Lecture 01/24/20

Dr. Syring

Point Estimation

Last time we discussed histogram estimates for the distribution P of a random variable X . Today we are interested in learning about some aspect of P such as the mean or the median, not necessarily P itself.

From MATH 3200 you learned that certain estimators are usually "good" for estimating certain charactersitics of a distribution (population).

For example, the sample mean $\bar{X}$ of iid data is good for estimating the population mean μ whenever μ exists (is finite). By "good" we might mean, for example, unbiased.

$$E(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu,$$

since the X_i are identically distributed (they have the same mean.)

But, how, in general, do we come up with "good" point estimators?

Estimating equations and likelihood

Often times we can derive point estimators as the minimizer (or maximizer, or root) of a function $\ell(\theta; x)$ that depends on data and a parameter. Sometimes we can use intuition to come up with $\ell(\theta; x)$ or we define it by considering a family of possible probability models.

One way to find estimating equations is to choose a particular class of probability models for P . For example, we might assume that $P \in \{\Phi(x; \mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$, that is, we assume P is a normal distribution. Denote the pdf of X by $f(x; \theta)$ which depends on some parameter θ . Then, the likelihood is $L(\theta; x) = f(x; \theta)$; it is simply the density function, but while the density is viewed as a funciton of x for fixed θ the likelihood is viewed as a function of the parameter for fixed data. Often times we would rather work with the loglikelihood $\ell(\theta; x) = \log L(\theta; x)$. It is important to note that since the logarithm is a monotonic and one-to-one transformation the loglikelihood and likelihood are maximized at the same point.

Maximum likelihood estimation

Suppose $f(x;\theta^\star)$ is the pdf of a r.v. X and denote $\ell(\theta;x) = \log f(x;\theta)$ for a generic θ . The "maximum likelihood" estimation technique says to use

$$\hat{\theta} := \arg\max_{\theta} \ell(\theta;x)$$

for observed (non-random) x . Typically, $\ell(\theta;x)$ is differentiable so thatt he maximizer can be found by solving

$$\frac{\partial}{\partial\theta}\ell(\theta;x) = 0$$

in θ . This is called the "estimating equation" in your textbook.

Some support for MLE

So, why would this technique work?

Since $\log(t) \leq t - 1$ we have

$$E \log\left(\frac{f(x;\theta)}{f(x;\theta^\star)} \right) = \int \log\left(\frac{f(x;\theta)}{f(x;\theta^\star)} \right) f(x;\theta^\star) dx$$

$$\leq \int \left(\frac{f(x;\theta)}{f(x;\theta^\star)} - 1 \right) f(x;\theta^\star) dx$$

$$= \int f(x;\theta) - f(x;\theta^\star) dx = 0.$$

Therefore, $E \log\left(\frac{f(x;\theta)}{f(x;\theta^\star)} \right)$ is maximized at $\theta = \theta^\star$ which implies that estimates

$$\hat{\theta} = \arg\max_\theta \sum_i \log\left(\frac{f(x_i;\theta)}{f(x_i;\theta^\star)} \right)$$

are sensible for estimating θ .

MLE Examples

Poisson distribution.

We want to maximize the loglikelihood

$$\sum_i \log\left(\frac{\lambda^{x_i} e^{-\lambda}}{x_i!}\right).$$

Take the derivative and set equal to zero

$$\frac{\partial}{\partial\lambda}\sum_i \log\left(\frac{\lambda^{x_i} e^{-\lambda}}{x_i!}\right) = \frac{\partial}{\partial\lambda}\sum_i [x_i \log\lambda - \lambda - \log x_i!]$$

$$= \frac{1}{\lambda} n\bar{x} - n$$

which implies $\hat{\lambda} = \bar{x}$.

MLE Examples

Laplace distribution $f(x) = \frac{1}{2\sigma} e^{-\frac{|x-\mu|}{\sigma}}$, $x \in \mathbb{R}$, $\sigma > 0$.

The loglikelihood is

$$-n \log 2\sigma - \sum_i \frac{|x_i - \mu|}{\sigma}.$$

For a fixed σ the loglikelihood is maximized in μ by taking μ to be any value such that $\#\{x_i > \mu\} = \#\{x_i < \mu\}$, which means $\hat{\mu} = \tilde{x}$, the sample median (not unique). Tricky to see why. $|x_i - \mu|$ is differentiable everywhere except at the points $x_1, x_2, \ldots, x_n$ with derivative $-sign(x_i - \mu)$. The sum of these is zero at the samle median.

To maximize with respect to σ we can take the partial derivative and set equal to zero,

$$\frac{\partial}{\partial \sigma}[-n \log 2\sigma - \sum_i \frac{|x_i - \hat{\mu}|}{\sigma}]$$

$$= -n/\sigma + \sum_i |x_i - \hat{\mu}|/\sigma^2$$

which implies $\hat{\sigma} = \sum_i |x_i - \hat{\mu}|/n$.

An intuitive estimating equation for a mean

Suppose X has a finite mean and variance. Consider the function $\ell(x;\theta) = (x-\theta)^2$ for some generic value θ , and consider $\ell(x;\mu) = (x-\mu)^2$ where $\mu = E(X)$.

Why might this be a good function to use to estimate μ ? Well, the mean is a measure of the "center" of the distribution of X . So, if we pick θ to be near the mean, then the average squared distance from θ to the data points should be small, whereas if we pick θ far from the mean, it will probably be far from many or all of the data points, and $\ell(x;\theta)$ will tend to be large by comparison.

With a little algebra

$$E[\ell(X;\theta) - \ell(X;\mu)] = E(X^2) - 2E(X)\theta + \theta^2 - E(X^2) + 2E(X)\mu - \mu^2$$

$$= \theta^2 - 2\mu\theta + \mu^2$$

$$= (\theta - \mu)^2 > 0,$$

which implies $E[\ell(X;\theta)]$ is minimized when $\theta = \mu$.

Therefore, it is sensible to estimate μ by minimizing the function

$$\frac{1}{n}\sum_i (x_i - \theta)^2$$

with respect to θ , for some data $x_1, x_2, \ldots, x_n$.

To find the point estimator, take the partial derivative of $\frac{1}{n}\sum_i (x_i - \theta)^2$ with respect to θ , set it equal to zero and solve:

$$\frac{\partial}{\partial \theta}\left[\frac{1}{n}\sum_i (x_i - \theta)^2\right]$$

$$= -\frac{2}{n}\sum_i (x_i - \theta).$$

The estimating equation is $-\frac{2}{n}\sum_i (x_i - \theta) = 0$, which means that $\hat{\theta} = \bar{x}$. And, we know $\bar{x}$ is a good point estimator for the mean.