LECTURE 3 NOTES

1. The likelihood function.

DEFINITION 1.1 (likelihood function). *The likelihood function* $l : \Theta \to [0, \infty)$ *of a sample* $x \in \mathcal{X}$ *is given by*

$$l_x(\theta) = f_\theta(x).$$

As we shall see, the likelihood function plays a crucial role in statistical inference. One possible explanation for its imperativeness is its connection to the minimal sufficient partition.

THEOREM 1.2. *A partition* $\{\mathcal{A}_t\}_{t\in\mathcal{T}}$ *is the minimal sufficient partition of* $\mathcal{X}$ *if the ratio* $\frac{l_{x_1}(\theta)}{l_{x_2}(\theta)}$ *is constant in* θ *if and only if* $x_1, x_2 \in \mathcal{A}_t$.

PROOF. The theorem is a restatement of Lecture 2, Theorem 2.2. $\square$

Thus knowledge of any sufficient statistic determines the likelihood function up to a constant. We remark that

1. the likelihood is a *random* function. It depends on the (random) observations.
2. the likelihood is not a density. It is a function of θ, not of x.

Often, we work with the *log-likelihood function* $\ell_x : \Theta \to \mathbf{R}$ given by

$$\ell_x(\theta) = \log l_x(\theta).$$

If the observations $\mathbf{x} = (\mathbf{x}_i)_{i\in[n]}$ consists of *i.i.d.* random variables $\mathbf{x}_i$, the (joint) likelihood is a product of likelihoods:

$$l_x(\theta) = \textstyle\prod_{i\in[n]} f_\theta^1(x_i),$$

where $f_\theta^1(x)$ is the density of $\mathbf{x}_1$, and the log-likelihood is a sum of log likelihoods:

$$\ell_x(\theta) = \textstyle\sum_{i=1}^n \log l_x(\theta).$$

EXAMPLE 1.3. *Let* $\mathbf{x}_i \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(p)$. *The log-likelihood function is*

$$\begin{aligned}\ell_{\mathbf{x}}(p) &= \textstyle\sum_{i\in[n]} \ell_{\mathbf{x}_i}(p)\\ &= \textstyle\sum_{i\in[n]} \mathbf{x}_i \log(p) + (1-\mathbf{x}_i)\log(1-p)\\ &= \mathbf{t}\log(p) + (n-\mathbf{t})\log(1-p).\end{aligned}$$

where $\mathbf{t} = \sum_{i\in[n]} \mathbf{x}_i$.

Example 1.4. *Let* $\mathbf{x}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, *where* σ^2 *is known. The log likelihood function is*

$$\ell_{\mathbf{x}}(\mu) = \textstyle\sum_{i\in[n]} -\frac{1}{2\sigma^2}(\mathbf{x}_i - \mu)^2 - \log\sigma - \frac{1}{2}\log(2\pi).$$

Dropping the terms that do not depend on μ,

$$\propto -\tfrac{n}{2\sigma^2}(\bar{\mathbf{x}} - \mu)^2.$$

Example 1.5. *Let* $\mathbf{x}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \Sigma)$. *The log-likelihood function is*

$$\begin{aligned}\ell_{\mathbf{x}}(\mu, \Sigma) &= \sum_{i\in[n]} -\frac{1}{2}\|\mathbf{x}_i - \mu\|^2_{\Sigma^{-1}} - \frac{1}{2}\log\det(\Sigma) - \frac{1}{2}\log(2\pi) \\ &\propto -\frac{1}{2}\sum_{i\in[n]} \|\mathbf{x}_i - \mu\|^2_{\Sigma^{-1}} - \frac{n}{2}\log\det(\Sigma).\end{aligned}$$

It is possible to show that

$$\sum_{i\in[n]} \|\mathbf{x}_i - \mu\|^2_{\Sigma^{-1}} = \sum_{i\in[n]} \|\mathbf{x}_i - \bar{\mathbf{x}}\|^2_{\Sigma^{-1}} + n\|\bar{\mathbf{x}} - \mu\|^2_{\Sigma^{-1}}.$$

By the properties of the tr,

$$\begin{aligned}\sum_{i\in[n]} \|\mathbf{x}_i - \bar{\mathbf{x}}\|^2_{\Sigma^{-1}} &= \sum_{i\in[n]} \operatorname{tr}\big((\mathbf{x}_i - \bar{\mathbf{x}})^T\Sigma^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})\big) \\ &= \sum_{i\in[n]} \operatorname{tr}\big((\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T\Sigma^{-1}\big) \\ &= n\operatorname{tr}(\mathbf{S}\Sigma^{-1}),\end{aligned}$$

where $\mathbf{S} := \frac{1}{n}\sum_{i\in[n]}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$ *is the sample covariance matrix. Thus the log-likelihood is*

$$\ell_{\mathbf{x}}(\mu, \Sigma) \propto -\tfrac{n}{2}\operatorname{tr}(\mathbf{S}\Sigma^{-1}) - n\|\bar{\mathbf{x}} - \mu\|^2_{\Sigma^{-1}} - \tfrac{n}{2}\log\det(\Sigma).$$

2. The maximum likelihood estimator.

Definition 2.1. *A maximum likelihood estimator (MLE) of* $\theta^* \in \Theta$ *on a sample* $x \in \mathcal{X}$ *is given by* $\arg\max_{\theta\in\Theta} \ell_{\mathbf{x}}(\theta)$.

In the rest of the notes, we assume the MLE is unique; i.e. the $\arg\max$ in Definition 2.1 is attained at a unique $\hat{\theta}$. When the MLE is not unique, it suggests either the model is unidentifiable (e.g. a non-minimal exponential family) or the data is insufficient.

Intuitively, the MLE a parameter point at which the observed sample is most likely. As we shall see, the MLE is generally a good point estimator, possessing some of the optimality properties that we shall discuss later. The main drawback to the MLE is the hardness of finding the *global* maximizer of the likelihood function, especially when the likelihood is non-concave.

EXAMPLE 2.2 (Example 1.3 continued). *Recall the log-likelihood function is*

$$\begin{aligned}\ell_{\mathbf{x}}(p) &= \sum_{i\in[n]} \mathbf{x}_i \log p + (1-\mathbf{x}_i)\log(1-p) \\ &= \mathbf{t}\log p + (n-\mathbf{t})\log(1-p),\end{aligned}$$

where $t = \sum_{i\in[n]} \mathbf{x}_i$. *By the optimality of the MLE* $\hat{p}_{\mathrm{ML}}$,

$$0 = \nabla \ell_{\mathbf{x}}(\hat{p}_{\mathrm{ML}}) = \frac{\mathbf{t}}{\hat{p}_{\mathrm{ML}}} - \frac{n-\mathbf{t}}{1-\hat{p}_{\mathrm{ML}}}.$$

We solve for $\hat{p}_{\mathrm{ML}}$ *to obtain* $\hat{p}_{\mathrm{ML}} = \frac{\mathbf{t}}{n}$.

We observe that $\mathbf{t}$ is a sufficient statistic for the *i.i.d.* Bernoulli model, and $\hat{p}$ depends only on $\mathbf{x}$ through $\mathbf{t}$. This is not a coincidence: the MLE generally depends only on the data through a sufficient statistic. Indeed, by the factorization theorem, we have

$$\begin{aligned}\arg\max_{\theta\in\Theta} \ell_{\mathbf{x}}(\theta) &= \arg\max_{\theta\in\Theta} \log f_\theta(\mathbf{x}) \\ &= \arg\max_{\theta\in\Theta} \log g_\theta(\phi(\mathbf{x})) + \log h(\mathbf{x}) \\ &= \arg\max_{\theta\in\Theta} \log g_\theta(\phi(\mathbf{x})).\end{aligned}$$

EXAMPLE 2.3. *Let* $\mathbf{x}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$. *The log-likelihood function is*

$$\begin{aligned}\ell_{\mathbf{x}}(\mu, \sigma^2) &= \frac{1}{2\sigma^2} \sum_{i\in[n]} \left[-(\mathbf{x}_i - \mu)^2 - \log\sigma - \log\sqrt{2\pi}\right] \\ &\propto -\frac{1}{2\sigma^2} \sum_{i\in[n]} (\mathbf{x}_i - \mu)^2 - n\log\sigma.\end{aligned}$$

By the optimality of $\hat{\mu}_{\text{ML}}$*,*

$$0 = -\frac{1}{\hat{\sigma}_{\text{ML}}^2}\sum_{i\in[n]}\left[\hat{\mu}_{\text{ML}} - \mathbf{x}_i\right]$$
$$= -\frac{1}{\hat{\sigma}_{\text{ML}}^2}\Big(n\hat{\mu}_{\text{ML}} - \textstyle\sum_{i\in[n]}\mathbf{x}_i\Big).$$

We solve for $\hat{\mu}$ *to obtain* $\hat{\mu} = \frac{1}{n}\sum_{i\in[n]}\mathbf{x}_i$*. By a similar argument, it is possible to show* $\hat{\sigma}_{\text{ML}}^2 = \frac{1}{n}\sum_{i\in[n]}(\mathbf{x}_i - \hat{\mu})^2$*. Indeed, by the optimality of* $\hat{\sigma}_{\text{ML}}$*,*

$$0 = \frac{1}{\hat{\sigma}_{\text{ML}}^3}\sum_{i\in[n]}(\mathbf{x}_i - \mu)^2 - \frac{n}{\hat{\sigma}_{\text{ML}}}.$$

We solve for $\hat{\sigma}_{\text{ML}}^2$ *to obtain* $\hat{\sigma}_{\text{ML}}^2 = \frac{1}{n}\sum_{i\in[n]}(\mathbf{x}_i - \hat{\mu})^2$*.*

Example 2.4. *Let* $\mathbf{x}_i \overset{\text{i.i.d.}}{\sim} \text{unif}(0,\theta)$*. The likelihood function is*

$$l(\theta) = \theta^{-1}\prod_{i\in[n]}\mathbf{1}_{[0,\theta]}(\mathbf{x}_i) = \theta^{-n}\prod_{i\in[n]}\mathbf{1}_{[0,\theta]}(\mathbf{x}_i).$$

If θ *is smaller than any observation, the likelihood vanishes. Thus* $\hat{\theta}_{\text{ML}}$ *is at least* $\max_{i\in[n]}\mathbf{x}_i$*. But* θ^{-n} *is larger at smaller values of* θ*. Thus*

$$\hat{\theta}_{\text{ML}} := \max\nolimits_{i\in[n]}\mathbf{x}_i.$$

Before moving on to other approaches to point estimation, we mention that the MLE is *equivariant*: if $\hat{\theta}$ is the MLE of a parameter θ^*, then $\tau(\hat{\theta})$ is the MLE of $\tau(\theta^*)$. It is a consequence of the equivariance of optimization to re-parametrization.

Lemma 2.5. *If* $\hat{\theta}_{\text{ML}}$ *is the MLE of* θ*,* $g(\hat{\theta}_{\text{ML}})$ *is the MLE of* $\eta = g(\theta)$*.*

Proof. Let $g^{-1}(\eta) := \{\theta \in \Theta : g(\theta) = \eta\}$. We remark that g^{-1} is not a function; it is a set-valued mapping. The reparametrized likelihood is

$$l'_{\mathbf{x}}(\eta) = \sup\nolimits_{\theta\in g^{-1}(\eta)} l_{\mathbf{x}}(\theta),$$

where $l_{\mathbf{x}}(\theta)$ is the (original) likelihood. By the optimality of $\hat{\theta}_{\text{ML}}$,

$$l'_{\mathbf{x}}(g(\hat{\theta}_{\text{ML}})) = \sup\nolimits_{\theta\in g^{-1}(g(\hat{\theta}_{\text{ML}}))} l_{\mathbf{x}}(\theta)$$
$$= l_{\mathbf{x}}(\hat{\theta}_{\text{ML}})$$
$$\geq l_{\mathbf{x}}(\theta)$$

for any $\theta \in \Theta$, where the second equality is by

1. $\hat{\theta}_{\mathrm{ML}} \in g^{-1}(g(\hat{\theta}_{\mathrm{ML}})) \subset \Theta$ by the definition of g^{-1},
2. $l_{\mathbf{x}}(\hat{\theta}_{\mathrm{ML}}) \ge l_{\mathbf{x}}(\theta)$ for any $\theta \in g^{-1}(g(\hat{\theta}_{\mathrm{ML}}))$.

Thus

$$l'_{\mathbf{x}}(g(\hat{\theta}_{\mathrm{ML}})) \ge \sup_{\theta \in g^{-1}(\eta)} l_{\mathbf{x}}(\theta) = l'_{\mathbf{x}}(\eta)$$

for any $\eta = g(\theta)$ for some $\theta \in \Theta$. $\square$

EXAMPLE 2.6 (Example 1.3 continued). *Going back to Example 1.3, if the parameter of interest is the odds ratio $\frac{p}{1-p}$, by the equivariance of the MLE, the MLE of the odds ratio is $\frac{\hat{p}_{\mathrm{ML}}}{1-\hat{p}_{\mathrm{ML}}}$.*

3. The method of moments. An older approach to point estimation is the method of moments (MoM). Let $\mathbf{x} \in \mathbf{R}^n$ consist of *i.i.d.* random variables $\mathbf{x}_i \in \mathbf{R}$, $i \in [n]$. In its most simple form, the MoM

1. expresses the first m moments of $\mathbf{x}_1$ in terms of θ:
$$\mu_k(\theta) = \mathbf{E}_\theta\big[\mathbf{x}_1^k\big],\ k \in [m];$$
2. plugs in the first m sample moments and solve for $\hat{\theta}$:
$$\frac{1}{n}\sum_{i\in[n]} \mathbf{x}_i^k - \mu_k(\hat{\theta}_{\mathrm{MoM}}) = 0,\ k \in [m].$$

EXAMPLE 3.1 (Example 1.3 continued). *The first moment of $\mathbf{x}_1$ is p. We plug in the first sample moment $\hat{\mu}_1 = \frac{1}{n}\sum_{i\in[n]} \mathbf{x}_i$ to obtain*

$$\hat{p}_{\mathrm{MoM}} = \frac{1}{n}\sum_{i\in[n]} \mathbf{x}_i$$

Thus, in the coin tossing example, the MLE and the MoM are the same! As we shall see, this is no mere coincidence: the two approaches are generally equivalent when the model is an exponential family.

EXAMPLE 3.2. *Let $\mathbf{x}_i \overset{\text{i.i.d.}}{\sim} \mathrm{Bin}(n,p)$, where both n and p are unknown. By the properties of the binomial distribution,*

$$\mu_1(n,p) = np,$$
$$\mu_2(n,p) = np(1-p) + (np)^2.$$

We plug in the first two sample moments and solve for n and p to obtain

$$\hat{n} = \frac{\hat{\mu}_2^2}{\hat{\mu}_1 - \hat{\mu}_2 + \hat{\mu}_1^2},\ \hat{p} = \frac{\hat{\mu}_1}{\hat{n}}.$$