# Stat 2911 Lecture Notes

Class 12, 2017

Uri Keich



MLE: Bernoulli, Poisson, Binomial(p), Binomial(m,p), Multinomial (Rice 8.5)

# Maximum Likelihood Estimation (MLE)

Given a sample $x_1, \ldots, x_n$ from a dist. with pmf $f_\theta(x)$ we want to estimate $\theta$.

The likelihood function gives the probs. of the observed sample assuming $\theta$ is known:

$$L(\theta) = P_\theta(X_1 = x_1, \ldots, X_n = x_n)$$

If the sample is modeled as an iid then

$$L(\theta) = \prod_1^n f_\theta(x_i)$$

The MLE is defined as

$$\hat{\theta} = \operatorname*{argmax}_\theta L(\theta)$$

Since log is a monotone increasing function

$$\hat{\theta} = \operatorname*{argmax}_\theta \ell(\theta),$$

where $\ell(\theta) = \log L(\theta) \overset{\text{in iid case}}{=} \sum_1^n \log f_\theta(x_i)$

Example. $X_i \sim \text{Bernoulli}(\theta)$

$$L(\theta) = \theta^{\sum x_i} (1-\theta)^{n - \sum x_i}$$

Examples 1) $X_i \sim$ Bernoulli $(\theta) \Rightarrow L(\theta) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$

$\Rightarrow \ell(\theta) = s \cdot \log\theta + (n-s)\log(1-\theta)$, where $s = \sum_1^n x_i$

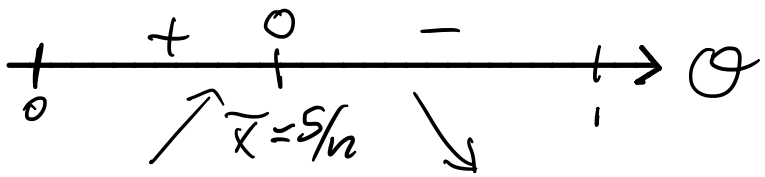
How shall we maximize $\ell(\theta)$?

$$\ell'(\theta) = \frac{s}{\theta} - \frac{n-s}{1-\theta} \qquad \theta \in (0,1)$$

$$= \frac{s - n\theta}{\theta(1-\theta)}$$

$$= \frac{n}{\theta(1-\theta)}(s/n - \theta)$$

$\operatorname{Sgn}(\ell')$:




$\Rightarrow$ Assuming $\bar{x} \in (0,1)$, $\theta = s/n = \bar{x}$ is a unique critical point which is a global max for $\theta \in [0,1]$ (why?), so the MLE is $\hat{\theta} = \bar{X}$. (moment estimator?) What if $\bar{x} \in \{0,1\}$?

If $\bar{x} = 0$, $s = 0$ so $\ell(\theta) = n\log(1-\theta)$ which is clearly maximized for $\theta =$ = .

$\Rightarrow \hat{\theta} = \bar{x}$ in this case as well.

Finally if $\bar{x} = 1$, $s = n$ so $\ell(\theta) = n\log(\theta)$ which is maximized for $\hat{\theta} =$ = .

$\Rightarrow \hat{\theta} = s/n = \bar{x}$ is the Bernoulli $(\theta)$ MLE.

Note that the above analysis shows that for $x_i \geq 0$ and $S = \sum_1^n x_i$,

$$\operatorname*{argmax}_{\theta \in [0,1]} \theta^S (1-\theta)^{n-S} = \frac{1}{n} S .$$

2) $X_i \sim \text{Pois}(\lambda)$ $\quad \lambda \geq 0$

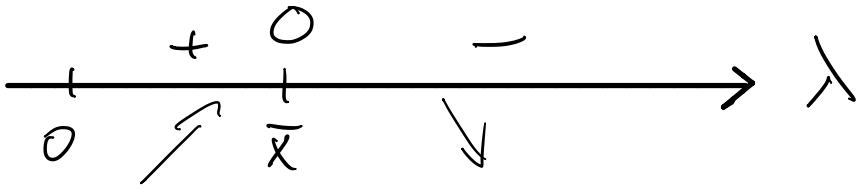
$$\ell(\lambda) = \sum_1^n \log\left(e^{-\lambda} \lambda^{x_i} / x_i!\right)$$

$$= -n\lambda + \left(\sum x_i\right) \log \lambda - \sum \log(x_i!)$$

$$\Rightarrow \ell'(\lambda) = -n + \frac{1}{\lambda} \sum x_i \qquad \lambda > 0$$

$$= \frac{n}{\lambda}(\bar{x} - \lambda)$$

$\text{sgn}(\ell')$:




If $\bar{x} > 0$ then $\lambda = \bar{x}$ is a unique critical point of $\ell$ which is a global maximum (why?)

$\Rightarrow \hat{\lambda} = \bar{x}$ is the MLE (moment estimator?)

If $\bar{x} = 0$, then $\sum x_i = 0$ so $\ell(\lambda) = -n\lambda$ which is clearly maximized for $\lambda = 0$, so again $\hat{\lambda} = \bar{x}$.

$\Rightarrow$ $\hat{\lambda} = \bar{x}$ is always the MLE

3) Binomial$(m,p)$ where $m$ is known and $p$ is not.

$$L(p) = \prod_1^n \binom{m}{x_i} p^{x_i}(1-p)^{m-x_i}$$

$$= p^{\sum_1^n x_i} (1-p)^{mn - \sum_1^n x_i} \prod_1^n \binom{m}{x_i}$$

If we denote by $N = m \cdot n$ (total number of iid Bernoulli trials), then

$$L(p) = C(m; x_1, \dots, x_n)\, p^{\sum_1^n x_i} (1-p)^{N - \sum_1^n x_i}$$

So ignoring $C(\,)$ this is the same function we maximized in the Bernoulli case (check that all we used there is that $x_i \geq 0$), and therefore the MLE is:

$$\hat{p} = \frac{1}{N} \sum_1^n x_i = \frac{1}{m \cdot n} \sum_1^n x_i = \frac{\bar{x}}{m}$$

(moment estimator?)

4) Binomial $(m,p)$ both $m$ and $p$ are unknown

Moments estimators:

$\tilde{m} = \bar{X}^2/(\bar{X}-\hat{\sigma}^2)$, $\tilde{p} = \bar{X}/\tilde{m} = (\bar{X}-\hat{\sigma}^2)/\bar{X}$ $(\tilde{m}\tilde{p} = \bar{X})$

If $m$ is known then we saw the MLE is $\hat{p} = \bar{X}/m$.

In other words, if $\ell(p,m) = \sum_{k=1}^{n} \log\left[\binom{m}{x_k} p^{x_k}(1-p)^{m-x_k}\right]$,

then $\operatorname{argmax}_p \ell(p,m) = \bar{X}/m =: \hat{p}(m)$. (fixed sample)

$$\Rightarrow \max_{p,m} \ell(p,m) = \max_m \max_p \ell(p,m)$$

$$= \max_m \ell(\hat{p}(m), m)$$

$$= \max_m \sum_{k=1}^{n} \log\left[\binom{m}{x_k}\left(\frac{\bar{x}}{m}\right)^{x_k}\left(1-\frac{\bar{x}}{m}\right)^{m-x_k}\right]$$

This maximization can be partially achieved numerically:

For each value $m =$ , , ... compute $\ell(\bar{x}/m, m)$ and choose $m$ that gives the value.

How far should the "..." extend to?

Unfortunately, we cannot prescribe an upper bound on $m$ (example in next tutorial). In practice we bound $m$.

Note that the MLE and moment estimators differ!

can be negative!

5) A sample $(x_1, \dots, x_r)$ is drawn from a multinomial $(n; p_1, \dots, p_r)$ distribution where $n$ is known but $\underline{\theta} := (p_1, \dots, p_r)$ is not.

E.g. in $n=1000$ rolls of a die we observed

$$\underline{x} = (105, 207, 253, 152, 182, 101)$$

How would you estimate $p_i$?

$$L(\underline{\theta}) = P_{\underline{\theta}}(\underline{X}_1 = x_1, \dots, \underline{X}_r = x_r)$$

$$= \binom{n}{x_1 \dots x_r} \prod_1^r \theta_i^{x_i}$$

$$\Rightarrow \ell(\underline{\theta}) = \underbrace{\log \binom{n}{x_1, \dots, x_r}}_{\text{const.}} + \underbrace{\sum_1^r x_i \log \theta_i}_{g(\underline{\theta})} \quad (x_i \text{ are fixed})$$

$$\hat{\underline{\theta}} = \underset{\underline{\theta} \in \mathbb{R}^r : \theta_i \geq 0, \sum_1^r \theta_i = 1}{\arg\max} g(\underline{\theta})$$

General method for constrained optimization: Lagrange Multipliers. Alternatively, use ad-hoc optimization.

Consider $\underline{x} = (5, 0, 0) \Rightarrow g(\underline{\theta}) = 5 \log \theta_1 \Rightarrow \hat{\underline{\theta}} = (\ , \ , \ )$

$\underline{x} = (4, 6, 0) \Rightarrow g(\underline{\theta}) = 4 \log \theta_1 + 6 \log \theta_2 \Rightarrow \hat{\underline{\theta}} = (\ , \ , \ )$

WLOG (without loss of generality) all $x_i > 0$: otherwise, if $x_i = 0$ the maximum is attained at $\theta_i = 0$ and we can ignore the $i^{\text{th}}$ coord.

Note that we optimize over $\underline{\theta} \in \mathbb{R}^r$ with $\sum_1^r \theta_i = 1$ (and $\theta_i \geq 0$), therefore for those $\underline{\theta}$,

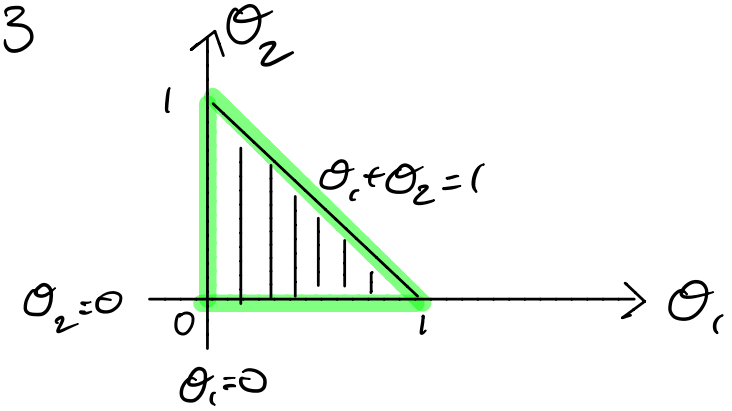
$$g(\theta_1, \dots, \theta_r) = g(\theta_1, \dots, \theta_{r-1}, 1 - \sum_1^{r-1} \theta_i)$$
$$= \sum_1^{r-1} x_i \log \theta_i + x_r \log\left(1 - \sum_1^{r-1} \theta_i\right)$$

Therefore, with $h(\theta_1, \dots, \theta_{r-1}) := g(\theta_1, \dots, \theta_{r-1}, 1 - \sum_1^{r-1} \theta_i)$

$$\hat{\theta} = \operatorname*{argmax}_{\underline{\theta} \in \mathbb{R}^{r-1};\ \theta_i \geq 0,\ \sum_1^{r-1} \theta_i \leq 1} h(\theta_1, \dots, \theta_{r-1})$$

The latter is essentially an unconstrained optimization.

Example. $r = 3$




In particular, the maximum of $h$ on the simplex $\{\underline{\theta} \in \mathbb{R}^{r-1} : \theta_i \geq 0, \sum_1^{r-1} \theta_i \leq 1\}$ is attained either at the boundary ($\theta_i = 0$ for $i = 1, 2, \dots, r-1$, r), or at a critical point ($\sum_1^{r-1} \theta_i = 1 \Updownarrow \theta_r = 1$)

$$\frac{\partial h}{\partial \theta_i} = 0 \qquad i = 1, 2, \dots, r-1$$

Recall $h(\theta_1,\dots,\theta_{r-1}) = \sum_1^{r-1} x_i \log\theta_i + x_r \log(1-\sum_1^{r-1}\theta_i)$ ($\theta_r$ = $1-\sum_1^{r-1}\theta_i$)

Since $x_i > 0$ $\forall i$ the maximum cannot be attained at the boundary. (why?)

At a critical point $\frac{\partial h}{\partial \theta_k} = 0$ for $k=1,2,\dots,r-1$.

$$\iff \frac{x_k}{\theta_k} - \frac{x_r}{1-\sum_1^{r-1}\theta_i} = 0 \quad \text{for } k=1,2,\dots,r-1.$$

$$\iff \frac{x_k}{\theta_k} = \frac{x_r}{\theta_r} = c \qquad k=1,\dots,r-1.$$

But

$$n = \sum_1^r x_k = \sum_1^r c\cdot\theta_k = c$$

$\Rightarrow$ At a critical point $\theta_k = \frac{x_k}{c} = \frac{x_k}{n}$.

$\Rightarrow$ The critical point is unique and therefore has to be a global maximum. (why?)

$\Rightarrow \hat\theta_k = \frac{x_k}{n}$ $k=1,\dots,r$ is the MLE

Same for the moment estimate (tutorial).