Stat 2911 Lecture Notes
Class 30, 2017

Uri Keich

Markov, Chebyshev and the LLN revisited, Estimation in the continuous case, Estimating the normal parameters, Conditional expectation and variance (Rice 4.3-4.4, 8.3-8.5)

Markov's Inequality: if $X \geq 0$ and $t > 0$ then

$$P(X \geq t) \leq E(X)/t \,. \quad (p \to f, \textstyle\sum \to \int)$$

Chebyshev's Inequality: if $X \in L^2$, $t > 0$ then

$$P(|X - E(X)| \geq t) \leq V(X)/t^2 \,.$$

Proof using Markov's Inequality: as-is.

WLLN: if X_i are iid L^1-RVs with $\mu = E(X_i)$ then

$$\forall \varepsilon > 0 \quad P\left(\left|\frac{1}{n}\sum_1^n X_i - \mu\right| > \varepsilon\right) \xrightarrow[n]{} 0 \,.$$

Convergence in prob.; proof for $X_i \in L^2$ as-is:

$$V\left(\frac{1}{n}\sum_1^n X_i\right) = \frac{1}{n} V(X_i) \,.$$

SLLN: $$P\left(\frac{1}{n}\sum_1^n X_i \xrightarrow[n]{} \mu\right) = 1 \,.$$

Estimation (cont. dist.)

Given a sample $x_1, \dots, x_n$ from a dist. F_θ, estimate θ.

Examples:

$N(\theta,1)$ $\quad \theta \in \mathbb{R}$ $\quad f_\theta(x) = \frac{1}{\sqrt{2\pi}} e^{-(x-\theta)^2/2}$

$N(\mu,\sigma^2)$ $\quad \underline{\theta} = (\mu,\sigma^2)$, $\mu \in \mathbb{R}$, $\sigma > 0$ $\quad f_{\underline{\theta}}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

$\mathrm{Exp}(\theta)$ $\quad \theta > 0$ $\quad f_\theta(x) = \theta e^{-\theta x} \cdot \mathbb{1}_{x>0}$

scaled Cauchy $\quad \theta > 0$ $\quad f_\theta(x) = \frac{1}{\pi\theta} \frac{1}{1+(x/\theta)^2}$

($\theta \cdot$ Cauchy)

Method of moments

(exactly as in the discrete case)

Find $\underline{\theta} = g(\mu_1, \dots, \mu_k)$ where $\mu_i = E(\underline{X}^i)$, then

$$\tilde{\underline{\theta}} := g(\tilde{\mu}_1, \dots, \tilde{\mu}_k),$$

where

$$\tilde{\mu}_i = \sum_{j=1}^{n} x_j^i \cdot \frac{1}{n}$$ (moments of the empirical dist.)

Examples. 1) $x_1, \dots, x_n$ is a sample from $N(\theta,1)$

$$\theta = E(X_j) = \mu_1$$

$$\Rightarrow \quad \tilde{\theta} = \tilde{\mu}_1 =$$

2) $N(\mu, \sigma^2) \Rightarrow \mu = E(X_j) = \mu_1$

$$\sigma^2 = V(X_j) = \mu_2 - \mu_1^2$$

$$\Rightarrow \tilde{\mu} = \tilde{\mu}_1 = \bar{x}$$

$$\tilde{\sigma}^2 = \tilde{\mu}_2 - \tilde{\mu}_1^2$$

$$\overset{ex}{=} \sum_j (x_j - \bar{x})^2 \cdot \frac{1}{n}.$$

Considering $\sum_j (X_j - \bar{X})^2 \frac{1}{n}$, this estimator is ...

(use $\frac{1}{n-1}$ to get an unbiased estimator)

3) $\mathrm{Exp}(\theta)$ $\quad E(X_j) =$

$$\Rightarrow \theta = \frac{1}{\mu_1}$$

$$\Rightarrow \tilde{\theta} = \frac{1}{\bar{x}}.$$

4) Scaled Cauchy

$$E(X_j) = \int_{-\infty}^{\infty} x \cdot \frac{1}{\pi\theta} \frac{1}{1 + (x/\theta)^2}\, dx$$

OOPS! $X_j \notin L^1$ (otherwise X_j/θ which is Cauchy would have been L^1) so the scaled Cauchy dist. has no moments!

MLE

In the discrete case given $\underline{x} = (x_1, \dots, x_n)$ we defined the likelihood function $L(\underline{\theta}; \underline{x}) = P_{\underline{\theta}}(\underline{X} = \underline{x})$.

We mostly worked with

$$\ell(\underline{\theta}; \underline{x}) = \log L(\underline{\theta}; \underline{x}) \overset{iid}{=} \sum_j \log P_{\underline{X};\underline{\theta}}(x_j).$$

$P_{\underline{\theta}}(\underline{X} = \underline{x}) = 0$ in the cont. case, so instead we use

$$\ell(\underline{\theta}; \underline{x}) = \log f_{\underline{X};\underline{\theta}}(\underline{x}) \overset{iid}{=} \sum_j \log f_{\underline{X};\underline{\theta}}(x_j).$$

Either way, the MLE is

$$\hat{\underline{\theta}} = \underset{\underline{\theta}}{\arg\max}\ \ell(\underline{\theta}; \underline{x}).$$

Examples. 1) x_i drawn from $N(\theta, 1)$; $f_\theta(x) = \frac{1}{\sqrt{2\pi}} e^{-(x-\theta)^2/2}$

$$\ell(\theta) = \sum_{j=1}^{n} \left[\log \frac{1}{\sqrt{2\pi}} + \left(- \frac{(x_j - \theta)^2}{2}\right)\right]$$

$$\ell'(\theta) = \sum_{j=1}^{n} (x_j - \theta)$$

$$= n(\bar{x} - \theta)$$

$\operatorname{sgn} \ell'$: $+$ 0 $-$ along the θ-axis, with 0 at $\bar{x}$

$\Rightarrow \ell$ has a global max at $\theta = \bar{x} \Rightarrow \hat{\theta} = \bar{x} = \tilde{\theta}$.

30,5

2) $N(\mu,\sigma^2)$ $\qquad f_{\underline{\theta}}(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/2\sigma^2}$ $\quad \begin{pmatrix}\sigma>0\\ \mu\in\mathbb{R}\end{pmatrix}$

$$l(\mu,\sigma^2)=\sum_{j=1}^{n}\left[\log\frac{1}{\sqrt{2\pi}}-\log\sigma-\frac{(x_j-\mu)^2}{2\sigma^2}\right]$$

$$\Rightarrow \qquad \frac{\partial l}{\partial\mu}=\sum_1^n\frac{(x_j-\mu)}{\sigma^2}=\frac{1}{\sigma^2}n(\bar{x}-\mu)$$

Therefore, $\forall\sigma>0$, $\quad \operatorname{sgn}\frac{\partial l}{\partial\mu}$: $\quad \overset{+ \quad 0 \quad -}{\underset{\bar{x}}{\longrightarrow}}\ \mu$

$\Rightarrow$ for each fixed σ, l is maximized for $\mu=\bar{x}$

$\Rightarrow$ Maximizing l over (μ,σ^2) is equivalent to maximizing

$$\varphi(\sigma):=l(\mu=\bar{x},\sigma^2)=\sum_{j=1}^{n}\left[\log\frac{1}{\sqrt{2\pi}}-\log\sigma-\frac{(x_j-\bar{x})^2}{2\sigma^2}\right]$$

$$\varphi'(\sigma)=\frac{\partial l}{\partial\sigma}(\bar{x},\sigma^2)=\sum_{j=1}^{n}\left[-\frac{1}{\sigma}+\frac{(x_j-\bar{x})^2}{\sigma^3}\right]$$

$$=\frac{-n\sigma^2+\sum_1^n(x_j-\bar{x})^2}{\sigma^3}$$

$$=\frac{n\left[\sum_1^n(x_j-\bar{x})^2\cdot\frac{1}{n}-\sigma^2\right]}{\sigma^3}$$

$\operatorname{sgn}(\varphi'(\sigma))$: $\quad \overset{0 \quad + \quad\quad -}{\underset{\frac{1}{n}\sum_1^n(x_j-\bar{x})^2}{\longrightarrow}}\ \sigma^2$

$\Rightarrow \varphi$ has a global max for $\sigma^2=\frac{1}{n}\sum_1^n(x_j-\bar{x})^2$

Therefore, the MLE is:

$$\hat{\mu} = \bar{x} = \tilde{\mu}$$

$$\hat{\sigma}^2 = \sum_{i=1}^{n} (x_j - \bar{x})^2 \cdot \frac{1}{n} = \tilde{\sigma}^2$$

Clearly, $\hat{\mu} \in \mathbb{R}$ and $\hat{\sigma}^2 \geq 0$ but can $\hat{\sigma}^2 = 0$?

Claim. $P\left(\sum_{i=1}^{n} (X_j - \bar{X})^2 = 0\right) = 0$

Proof. $\sum_{i=1}^{n} (X_j - \bar{X})^2 = 0 \iff X_j = \bar{X} \;\; \forall j$

but $\forall i \neq j \quad P(X_i = X_j) = 0$ (why?),

so $P(X_1 = \dots = X_n) = 0$.

Conditional Expectation

Discrete case: for X with pmf $p_X(x) > 0$

$$E(Y|X=x) = \sum_j y_j \, p_{Y|X}(y_j|x)$$

If X and Y have a joint density f_{XY} which is cont. then for any x with $f_X(x) > 0$,

$$f_{Y|X}(y|x) := \frac{f_{XY}(x,y)}{f_X(x)} .$$

This defines the conditional dist. of Y given $X=x$, and we can study features of this dist.

Claim. If $Y \in L^1$ then for "almost every" x s.t. $f_X(x) > 0$,

$$\int_{-\infty}^{\infty} |y| \, f_{Y|X}(y|x) \, dy < \infty .$$

In particular, for such x's we can define the conditional expectation of Y given $X=x$ as

$$E(Y|X=x) = \int_{-\infty}^{\infty} y \, f_{Y|X}(y|x) \, dy .$$

Proof. (sketch)

Let $g(x) = \int_{-\infty}^{\infty} |y| \, f_{Y|X}(y|x) \, dy$, so $g : X(\Omega) \to \mathbb{R}^+ \cup \{\infty\}$.

$$\int_{\mathbb{R}} g(x) f_X(x) \, dx = \int_{\mathbb{R}}\int_{\mathbb{R}} |y| \, f_{Y|X}(y|x) \, dy \, f_X(x) \, dx = \int_{\mathbb{R}}\int_{\mathbb{R}} |y| \, f_{XY}(x,y) \, dy \, dx$$

$$\overset{?}{=} \int_{\mathbb{R}} |y| \int_{\mathbb{R}} f_{XY}(x,y)\,dx\,dy = \int |y|\, f_Y(y)\,dy \overset{?}{<} \infty$$

$\Rightarrow g(x) < \infty$ for almost every (a.e.) x for which $f_X(x) > 0$ ($\forall x$ with $f_X(x)>0$, if g and f_X are cont.) $\square$

More generally, if $h: \mathbb{R} \to \mathbb{R}$ and $h(Y) \in L^1$ then

$$E[h(Y) | X = x] = \int_{-\infty}^{\infty} h(y)\, f_{Y|X}(y|x)\,dy$$

is well-defined (for a.e. x with $f_X(x) > 0$). (same proof)

In particular, if $Y \in L^2$ then for a.e. x with $f_X(x) > 0$,

$$E(Y^2 | X = x) = \int_{-\infty}^{\infty} y^2 f_{Y|X}(y|x)\,dy < \infty.$$

Hence, we can define

$$V(Y|X=x) = \int_{-\infty}^{\infty} [y - E(Y|X=x)]^2 f_{Y|X}(y|x)\,dy$$

$$\overset{ex}{=} E(Y^2|X=x) - E^2(Y|X=x).$$

As in the discrete case, with $h(x) = E(Y|X=x)$ we can define the <u>RV</u>

$$E(Y|X) = h(X).$$

Equivalently,

$$E(Y|X)(\omega) = E(Y|X=x) \quad \text{where } x = X(\omega).$$

<u>Claim.</u> If X and Y are RVs defined on Ω and $Y \in \mathcal{L}'$

$$E[E(Y|X)] = \qquad .$$

<u>Proof.</u> If X and Y are jointly cont. with density f_{XY} and if $f_X(x) > 0 \ \forall x \in \mathbb{R}$ then the proof in the discrete case applies with $p \to f$, $\Sigma \to \int$.

The condition $f_X(x) > 0 \ \forall x \in \mathbb{R}$ only serves to simplify notations, but existence of f_{XY} is critical for <u>this proof</u>.

So far we studied jointly distributed RVs that are either discrete (p_{XY}) or jointly cont. (f_{XY}). However, we often encounter mixed dists.

Examples.

1) $U \sim U(0,1)$ $\qquad X | U = p \sim \text{Binom}(m, p)$