STAT3610 Maximum Likelihood Estimation   Reference: Chapter 6 of Devore

S. Maghsoodloo

Statistical Inference

By statistical inference we mean Estimation and Test of Hypothesis. Estimation consists of point and interval estimation. The entire Chapters 6, 7, 8 & 9 of Devore (8e) are devoted to point estimators and (confidence) interval estimation. A point estimator of a population vector parameter θ (such as the population mean μ and standard deviation σ) is a sample statistic, $\hat{\theta}$, which is a rv with a frequency function (or SMD = Sampling Distribution) that depends on the underlying distribution, $f(x;\theta)$, which is also called the parent population. If the vector $\theta = [\theta_1 \quad \theta_2 \quad \theta_3 \ldots \quad \theta_m]'$, appearing in the expression of $f(x;\theta)$, then the underlying population has m unknown parameters to be estimated. In Statistical applications, the value of m rarely exceeds 3. For the exponential density $\lambda e^{-\lambda t}$, $\theta = \lambda$, is a scalar, i.e., m = 1. For a normal parent population $\theta = [\mu \quad \sigma]'$ is a 2×1 vector so that m = 2, and for a Weibull TTF (Time-to-Failure) the vector-parameter $[\delta \quad \theta \quad \beta]'$ is 3×1 for which m = 3, where θ now is the characteristic-life and β is the slope (or shape). Our objective in point estimation is to use sample data to obtain an "accurate" vector point estimator of θ, denoted by $\hat{\theta}$. The Accuracy of a single point estimator can be measured through several properties such as bias, consistency, mean square error (MSE), and efficiency. All these properties are well defined and explained in my Chapter 6 notes. There are several methods of obtaining point estimators :(1) Method of Moments, (2) Maximum Likelihood estimation (MLE), and (3) Least-squares Estimation. In Reliability Engineering the MLEs are often used because all MLEs in the universe have the asymptotic property that their SMD approaches normality as the sample size $n \to \infty$, and they are also asymptotically unbiased. Further, they have the very nice property that if $\hat{\theta}$ is a MLE of θ, then $h(\hat{\theta})$ is a MLE of $h(\theta)$ for any functional form h. Thus, we will show how to obtain ML estimates for population parameters in the next section.

Maximum Likelihood Estimators

Let $f(x;\theta)$ represent the frequency function of the population from which a random sample of size n is drawn. The occurrence of the sample ($x_1, x_2, ..., x_n$) has a likelihood (or a Pr of) $[f(x_1; \theta) dx_1]\times[f(x_2; \theta)dx_2]\times ... \times [f(x_n; \theta)dx_n] = \prod_{i=1}^{n}[f(x_i;\theta)\,dx_i]$. The quantity $\prod_{i=1}^{n} f(x_i;\theta)$ in this last Pr statement is called the likelihood function (LF), and after the sample is drawn the sample values $x_1, x_2, ..., x_n$ are known numbers (no longer rvs), then the likelihood $\prod_{i=1}^{n} f(x_i;\theta)$ is only a function of θ which I denote by $\mathsf{L}(\theta)$. That is, $\mathsf{L}(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$, L for likelihood. The maximum likelihood estimator (MLE) of the vector θ is obtained by maximizing $\mathsf{L}(\theta)$ with respect to (wrt) all the m parameters in $\mathsf{L}(\theta)$. Further, for notational convenience, let $L(\theta) = \ln[\mathsf{L}(\theta)]$, i.e., $L(\theta)$ is the natural logarithm of the likelihood function. Below I will show that maximizing $L(\theta)$ is equivalent to maximizing $\mathsf{L}(\theta)$:

$\partial L(\theta)/\partial\theta_k = \frac{1}{\mathsf{L}(\theta)}\,\partial \mathsf{L}(\theta)/\partial\theta_k$; Since $0 < \mathsf{L}(\theta) < \infty$, i.e., $\mathsf{L}(\theta)$ is finite, then $\partial L(\theta)/\partial\theta_k = 0$ iff $\partial \mathsf{L}(\theta)/\partial\theta_k = 0$. Note that most authors will use the notation $L(\theta)$ for the likelihood function itself, but I am departing a bit from tradition only because of notational convenience; further, nearly always it is the $L(\theta) = \ln[\mathsf{L}(\theta)] = \log_e$-Likelihood that is maximized instead of $\mathsf{L}(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$.

MLE for the Two-Parameter Exponential Underlying Distribution

Suppose the Time-Headway in a traffic flow [see the Example 4.5 on p. 141 of Devore(8e)] has the underlying distribution given by $f(t) = \lambda e^{-\lambda(t-\delta)}$, $t \ge \delta \ge 0$, where λ is the flow-rate per second and $\delta \ge 0$ is the minimum Headway measured in seconds. For a complete observed sample of size n, the likelihood function is given by $\mathsf{L}(\theta) = \mathsf{L}(\lambda, \delta) = \prod_{i=1}^{n} f(t_i;\lambda,\delta) = \prod_{i=1}^{n} \lambda e^{-\lambda(t_i-\delta)} = \lambda^n e^{-\lambda\sum_{i=1}^{n}(t_i-\delta)} \rightarrow$

$L(\theta) = L(\lambda, \delta) = \ln[\lambda^n e^{-\lambda\sum_{i=1}^{n}(t_i-\delta)}) = n\times\ln\lambda - \lambda\sum_{i=1}^{n}(t_i-\delta) \rightarrow \partial L(\lambda,\delta)/\partial\lambda = n/\lambda - \sum_{i=1}^{n}(t_i-\delta) \xrightarrow{\text{Set to}} 0$

$\rightarrow n/\hat{\lambda} = \sum_{i=1}^{n}(t_i-\hat{\delta})$. This last relationship shows that the MLE of λ given by $\hat{\lambda} = n/\sum_{i=1}^{n}(t_i-\hat{\delta})$ because

$\partial^2 L(\lambda,\delta)/\partial\lambda^2 = -n/\lambda^2 < 0$, implying that the LF (likelihood function) is strictly concave, and hence the optimum is a point of maximum. Cleary, the point estimate of the rate-parameter λ, $n/\sum_{i=1}^{n}(t_i - \hat{\delta})$, depends on the MLE of Minimum-life δ. So, we now differentiate $L(\lambda, \delta) = n\times\ln\lambda - \lambda\sum_{i=1}^{n}(t_i - \delta)$ wrt δ in order to obtain the MLE of δ. Partial differentiation yields $\partial L(\lambda,\delta)/\partial\delta = -\lambda\sum_{i=1}^{n}(0-1) \xrightarrow{\text{Set to}} 0$; but this last does not yield any estimator for δ. Next, we examine $L(\lambda, \delta) = n\times\ln\lambda - \lambda\sum_{i=1}^{n}(t_i - \delta) = n\times\ln\lambda - \lambda\sum_{i=1}^{n}t_i + n\lambda\delta$; this last, unfortunately shows that the sample likelihood is maximum when δ is maximum, i.e., δ should be estimated by $x_{(n)}$. However, this is impossible and contradictory because all observed sample values $t_1, t_2, \ldots, t_n$ must be at least as large as δ (recall δ is minimum-life). Put differently, if we let $\hat{\delta} = x_{(n)}$, then the LF becomes identically equal to zero. Thus, δ being the minimum-life (in this case, the Min-Headway), its MLE has to be the value of the 1st-order statistic $x_{(1)}$. For practical applications, the above MLE, $x_{(1)}$, should be modified to $[0.85\times x_{(1)}, 0.90\times x_{(1)}]$.

Exercise 29 on page 265 of Devore(8e). In this example, n = 10 Headway-times are obtained with values 3.11, 0.64, 2.55, 2.20, 5.44, 3.42, 10.39, 8.93, 17.82, and 1.30 seconds. Then, $\hat{\delta} = x_{(1)} = 0.64$, $\sum_{i=1}^{10} t_i = 55.80$, $\sum_{i=1}^{n}(t_i - \hat{\delta}) = 55.80 - 6.4 = 49.40$, $\hat{\lambda} = n/\sum_{i=1}^{n}(t_i - \hat{\delta}) = 10/49.40 = 0.20243$ per second. That is, $\hat{\lambda} = 0.20243$/second. The $cv_X = S_x/5.580 = 5.35713024/5.580 = 95.01\%$, which is close to that of the exponential of 100%, i.e., the cv_X does not contradict the exponentiality of the data. A more practical point estimator is $\hat{\delta} = [0.544, 0.90\times x_{(1)} = 0.5760$ seconds].

For MLE of Normal and Poisson Parameters study page 260 of Devore (8e).

Bonus HW: Work Exercise 22 on p. 264 of Devore(8e). My Answers: **(a)** $\hat{\theta}_{\text{Moment}}$ **= 3.000;** $\hat{\theta}_{\text{MLE}}$ **= 3.11607.**