STA111 - Lecture 12

Introduction to Maximum Likelihood Estimation

1 Recap and Motivation

So far we have been finding estimators using our intuition and evaluated their properties (bias, variance, MSE) on a case-by-case basis. In (almost all of) our simple examples, the parameters were population averages (expectations) of some sort, which we estimated using sample means.

Properties of sample means: Suppose that $X_1, X_2, \ldots, X_n$ are iid from some distribution with $E(X_1) = \mu$ and $V(X_1) = \sigma^2$, and let $\overline{X}_n = (X_1 + X_2 + \cdots + X_n)/n$ be the sample mean. We know that

1. The sample mean is **unbiased**: $E(\overline{X}_n) = \mu$.
2. The variance of the sample mean is σ^2/n: $V(\overline{X}_n) = \sigma^2/n$.
3. The sample mean is **consistent**: $\overline{X}_n \to \mu$ as $n \to \infty$ (with probability 1).
4. The sample mean is **approximately Normal** if the sample size is big enough: $\overline{X}_n \approx \text{Normal}(\mu, \sigma^2/n)$.

Remark: μ and σ^2 are shorthand notation for "population mean" and "population variance". We are **not** assuming that the random variables are Normal(μ, σ^2): the results above apply to all the models we have seen in the course.

The sample mean has great properties, but sometimes we want to estimate parameters that are not population averages. We have already seen a few examples, like the population variance or M in the Hypergeometric(N, M, n) model (this is the example where we want to estimate the total number of students who are in favor of a policy, see Lecture Notes 10). An important result that we managed to derive is that the **sample variance**

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

is an **unbiased** estimator of the population variance σ^2 (remark: "sample variance" isn't the same as "variance of the sample mean": the sample variance is the name of s_n^2, an unbiased estimator of the population variance σ^2; the variance of the sample mean is a property of $\overline{X}_n$, which is an unbiased estimator of μ).

What should we do if we work with complicated models for which we can't come up with intuitive estimators? A possible answer to this question is using Maximum Likelihood Estimation, which is a systematic method for finding estimators (given a model) with good theoretical properties.

2 Introduction to Maximum Likelihood Estimation

(Loosely based on: https://onlinecourses.science.psu.edu/stat414/node/191)

Suppose we have a random sample of iid random variables $X_1, X_2, \ldots, X_n$ with a PMF or PDF $f_\theta(x)$ which depends on a parameter θ. The joint PMF/PDF is

$$f_\theta(x_1, x_2, \ldots, x_n) = f_\theta(x_1)f_\theta(x_2)\cdots f_\theta(x_n) = \prod_{i=1}^{n} f_\theta(x_i).$$

Upon observing the data, we can substitute $x_1, x_2, \dots, x_n$ by the actual values in the sample, so $f_\theta(x_1, x_2, \dots, x_n)$ becomes a function of θ alone. Seeing $f_\theta(x_1, x_2, \dots, x_n)$ as a function of θ, we write

$$\mathcal{L}(\theta) = f_\theta(x_1, x_2, \dots, x_n),$$

and call $\mathcal{L}(\theta)$ the **likelihood** of the data. The Maximum Likelihood Estimator of θ (MLE) is the value $\widehat{\theta}$ that maximizes the likelihood. Products are typically hard to maximize, so we usually take logarithms and maximize the log-likelihood $\ell(\theta) = \log \mathcal{L}(\theta)$ instead.

MLEs have very good theoretical properties. Under some regularity conditions (to be discussed in STA250), one can show that:

1. MLEs are **consistent**.
2. Their **bias goes to zero** as $n \to \infty$.
3. They are **approximately Normal** with known parameters.
4. Their **variance is optimal** in some sense (to be discussed in STA250).
5. If $\widehat{\theta}$ is the MLE for θ, $g(\widehat{\theta})$ is the MLE for $g(\theta)$. For example, if $\widehat{\mu}$ is the MLE of μ, $e^{\sin(\widehat{\mu}^2)}$ is the MLE of $e^{\sin(\mu^2)}$.

Examples:

- Let $X_1, X_2, \dots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$ with p unknown, and suppose that $x_1, x_2, \dots, x_n$ have been observed, then:

  $$\mathcal{L}(p) = \prod_{i=1}^{n} P(X_i = x_i) = \prod_{i=1}^{n} p^{x_i}(1-p)^{x_i} = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i} = p^{S_n}(1-p)^{n-S_n},$$

  where $S_n = \sum_{i=1}^n x_i$. The likelihood can be interpreted at fixed values of p as follows: if we evaluate it at, say, $p = 0.35$, $\mathcal{L}(0.35)$ is the probability of observing the data if the true value of p were 0.35. Since the MLE $\widehat{p}$ is the value of p that maximizes the likelihood, we can say that the maximum likelihood estimator is the value of p that is "most likely" to have generated the data. The log-likelihood is

  $$\ell(p) = S_n \log p + (n - S_n)\log(1-p).$$

  We can find the MLE by differentiating $\ell(\theta)$

  $$\ell'(p) = S_n/p - (n - S_n)/(1-p)$$

  and setting it to 0, and we find that the MLE is the sample proportion:

  $$\widehat{p} = \frac{S_n}{n} = \overline{X}_n,$$

  which is a maximum since $\ell''(\widehat{p}) < 0$.

- Let $X_1, X_2, \dots, X_n \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$, then

  $$\mathcal{L}(\lambda) = \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda_i^x}{x_i!} = C\,\lambda^{S_n} e^{-n\lambda},$$

where $S_= \sum_{i=1}^{n} x_i$ and $C = 1/\prod_{i=1}^{n} x_i!$. The log-likelihood function is

$$\ell(\lambda) = \log(C) + S_n \log\lambda - n\lambda,$$

so the derivative of the log-likelihood is

$$\ell'(\lambda) = S_n/\lambda - n$$

Therefore, we can find $\widehat{\lambda} = S_n/n = \overline{X}_n$. It is easy to check that $\ell''(\widehat{\lambda}) < 0$, so $\widehat{\lambda}$ is indeed a maximum.

- The MLE of μ and σ^2 of Normal(μ, σ^2) are $\widehat{\mu} = \overline{X}_n$ and $\widehat{\sigma^2} = \sum_{i=1}^{n}(X_i - \overline{X}_n)^2/n$, respectively. Note that $\widehat{\sigma^2}$ is biased.
- Let $X_1, X_2, \ldots, X_n$ be a random sample from a discrete distribution with support $\{0, 1, 2\}$. Suppose that θ, the parameter of interest, can only take on the values $\theta = 0$ and $\theta = 1$. The PMFs for $\theta = 0$ and $\theta = 1$ are:

	$\theta = 0$	$\theta = 1$
$X = 0$	0.1	0.2
$X = 1$	0.3	0.5
$X = 2$	0.6	0.3

Suppose we observe a random sample of size 4 and the values are 0, 0, 1, 2. What is the MLE of θ? The likelihood at $\theta = 0$ is

$$\begin{aligned}\mathcal{L}(0) &= P_{\theta=0}(X_1 = 0, X_2 = 0, X_3 = 1, X_4 = 2)\\ &= P_{\theta=0}(X = 0)^2 P_{\theta=0}(X = 1) P_{\theta=0}(X = 2)\\ &= 0.1^2 \cdot 0.3 \cdot 0.6 = 0.0018,\end{aligned}$$

and analogously

$$\mathcal{L}(1) = 0.2^2 \cdot 0.5 \cdot 0.3 = 0.006,$$

so the MLE in this case is $\widehat{\theta} = 1$.

Exercise 1. *Let $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim}$ Exponential(λ). Find the MLE of λ.*

Exercise 2. *Suppose you have 3 coins with probability of heads p equal to 0.4, 0.5 and 0.6, respectively. Bobby picks one of the coins (all of them are equally likely) and gives it to you. After flipping the coin 100 times, you get heads 53 times. Find the MLE of p in this experiment.*

Exercise 3. *Let $X_1, X_2, \ldots, X_n$ be a random sample from a discrete distribution with support equal to $\{0, 1, 2, 3\}$. Suppose that θ can only take on the values $\theta = 0$ and $\theta = 1$. The PMFs for $\theta = 0$ and $\theta = 1$ are:*

	$\theta = 0$	$\theta = 1$
$X = 0$	0.1	0.2
$X = 1$	0.3	0.4
$X = 2$	0.3	0.3
$X = 3$	0.3	0.1

Suppose that $n = 6$ and the data is 0,3,1,2,0,3. Find the MLE of θ.

Exercise 4. *Let $X_1, X_2, \dots, X_n$ be a random sample from a random variable with pdf*

$$f_\alpha(x) = (2\alpha + 1)x^{2\alpha}, \ 0 \leq x \leq 1,$$

and 0 otherwise. Find the MLE of α.