UNIVERSITY OF LONDON

Goldsmiths
UNIVERSITY OF LONDON

Oct. 10^{th} 2021

Estimating Parameters: Maximum Likelihood and Moments

Study Notes | Written by Larry Cui

If a phenomenon is likely to be described by a kind of distribution function, we might want to know the best parameters for the function. There're two ways to estimate the parameters based on a collection of samples, the method of maximum likelihood and the method of moments.

1 The Method of Maximum Likelihood

This method relies on an assumption. When we times together all probabilities of the items in a sample of size n, we get the product of those probabilities as a function of unknown parameter(s). The assumption is that the sample is best reflecting the population distribution, thus what we get in the sample comes with the most likelihood in the population. So this method is to find the parameter(s) that will give the maximum value of the product function.

1.1 First Derivative Test for Single

As a convention, we use L to denote the product of probabilities, and if there's only one unknown parameter, the task becomes finding the parameter when the first derivative of L or $\ln L$ equals to 0.

Example (1): estimate λ of a Poisson distribution if a size of n sample is given.

We know a Poisson distribution is $p_X(k) = e^{\lambda}\lambda^k/k!$, $k = 0, 1, 2, \ldots$, so

$$L(\lambda) = \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{k_i}}{k_i!} = e^{-n\lambda}\lambda^{\sum_{i=1}^{n} k_i}\frac{1}{\prod_{i=1}^{n} k_i!}$$

$$\ln L(\lambda) = -n\lambda + \left(\sum_{i=1}^{n} k_i\right)\ln\lambda - \ln\prod_{i=1}^{n} k_i!$$

and

$$\frac{d\,\ln L(\lambda)}{d\lambda} = -n + \sum_{i=1}^{n} k_i/\lambda$$

If we use $\overline{k}$ to denote $\sum_{i=1}^{n} k_i/n$, then $\lambda = \overline{k}$ when the above first derivative equals 0.

1.2 Partial Derivative Test for Double or More

If there're two or more unknown parameters, it is simply a task of finding local extremes by partial first derivative for all unknown variables.

Example (2): suppose a random sample of size n is drawn, find μ and σ^2 of a normal pdf.

We start by finding L and $\ln L$:

$$\begin{aligned} L(\mu, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\frac{(y_i-\mu)^2}{\sigma^2}\right] \\ &= (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right] \end{aligned}$$

and

$$\ln L(\mu, \sigma^2) = -\frac{n}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2$$

Now we can differentiate with μ and σ respectively:

$$\frac{\partial \ln L(\mu, \sigma^2)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\mu)$$

and

$$\frac{\partial \ln L(\mu, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i-\mu)^2$$

A local maximum is where both partial derivatives are zero, so the next task is to find μ and σ from below two equations:

$$\begin{cases} \dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{n}(y_i-\mu) = 0 \quad \rightarrow \quad \sum_{i=1}^{n}(y_i-\mu) = 0 \quad \rightarrow \quad \sum_{i=1}^{n} y_i - n\mu = 0 \\ -\dfrac{n}{2\sigma^2} + \dfrac{1}{2\sigma^4}\displaystyle\sum_{i=1}^{n}(y_i-\mu)^2 = 0 \quad \rightarrow \quad -n\sigma^2 + \sum_{i=1}^{n}(y_i-\mu)^2 = 0 \end{cases}$$

so we have:

Estimator for Normal Distribution Given a sample of size n, estimated μ and σ^2:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} y_i = \overline{y} \quad \text{and} \quad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2$$

1.3 When the Derivative Test Fails

Sometimes the derivative test just fails to produce the result for parameters, when a close observation on order statistics technique kicks in.

Example (3): a size of n sample is obtained for a pdf $f_Y(y;\theta) = e^{-(y-\theta)}, y \geqslant \theta$, find θ.

From the information above, we have

$$L(\theta) = \prod_{i=1}^{n} e^{-(y-\theta)} = \exp\left[-\sum_{i=1}^{n} y_i + n\theta\right]$$

and

$$\ln L(\theta) = -\sum_{i=1}^{n} y_i + n\theta$$

But when we differentiate $\ln L(\theta)$, we get n, which is pointless for the test. Take a second look at the equation above, we can see that in order to have $\ln L(\theta)$, we need to make the θ as big as possible. Taking into the account that θ must be less or equal to y_i, the conclusion is $\theta = y_{\min}$.

2 The Method of Moments

This procedure for estimating parameters was proposed near the turn of the twentieth century by British statistician, Karl Pearson. Suppose that Y is a continuous random variable and that its pdf is a function of unknown parameters, $\theta_1, \theta_2, \ldots, \theta_s$. The expected value of its moments can be listed as follows:

$$\begin{aligned} E(Y^1) &= \int_{-\infty}^{\infty} y^1 \cdot f_Y(y; \theta_1, \theta_2, \ldots, \theta_s)\, dy \\ E(Y^2) &= \int_{-\infty}^{\infty} y^2 \cdot f_Y(y; \theta_1, \theta_2, \ldots, \theta_s)\, dy \\ &\vdots \\ E(Y^s) &= \int_{-\infty}^{\infty} y^s \cdot f_Y(y; \theta_1, \theta_2, \ldots, \theta_s)\, dy \end{aligned}$$

Apparently, it's an easy task of solving s parameters from s equations. By intuition, we can approximate $E(Y^k)$ by $\dfrac{1}{n}\sum_{i=1}^{n} Y^k$, for $k = 1, 2, \ldots, s$.

Example (4): find parameters r and p of a negative binomial distribution.

We have the function form

$$p_X(k; p, r) = \binom{k+r-1}{k}(1-p)^k p^r, k = 0, 1, 2, \ldots$$

and

$$E(X) = \frac{r(1-p)}{p} \quad \text{and} \quad \text{Var}\,(X) = \frac{r(1-p)}{p^2}$$

So $E(X^2)$ is a straight forward calculation

$$E(X^2) = \text{Var}\,(X) + E(X)^2 = \frac{r(1-p) - r^2(1-p)^2}{p^2} = \frac{r(1-p)(1-r+rp)}{p^2}$$

Two equations for the parameters (solution omitted):

$$\begin{cases} \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} k_i = \overline{k} = \dfrac{r(1-p)}{p} \\ \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} k_i^2 = \overline{k^2} = \dfrac{r(1-p)(1-r+rp)}{p^2} \end{cases}$$

3 Properties of Estimators - sample standard deviation

The method of maximum likelihood and moments do not always yield the same answer. It comes naturally to people the question: which one should we use over the other? Or what qualities should a "good" estimator have? No matter which method we are going to use, the estimator, $\hat{\theta}$, is a function of random variables, let's say Y_i, of sample size n. As such, any $\hat{\theta}$ is also a random variable, and usually has its own pdf, expected value and variance. We define an estimator $\hat{\theta}$ is "unbiased" if $E(\hat{\theta}) = \theta$ for all θ.

It should be noted that it's quite difficult to get the expected value and variance of $\hat{\theta}$ by direct summation, since we don't have its pdf on hand. However, we can go around this problem by dividing $\hat{\theta}$ into the function of variables Y_i, and conquering each and every expected value and variance of Y_i get us the desired result for $E(\hat{\theta})$.

Example (5): Given a random sample $Y_1, Y_2, \ldots, Y_n$ from a normal distribution whose parameters μ and σ^2 are both unknown, the estimator for σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

Use expected value to check if $\hat{\sigma}^2$ is unbiased.

First of all, we have two conclusions about $\overline{Y}$:

(a) $E(\overline{Y}) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\cdot E(Y_1 + Y_2 + \cdots + Y_n) = E(Y_i)$

(b) $\text{Var}\,(\overline{Y}) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\text{Var}\,(Y_1 + Y_2 + \cdots + Y_n) = \frac{1}{n}\text{Var}\,(Y_i)$

Now we write the $\hat{\theta}$ in function form to find the expected value as a function of real θ (here the θ refers to σ^2):

$$\begin{aligned}
E(\hat{\sigma}^2) &= E\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2\right] \\
&= E\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i^2 - 2Y_i\overline{Y} + \overline{Y}^2)\right] \\
&= E\left[\frac{1}{n}\left(\sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2\right)\right] && \text{note: } \sum_{i=1}^{n} Y_i = n\overline{Y} \\
&= \frac{1}{n}\left[\sum_{i=1}^{n} E(Y_i^2) - nE(\overline{Y}^2)\right] && \text{note: } E(\overline{Y}^2) = \text{Var}\,(\overline{Y}) + E(\overline{Y})^2 = \frac{1}{n}\sigma^2 + \mu^2 \\
&= \frac{1}{n}\left[\sum_{i=1}^{n}(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right] \\
&= \frac{n-1}{n}\cdot\sigma^2
\end{aligned}$$

Since $E(\hat{\sigma}^2) \neq \sigma^2$, we say $\hat{\sigma}^2$ is "biased". The way to correct its "biasedness" is quite straight forward, we simply times $\hat{\sigma}^2$ by $n/(n-1)$.

By convention, this unbiased version of estimator in a normal distribution has a special

name, S^2, and is referred to as the *sample variance*:

$$S^2 = \frac{n}{n-1} \cdot \frac{1}{n} \sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

A related concept is *sample standard deviation*, though $E(S) \neq \sigma$:

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y})^2}$$

4 Common Distributions and Their Properties

Distribution	Function	Mean	Variance	M.G.F.
Uniform	$f(y) = \frac{1}{\theta_2 - \theta_1}; \theta_1 \leqslant y \leqslant \theta_2$	$\frac{\theta_1 + \theta_2}{2}$	$\frac{(\theta_1 + \theta_2)^2}{12}$	$\frac{e^{t\theta_2} - e^{t\theta_1}}{t(\theta_2 - \theta_1)}$
Normal	$f(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\frac{(y_i - \mu)^2}{\sigma^2}\right]$ $-\infty < y < \infty$	μ	σ^2	$\exp\left(\mu t + \frac{t^2\sigma^2}{2}\right)$
Exponential	$\lambda e^{-\lambda y}; y > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{1}{1 - t/\lambda}$
Gamma	$f_Y(y) = \frac{\lambda^r}{\Gamma(r)} y^{r-1} e^{-\lambda y}, y \geqslant 0$ where $\Gamma(r) = \int_0^\infty y^{r-1} e^{-y}\, dy$	$\frac{r}{\lambda}$	$\frac{r}{\lambda^2}$	$\frac{1}{(1 - t/\lambda)^r}$
Binomial	$p(k) = \binom{n}{k} p^k (1-p)^{n-k}$ $k = 0, 1, 2, \ldots$	np	$np(1-p)$	$[pe^t + (1-p)]^n$
Hpyer Geometric	$p(k) = \frac{\binom{r}{k}\binom{N-r}{n-k}}{\binom{N}{n}}$ $k = 0, 1, \ldots, n$ if $n \leqslant r$ $k = 0, 1, \ldots, r$ if $n > r$	$\frac{nr}{N}$	$np(1-p)\left(\frac{N-n}{N-1}\right)$ let $p = \frac{r}{N}$	
Geometric	$p(k) = p(1-p)^{k-1}; k = 1, 2, \ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{pe^t}{1 - (1-p)e^t}$
Negative Binomial	$p(k) = \binom{k-1}{r-1} p^r (1-p)^{k-r};$ $k = r, r+1, \ldots$	$\frac{r}{p}$	$\frac{r(1-p)}{p^2}$	$\left[\frac{pe^t}{1 - (1-p)e^t}\right]^r$
Poisson	$p(k) = \frac{\lambda^k e^{-\lambda}}{k!}; k = 0, 1, 2, \ldots$	λ	λ	$\exp\left[\lambda(e^t - 1)\right]$