Chapter 5

Point Estimation

5.1 Maximum Likelihood Estimators

A point estimator gives a single value as an estimate of a parameter. For example, $\overline{Y} = 10.54$ is a point estimate of the population mean μ. An interval estimator gives a range (L_n, U_n) of reasonable values for the parameter. Confidence intervals, studied in Chapter 9, are interval estimators. The most widely used point estimators are the maximum likelihood estimators.

Definition 5.1. Let $f(\boldsymbol{y}|\boldsymbol{\theta})$ be the pmf or pdf of a sample $\boldsymbol{Y}$ with parameter space Θ. If $\boldsymbol{Y} = \boldsymbol{y}$ is observed, then the **likelihood function** $L(\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta}|\boldsymbol{y}) = f(\boldsymbol{y}|\boldsymbol{\theta})$. For each sample point $\boldsymbol{y} = (y_1, ..., y_n)$, let $\hat{\boldsymbol{\theta}}(\boldsymbol{y}) \in \Theta$ be the parameter value at which $L(\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta}|\boldsymbol{y})$ attains its maximum as a function of $\boldsymbol{\theta}$ with $\boldsymbol{y}$ held fixed. Then the maximum likelihood estimator (**MLE**) of the parameter $\boldsymbol{\theta}$ based on the sample $\boldsymbol{Y}$ is $\hat{\boldsymbol{\theta}}(\boldsymbol{Y})$.

The following remarks are important. I) It is crucial to observe that the likelihood function is a function of $\boldsymbol{\theta}$ (and that $y_1, ..., y_n$ act as fixed constants). Note that the pdf or pmf $f(\boldsymbol{y}|\boldsymbol{\theta})$ is a function of n variables while $L(\boldsymbol{\theta})$ is a function of k variables if $\boldsymbol{\theta}$ is a $k \times 1$ vector. Often $k = 1$ or $k = 2$ while n could be in the hundreds or thousands.

II) If $Y_1, ..., Y_n$ is an independent sample from a population with pdf or pmf $f(y|\boldsymbol{\theta})$, then the likelihood function

$$L(\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta}|y_1, ..., y_n) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta}). \tag{5.1}$$

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f_i(y_i|\boldsymbol{\theta})$$

if the Y_i are independent but have different pdfs or pmfs.

III) If the MLE $\hat{\boldsymbol{\theta}}$ exists, then $\hat{\boldsymbol{\theta}} \in \Theta$. Hence if $\hat{\boldsymbol{\theta}}$ is not in the parameter space Θ, then $\hat{\boldsymbol{\theta}}$ is not the MLE of $\boldsymbol{\theta}$.

IV) If the MLE is unique, then the MLE is a function of the minimal sufficient statistic. See Levy (1985) and Moore (1971). This fact is useful since exponential families tend to have a tractable log likelihood and an easily found minimal sufficient statistic.

Theorem 5.1: Invariance Principle. If $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$, then $h(\hat{\boldsymbol{\theta}})$ is the MLE of $h(\boldsymbol{\theta})$ where h is a function with domain Θ.

This theorem will be proved in Section 5.4.

There are **four commonly used techniques** for finding the MLE.

- Potential candidates can be found by differentiating log $L(\boldsymbol{\theta})$, the log likelihood.
- Potential candidates can be found by differentiating the likelihood $L(\boldsymbol{\theta})$.
- The MLE can sometimes be found by direct maximization of the likelihood $L(\boldsymbol{\theta})$.
- **Invariance Principle**: If $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$, then $h(\hat{\boldsymbol{\theta}})$ is the MLE of $h(\boldsymbol{\theta})$.

The one parameter case can often be solved by hand with the following technique. To show that $\hat{\theta}$ is the MLE of θ is equivalent to showing that $\hat{\theta}$ is the global maximizer of $\log L(\theta)$ on Θ where Θ is an interval with endpoints a and b, not necessarily finite. Show that $\log L(\theta)$ is differentiable on (a, b). Then show that $\hat{\theta}$ is the unique solution to the equation $\frac{d}{d\theta} \log L(\theta) = 0$ and that the 2nd derivative evaluated at $\hat{\theta}$ is negative: $\left.\frac{d^2}{d\theta^2} \log L(\theta)\right|_{\hat{\theta}} < 0$. See Remark 5.1V below.

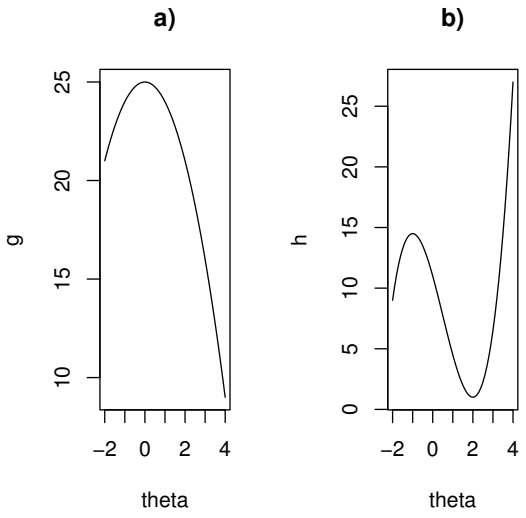

Figure 5.1: The local max in a) is a global max, but not for b).

Remark 5.1. From calculus, recall the following facts. I) If the function h is continuous on an interval $[a, b]$ then both the max and min of h exist. Suppose that h is continuous on an interval $[a, b]$ and differentiable on (a, b). Solve $h'(\theta) \equiv 0$ and find the places where $h'(\theta)$ does not exist. These values are the **critical points**. Evaluate h at a, b, and the critical points. One of these values will be the min and one the max.

II) Assume h is continuous. Then a critical point θ_o is a local max of $h(\theta)$ if h is increasing for $\theta < \theta_o$ in a neighborhood of θ_o and if h is decreasing for $\theta > \theta_o$ in a neighborhood of θ_o (and θ_o is a global max if you can remove the phrase "in a neighborhood of θ_o"). The first derivative test is often used.

III) If h is strictly concave ($\frac{d^2}{d\theta^2} h(\theta) < 0$ for all $\theta \in \Theta$), then any local max of h is a global max.

IV) Suppose $h'(\theta_o) = 0$. The 2nd derivative test states that if $\frac{d^2}{d\theta^2} h(\theta_o) < 0$, then θ_o is a local max.

V) If $h(\theta)$ is a continuous function on an interval with endpoints $a < b$ (not necessarily finite), differentiable on (a, b) and if the **critical point is unique**,

then the critical point is a **global maximum** if it is a local maximum. To see this claim, note that if the critical point is not the global max then there would be a local minimum and the critical point would not be unique. Also see Casella and Berger (2002, p. 317). Let $a = -2$ and $b = 4$. In Figure 5.1 a), the critical point for $g(\theta) = -\theta^2 + 25$ is at $\theta = 0$, is unique, and is both a local and global maximum. In Figure 5.1 b), $h(\theta) = \theta^3 - 1.5\theta^2 - 6\theta + 11$, the critical point $\theta = -1$ is not unique and is a local max but not a global max.

VI) If h is strictly convex ($\frac{d^2}{d\theta^2}h(\theta) > 0$ for all $\theta \in \Theta$), then any local min of h is a global min. If $h'(\theta_o) = 0$, then the 2nd derivative test states that if $\frac{d^2}{d\theta^2}h(\theta_o) > 0$, then θ_o is a local min.

Tips: a) $\exp(a) = e^a$ and $\log(y) = \ln(y) = \log_e(y)$ is the **natural logarithm**.
b) $\log(a^b) = b\log(a)$ and $\log(e^b) = b$.
c) $\log(\prod_{i=1}^n a_i) = \sum_{i=1}^n \log(a_i)$.
d) $\log L(\theta) = \log(\prod_{i=1}^n f(y_i|\theta)) = \sum_{i=1}^n \log(f(y_i|\theta))$.
e) If t is a differentiable function and $t(\theta) \neq 0$, then $\frac{d}{d\theta}\log(|t(\theta)|) = \frac{t'(\theta)}{t(\theta)}$ where $t'(\theta) = \frac{d}{d\theta}t(\theta)$. In particular, $\frac{d}{d\theta}\log(\theta) = 1/\theta$.
f) Anything that does not depend on θ is treated as a constant with respect to θ and hence has derivative 0 with respect to θ.

Showing that $\hat{\boldsymbol{\theta}}$ is the global maximum of $\log(L(\boldsymbol{\theta}))$ is much more difficult in the multiparameter case. To show that $\hat{\boldsymbol{\theta}}$ is a local max often involves using a Hessian matrix of second derivatives. Calculations involving the Hessian matrix are often too difficult for exams. Often there is no closed form solution for the MLE and a computer needs to be used. For hand calculations, Remark 5.2 and Theorem 5.2 can often be used to avoid using the Hessian matrix.

Definition 5.2. Let the data be $Y_1, ..., Y_n$ and suppose that the parameter $\boldsymbol{\theta}$ has components $(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_k)$. Then $\hat{\boldsymbol{\theta}}_i$ will be called the MLE of $\boldsymbol{\theta}_i$. Without loss of generality, assume that $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, that the MLE of $\boldsymbol{\theta}$ is $(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)$ and that $\hat{\boldsymbol{\theta}}_2$ is known. The **profile likelihood function** is $L_P(\boldsymbol{\theta}_1) = L(\boldsymbol{\theta}_1, \hat{\boldsymbol{\theta}}_2(\boldsymbol{y}))$ with domain $\{\boldsymbol{\theta}_1 : (\boldsymbol{\theta}_1, \hat{\boldsymbol{\theta}}_2) \in \Theta\}$.

Remark 5.2. Since $L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ is maximized over Θ by $(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)$, the maximizer of the profile likelihood function and of the log profile likelihood func-

tion is $\hat{\boldsymbol{\theta}}_1$. The log profile likelihood function can often be maximized using calculus if $\boldsymbol{\theta}_1 = \theta_1$ is a scalar.

Theorem 5.2: Existence of the MLE for a REF (Barndorff–Nielsen 1982): Assume that the natural parameterization of the k-parameter REF is used so that Ω is an open k-dimensional convex set (usually an open interval or cross product of open intervals). Then the log likelihood function $\log L(\boldsymbol{\eta})$ is a strictly concave function of $\boldsymbol{\eta}$. Hence if $\hat{\boldsymbol{\eta}}$ is a critical point of $\log \mathrm{L}(\boldsymbol{\eta})$ and if $\hat{\boldsymbol{\eta}} \in \Omega$ then $\hat{\boldsymbol{\eta}}$ is the unique MLE of $\boldsymbol{\eta}$. Hence the Hessian matrix of 2nd derivatives does not need to be checked!

Remark 5.3. A nice proof of this result would be useful to show that the result is true and not just part of the statistical folklore. For k-parameter exponential families with $k > 1$, it is usually easier to verify that the family is regular than to calculate the Hessian matrix. For 1P–REFs, check that the critical point is a global maximum using standard calculus techniques such as calculating the second derivative of the log likelihood $\log L(\boldsymbol{\theta})$. For a 1P–REF, verifying that the family is regular is often more difficult than using calculus. Also, often the MLE is desired for a parameter space Θ_U which is not an open set (eg for $\Theta_U = [0, 1]$ instead of $\Theta = (0, 1)$).

Remark 5.4, (Barndorff–Nielsen 1982). The MLE does not exist if $\hat{\boldsymbol{\eta}}$ is not in Ω, an event that occurs with positive probability for discrete distributions. If $\boldsymbol{T}$ is the complete sufficient statistic and C is the closed convex hull of the support of $\boldsymbol{T}$, then the MLE exists iff $\boldsymbol{T} \in int\ C$ where $int\ C$ is the interior of C.

Remark 5.5. As illustrated in the following examples, the 2nd derivative is evaluated at $\hat{\theta}(\boldsymbol{y})$. The MLE is a statistic and $T_n(\boldsymbol{y}) = \hat{\theta}(\boldsymbol{y})$ is the observed value of the MLE $T_n(\boldsymbol{Y}) = \hat{\theta}(\boldsymbol{Y})$. Often $\boldsymbol{y}$ and $\boldsymbol{Y}$ are suppressed.

Example 5.1. Suppose that $Y_1, ..., Y_n$ are iid Poisson (θ). This distribution is a 1P–REF with $\Theta = (0, \infty)$. The likelihood

$$L(\theta) = c\ e^{-n\theta} \exp[\log(\theta) \sum y_i]$$

where the constant c does not depend on θ, and the log likelihood

$$\log(L(\theta)) = d - n\theta + \log(\theta) \sum y_i$$

where $d = \log(c)$ does not depend on θ. Hence

$$\frac{d}{d\theta} \log(L(\theta)) = -n + \frac{1}{\theta} \sum y_i \stackrel{set}{=} 0,$$

or $\sum y_i = n\theta$, or

$$\hat{\theta} = \overline{y}.$$

Notice that $\hat{\theta}$ is the unique solution and

$$\frac{d^2}{d\theta^2} \log(L(\theta)) = \frac{-\sum y_i}{\theta^2} < 0$$

unless $\sum y_i = 0$. Hence for $\sum y_i > 0$ the log likelihood is strictly concave and $\overline{Y}$ is the MLE of θ. The MLE does not exist if $\sum_{i=1}^{n} Y_i = 0$ since 0 is not in Θ.

Now suppose that $\Theta = [0, \infty)$. This family is not an exponential family since the same formula for the pmf needs to hold for all values of $\theta \in \Theta$ and 0^0 is not defined. Notice that

$$f(y|\theta) = \frac{e^{-\theta}\theta^y}{y!} I[\theta > 0] + 1 \; I[\theta = 0, y = 0].$$

Now

$$I_A(\theta) I_B(\theta) = I_{A \cap B}(\theta)$$

and $I_{\emptyset}(\theta) = 0$ for all θ. Hence the likelihood

$$L(\theta) = e^{-n\theta} \exp[\log(\theta) \sum_{i=1}^{n} y_i] \frac{1}{\prod_{i=1}^{n} y_i!} I[\theta > 0] + 1 \; I[\theta = 0, \sum_{i=1}^{n} y_i = 0].$$

If $\sum y_i \neq 0$, then $\overline{y}$ maximizes $L(\theta)$ by the work above. If $\sum y_i = 0$, then $L(\theta) = e^{-n\theta} I(\theta > 0) + I(\theta = 0) = e^{-n\theta} I(\theta \geq 0)$ which is maximized by $\theta = 0 = \overline{y}$. Hence $\overline{Y}$ is the MLE of θ if $\Theta = [0, \infty)$.

By invariance, $t(\overline{Y})$ is the MLE of $t(\theta)$. Hence $(\overline{Y})^2$ is the MLE of θ^2. $\sin(\overline{Y})$ is the MLE of $\sin(\theta)$, et cetera.

Example 5.2. Suppose that $Y_1, ..., Y_n$ are iid $N(\mu, \sigma^2)$ where $\sigma^2 > 0$ and $\mu \in \Re = (-\infty, \infty)$. Then

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

Notice that

$$\frac{d}{d\mu}\sum_{i=1}^{n}(y_i-\mu)^2=\sum_{i=1}^{n}-2(y_i-\mu)\stackrel{\text{set}}{=}0$$

or $\sum_{i=1}^{n} y_i = n\mu$ or $\hat{\mu}=\overline{y}$. Since $\hat{\mu}$ is the unique solution and

$$\frac{d^2}{d\mu^2}\sum_{i=1}^{n}(y_i-\mu)^2=2n>0,$$

$\hat{\mu}=\overline{y}$ is the minimizer of $h(\mu)=\sum_{i=1}^{n}(y_i-\mu)^2$. Hence $\overline{y}$ is the maximizer of

$$\exp\left[\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right]$$

regardless of the value of $\sigma^2>0$. Hence $\hat{\mu}=\overline{Y}$ is the MLE of μ and the MLE of σ^2 can be found by maximizing the profile likelihood

$$L_P(\sigma^2)=L(\hat{\mu}(\boldsymbol{y}),\sigma^2)=\left(\frac{1}{\sqrt{2\pi}}\right)^n\frac{1}{(\sigma^2)^{n/2}}\exp\left[\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\overline{y})^2\right].$$

Writing $\tau=\sigma^2$ often helps prevent calculus errors. Then

$$\log(L_p(\tau))=d-\frac{n}{2}\log(\tau)+\frac{-1}{2\tau}\sum_{i=1}^{n}(y_i-\overline{y})^2$$

where the constant d does not depend on τ. Hence

$$\frac{d}{d\tau}\log(L_p(\tau))=\frac{-n}{2}\frac{1}{\tau}+\frac{1}{2\tau^2}\sum_{i=1}^{n}(y_i-\overline{y})^2\stackrel{\text{set}}{=}0,$$

or

$$n\tau=\sum_{i=1}^{n}(y_i-\overline{y})^2$$

or

$$\hat{\tau}=\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2$$

and the solution $\hat{\tau}$ is the unique critical point. Note that

$$\frac{d^2}{d\mu^2}\log(L_P(\tau))=\frac{n}{2(\tau)^2}-\frac{\sum(y_i-\overline{y})^2}{(\tau)^3}\bigg|_{\tau=\hat{\tau}}=\frac{n}{2(\hat{\tau})^2}-\frac{n\hat{\tau}}{(\hat{\tau})^3}\frac{2}{2}$$

$$= \frac{-n}{2(\hat{\tau})^2} < 0.$$

Hence $\hat{\sigma}^2 = \hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ is the MLE of σ^2 by Remark 5.1 V). Thus $(\overline{Y}, \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2)$ is the MLE of (μ, σ^2).

Example 5.3. Following Pewsey (2002), suppose that $Y_1, ..., Y_n$ are iid HN(μ, σ^2) where μ and σ^2 are both unknown. Let the ith order statistic $Y_{(i)} \equiv Y_{i:n}$. Then the likelihood

$$L(\mu, \sigma^2) = cI[y_{1:n} \geq \mu]\frac{1}{\sigma^n}\exp\left[(\frac{-1}{2\sigma^2})\sum(y_i - \mu)^2\right].$$

For any fixed $\sigma^2 > 0$, this likelihood is maximized by making $\sum(y_i - \mu)^2$ as small as possible subject to the constraint $y_{1:n} \geq \mu$. Notice that for any $\mu_o < y_{1:n}$, the terms $(y_i - y_{1:n})^2 < (y_i - \mu_o)^2$. Hence the MLE of μ is

$$\hat{\mu} = Y_{1:n}$$

and the MLE of σ^2 is found by maximizing the log profile likelihood

$$\log(L_P(\sigma^2)) = \log(L(y_{1:n}, \sigma^2)) = d - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum(y_i - y_{1:n})^2,$$

and

$$\frac{d}{d(\sigma^2)}\log(L(y_{1:n}, \sigma^2)) = \frac{-n}{2(\sigma^2)} + \frac{1}{2(\sigma^2)^2}\sum(y_i - y_{1:n})^2 \stackrel{set}{=} 0.$$

Or $\sum(y_i - y_{1:n})^2 = n\sigma^2$. So

$$\hat{\sigma}^2 \equiv w_n = \frac{1}{n}\sum(y_i - y_{1:n})^2.$$

Since the solution $\hat{\sigma}^2$ is unique and

$$\frac{d^2}{d(\sigma^2)^2}\log(L(y_{1:n}, \sigma^2)) =$$

$$\frac{n}{2(\sigma^2)^2} - \frac{\sum(y_i - \mu)^2}{(\sigma^2)^3})\Big|_{\sigma^2=\hat{\sigma}^2} = \frac{n}{2(\hat{\sigma}^2)^2} - \frac{n\hat{\sigma}^2}{(\hat{\sigma^2})^3}\frac{2}{2} = \frac{-n}{2\hat{\sigma}^2} < 0,$$

$(\hat{\mu}, \hat{\sigma}^2) = (Y_{1:n}, W_n)$ is MLE of (μ, σ^2).

Example 5.4. Suppose that the random vectors $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid from a multivariate normal $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution where $\boldsymbol{\Sigma}$ is a positive definite matrix. To find the MLE of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ we will use three results proved in Anderson (1984, p. 62).

$$\text{i)} \sum_{\text{i=1}}^{\text{n}} (\boldsymbol{x}_{\text{i}} - \boldsymbol{\mu})^{\text{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_{\text{i}} - \boldsymbol{\mu}) = \text{tr}(\boldsymbol{\Sigma}^{-1} \boldsymbol{A}) + \text{n}(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^{\text{T}} \boldsymbol{\Sigma}^{-1} (\overline{\boldsymbol{x}} - \boldsymbol{\mu})$$

where

$$\boldsymbol{A} = \sum_{i=1}^{n} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T.$$

ii) Let $\boldsymbol{C}$ and $\boldsymbol{D}$ be positive definite matrices. Then $\boldsymbol{C} = \frac{1}{n}\boldsymbol{D}$ maximizes

$$h(\boldsymbol{C}) = -n \log(|\boldsymbol{C}|) - tr(\boldsymbol{C}^{-1}\boldsymbol{D})$$

with respect to positive definite matrices.
iii) Since $\boldsymbol{\Sigma}^{-1}$ is positive definite, $(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\overline{\boldsymbol{x}} - \boldsymbol{\mu}) \geq 0$ as a function of $\boldsymbol{\mu}$ with equality iff $\boldsymbol{\mu} = \overline{\boldsymbol{x}}$.

Since

$$f(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right],$$

the likelihood function

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^{n} f(\boldsymbol{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$= \frac{1}{(2\pi)^{np/2}|\boldsymbol{\Sigma}|^{n/2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu})\right],$$

and the log likelihood $\log(L(\boldsymbol{\mu}, \boldsymbol{\Sigma})) =$

$$-\frac{np}{2}\log(2\pi) - \frac{n}{2}\log(|\boldsymbol{\Sigma}|) - \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu})$$

$$= -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log(|\boldsymbol{\Sigma}|) - \frac{1}{2}tr(\boldsymbol{\Sigma}^{-1}\boldsymbol{A}) - \frac{n}{2}(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\overline{\boldsymbol{x}} - \boldsymbol{\mu})$$

by i). Now the last term is maximized by $\boldsymbol{\mu} = \overline{\boldsymbol{x}}$ by iii) and the middle two terms are maximized by $\frac{1}{n}\boldsymbol{A}$ by ii) since $\boldsymbol{\Sigma}$ and $\boldsymbol{A}$ are both positive definite. Hence the MLE of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

$$(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = (\overline{\boldsymbol{X}}, \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \overline{\boldsymbol{X}})(\boldsymbol{X}_i - \overline{\boldsymbol{X}})^T).$$

Example 5.5. Let $X_1, ..., X_n$ be independent identically distributed random variables from a lognormal (μ, σ^2) distribution with pdf

$$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(\log(x) - \mu)^2}{2\sigma^2}\right)$$

where $\sigma > 0$ and $x > 0$ and μ is real. **Assume that σ is known.**

a) Find the maximum likelihood estimator of μ.

b) What is the maximum likelihood estimator of μ^3? Explain.

Solution: a)

$$\hat{\mu} = \frac{\sum \log(X_i)}{n}$$

To see this note that

$$L(\mu) = \left(\prod \frac{1}{x_i\sqrt{2\pi\sigma^2}}\right) \exp\left(\frac{-\sum(\log(x_i) - \mu)^2}{2\sigma^2}\right.$$

So

$$\log(L(\mu)) = \log(c) \; - \; \frac{\sum(\log(x_i) - \mu)^2}{2\sigma^2}$$

and the derivative of the log likelihood wrt μ is

$$\frac{\sum 2(\log(x_i) - \mu)}{2\sigma^2}.$$

Setting this quantity equal to 0 gives $n\mu = \sum \log(x_i)$ and the solution is unique. The second derivative is $-n/\sigma^2 < 0$, so $\hat{\mu}$ is indeed the global maximum.

b)

$$\left(\frac{\sum \log(X_i)}{n}\right)^3$$

by invariance.

Example 5.6. Suppose that the joint probability distribution function of $X_1, ..., X_k$ is

$$f(x_1, x_2, ..., x_k|\theta) = \frac{n!}{(n-k)!\theta^k} \exp\left(\frac{-[(\sum_{i=1}^k x_i) + (n-k)x_k]}{\theta}\right)$$

where $0 \leq x_1 \leq x_2 \leq \cdots \leq x_k$ and $\theta > 0$.

a) Find the maximum likelihood estimator (MLE) for θ.

b) What is the MLE for θ^2? Explain briefly.

Solution: a) Let $t = [(\sum_{i=1}^k x_i) + (n-k)x_k]$. $L(\theta) = f(\boldsymbol{x}|\theta)$ and $\log(L(\theta)) = \log(f(\boldsymbol{x}|\theta)) =$

$$d - k\log(\theta) - \frac{t}{\theta}.$$

Hence

$$\frac{d}{d\theta}\log(L(\theta)) = \frac{-k}{\theta} + \frac{t}{\theta^2} \stackrel{set}{=} 0.$$

Hence

$$k\theta = t$$

or

$$\hat{\theta} = \frac{t}{k}.$$

This is a unique solution and

$$\frac{d^2}{d\theta^2}\log(L(\theta)) = \frac{k}{\theta^2} - \frac{2t}{\theta^3}\bigg|_{\theta=\hat{\theta}} = \frac{k}{\hat{\theta}^2} - \frac{2k\hat{\theta}}{\hat{\theta}^3} = -\frac{k}{\hat{\theta}^2} < 0.$$

Hence $\hat{\theta} = T/k$ is the MLE where $T = [(\sum_{i=1}^k X_i) + (n-k)X_k]$.

b) $\hat{\theta}^2$ by the invariance principle.

Example 5.7. Let $X_1, ..., X_n$ be independent identically distributed random variables with pdf

$$f(x) = \frac{1}{\lambda} x^{\frac{1}{\lambda}-1},$$

where $\lambda > 0$ and $0 < x \leq 1$.

a) Find the maximum likelihood estimator of λ.

b) What is the maximum likelihood estimator of λ^3? Explain.

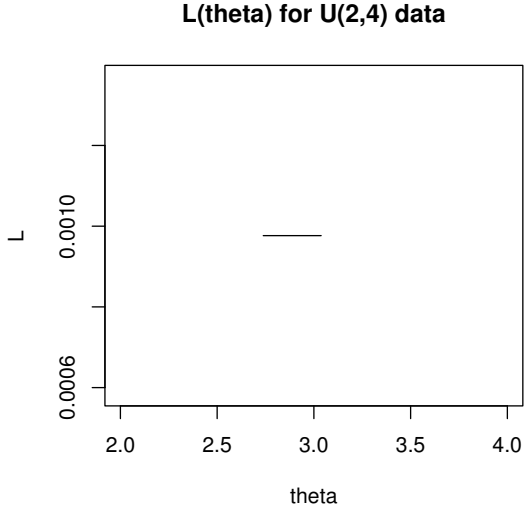

Figure 5.2: Sample Size $n = 10$

Solution: a) For $0 < x \leq 1$

$$f(x) = \frac{1}{\lambda} \exp\left[(\frac{1}{\lambda} - 1)\log(x)\right].$$

Hence the likelihood

$$L(\lambda) = \frac{1}{\lambda^n} \exp\left[(\frac{1}{\lambda} - 1)\sum \log(x_i)\right],$$

and the log likelihood

$$\log(L(\lambda)) = -n\log(\lambda) + (\frac{1}{\lambda} - 1)\sum \log(x_i).$$

Hence

$$\frac{d}{d\lambda}\log(L(\lambda)) = \frac{-n}{\lambda} - \frac{\sum \log(x_i)}{\lambda^2} \stackrel{set}{=} 0,$$

or $-\sum \log(x_i) = n\lambda$, or

$$\hat{\lambda} = \frac{-\sum \log(x_i)}{n}.$$

Notice that $\hat{\lambda}$ is the unique solution and that

$$\frac{d^2}{d\lambda^2}\log(L(\lambda)) = \frac{n}{\lambda^2} + \frac{2\sum\log(x_i)}{\lambda^3}\Big|_{\lambda=\hat{\lambda}}$$

$$= \frac{n}{\hat{\lambda}^2} - \frac{2n\hat{\lambda}}{\hat{\lambda}^3} = \frac{-n}{\hat{\lambda}^2} < 0.$$

Hence $\hat{\lambda} = -\sum\log(X_i)/n$ is the MLE of λ.

b) By invariance, $\hat{\lambda}^3$ is the MLE of λ.

Example 5.8. Suppose $Y_1, ..., Y_n$ are iid $U(\theta - 1, \theta + 1)$. Then

$$L(\theta) = \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} \frac{1}{2} I(\theta - 1 \leq y_i \leq \theta + 1) = \frac{1}{2^n} I(\theta - 1 \leq \text{ all } y_i \leq \theta + 1)$$

$$= \frac{1}{2^n} I(\theta - 1 \leq y_{(1)} \leq y_{(n)} \leq \theta + 1) = \frac{1}{2^n} I(y_{(n)} - 1 \leq \theta \leq y_{(1)} + 1).$$

Let $0 \leq c \leq 1$. Then any estimator of the form $\hat{\theta}_c = c[Y_{(n)} - 1] + (1-c)[Y_{(1)} + 1]$ is an MLE of θ. Figure 5.2 shows $L(\theta)$ for $U(2, 4)$ data with $n = 10, y_{(1)} = 2.0375$ and $y_{(n)} = 3.7383$.

5.2 Method of Moments Estimators

The method of moments is another useful way for obtaining point estimators. Let $Y_1, ..., Y_n$ be an iid sample and let

$$\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n} Y_i^j \text{ and } \mu_j \equiv \mu_j(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}(Y^j) \tag{5.2}$$

for $j = 1, ..., k$. So $\hat{\mu}_j$ is the jth sample moment and μ_j is the jth population moment. Fix k and assume that $\mu_j = \mu_j(\theta_1, ..., \theta_k)$. Solve the system

$$\begin{aligned} \hat{\mu}_1 &\stackrel{\text{set}}{=} \mu_1(\theta_1, ..., \theta_k) \\ &\vdots \qquad\qquad \vdots \\ \hat{\mu}_k &\stackrel{\text{set}}{=} \mu_k(\theta_1, ..., \theta_k) \end{aligned}$$

for $\tilde{\boldsymbol{\theta}}$.

Definition 5.3. The solution $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1, ..., \tilde{\theta}_k)$ is the **method of moments estimator** of $\boldsymbol{\theta}$. If g is a continuous function of the first k moments and $h(\boldsymbol{\theta}) = g(\mu_1(\boldsymbol{\theta}), ..., \mu_k(\boldsymbol{\theta}))$, then the method of moments estimator of $h(\boldsymbol{\theta})$ is

$$g(\hat{\mu}_1, ..., \hat{\mu}_k).$$

Sometimes the notation $\hat{\theta}_{MLE}$ and $\hat{\theta}_{MM}$ will be used to denote the MLE and method of moments estimators of θ, respectively.

Example 5.9. Let $Y_1, ..., Y_n$ be iid from a distribution with a given pdf or pmf $f(y|\theta)$.

a) If $E(Y) = h(\theta)$, then $\hat{\theta}_{MM} = h^{-1}(\overline{Y})$.

b) The method of moments estimator of $E(Y) = \mu_1$ is $\hat{\mu}_1 = \overline{Y}$.

c) The method of moments estimator of $\text{VAR}_\theta(Y) = \mu_2(\theta) - [\mu_1(\theta)]^2$ is

$$\hat{\sigma}^2_{MM} = \hat{\mu}_2 - \hat{\mu}_1^2 = \frac{1}{n}\sum_{i=1}^n Y_i^2 - (\overline{Y})^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \overline{Y})^2 \equiv S^2_M.$$

Method of moments estimators need not be unique. For example both $\overline{Y}$ and S^2_M are method of moment estimators of θ for iid Poisson(θ) data. Generally the method of moments estimators that use small j for $\hat{\mu}_j$ are preferred, so use $\overline{Y}$ for Poisson data.

Proposition 5.3. Let $S^2_M = \frac{1}{n}\sum_{i=1}^n (Y_i - \overline{Y})^2$ and suppose that $E(Y) = h_1(\theta_1, \theta_2)$ and $V(Y) = h_2(\theta_1, \theta_2)$. Then solving

$$\begin{aligned} \overline{Y} &\stackrel{\text{set}}{=} h_1(\theta_1, \theta_2) \\ S^2_M &\stackrel{\text{set}}{=} h_2(\theta_1, \theta_2) \end{aligned}$$

for $\tilde{\boldsymbol{\theta}}$ is a method of moments estimator.

Proof. Notice that $\mu_1 = h_1(\theta_1, \theta_2) = \mu_1(\theta_1, \theta_2)$ while $\mu_2 - [\mu_1]^2 = h_2(\theta_1, \theta_2)$. Hence $\mu_2 = h_2(\theta_1, \theta_2) + [h_1(\theta_1, \theta_2)]^2 = \mu_2(\theta_1, \theta_2)$. Hence the method of moments estimator is a solution to $\overline{Y} \stackrel{\text{set}}{=} \mu_1(\theta_1, \theta_2)$ and $\frac{1}{n}\sum_{i=1}^n Y_i^2 \stackrel{\text{set}}{=} h_2(\theta_1, \theta_2) + [\mu_1(\theta_1, \theta_2)]^2$. Equivalently, solve $\overline{Y} \stackrel{\text{set}}{=} h_1(\theta_1, \theta_2)$ and $\frac{1}{n}\sum_{i=1}^n Y_i^2 - [\overline{Y}]^2 = S^2_M \stackrel{\text{set}}{=} h_2(\theta_1, \theta_2)$. QED

Example 5.10. Suppose that $Y_1, ..., Y_n$ be iid gamma (ν, λ). Then $\hat{\mu}_1 \stackrel{\text{set}}{=} E(Y) = \nu\lambda$ and $\hat{\mu}_2 \stackrel{\text{set}}{=} E(Y^2) = \text{VAR}(Y) + [E(Y)]^2 = \nu\lambda^2 + \nu^2\lambda^2 = \nu\lambda^2(1+\nu)$.

Substitute $\nu = \hat{\mu}_1/\lambda$ into the 2nd equation to obtain

$$\hat{\mu}_2 = \frac{\hat{\mu}_1}{\lambda}\lambda^2(1 + \frac{\hat{\mu}_1}{\lambda}) = \lambda\hat{\mu}_1 + \hat{\mu}_1^2.$$

Thus

$$\tilde{\lambda} = \frac{\hat{\mu}_2 - \hat{\mu}_1^2}{\hat{\mu}_1} = \frac{S_M^2}{\overline{Y}} \text{ and } \tilde{\nu} = \frac{\hat{\mu}_1}{\tilde{\lambda}} = \frac{\hat{\mu}_1^2}{\hat{\mu}_2 - \hat{\mu}_1^2} = \frac{[\overline{Y}]^2}{S_M^2}.$$

Alternatively, solve $\overline{Y} \stackrel{\text{set}}{=} \nu\lambda$ and $S_M^2 \stackrel{\text{set}}{=} \nu\lambda^2 = (\nu\lambda)\lambda$. Hence $\tilde{\lambda} = S_M^2/\overline{Y}$ and

$$\tilde{\nu} = \frac{\overline{Y}}{\tilde{\lambda}} = \frac{[\overline{Y}]^2}{S_M^2}.$$

5.3 Summary

A) Let $Y_1, ..., Y_n$ be iid with pdf or pmf $f(y|\theta)$. Then $L(\theta) = \prod_{i=1}^n f(y_i|\theta)$. To find the MLE,
i) find $L(\theta)$ and then find the log likelihood $\log L(\theta)$.
ii) Find the derivative $\frac{d}{d\theta}\log L(\theta)$, set the derivative equal to zero and solve for θ. The solution is a candidate for the MLE.
iii) **Invariance Principle:** If $\hat{\theta}$ is the MLE of θ, then $\tau(\hat{\theta})$ is the MLE of $\tau(\theta)$.
iv) Show that $\hat{\theta}$ is the MLE by showing that $\hat{\theta}$ is the global maximizer of $\log L(\theta)$. Often this is done by noting that $\hat{\theta}$ is the unique solution to the equation $\frac{d}{d\theta}\log L(\theta) = 0$ and that the 2nd derivative evaluated at $\hat{\theta}$ is negative: $\frac{d^2}{d\theta^2}\log L(\theta)|_{\hat{\theta}} < 0$.

B) If $\log L(\theta)$ is strictly concave ($\frac{d^2}{d\theta^2}\log L(\theta) < 0$ for all $\theta \in \Theta$), then any local max of $\log L(\theta)$ is a global max.

C) Know how to find the MLE for the normal distribution (including when μ or σ^2 is known). Memorize the MLEs

$$\overline{Y},\ S_M^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \overline{Y})^2,\ \frac{1}{n}\sum_{i=1}^n (Y_i - \mu)^2$$

for the normal and for the uniform distribution. Also $\overline{Y}$ is the MLE for several brand name distributions. Notice that S_M^2 is the method of moments estimator for $V(Y)$ and is the MLE for $V(Y)$ if the data are iid $N(\mu, \sigma^2)$.

D) **On qualifying exams**, the $N(\mu, \mu)$ and $N(\mu, \mu^2)$ distributions are common. See Problem 5.35.

E) Indicators are useful. For example, $\prod_{i=1}^{n} I_A(y_i) = I(\text{all } y_i \in A)$ and $\prod_{j=1}^{k} I_{A_j}(y) = I_{\cap_{j=1}^{k} A_j}(y)$. Hence $I(0 \leq y \leq \theta) = I(0 \leq y)I(y \leq \theta)$, and $\prod_{i=1}^{n} I(\theta_1 \leq y_i \leq \theta_2) = I(\theta_1 \leq y_{(1)} \leq y_{(n)} \leq \theta_2) = I(\theta_1 \leq y_{(1)})I(y_{(n)} \leq \theta_2)$.

F) Let $\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n} Y_i^j$, let $\mu_j = E(Y^j)$ and assume that $\mu_j = \mu_j(\theta_1, ..., \theta_k)$. Solve the system

$$\begin{array}{ccc} \hat{\mu}_1 & \stackrel{\text{set}}{=} & \mu_1(\theta_1, ..., \theta_k) \\ \vdots & & \vdots \\ \hat{\mu}_k & \stackrel{\text{set}}{=} & \mu_k(\theta_1, ..., \theta_k) \end{array}$$

for the method of moments estimator $\tilde{\boldsymbol{\theta}}$.

G) If g is a continuous function of the first k moments and $h(\boldsymbol{\theta}) = g(\mu_1(\boldsymbol{\theta}), ..., \mu_k(\boldsymbol{\theta}))$, then the method of moments estimator of $h(\boldsymbol{\theta})$ is $g(\hat{\mu}_1, ..., \hat{\mu}_k)$.

5.4 Complements

Optimization theory is also known as nonlinear programming and shows how to find the global max and min of a multivariate function. Peressini, Sullivan and Uhl (1988) is an undergraduate text. Also see Sundaram (1996) and Bertsekas (1999).

Maximum likelihood estimation is widely used in statistical models. See Pawitan (2001) and texts for Categorical Data Analysis, Econometrics, Multiple Linear Regression, Generalized Linear Models, Multivariate Analysis and Survival Analysis.

Suppose that $Y = t(W)$ and $W = t^{-1}(Y)$ where W has a pdf with parameters $\boldsymbol{\theta}$, the transformation t does not depend on any unknown parameters, and the pdf of Y is

$$f_Y(y) = f_W(t^{-1}(y)) \left| \frac{dt^{-1}(y)}{dy} \right|.$$

If $W_1, ..., W_n$ are iid with pdf $f_W(w)$, assume that the MLE of $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}}_W(\boldsymbol{w})$ where the w_i are the observed values of W_i and $\boldsymbol{w} = (w_1, ..., w_n)$. If $Y_1, ..., Y_n$

are iid and the y_i are the observed values of Y_i, then the likelihood is

$$L_Y(\boldsymbol{\theta}) = \left(\prod_{i=1}^{n} \left| \frac{dt^{-1}(y_i)}{dy} \right| \right) \prod_{i=1}^{n} f_W(t^{-1}(y_i)|\boldsymbol{\theta}) = c \prod_{i=1}^{n} f_W(t^{-1}(y_i)|\boldsymbol{\theta}).$$

Hence the log likelihood is $\log(L_Y(\boldsymbol{\theta})) =$

$$d + \sum_{i=1}^{n} \log[f_W(t^{-1}(y_i)|\boldsymbol{\theta})] = d + \sum_{i=1}^{n} \log[f_W(w_i|\boldsymbol{\theta})] = d + \log[L_W(\boldsymbol{\theta})]$$

where $w_i = t^{-1}(y_i)$. Hence maximizing the $\log(L_Y(\boldsymbol{\theta}))$ is equivalent to maximizing $\log(L_W(\boldsymbol{\theta}))$ and

$$\hat{\boldsymbol{\theta}}_Y(\boldsymbol{y}) = \hat{\boldsymbol{\theta}}_W(\boldsymbol{w}) = \hat{\boldsymbol{\theta}}_W(t^{-1}(y_1), ..., t^{-1}(y_n)). \tag{5.3}$$

Compare Meeker and Escobar (1998, p. 175).

Example 5.11. Suppose $Y_1, ..., Y_n$ are iid lognormal (μ, σ^2). Then $W_i = \log(Y_i) \sim N(\mu, \sigma^2)$ and the MLE $(\hat{\mu}, \hat{\sigma}^2) = (\overline{W}, \frac{1}{n}\sum_{i=1}^{n}(W_i - \overline{W})^2)$.

One of the most useful properties of the maximum likelihood estimator is the invariance property: if $\hat{\theta}$ is the MLE of θ, then $\tau(\hat{\theta})$ is the MLE of $\tau(\theta)$. Olive (2004) is a good discussion of the MLE invariance principle. Also see Pal and Berry (1992). Many texts either define the MLE of $\tau(\theta)$ to be $\tau(\hat{\theta})$, say that the property is immediate from the definition of the MLE, or quote Zehna (1966). A little known paper, Berk (1967), gives an elegant proof of the invariance property that can be used in introductory statistical courses. The next subsection will show that Berk (1967) answers some questions about the MLE which can not be answered using Zehna (1966).

5.4.1 Two "Proofs" of the Invariance Principle

"Proof" I) The following argument of Zehna (1966) also appears in Casella and Berger (2002, p. 320). Let $\boldsymbol{\theta} \in \Theta$ and let $h : \Theta \to \Lambda$ be a function. Since the MLE

$$\hat{\boldsymbol{\theta}} \in \Theta, \quad h(\hat{\boldsymbol{\theta}}) = \hat{\boldsymbol{\lambda}} \in \Lambda.$$

If h is not one to one, then many values of $\boldsymbol{\theta}$ may be mapped to $\boldsymbol{\lambda}$. Let

$$\Theta_{\boldsymbol{\lambda}} = \{\boldsymbol{\theta} : h(\boldsymbol{\theta}) = \boldsymbol{\lambda}\}$$

and define the induced likelihood function $M(\boldsymbol{\lambda})$ by

$$M(\boldsymbol{\lambda}) = \sup_{\boldsymbol{\theta} \in \Theta_{\boldsymbol{\lambda}}} L(\boldsymbol{\theta}). \tag{5.4}$$

Then for any $\boldsymbol{\lambda} \in \Lambda$,

$$M(\boldsymbol{\lambda}) = \sup_{\boldsymbol{\theta} \in \Theta_{\boldsymbol{\lambda}}} L(\boldsymbol{\theta}) \leq \sup_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}) = L(\hat{\boldsymbol{\theta}}) = M(\hat{\boldsymbol{\lambda}}). \tag{5.5}$$

Hence $h(\hat{\boldsymbol{\theta}}) = \hat{\boldsymbol{\lambda}}$ maximizes the induced likelihood $M(\boldsymbol{\lambda})$. Zehna (1966) says that since $h(\hat{\boldsymbol{\theta}})$ maximizes the induced likelihood, we should call $h(\hat{\boldsymbol{\theta}})$ the MLE of $h(\boldsymbol{\theta})$, but the definition of MLE says that we should be maximizing a genuine likelihood.

This argument raises two important questions.

- If we call $h(\hat{\boldsymbol{\theta}})$ the MLE of $h(\boldsymbol{\theta})$ and h is not one to one, does $h(\hat{\boldsymbol{\theta}})$ maximize a likelihood or should $h(\hat{\boldsymbol{\theta}})$ be called a maximum induced likelihood estimator?
- If $h(\hat{\boldsymbol{\theta}})$ is an MLE, what is the likelihood function $K(h(\boldsymbol{\theta}))$?

Some examples might clarify these questions.

- If the population come from a $N(\mu, \sigma^2)$ distribution, the invariance principle says that the MLE of μ/σ is $\overline{X}/S_M$ where

  $$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

  and

  $$S_M^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

  are the MLEs of μ and σ^2. Since the function $h(x, y) = x/\sqrt{y}$ is not one to one (eg $h(x, y) = 1$ if $x = \sqrt{y}$), what is the likelihood $K(h(\mu, \sigma^2)) = K(\mu/\sigma)$ that is being maximized?
- If X_i comes from a Bernoulli(ρ) population, why is $\overline{X}_n(1 - \overline{X}_n)$ the MLE of $\rho(1 - \rho)$?

Proof II) Examining the invariance principle for one to one functions h is also useful. When h is one to one, let $\boldsymbol{\eta} = h(\boldsymbol{\theta})$. Then the inverse function h^{-1} exists and $\boldsymbol{\theta} = h^{-1}(\boldsymbol{\eta})$. Hence

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = f(\boldsymbol{x}|h^{-1}(\boldsymbol{\eta})) \tag{5.6}$$

is the joint pdf or pmf of $\boldsymbol{x}$. So the likelihood function of $h(\boldsymbol{\theta}) = \eta$ is

$$L^*(\boldsymbol{\eta}) \equiv K(\boldsymbol{\eta}) = L(h^{-1}(\boldsymbol{\eta})). \tag{5.7}$$

Also note that

$$\sup_{\boldsymbol{\eta}} K(\boldsymbol{\eta}|\boldsymbol{x}) = \sup_{\boldsymbol{\eta}} L(h^{-1}(\boldsymbol{\eta})|\boldsymbol{x}) = L(\hat{\boldsymbol{\theta}}|\boldsymbol{x}). \tag{5.8}$$

Thus

$$\hat{\boldsymbol{\eta}} = h(\hat{\boldsymbol{\theta}}) \tag{5.9}$$

is the MLE of $\boldsymbol{\eta} = h(\boldsymbol{\theta})$ when h is one to one.

If h is not one to one, then the new parameters $\boldsymbol{\eta} = h(\boldsymbol{\theta})$ do not give enough information to define $f(\boldsymbol{x}|\boldsymbol{\eta})$. Hence we cannot define the likelihood. That is, a $N(\mu, \sigma^2)$ density cannot be defined by the parameter μ/σ alone. Before concluding that the MLE does not exist if h is not one to one, note that if $X_1, ..., X_n$ are iid $N(\mu, \sigma^2)$ then $X_1, ..., X_n$ remain iid $N(\mu, \sigma^2)$ even though the investigator did not rename the parameters wisely or is interested in a function $h(\mu, \sigma) = \mu/\sigma$ that is not one to one. Berk (1967) said that if h is not one to one, define

$$w(\boldsymbol{\theta}) = (h(\boldsymbol{\theta}), u(\boldsymbol{\theta})) = (\boldsymbol{\eta}, \boldsymbol{\gamma}) = \boldsymbol{\xi} \tag{5.10}$$

such that $w(\boldsymbol{\theta})$ is one to one. Note that the choice

$$w(\boldsymbol{\theta}) = (h(\boldsymbol{\theta}), \boldsymbol{\theta})$$

works. In other words, we can always take u to be the identity function.

The choice of w is not unique, but the inverse function

$$w^{-1}(\boldsymbol{\xi}) = \boldsymbol{\theta}$$

is unique. Hence the likelihood is well defined, and $w(\hat{\boldsymbol{\theta}})$ is the MLE of $\boldsymbol{\xi}$. QED

Example 5.12. Following Lehmann (1999, p. 466), let

$$f(x|\sigma) = \frac{1}{\sqrt{2\pi}\ \sigma} \exp(\frac{-x^2}{2\sigma^2})$$

where x is real and $\sigma > 0$. Let $\eta = \sigma^k$ so $\sigma = \eta^{1/k} = h^{-1}(\eta)$. Then

$$f^*(x|\eta) = \frac{1}{\sqrt{2\pi}\ \eta^{1/k}} \exp(\frac{-x^2}{2\eta^{2/k}}) = f(x|\sigma = h^{-1}(\eta)).$$

Notice that calling $h(\hat{\boldsymbol{\theta}})$ the MLE of $h(\boldsymbol{\theta})$ is analogous to calling $\overline{X}_n$ the MLE of μ when the data are from a $N(\mu, \sigma^2)$ population. It is often possible to choose the function u so that if $\boldsymbol{\theta}$ is a $p \times 1$ vector, then so is $\boldsymbol{\xi}$. For the $N(\mu, \sigma^2)$ example with $h(\mu, \sigma^2) = h(\boldsymbol{\theta}) = \mu/\sigma$ we can take $u(\boldsymbol{\theta}) = \mu$ or $u(\boldsymbol{\theta}) = \sigma^2$. For the Ber$(\rho)$ example, $w(\rho) = (\rho(1-\rho), \rho)$ is a reasonable choice.

To summarize, Berk's proof should be widely used to prove the invariance principle, and

I) changing the names of the parameters does not change the distribution of the sample, eg, if $X_1, ..., X_n$ are iid $N(\mu, \sigma^2)$, then $X_1, ..., X_n$ remain iid $N(\mu, \sigma^2)$ regardless of the function $h(\mu, \sigma^2)$ that is of interest to the investigator.

II) The invariance principle holds as long as $h(\hat{\boldsymbol{\theta}})$ is a random variable or random vector: h does not need to be a one to one function. If there is interest in $\boldsymbol{\eta} = h(\boldsymbol{\theta})$ where h is not one to one, then additional parameters $\boldsymbol{\gamma} = u(\boldsymbol{\theta})$ need to be specified so that $w(\boldsymbol{\theta}) = \boldsymbol{\xi} = (\boldsymbol{\eta}, \boldsymbol{\gamma}) = (h(\boldsymbol{\theta}), u(\boldsymbol{\theta}))$ has a well defined likelihood $K(\boldsymbol{\xi}) = L(w^{-1}(\boldsymbol{\xi}))$. Then by Definition 5.2, the MLE is $\hat{\boldsymbol{\xi}} = w(\hat{\boldsymbol{\theta}}) = w(h(\hat{\boldsymbol{\theta}}), u(\hat{\boldsymbol{\theta}}))$ and the MLE of $\boldsymbol{\eta} = h(\boldsymbol{\theta})$ is $\hat{\boldsymbol{\eta}} = h(\hat{\boldsymbol{\theta}})$.

III) Using the identity function $\boldsymbol{\gamma} = u(\boldsymbol{\theta}) = \boldsymbol{\theta}$ always works since $\boldsymbol{\xi} = w(\boldsymbol{\theta}) = (h(\boldsymbol{\theta}), \boldsymbol{\theta})$ is a one to one function of $\boldsymbol{\theta}$. However, using $u(\boldsymbol{\theta})$ such that $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$ have the same dimension is often useful.

5.5 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

Refer to Chapter 10 for the pdf or pmf of the distributions in the problems below.

5.1*. Let $Y_1, ..., Y_n$ be iid binomial $(k = 1, \rho)$.

a) Assume that $\rho \in \Theta = (0, 1)$ and that $0 < \sum_{i=1}^{n} y_i < n$. Show that the MLE of ρ is $\hat{\rho} = \overline{Y}$.

b) Now assume that $\rho \in \Theta = [0, 1]$. Show that $f(y|\rho) = \rho^y (1-\rho)^{1-y} I(0 < \rho < 1) + I(\rho = 0, y = 0) + I(\rho = 1, y = 1)$. Then show that

$$L(\rho) = \rho^{\sum y}(1-\rho)^{n-\sum y} I(0 < \rho < 1) + I(\rho = 0, \sum y = 0) + I(\rho = 1, \sum y = n).$$

If $\sum y = 0$ show that $\hat{\rho} = 0 = \overline{y}$. If $\sum y = n$ show that $\hat{\rho} = 1 = \overline{y}$. Then explain why $\hat{\rho} = \overline{Y}$ if $\Theta = [0, 1]$.

5.2. (1989 Univ. of Minn. and Aug. 2000 SIU QUAL): Let (X, Y) have the bivariate density

$$f(x, y) = \frac{1}{2\pi} \exp(\frac{-1}{2}[(x - \rho\cos\theta)^2 + (y - \rho\sin\theta)^2]).$$

Suppose that there are n independent pairs of observations (X_i, Y_i) from the above density and that ρ is known. Assume that $0 \leq \theta \leq 2\pi$. Find a candidate for the maximum likelihood estimator $\hat{\theta}$ by differentiating the log likelihood $L(\theta)$. (Do not show that the candidate is the MLE, it is difficult to tell whether the candidate, 0 or 2π is the MLE without the actual data.)

5.3*. Suppose a single observation $X = x$ is observed where X is a random variable with pmf given by the table below. Assume $0 \leq \theta \leq 1$, and find the MLE $\hat{\theta}_{MLE}(x)$. (Hint: drawing $L(\theta) = L(\theta|x)$ for each of the four values of x may help.)

x	1	2	3	4
$f(x\|\theta)$	1/4	1/4	$\frac{1+\theta}{4}$	$\frac{1-\theta}{4}$

5.4. Let $X_1, ..., X_n$ be iid normal $N(\mu, \gamma_o^2 \mu^2)$ random variables where $\gamma_o^2 > 0$ is **known** and $\mu > 0$. Find the log likelihood $\log(\mathrm{L}(\mu|x_1, ..., x_n))$ and solve

$$\frac{d}{d\mu} \log(L(\mu|x_1, ..., x_n)) = 0$$

for $\hat{\mu}_o$, a potential candidate for the MLE of μ.

5.5. Suppose that $X_1, ..., X_n$ are iid uniform $U(0, \theta)$. Use the factorization theorem to write $f(\boldsymbol{x}|\theta) = g(T(\boldsymbol{x})|\theta)$ (so $h(\boldsymbol{x}) \equiv 1$) where $T(\boldsymbol{x})$ is a one

dimensional sufficient statistic. Then plot the likelihood function $L(\theta) = g(T(\boldsymbol{x})|\theta)$ and find the MLE of θ.

5.6. Let $Y_1, ..., Y_n$ be iid Burr(λ, ϕ) with ϕ known. Find the MLE of λ.

5.7. Let $Y_1, ..., Y_n$ be iid chi(p, σ) with p known. Find the MLE of σ^2.

5.8. Let $Y_1, ..., Y_n$ iid double exponential $DE(\theta, \lambda)$ with θ known. Find the MLE of λ.

5.9. Let $Y_1, ..., Y_n$ be iid exponential EXP(λ). Find the MLE of λ.

5.10. If $Y_1, ..., Y_n$ are iid gamma $G(\nu, \lambda)$ with ν known, find the MLE of λ.

5.11. If $Y_1, ..., Y_n$ are iid geometric geom(ρ), find the MLE of ρ.

5.12. If $Y_1, ..., Y_n$ are iid inverse Gaussian $IG(\theta, \lambda)$ with λ known, find the MLE of θ.

5.13. If $Y_1, ..., Y_n$ are iid inverse Gaussian $IG(\theta, \lambda)$ with θ known, find the MLE of λ.

5.14. If $Y_1, ..., Y_n$ are iid largest extreme value LEV(θ, σ) where σ is known, find the MLE of θ.

5.15. If $Y_1, ..., Y_n$ are iid negative binomial $NB(r, \rho)$ with r known, find the MLE of ρ.

5.16. If $Y_1, ..., Y_n$ are iid Rayleigh $R(\mu, \sigma)$ with μ known, find the MLE of σ^2.

5.17. If $Y_1, ..., Y_n$ are iid Weibull $W(k, \rho)$ with k known, find the MLE of ρ.

5.18. If $Y_1, ..., Y_n$ are iid binomial $BIN(\phi, \lambda)$ with ϕ known, find the MLE of λ.

5.19. Suppose $Y_1, ..., Y_n$ are iid two parameter exponential EXP(θ, λ).

a) Show that for any fixed $\lambda > 0$, the log likelihood is maximized by $y_{(1)}$. Hence the MLE $\hat{\theta} = Y_{(1)}$.

b) Find $\hat{\lambda}$ by maximizing the profile likelihood.

5.20. Suppose $Y_1, ..., Y_n$ are iid truncated extreme value TEV(λ). Find the MLE of λ.

Problems from old quizzes and exams.

Note: Problem 5.21 would be better if it replaced "$\lambda \geq 0$" by "$\lambda > 0$, and assume $\sum x_i > 0$." But problems like 5.21 are extremely common on exams and in texts.

5.21. Suppose that $X_1, ..., X_n$ are iid Poisson with pmf

$$f(x|\lambda) = P(X = x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$$

where $x = 0, 1, ...$ and $\lambda \geq 0$.

a) Find the MLE of λ. (Make sure that you prove that your estimator maximizes the likelihood).

b) Find the MLE of $(1 - \lambda)^2$.

5.22. Suppose that $X_1, ..., X_n$ are iid $U(0, \theta)$. Make a plot of $L(\theta|x_1, ..., x_n)$.

a) If the uniform density is $f(x) = \frac{1}{\theta} I(0 \leq x \leq \theta)$, find the MLE of θ if it exists.

b) If the uniform density is $f(x) = \frac{1}{\theta} I(0 < x < \theta)$, find the MLE of θ if it exists.

5.23. (Jan. 2001 Qual): Let $X_1, ..., X_n$ be a random sample from a normal distribution with **known** mean μ and unknown variance τ.

a) Find the maximum likelihood estimator of the variance τ.

b) Find the maximum likelihood estimator of the standard deviation $\sqrt{\tau}$. Explain how the MLE was obtained.

5.24. Suppose a single observation $X = x$ is observed where X is a random variable with pmf given by the table below. Assume $0 \leq \theta \leq 1$. and find the MLE $\hat{\theta}_{MLE}(x)$. (Hint: drawing $L(\theta) = L(\theta|x)$ for each of the values of x may help.)

x	0	1
$f(x\|\theta)$	$\frac{1+\theta}{2}$	$\frac{1-\theta}{2}$

5.25. Suppose that X is a random variable with pdf $f(x|\theta) = (x-\theta)^2/3$ for $\theta - 1 \leq x \leq 2 + \theta$. Hence $L(\theta) = (x-\theta)^2/3$ for $x - 2 \leq \theta \leq x + 1$. Suppose that one observation $X = 7$ was observed. Find the MLE $\hat{\theta}$ for θ. (Hint: evaluate the likelihood at the critical value and the two endpoints. One of these three values has to be the MLE.)

5.26. Let $X_1, ..., X_n$ be iid from a distribution with pdf

$$f(x|\theta) = \theta x^{-2}, \quad 0 < \theta \leq x < \infty.$$

a) Find a minimal sufficient statistic for θ.

b) Find the MLE for θ.

5.27. Let $Y_1, ..., Y_n$ be iid from a distribution with probability mass function

$$f(y|\theta) = \theta(1-\theta)^y, \quad \text{where y} = 0, 1, ... \text{ and } 0 < \theta < 1.$$

Assume $0 < \sum y_i < n$.

a) Find the MLE of θ. (Show that it is the global maximizer.)

c) What is the MLE of $1/\theta^2$? Explain.

5.28. (Aug. 2002 QUAL): Let $X_1, ..., X_n$ be independent identically distributed random variables from a half normal HN(μ, σ^2) distribution with pdf

$$f(x) = \frac{2}{\sqrt{2\pi}\ \sigma} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

where $\sigma > 0$ and $x > \mu$ and μ is real. **Assume that μ is known.**

a) Find the maximum likelihood estimator of σ^2.

b) What is the maximum likelihood estimator of σ? Explain.

5.29. (Jan. 2003 QUAL): Let $X_1, ..., X_n$ be independent identically distributed random variables from a lognormal (μ, σ^2) distribution with pdf

$$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(\log(x)-\mu)^2}{2\sigma^2}\right)$$

where $\sigma > 0$ and $x > 0$ and μ is real. **Assume that σ is known.**

a) Find the maximum likelihood estimator of μ.

b) What is the maximum likelihood estimator of μ^3? Explain.

5.30. (Aug. 2004 QUAL): Let X be a single observation from a normal distribution with mean θ and with variance θ^2, where $\theta > 0$. Find the maximum likelihood estimator of θ^2.

5.31. (Sept. 2005 QUAL): Let $X_1, ..., X_n$ be independent identically distributed random variables with probability density function

$$f(x) = \frac{\sigma^{1/\lambda}}{\lambda} \exp\left[-(1 + \frac{1}{\lambda}) \log(x)\right] I[x \geq \sigma]$$

where $x \geq \sigma$, $\sigma > 0$, and $\lambda > 0$. The indicator function $I[x \geq \sigma] = 1$ if $x \geq \sigma$ and 0, otherwise. Find the maximum likelihood estimator (MLE) $(\hat{\sigma}, \hat{\lambda})$ of (σ, λ) with the following steps.

a) Explain why $\hat{\sigma} = X_{(1)} = \min(X_1, ..., X_n)$ is the MLE of σ regardless of the value of $\lambda > 0$.

b) Find the MLE $\hat{\lambda}$ of λ if $\sigma = \hat{\sigma}$ (that is, act as if $\sigma = \hat{\sigma}$ is known).

5.32. (Aug. 2003 QUAL): Let $X_1, ..., X_n$ be independent identically distributed random variables with pdf

$$f(x) = \frac{1}{\lambda} \exp\left[-(1 + \frac{1}{\lambda}) \log(x)\right]$$

where $\lambda > 0$ and $x \geq 1$.

a) Find the maximum likelihood estimator of λ.

b) What is the maximum likelihood estimator of λ^8 ? Explain.

5.33. (Jan. 2004 QUAL): Let $X_1, ..., X_n$ be independent identically distributed random variables with probability mass function

$$f(x) = e^{-2\theta} \frac{1}{x!} \exp[\log(2\theta)x],$$

for $x = 0, 1, \ldots,$ where $\theta > 0$. Assume that at least one $X_i > 0$.

a) Find the maximum likelihood estimator of θ.

b) What is the maximum likelihood estimator of $(\theta)^4$? Explain.

5.34. (Jan. 2006 QUAL): Let $X_1, ..., X_n$ be iid with one of two probability density functions. If $\theta = 0$, then

$$f(x|\theta) = \begin{cases} 1, & 0 \leq x \leq 1 \\ 0, & \text{otherwise.} \end{cases}$$

If $\theta = 1$, then

$$f(x|\theta) = \begin{cases} \frac{1}{2\sqrt{x}}, & 0 \leq x \leq 1 \\ 0, & \text{otherwise.} \end{cases}$$

Find the maximum likelihood estimator of θ.

Warning: Variants of the following question often appears on qualifying exams.

5.35. (Aug. 2006 Qual): Let $Y_1, ..., Y_n$ denote a random sample from a $N(a\theta, \theta)$ population.

a) Find the MLE of θ when $a = 1$.

b) Find the MLE of θ when a is known but arbitrary.

5.36. Suppose that $X_1, ..., X_n$ are iid random variable with pdf

$$f(x|\theta) = (x - \theta)^2/3$$

for $\theta - 1 \leq x \leq 2 + \theta$.

a) Assume that $n = 1$ and that $X = 7$ was observed. Sketch the log likelihood function $L(\theta)$ and find the maximum likelihood estimator (MLE) $\hat{\theta}$.

b) Again assume that $n = 1$ and that $X = 7$ was observed. Find the MLE of

$$h(\theta) = 2\theta - \exp(-\theta^2).$$

5.37. (Aug. 2006 Qual): Let $X_1, ..., X_n$ be independent identically distributed (iid) random variables with probability density function

$$f(x) = \frac{2}{\lambda\sqrt{2\pi}} e^x \exp\left(\frac{-(e^x - 1)^2}{2\lambda^2}\right)$$

where $x > 0$ and $\lambda > 0$.

a) Find the maximum likelihood estimator (MLE) $\hat{\lambda}$ of λ.

b) What is the MLE of λ^2? Explain.

5.38. (Jan. 2007 Qual): Let $X_1, ..., X_n$ be independent identically distributed random variables from a distribution with pdf

$$f(x) = \frac{2}{\lambda\sqrt{2\pi}} \frac{1}{x} \exp\left[\frac{-(\log(x))^2}{2\lambda^2}\right]$$

where $\lambda > 0$ where and $0 \leq x \leq 1$.

a) Find the maximum likelihood estimator (MLE) of λ.

b) Find the MLE of λ^2.