Maximum Likelihood for IID Data

Prof. Dennis Sun

Data 401

Where We Left Off

• We calculated the maximum likelihood estimate (MLE) of $\theta$ based on a single observation $X \sim p_\theta(x)$, using two approaches: graphical and analytic.

• What if we have multiple observations $X_1, ..., X_n$, independent and identically distributed (i.i.d.)?

For example, in the store purchases example, we might record the times of $n$ purchases, giving us $n - 1$ interarrival times to estimate $\lambda$:

$T_2, ..., T_n \overset{i.i.d.}{\sim} \text{Expo}(\lambda)$. 
Maximum Likelihood for IID Data

We can still apply the maximum likelihood principle. We now have to write down the likelihood of the data \( x_1, \ldots, x_n \):

\[
L_{x_1, \ldots, x_n} (\theta) = p_\theta (x_1, \ldots, x_n) = \prod_{i=1}^{n} p_\theta (x_i).
\]

Would the graphical approach still work here? Why or why not?

But this is still just a one-dimensional optimization problem (over \( \theta \)), for fixed values \( x_1, \ldots, x_n \).

Store Purchases Example

Suppose we observe interarrival times \( T_2, \ldots, T_n \overset{i.i.d}{\sim} \text{Expo}(\lambda) \). Find the maximum likelihood estimator of \( \lambda \). Then, use this to find the estimate when the first 5 purchases happen at:

8:01:20, 8:04:26, 8:04:56, 8:06:50, 8:11:20.
Store Purchases Example

More Practice

Calculate the maximum likelihood estimator for the following parameters. (You may need to look up the p.m.f. / p.d.f. of these distributions online.)

- $X_1, ..., X_n \sim \text{Poisson}(\lambda)$. What is $\hat{\lambda}_{\text{MLE}}$?
- $X_1, ..., X_n \sim \text{Normal}(\mu, 1)$. What is $\hat{\mu}_{\text{MLE}}$?
- $X_1, ..., X_n \sim \text{Normal}(0, \sigma^2)$. What is $\hat{\sigma}^2_{\text{MLE}}$? (Hint: To make the math simpler, treat $v = \sigma^2$ as the parameter, rather than trying to differentiate with respect to $\sigma$.)
**Multiparameter Problems**

What if $X_1, ..., X_n \sim \text{Normal}(\mu, \sigma^2)$ with both $\mu$ and $\sigma^2$ unknown?

- **Method 1:** Take partial derivatives of the log-likelihood with respect to $\mu$ and $\sigma^2$, and solve the system of equations:
  
  $$
  \frac{\partial}{\partial \mu} \log L_{X_1,\ldots,X_n}(\mu, \sigma^2) = 0
  $$
  
  $$
  \frac{\partial}{\partial \sigma^2} \log L_{X_1,\ldots,X_n}(\mu, \sigma^2) = 0
  $$

- **Method 2:** First, for any fixed $\mu$, calculate the MLE of $\sigma^2$ as a function of $\mu$: $\hat{\sigma}^2_{\text{MLE}}(\mu)$. Then, maximize what statisticians call the **profile likelihood**:

  $$
  \hat{\mu}_{\text{MLE}} = \arg\max_{\mu} \log L(\mu, \hat{\sigma}^2_{\text{MLE}}(\mu)).
  $$

  Note that this is only a one-parameter optimization problem (in $\mu$).

---

**Some Theory of the MLE**

Suppose $X_1, ..., X_n \overset{i.i.d.}{\sim} p_{\theta_0}$. We calculate the MLE:

$$
\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \sum_{i=1}^{n} \log p_{\theta}(X_i).
$$

Why is this a good estimator?

- As $n \rightarrow \infty$, $\hat{\theta}_{\text{MLE}} \rightarrow \theta_0$. This property is called **consistency**.
- As $n \rightarrow \infty$, the distribution of $\hat{\theta}_{\text{MLE}}$ is approximately:

  $$
  \hat{\theta}_{\text{MLE}} \approx \text{Normal}(\theta_0, \frac{1}{nI(\theta_0)}),
  $$

  where $I(\theta_0)$ is a quantity called the Fisher information.
- The **Cramér-Rao Theorem** says that for any unbiased estimator $\hat{\theta}$,

  $$
  \text{Var}[\hat{\theta}] \geq \frac{1}{nI(\theta_0)},
  $$

  so as $n \rightarrow \infty$, the MLE achieves the smallest possible variance of any estimator! This property is called **efficiency**.