Lecture 7

13. Point Estimation

Let Θ be a parameter with parameter space Ω and $g : \Omega \to G$ a function of Θ. The objective of point estimation is to find a guess for the "true" value θ of Θ used to generate the data. Alternatively, to find the true value of $g(\Theta)$, some function of the parameter. For instance, if Θ is multidimensional we might be interested in just the first component of Θ, say.

Definition 21. Let $G' \supset G$. A measurable function $\phi : \mathcal{X} \to G'$ is called an *estimator* of $g(\Theta)$. It is called an *unbiased estimator* if $E_\theta[\phi(X)] = g(\theta)$ for all $\theta \in \Omega$. The *bias* is defined as

$$b_\phi(\theta) = E_\theta[\phi(X)] - g(\theta).$$

13.1. **Moment matching – An engineering approach.** The first approach we will consider is called *moment matching* or *method of moments*. This is sort of an engineering approach of fiddeling with the parameters until the sample moments matches the theoretical moments.

Let $X_1, \dots, X_n$ be an iid sample from $f_{X|\Theta}(x \mid \theta)$, let $\mu_k(\theta) = E_\theta[X_i^k]$ be the k:th moment and

$$m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

be the k:th sample moment. Find θ such that $\mu_i(\theta) = m_i$ for $i = 1, \dots, k$. Here one has to decide how many moments to fit. As a rule of thumb, the number of moments to fit should equal the dimension of Θ. Then you have as many equations as you have unknown variables.

Example 23. Suppose $X_1, \dots, X_n$ are iid $N(\mu, \sigma^2)$. Then $\mu_1 = \mu$ and $\mu_2 = \sigma^2 + \mu^2$ and to match the moments we need to solve

$$\mu = \bar{X}, \qquad \sigma^2 + \mu^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2,$$

which gives

$$\mu = \bar{X}, \qquad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

In this case it happens to result in a sensible estimator.

13.2. **Maximum likelihood estimation.** A very popular method to find estimators is the maximum likelihood method.

Definition 22. Let $f_{X|\Theta}(x \mid \theta)$ be the conditional density of X given $\Theta = \theta$. If x is observed the function $\theta \mapsto f_{X|\Theta}(x \mid \theta)$ is called the *likelihood function*. Any random quantity $\hat{\Theta}$ such that

$$\max_{\theta \in \Omega} f_{X|\Theta}(X \mid \theta) = f_{X|\Theta}(X \mid \hat{\Theta})$$

is called a *maximum likelihood estimator (MLE)* of Θ.

When interested in estimating a function $g(\Theta)$ we have an invariance property; the MLE of $g(\Theta)$ is equal to $g(\hat{\Theta})$. But wait! one has to be careful if g is not one-to-one.

Let $\Psi = g(\Theta)$ be the new parameter. If g is one-to-one we can write

$$f_{X|\Psi}(x \mid \psi) = f_{X|\Theta}(x \mid g^{-1}(\psi))$$

for the likelihood of Ψ. If $\hat{\psi}$ maximizes the left-hand-side we see that $g^{-1}(\hat{\psi}) = \hat{\theta}$ and hence $g(\hat{\theta}) = \hat{\psi}$.

If g is not one-to-one we could introduce the *induced likelihood of* ψ by

$$L^*(\psi) = \sup_{\{\theta: g(\theta)=\psi\}} f_{X|\Theta}(x \mid \theta)$$

and call the maximizer of L^* the MLE of Ψ. Then we have the following result

Theorem 17. *Let $g : \Omega \to G$ be a measurable function. If $\hat{\Theta}$ is an MLE of Θ, then $g(\hat{\Theta})$ is an MLE of $g(\Theta)$.*

Proof. Let $\hat{\psi}$ be the maximizer of L^*. We need to show that $L^*(\hat{\psi}) = L^*(g(\hat{\theta}))$, where $L(\theta) = f_{X|\Theta}(x \mid \theta)$. Note that

$$\sup_{\psi} L^*(\psi) = \sup_{\psi} \sup_{\{\theta: g(\theta)=\psi\}} L(\theta) = \sup_{\theta} L(\theta) = L(\hat{\theta}).$$

Then the claim follows since

$$L(\hat{\theta}) = \sup_{\{\theta: g(\theta)=g(\hat{\theta})\}} L(\theta) = L^*(g(\hat{\theta})).$$

□

13.3. **Bayesian decision theory and estimation.** Within the Bayesian methodology one obtains the posterior distribution of Θ given $X = x$. Thus we get an entire distribution, not only a particular value. The most common choices for deciding on a point estimate in the Bayesian context is using either

- MAP - the maximum of the posterior distribution, or
- the mean of the posterior distribution, or
- the median of the posterior distribution.

These choices seem intuitive if the posterior is unimodal and reasonably concentrated (by this I mean that you can imagine the posterior being a normal distribution, or deviate slightly from a normal).

The formal approach to Bayesian point estimation is through the language of decision theory. The problem of estimating $g(\Theta)$ can be viewed as a decision problem where the action space is G'. The decision rule is to take the action $\phi(X)$. The loss function is usually increasing as a function of the distance between $g(\Theta)$ and $\phi(X)$. The most common is the square loss with $L(\theta, a) = (g(\theta) - a)^2$. In Bayesian decision theory we look for a formal Bayes rule. For a quadratic loss function one should use the posterior mean.

Proposition 1. *Let $g : \Omega \to G$ and $\aleph = G$. Suppose the loss function is $L(\theta, a) = (g(\theta) - a)^2$. If the posterior variance is finite, then a formal Bayes rule is $E[g(\Theta) \mid X = x]$.*

Proof. For any decision rule δ,

$$\begin{aligned} r(\delta \mid x) &= \int_\Omega (g(\theta) - \delta(x))^2 \mu_{\Theta|X}(d\theta \mid x) \\ &= E[g(\Theta)^2 \mid X = x] - 2\delta(x)E[g(\Theta) \mid X = x] + \delta(x)^2. \end{aligned}$$

This is minimized by taking $\delta(x) = E[g(\Theta) \mid X = x]$. □

Another loss function that increases as a function of the distance is $L(\theta, a) = |\theta - a|$. This loss function suggests using the median of the posterior distribution. More generally, we have the following.

Theorem 18. *Suppose Θ has finite posterior mean. For the loss function*

$$L(\theta, a) = c(a - \theta)I_{\{a \geq \theta\}} + (1 - c)(\theta - a)I_{\{a < \theta\}},$$

a formal Bayes rule is the $1 - c$ quantile of the posterior distribution of Θ.

Proof. Suppose a' is the $1 - c$ quantile of $\mu_{\Theta|X}(\cdot \mid x)$. Then

$$\mu_{\Theta|X}((-\infty, a'] \mid x) \geq 1 - c, \quad \mu_{\Theta|X}([a', \infty) \mid x) \geq c.$$

If $a > a'$ then

$$\begin{aligned} L(\theta, a) - L(\theta, a') &= \begin{cases} c(a - a'), & a' \geq \theta, \\ c(a - a') - (\theta - a'), & a \geq \theta > a', \\ (1 - c)(a' - a), & \theta > a. \end{cases} \\ &= c(a - a') + \begin{cases} 0, & a' \geq \theta, \\ a' - \theta, & a \geq \theta > a', \\ a' - a, & \theta > a. \end{cases} \end{aligned}$$

Hence the difference in posterior risks is

$$\begin{aligned} r(a \mid x) - r(a' \mid x) &= c(a - a') + \underbrace{\int_{(a', a]} (a' - \theta)\mu_{\Theta|X}(d\theta \mid x)}_{\geq 0} + (a' - a)\mu_{\Theta|X}(a, \infty) \mid x) \\ &\geq c(a - a') + (a' - a)\mu_{\Theta|X}((a', \infty) \mid x) \\ &= (a - a')(c - \mu_{\Theta|X}((a', \infty) \mid x)). \end{aligned}$$

Since $\mu_{\Theta|X}((a', \infty) \mid x) \leq c$ we have $r(a \mid x) \geq r(a' \mid x)$. A similar computation with $a < a'$ gives also $r(a \mid x) \geq r(a' \mid x)$. Hence a' provides the minimum posterior risk. □

If we choose $c = 1/2$ then we get the median as a formal Bayes rule.

14. Point estimation and classical decision theory

The problem of estimating $g(\Theta)$ where $g : \Omega \to G$ can be viewed as a decision problem where the action space is $G' \supset G$. The decision rule is to take the action $\phi(X)$ where $\phi : \mathcal{X} \to G'$ is the *point estimator*. The loss function is usually increasing as a function of the distance between $g(\Theta)$ and $\phi(X)$. The most common is the square loss with $L(\theta, a) = (g(\theta) - a)^2$ which gives the risk function

$$R(\theta, \phi) = E_\theta[(g(\theta) - \phi(X))^2] = b_\phi^2(\theta) + \mathrm{Var}_\theta(X),$$

where $b_\phi(\theta)$ is the bias $E_\theta[\phi(X)] - g(\theta)$. It is convenient to use *unbiased* estimators ($b_\phi(\theta) = 0$ for all θ) as they do "on average" a good job of estimating the unknown parameter.

Here is a natural optimality criteria for unbiased estimators.

Definition 23. An unbiased estimator ϕ is called uniformly minimum variance unbiased estimator (UMVUE) if ϕ has finite variance and for every unbiased estimator ψ, $\mathrm{Var}_\theta(\phi(X)) \leq \mathrm{Var}_\theta(\psi(X))$.

With this definition in mind we would like to check if a suggested estimator has as low variance as possible. This will be easier if we require some regularity on the distributions at hand. These conditions are satisfied for most examples we encounter in practise.

Definition 24. Suppose Θ is k-dimensional and for each θ, P_θ has density $f_{X|\Theta}(x \mid \theta)$ with respect to ν. Suppose

(i) the derivative $\frac{\partial}{\partial\theta_i} f_{X|\Theta}(x \mid \theta)$ exists for all θ, each i, and every x in a set B with $\nu(B^c) = 0$,
(ii) $\int_{\mathcal{X}} f_{X|\Theta}(x \mid \theta)\nu(dx)$ can be differentiated under the integral sign with respect to each coordinate of θ, and
(iii) the set $C = \{x : f_{X|\Theta}(x \mid \theta) > 0\}$ does not depend on θ.

Then $f_{X|\Theta}$ is said to satisfy the Fisher information (FI) regularity conditions.

Definition 25. Suppose $f_{X|\Theta}$ satisfies the FI regularity conditions. The random function $U(X) = (U_1(X), \ldots, U_k(X))$ given by

$$U_i(X) = \frac{\partial}{\partial\theta_i} \log f_{X|\Theta}(X \mid \theta)$$

is called the *score function*. The $k \times k$-matrix $\mathcal{I}_X(\theta)$ with entries

$$(\mathcal{I}_X(\theta))_{ij} = \mathrm{cov}_\theta(U_i(X), U_j(X))$$

is called the *Fisher information matrix about* Θ *based on* X.

If T is a statistic the *conditional score function* is given by

$$U_i(X \mid t) = \frac{\partial}{\partial\theta_i} \log f_{X|T,\Theta}(X \mid t, \theta)$$

and the *conditional Fisher information matrix* $\mathcal{I}_{X|T}(\theta \mid t)$ is given by

$$(\mathcal{I}_{X|T}(\theta \mid t))_{ij} = \mathrm{cov}_\theta(U_i(X \mid t), U_j(X \mid t)).$$

Example 24. Suppose σ^2 is known and $X \sim N(\theta, \sigma^2)$ given $\Theta = \theta$. Then the FI regularity conditions are satisfied and

$$\begin{aligned} f_{X|\Theta}(x \mid \theta) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\theta)^2}{2\sigma^2}}, \\ U(X) &= \frac{X-\theta}{\sigma^2}, \\ \mathcal{I}_X(\theta) &= \frac{1}{\sigma^2}. \end{aligned}$$

Hence, if the variance is small there is a lot of information about Θ. Similarly, if $X = (X_1, \ldots, X_n)$ are conditionally IID $N(\theta, \sigma^2)$ given $\Theta = \theta$. Then

$$\begin{aligned} f_{X|\Theta}(x \mid \theta) &= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{\sum_{i=1}^n (x_i-\theta)^2}{2\sigma^2}}, \\ U(X) &= \frac{\sum_{i=1}^n (X_i - \theta)}{\sigma^2}, \\ \mathcal{I}_X(\theta) &= \frac{n}{\sigma^2}. \end{aligned}$$

Hence, more data also gives more information.

For any estimator (biased or unbiased) the next theorem give a lower bound on the variance when the FI regularity conditions hold.

Theorem 19 (Cramér-Rao lower bound)**.** *Suppose the FI regularity conditions hold and let $\mathcal{I}_X(\theta)$ be the Fisher information. Suppose that $\mathcal{I}_X(\theta) > 0$ for all θ. Let $\phi(X)$ be a one-dimensional statistic with $E_\theta[|\phi(X)|] < \infty$ for all θ. Suppose also that $\int \phi(x) f_{X|\Theta}(x \mid \theta)\nu(dx)$ can be differentiated under the integrals sign with respect to θ. Then*

$$\mathrm{Var}_\theta(\phi(X)) \geq \frac{(\partial_\theta E_\theta[\phi(X)])^2}{\mathcal{I}_X\theta)}.$$

Proof. Let B be the set with $\nu(B) = 0$ such that for all θ, $\partial_{\theta_i} f_{X|\Theta}(x \mid \theta)$ exists for $x \notin B$. Let $C = \{x : f_{X|\Theta}(x \mid \theta) > 0\}$. Put $D = C \cap B^c$ so that for all θ, $P_\theta(D) = 1$, and $\int_D f_{X|\Theta}(x \mid \theta)\nu(dx) = 1$. Take the derivative w.r.t. θ gives

$$\begin{aligned} 0 &= \int_D \partial_\theta f_{X|\Theta}(x \mid \theta)\nu(dx) \\ &= \int_D \frac{\partial_\theta f_{X|\Theta}(x \mid \theta)}{f_{X|\Theta}(x \mid \theta)} f_{X|\Theta}(x \mid \theta)\nu(dx) \\ &= E_\theta[\partial_\theta \log f_{X|\Theta}(x \mid \theta)]. \end{aligned}$$

Similarly by differentiating $E_\theta[\phi(X)]$ we get

$$\begin{aligned} \partial_\theta E_\theta[\phi(X)] &= \int \phi(x)\partial_\theta f_{X|\Theta}(x \mid \theta)\nu(dx) \\ &= E_\theta[\phi(X)\partial_\theta \log f_{X|\Theta}(X \mid \theta) \\ &= E_\theta[(\phi(X) - \underbrace{E_\theta[\phi(X)]}_{=0})\partial_\theta \log f_{X|\Theta}(X \mid \theta). \end{aligned}$$

Using Cauchy-Schwarts inequality

$$\begin{aligned} |\partial_\theta E_\theta[\phi(X)]| &\leq \Big(E_\theta[(\phi(X) - E_\theta[\phi(X)])^2]\Big)^{1/2} \Big(E_\theta[(\partial_\theta \log f_{X|\Theta}(X \mid \theta))^2]\Big)^{1/2} \\ &= \sqrt{\mathrm{Var}_\theta\, \phi(X)}\sqrt{\mathcal{I}_X(\theta)}. \end{aligned}$$

□

Note that the only inequality used in the proof is the Cauchy-Schwartz inequality. Hence, a necessary and sufficient condition for the Cramer-Rao lower bound to be achieved is that the inequality becomes an equality. This happens if and only if the two quantities are linearly related, i.e. if $EX = 0$ and $EY = 0$ then $|EXY| =$

$(EX^2)^{1/2}(EY^2)^{1/2}$ iff there is $a \neq 0$ such that $X = aY$. Thus in our case the inequality becomes equality iff there is a function $a(\theta)$ such that

$$\partial_\theta \log f_{X|\Theta}(x \mid \theta) = a(\theta)(\phi(x) - \underbrace{E_\theta[\phi(X)]}_{d(\theta)}).$$

Solving this differential equation we see that

$$f_{X|\Theta}(x \mid \theta) = c(\theta)h(x)\exp\{\pi(\theta)\phi(x)\},$$

with $c(\theta) = \exp\{-\int a(\theta)d(\theta)d\theta\}$ and $\pi(\theta) = \int a(\theta)d\theta$. That is, the Cramér-Rao lower bound is sharp only in a one-parameter exponential family with $\phi(x)$ being a sufficient statistic.

14.1. **Point estimation, sufficient statistics, and nonrandomized decision rules.** Recall that in decision theory decisions only need to be based on sufficient statistics. We showed this in Theorem 9 (Lecture 6) where we showed that if δ is a decision rule and T is a sufficient statistic then there is a decision rule $\delta_1(t; A) = E[\delta(X; A) \mid T = t]$ with $R(\theta, \delta_1) = R(\theta, \delta)$ for all θ. However, the δ_1 may be randomized even if δ is not. It is not very nice to use randomized point estimators so a question is if it can be avoided. The next two results come up with a solution in the case of point estimation.

Proposition 2. *Suppose $\aleph \subset \mathbb{R}^m$ is convex and for each θ, $a \mapsto L(\theta, a)$ is convex. Let δ be a randomized rule, $B = \{x : \int_\aleph |a|\delta(da; x) < \infty\}$ and put*

$$\delta_0(X) = \int_\aleph a\delta(da; x),$$

the mean of the randomized rule δ. Then $L(\theta, \delta_0(x)) \leq L(\theta, \delta(x))$ for each θ and $x \in B$.

Proof. By Jensen's inequality

$$L(\theta, \delta_0(x)) = L(\theta, \int_\aleph a\delta(da; x)) \leq \int_\aleph L(\theta, a)\delta(da; x) = L(\theta, \delta(x)).$$

□

An important result in this area is the next theorem. It says when we can find a good deterministic rule.

Theorem 20 (Rao-Blackwell theorem)**.** *Suppose $\aleph \subset \mathbb{R}^m$ is convex, for each θ, $a \mapsto L(\theta, a)$ is convex, T is a sufficient statistic, and δ_0 a deterministic rule with $E_\theta[\|\delta_0(X)\|] < \infty$. Put*

$$\delta_1(t) = E[\delta_0(X) \mid T = t],$$

then $R(\theta, \delta_1) \leq R(\theta, \delta_0)$.

Proof. Think of δ_0 as random by putting $\delta_3(A; x) = I_A(\delta_0(x))$. We also put

$$\delta_4(A; t) = E[\delta_3(A; X) \mid T = t],$$

$$\delta_2(t) = \int_\aleph a\delta_4(da; t).$$

Then,

$$R(\theta, \delta_2) \leq \{\text{Thm 2}\} \leq R(\theta, \delta_4) = \{\text{Thm 9}\} = R(\theta, \delta_3) = R(\theta, \delta_0).$$

It remains to show $\delta_2 = \delta_1$:

$$\delta_2(t) = \int_{\aleph} a\delta_4(da; t) = E[\int_{\aleph} a\delta_3(da; X) \mid T = t] = E[\delta_0(X) \mid T = t] = \delta_1(t).$$

□

Let's be explicit in the context if point estimation. If $\psi(X)$ is an unbiased estimator and $L(\theta, a) = (g(\theta) - a)^2$ is the loss function, then if T is a sufficient statistic

$$\phi(T) = E_\theta[\psi(X) \mid T] = E[\psi(X) \mid T]$$

satisfies $E_\theta[\phi(T)] = E_\theta[\psi(X)] = g(\theta)$, so ϕ is unbiased and $\mathrm{Var}_\theta(\phi(T)) = R(\theta, \phi) \leq R(\theta, \psi) = \mathrm{Var}_\theta(\psi(X))$. Thus,

Complete sufficient statistics play an important role for unbiased estimators.

Theorem 21. *If T is a complete statistic, then all unbiased estimators of $g(\Theta)$ that are functions of T alone, are equal P_θ-a.s. for all $\theta \in \Omega$. If there exists an unbiased estimator that is a function of a complete sufficient statistic, then it is UMVUE.*

Proof. Suppose $\phi_1(T)$ and $\phi_2(T)$ are unbiased estimators of $g(\Theta)$. Then $E_\theta[\phi_1(T) - \phi_2(T)] = 0$ for each θ and hence, by completeness, $\phi_1(T) = \phi_2(T)$ P_θ-a.s.

Suppose there is an unbiased ϕ with finite variance. Put $\phi_3(T) = E[\phi(X) \mid T]$. Then ϕ_3 is unbiased and the Rao-Blackwell theorem says $R(\theta, \phi_3) \leq R(\theta, \phi)$ for all θ. Since the risk function of unbiased estimators is the variance, this makes ϕ_3 UMVUE. □