Lecture 4. Maximum Likelihood Estimation

Likelihood

Maximum likelihood estimation is one of the most important and widely used methods for finding estimators. Let $X_1, \ldots, X_n$ be rv's with joint pdf/pmf $f_{\mathbf{X}}(\mathbf{x} \mid \theta)$. We observe $\mathbf{X} = \mathbf{x}$.

Definition 4.1

The **likelihood** of θ is $\text{like}(\theta) = f_{\mathbf{X}}(\mathbf{x} \mid \theta)$, regarded as a function of θ. The **maximum likelihood estimator** (mle) of θ is the value of θ that maximises $\text{like}(\theta)$.

It is often easier to maximise the **log-likelihood**.

If $X_1, \ldots, X_n$ are iid, each with pdf/pmf $f_X(x \mid \theta)$, then

$$\begin{aligned} \text{like}(\theta) &= \prod_{i=1}^{n} f_X(x_i \mid \theta) \\ \text{loglike}(\theta) &= \sum_{i=1}^{n} \log f_X(x_i \mid \theta). \end{aligned}$$

Example 4.1

Let $X_1, \ldots, X_n$ be iid Bernoulli(p).

Then $l(p) = \text{loglike}(p) = (\sum x_i) \log p + (n - \sum x_i) \log(1-p)$.

Thus

$$dl/dp = \frac{\sum x_i}{p} - \frac{n - \sum x_i}{(1-p)}.$$

This is zero when $p = \sum x_i / n$, and the mle of p is $\hat{p} = \sum x_i / n$.

Since $\sum X_i \sim \text{Bin}(n, p)$, we have $\mathbb{E}(\hat{p}) = p$ so that $\hat{p}$ is unbiased.

Example 4.2

Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$. Then

$$l(\mu, \sigma^2) = \text{loglike}(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum (x_i - \mu)^2.$$

This is maximised when $\frac{\partial l}{\partial \mu} = 0$ and $\frac{\partial l}{\partial \sigma^2} = 0$. We find

$$\frac{\partial l}{\partial \mu} = -\frac{1}{\sigma^2}\sum (x_i - \mu), \quad \frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum (x_i - \mu)^2,$$

so the solution of the simultaneous equations is $(\hat{\mu}, \hat{\sigma}^2) = (\bar{x}, S_{xx}/n)$.

(writing $\bar{x}$ for $\frac{1}{n}\sum x_i$ and S_{xx} for $\sum (x_i - \bar{x})^2$)

Hence the maximum likelihood estimators are $(\hat{\mu}, \hat{\sigma}^2) = (\bar{X}, S_{XX}/n)$.

Example 4.3 - continued

We know $\hat{\mu} \sim N(\mu, \sigma^2/n)$ so $\hat{\mu}$ is unbiased.

We shall see later that $\frac{S_{xx}}{\sigma^2} = \frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}$.
Now $\mathbb{E}(\chi^2_{n-1}) = n-1$, and so

$$\mathbb{E}(\hat{\sigma}^2) = \mathbb{E}(\chi^2_{n-1} \times \tfrac{\sigma^2}{n}) = \tfrac{(n-1)\sigma^2}{n},$$

ie $\hat{\sigma}^2$ is biased.

Note that $\mathbb{E}(\hat{\sigma}^2 \times \frac{n}{n-1}) = \sigma^2$, and so $\frac{S_{xx}}{n-1}$ is unbiased.

This means that the classic sample variance estimator $\frac{\sum_i (x_i - \bar{x})^2}{n-1}$ with denominator $n-1$ is *unbiased*, MLE has denominator n is *biased*.]

However $\mathbb{E}(\hat{\sigma}^2) \to \sigma^2$ as $n \to \infty$, so $\hat{\sigma}^2$ is asymptotically unbiased.

Example 4.3

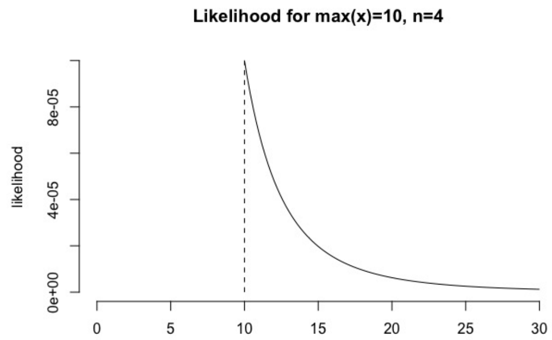
Let $X_1, \dots, X_n$ be iid $U[0, \theta]$. Then

$$\text{like}(\theta) = \frac{1}{\theta^n} 1_{\{\max_i x_i \leq \theta\}}(\max_i x_i).$$

For $\theta \geq \max x_i$, $\text{like}(\theta) = \frac{1}{\theta^n} > 0$ and is decreasing as θ increases, while for $\theta < \max x_i$, $\text{like}(\theta) = 0$.

Hence the value $\hat{\theta} = \max x_i$ maximises the likelihood.

Assume $\mathbf{x} = (4, 7, 2, 10)$, so that $n = 4, \max x_i = 10$.



Example 4.3 - continued

Is $\hat{\theta}$ unbiased? First we need to find the distribution of $\hat{\theta}$. For $0 \leq t \leq \theta$, the distribution function of $\hat{\theta}$ is

$$F_{\hat{\theta}}(t) = \mathbb{P}(\hat{\theta} \leq t) = \mathbb{P}(X_i \leq t, \text{ all } i) = (\mathbb{P}(X_i \leq t))^n = \left(\frac{t}{\theta}\right)^n,$$

where we have used independence at the second step.

Differentiating with respect to t, we find the pdf $f_{\hat{\theta}}(t) = \frac{nt^{n-1}}{\theta^n}$, $0 \leq t \leq \theta$. Hence

$$\mathbb{E}(\hat{\theta}) = \int_0^\theta t \frac{nt^{n-1}}{\theta^n} dt = \frac{n\theta}{n+1},$$

so $\hat{\theta}$ is biased, but asymptotically unbiased.

Properties of mle's

(i) If T is sufficient for θ, then the likelihood is $g(T(\mathbf{x}), \theta)h(\mathbf{x})$, which depends on θ only through $T(\mathbf{x})$.

To maximise this as a function of θ, we only need to maximise g, and so the mle $\hat{\theta}$ is a *function of the sufficient statistic*.

(ii) If $\phi = h(\theta)$ where h is injective $(1-1)$, then the mle of ϕ is $\hat{\phi} = h(\hat{\theta})$. This is called the invariance property of mle's. IMPORTANT.

(iii) It can be shown that, under regularity conditions, that $\sqrt{n}(\hat{\theta} - \theta)$ is asymptotically multivariate normal with mean 0 and 'smallest attainable variance' (see Part II Principles of Statistics).

(iv) Often there is no closed form for the mle, and then we need to find $\hat{\theta}$ numerically.

Example 4.4

Smarties come in k equally frequent colours, but suppose we do not know k.

[Assume there is a vast bucket of Smarties, and so the proportion of each stays constant as you sample. Alternatively, assume you sample with replacement, although this is rather unhygienic]

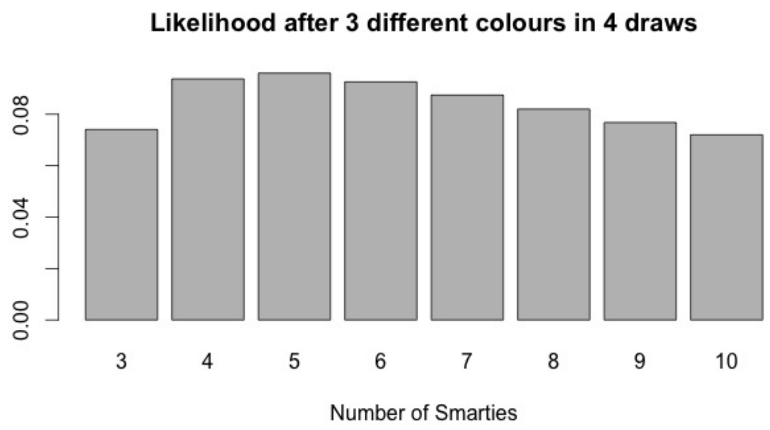
Our first four Smarties are Red, Purple, Red, Yellow.

The likelihood for k is (considered sequentially)

$$\begin{aligned} \text{like}(k) &= \mathbb{P}_k(\text{1st is a new colour})\,\mathbb{P}_k(\text{2nd is a new colour}) \\ &\quad \mathbb{P}_k(\text{3rd matches 1st})\,\mathbb{P}_k(\text{4th is a new colour}) \\ &= 1 \times \frac{k-1}{k} \times \frac{1}{k} \times \frac{k-2}{k} \\ &= \frac{(k-1)(k-2)}{k^3} \end{aligned}$$

(Alternatively, can think of Multinomial likelihood $\propto \frac{1}{k^4}$, but with $\binom{k}{3}$ ways of choosing those 3 colours.)

Can calculate this likelihood for different values of k:
$\text{like}(3) = 2/27, \text{like}(4) = 3/32, \text{like}(5) = 12/25, \text{like}(6) = 5/54$, maximised at $\hat{k} = 5$.



Fairly flat! Not a lot of information.