{x} = 0.$

(iii) $\frac{\partial}{\partial\theta}\int g(\underline{x})L_T(\underline{x};\theta)d\underline{x} = \int g(\underline{x})\frac{\partial L_T(\underline{x};\theta)}{\partial\theta}d\underline{x}$*, where g is any function which is not a function of θ (for example the estimator of θ).*

(iv) $\mathbb{E}\big(\frac{\partial\log L_T(\underline{X};\theta)}{\partial\theta}\big)^2 > 0.$

Theorem 1.1 (The Cramer-Rao bound) *Let $\tilde{\theta}(\underline{X})$ be an unbiased estimator of $\tilde{\theta}$. Suppose the likelihood $L_T(\underline{X};\theta)$ satisfies the regularity conditions (given in Assumption 1.1) and $\tilde{\theta}(\underline{X})$ is an unbiased estimator of θ, then we have*

$$var(\tilde{\theta}(\underline{X})) \geq \Big(\mathbb{E}\big(\frac{\partial\log L_T(\underline{X};\theta)}{\partial\theta}\big)^2\Big)^{-1} = \Big(\mathbb{E}\big(-\frac{\partial^2\log L_T(\underline{X};\theta)}{\partial\theta^2}\big)\Big)^{-1},$$

PROOF. Recall that $\tilde{\theta}(X)$ is an unbiased estimator of θ therefore

$$\int \tilde{\theta}(\underline{x}) L_T(\underline{x};\theta) d\underline{x} = \theta.$$

Differentiating both sides wrt to θ gives

$$\int \tilde{\theta}(\underline{x}) \frac{\partial L_T(\underline{x};\theta)}{\partial \theta} d\underline{x} = 1.$$

Since $\int \frac{\partial L_T(\underline{x};\theta)}{\partial \theta} d\underline{x} = 0$ we have

$$\int \left\{\theta - \tilde{\theta}(\underline{x})\right\} \frac{\partial L_T(\underline{x};\theta)}{\partial \theta} d\underline{x} = 1.$$

Multiplying and dividing by $L_T(\underline{x};\theta)$ gives

$$\int \left\{\tilde{\theta}(\underline{x}) - \theta\right\} \frac{1}{L_T(\underline{x};\theta)} \frac{\partial L_T(\underline{x};\theta)}{\partial \theta} L_T(\underline{x};\theta) dx = 1. \tag{3}$$

Hence (since $L_T(\underline{x};\theta)$ is the distribution of $\underline{X}$) we have

$$\mathbb{E}\left(\left\{\tilde{\theta}(X) - \theta\right\} \frac{1}{L_T(\underline{X};\theta)} \frac{\partial \log L_T(\underline{X};\theta)}{\partial \theta} \right) = 1.$$

Recalling that the Cauchy-Schwartz inequality is $\mathbb{E}(UV) \leq \mathbb{E}(U^2)^{1/2}\mathbb{E}(V^2)^{1/2}$ (where equality only arises if $U = aV + b$ (where a and b are constants)) and applying it to the above we have

$$\mathrm{var}(\tilde{\theta}(X))\mathbb{E}\Big(\left(\frac{\partial \log L_T(\underline{X};\theta)}{\partial \theta} \right)^2 \Big) \geq 1. \tag{4}$$

Thus giving us the Cramer-Rao inequality. Finally we need to prove that $\mathbb{E}\Big(\left(\frac{\partial \log L_T(\underline{X};\theta)}{\partial \theta} \right)^2 \Big) = \mathbb{E}\Big(- \frac{\partial^2 \log L_T(\underline{X};\theta)}{\partial \theta^2} \Big)$. To prove this result we use the fact that L_T is a density to obtain

$$\int L_T(\underline{x};\theta) d\underline{x} = 1.$$

Now by differentiating the above with respect to θ gives

$$\frac{\partial}{\partial \theta} \int L_T(\underline{x};\theta) d\underline{x} = 0.$$

By using Assumption 1.1(ii) we have

$$\int \frac{\partial L_T(\underline{x};\theta)}{\partial \theta} d\underline{x} = 0 \Rightarrow \quad \int \frac{\partial \log L_T(\underline{x};\theta)}{\partial \theta} L_T(\underline{x};\theta) d\underline{x} = 0$$

Differentiating again with respect to θ and taking the derivative inside gives

$$\begin{aligned}
& \int \frac{\partial^2 \log L_T(\underline{x};\theta)}{\partial \theta^2} L_T(\underline{x};\theta) d\underline{x} + \int \frac{\partial \log L_T(\underline{x};\theta)}{\partial \theta} \frac{\partial L_T(\underline{x};\theta)}{\partial \theta} d\underline{x} = 0 \\
\Rightarrow \quad & \int \frac{\partial^2 \log L_T(\underline{x};\theta)}{\partial \theta^2} L_T(\underline{x};\theta) d\underline{x} + \int \frac{\partial \log L_T(\underline{x};\theta)}{\partial \theta} \frac{1}{L_T(\underline{x};\theta)} \frac{\partial L_T(\underline{x};\theta)}{\partial \theta} L_T(\underline{x};\theta) d\underline{x} = 0 \\
\Rightarrow \quad & \int \frac{\partial^2 \log L_T(\underline{x};\theta)}{\partial \theta^2} L_T(\underline{x};\theta) d\underline{x} + \int \left(\frac{\partial \log L_T(\underline{x};\theta)}{\partial \theta} \right)^2 L_T(\underline{x};\theta) d\underline{x} = 0
\end{aligned}$$

Thus

$$-\mathbb{E}\Big(\frac{\partial^2 \log L_T(\underline{X};\theta)}{\partial\theta^2}\Big) = \mathbb{E}\Big(\frac{\partial \log L_T(\underline{X};\theta)}{\partial\theta}\frac{\partial L_T(\underline{X};\theta)}{\partial\theta}\Big).$$

Which gives us the required result. □

Corollary 1.1 (Estimators which attain the C-R bound) *Suppose Assumption 1.1 is satisfied. Then the estimator $\tilde{\theta}(\underline{X})$ attains the C-R bound only if it can be written as*

$$\hat{\theta}(\underline{X}) = a(\theta) + b(\theta)\frac{\partial \log L_T(\underline{X};\theta)}{\partial\theta}$$

for some functions $a(\cdot)$ and $b(\cdot)$.

PROOF. The proof is clear and follows from when the Cauchy-Schwartz inequality is an actual equality in the derivation of the C-R bound. □

We mention that there exists some well known distributions which do not satisfy Assumption 1.1. These are non-regular distributions. A classical example of a distribution which violates this assumption is the uniform distribution $f(x;\theta) = 1/\theta$, for $x \in [0,\theta]$ and zero elsewhere. Other examples, include distributions where the support of the distribution is a function of the parameter. The Cramer-Rao lower bound does hold or even exist for such distributions.

Example 1.2 (The classical example of the uniform) *Let us consider the example if the uniform distribution, which has the density $f(x;\theta) = \theta^{-1}\exp(-x/\theta)$. Given the iid uniform random variables $\{X_t\}$ the likelihood (it is easier to study the likelihood rather than the log-likelihood) is*

$$L_T(\underline{X}_T;\theta) = \frac{1}{\theta^T}\prod_{t=1}^{T} I_{[0,\theta]}(X_t).$$

Since the support of density involves the unknown parameter, then the derivative of $\log L_T(\underline{X}_T;\theta)$ is not well defined (what is the derivative of $\log I_{[0,\theta]}(X_t) = \log I_{[X_t,\infty)}(\theta)$ with respect to θ? - observe that at $\log 0$ is not well defined and the derivative at X_t is not well defined) and Assumption 1.1(ii) is not satisfied. This is a classical example of a density which does not satisfy the regularity conditions. This means that the inverse of the Fisher information does not give a lower bound for the variance estimator. And below we will show why.

In fact, using $L_T(\underline{X}_T;\theta)$, the maximum likelihood estimator of θ is $\hat{\theta}_T = \max_{1\leq t\leq T} X_t$ (you can see this by making a plot of $L_T(\underline{X}_T;\theta)$ against θ). It it well known that the distribution of $\max_{1\leq t\leq T} X_t$ is

$$P(\max_{1\leq t\leq T} X_t \leq x) \;=\; P(X_1 \leq x,\ldots,X_T \leq x) = \prod_{t=1}^{T} P(X_t \leq x) = \big(\frac{x}{\theta}\big)^T,$$

and the density of $\max_{1\leq t\leq T} X_t$ *is* $f_{\hat{\theta}_T}(x) = Tx^{T-1}/\theta^T$.

Exercise: Find the variance of $\hat{\theta}_T defined above.$

Often we want to estimate a function of θ, $\tau(\theta)$. The following corollary is a small generalisation of the Cramer-Rao bound.

Corollary 1.2 *Suppose the regularity conditions (Assumption 1.1) are satisfied and* $T(\underline{X})$ *is an unbiased estimator of* $\tau(\theta)$. *Then we have*

$$var(T(\underline{X})) \geq \frac{(\tau'(\theta))^2}{\mathbb{E}\big(\frac{\partial \log L_T(\underline{X};\theta)}{\partial\theta}\big)}.$$

We now define the notion of sufficiency, which gives us the ingredients for constructing a good estimator (see also Sections 4.2.2 7.1.1 and 7.1.3 Davison (2002)).

Definition 1.1 (Sufficiency and the factorisation theorem) *Suppose that* $\underline{X} = (X_1, \ldots, X_T)$ *is a random vector. The statistic* $s(\underline{X})$ *is called a sufficient statistic of the parameter* θ, *if the conditional distribution of* $\underline{X}$ *given* $s(\underline{X})$ *is not a function of* θ.

Normally it is extremely hard to obtain the sufficient statistic from its definition. However, the factorisation theorem gives us a way of obtaining the sufficient statistic.

The Factorisation Theorem *Suppose that the likelihood function can be partitioned as follows,* $L_T(\underline{X};\theta) = h(\underline{X})g(s(\underline{X});\theta)$, *where* $h(\underline{X})$ *is not a function of* θ, *then* $s(\underline{X})$ *is a sufficient statistic of* θ.

We see that a sufficient statistic contains all the ingredients about the parameter θ.

Theorem 1.2 (Rao-Blackwell Theorem) *Suppose* $s(\underline{X})$ *is a sufficient statistic and* $\theta(\underline{X})$ *is an unbiased estimator of* θ *then if we define the new unbiased estimator* $\mathbb{E}(\theta(\underline{X})|S(\underline{X}))$, *then*

$$var(\mathbb{E}(\tilde{\theta}(\underline{X})|S(\underline{X}))) \leq var(\tilde{\theta}(\underline{X})).$$

The Rao-Blackwell theorem tells us that estimators with the smallest variance must be a function of the sufficient statistic. Of course this begs the question is there a unique estimator with the minumum variance. For this we require completeness of the sufficient statistic. Uniqueness immediately follows from completeness.

Definition 1.2 (Completeness) *Let* $s(\underline{X})$ *be a sufficient statistic for* θ. *Suppose* $Z(\cdot)$ *is a function of* $s(\underline{X})$ *such that* $\mathbb{E}(Z(s(\underline{X}))) = 0$. $s(\underline{X})$ *is a complete sufficient statistic if and only if* $\mathbb{E}(Z(s(\underline{X}))) = 0$ *implies* $Z(t) = 0$ *for all* t.

Theorem 1.3 (Lehmann-Scheffe Theorem) *Suppose that $S(\underline{X})$ is a complete sufficient statistic and $\tilde{\theta}(S(\underline{X}))$ is an unbiased estimator estimator of θ then $\tilde{\theta}(S(\underline{X}))$ is the unique minimum variance unbiased estimator of θ.*

The theorems above are theoretical, in the sense that, under certain conditions they give a lower bound for the variance of a plausible estimator and practical in the sense that they tell us that the best estimator should be a function of sufficient statistic. The natural question to ask, is how to construct such estimators.

One of the most popular estimators in statistics are maximum likelihood estimators (mle). That is the mle of θ is $\hat{\theta}_T = \arg\max_{\theta\in\Theta} \mathcal{L}_T(\theta)$, where Θ is the parameter space which contains all values of θ with $\int f(x;\theta)dx = 1$. There are two reasons that they are so widely used (i) it can be shown for a wide range of probability distributions - including (under certain conditions) the exponential family of distributions, defined below, that the mle is a function of the sufficient statistic, hence the mle is often the minimum variance unbiased estimator (ii) asymptotically (at least) the mle under certain conditions attains the C-R bound.

Of course one can construct examples, where the regularity conditions are not satisfied and the mle is not the optimal estimator (examples include estimation of the range in the uniform distribution, where an estimator can be constructed which has a small variance than the mle). But for a mass majority of distributions the mle is optimal. It is also worth mentioning that there can exist biased estimators which have a smaller mean squared error than the MLE (this intriguing notion it called super-efficiency, which is beyond this course - see Stoica and Ottesten (1996) for a review).

1.3 Additional Notes

We will use various distributions in this course, it would be useful if you complied a list of these distributions and become familiar with them.

Example 1.3 (Useful transformations) *Question:*

The distribution function of the random variable X_t is $F_t(x) = 1 - \exp(-\lambda_t x)$.

(i) Give a transformation of X_t, such that the transformed variable is uniformly distributed on the interval $[0, 1]$.

(ii) Suppose that I observe the independent (but not necessarily identically distributed) random variables $\{X_t\}$, and I want to to check whether they have the distribution function $F_t(x) = 1 - \exp(-\lambda_t x)$. Using (i), suggest a method for checking this?

Answer:

(i) It is well known that if the random variable X_t has the distribution function $F_t(x)$, then the transformed random variable $Y_t = F_t(X_t)$ is uniformly distributed on the interval $[0, 1]$. To see this, note that the distribution of Y_t can be evaluated as

$$P(Y_t \leq y) = P(F_t(X_t) \leq y) = P(X_t \leq F_t^{-1}(y)) = F_t(F_t^{-1}(y)) = y, \qquad y \in [0, 1].$$

Thus to answer the question, we let $Y_t = 1 - \exp(-\lambda_t X_t)$, which as a uniform distribution.

(ii) If we want to check whether X_t follows the distribution $F_t(x) = 1 - \exp(-\lambda_t x)$, we can make the transformation $Y_t = 1 - \exp(-\lambda_t X_t)$, and use, for example, the Kolmogorov-Smirnov test to check whether $\{Y_t\}$ follows a uniform distribution.