Statistical Theory - Problem Sheet 2

Spring 2023. Email corrections to Kolyan Ray: kolyan.ray@ic.ac.uk

1. Define $\overline{X} := \frac{1}{n}\sum_{j=1}^n X_j$ and $\overline{X^2} := \frac{1}{n}\sum_{j=1}^n X_j^2$.

Distribution	$\hat\theta_{MLE}$	score equation	Fisher information
(a) Bernoulli(θ)	$\overline{X}$	$n(\overline{x}-\theta)/(\theta(1-\theta))=0$	$n/(\theta(1-\theta))$
(b) N(θ,1)	$\overline{X}$	$-n(\overline{x}-\theta)=0$	n
(c) N(0,θ)	$\overline{X^2}$	$n(\overline{x^2}-\theta)/(2\theta^2)=0$	$n/(2\theta^2)$
(d) N(μ,σ^2)	$\left(\overline{X}, \frac{1}{n}\sum_{j=1}^n (X_j-\overline{X})^2\right)^\top$	see below	see below
(e) Poisson(θ)	$\overline{X}$	$n(\overline{x}-\theta)/\theta=0$	n/θ
(f) $(1/\theta)e^{-x/\theta}$	$\overline{X}$	$n(\overline{x}-\theta)/\theta^2=0$	n/θ^2
(g) $\theta e^{-\theta x}$	$1/\overline{X}$	$n(1/\theta-\overline{x})=0$	n/θ^2

The score equation and the Fisher information for $N(\mu,\sigma^2)$ are

$$\begin{pmatrix} \frac{1}{\sigma^2}\sum_{j=1}^n (X_j-\mu) \\ -\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{j=1}^n (X_j-\mu)^2 \end{pmatrix} = 0 \quad \text{and} \quad \begin{pmatrix} n/\sigma^2 & 0 \\ 0 & n/(2\sigma^4) \end{pmatrix}.$$

2. The MLE is unbiased in (a), (b), (c), (e) and (f), and in all these cases, its variance attains the Cramer–Rao lower bound. This can be checked by either directly computing the variance or using Proposition 2.2 in the lecture notes, which provides conditions under which an estimator attains the Cramer-Rao bound.

For (d), we have $E_\theta[\hat\theta_{MLE}] = (\mu,(n-1)\sigma^2/n)^\top$ and thus the MLE is biased. We saw that (g) is biased in Q7 on Problem Sheet 1. For another proof, note that by Jensen's inequality and by (f) $E_\theta[\hat\theta_{MLE}] = E_\theta[1/\overline{X}] \geq 1/E_\theta[\overline{X}] = \theta$. The function $(0,\infty)\to(0,\infty), x\mapsto 1/x$ is strictly convex and $P(\hat\theta_{MLE}\neq\theta)>0$. We conclude that Jensen's inequality is strict and $E_\theta[\hat\theta_{MLE}]>\theta$. Consequently, the MLE is biased.

3. We saw in the last question that $\bar X_n$ attains the Cramer-Rao lower bound in (e). Since $E_\theta S_n^2 = \mathrm{Var}_\theta(X_1)=\theta$, S_n^2 is an unbiased estimator of the variance θ of the Poisson distribution. By the Cramer-Rao lower bound, $\frac{1}{nI_{X_1}(\theta)} = \mathrm{Var}(\bar X_n) \leq \mathrm{Var}(S_n^2)$.

4. Using the invariance of MLE, the MLEs are (i) $\bar X_n$, (ii) $\bar X_n(1-\bar X_n)$, (iii) $(\bar X_n)^2$.

5. Since $f_\theta(x_1,\dots,x_n)=\prod_{i=1}^n f_\theta(x_i)$ by the independence of the X_i's, it follows that $\log f_\theta(x_1,\dots,x_n) = \sum_{i=1}^n \log f_\theta(x_i)$ and

$$I_{(X_1,\dots,X_n)}(\theta) = E_\theta\left[\sum_{i=1}^n\sum_{j=1}^n \nabla_\theta \log f_\theta(X_i)\nabla_\theta \log f_\theta(X_j)^T\right].$$

For every X_i, we recall that $E_\theta[\nabla_\theta \log f_\theta(X_i)]=0$ (Lemma 2.2 in the notes). Using this and the independence, the last sum equals

$$I_{(X_1,\dots,X_n)}(\theta) = \sum_{i=1}^n E_\theta\left[\nabla_\theta \log f_\theta(X_i)\nabla_\theta \log f_\theta(X_i)^T\right] = nI_{X_1}(\theta).$$

6. Let $\Theta = \mathbb{R}$ and define

$$Q_n(\theta) := \begin{cases} 1/2 & \text{for } \theta = 0 \\ 1 & \text{for } \theta = n \\ 0 & \text{otherwise} \end{cases} \qquad \text{and} \qquad Q(\theta) := \frac{1}{2} 1_{\{0\}}(\theta).$$

Then $Q_n(\theta) \to Q(\theta)$ for all $\theta \in \Theta$ as $n \to \infty$, but $\hat{\theta}_n = n \nrightarrow \theta_0 = 0$.

7. (a) When there is no restriction on θ, we saw that the MLE is $\bar{X}_n$. However, if $\bar{X}_n < 0$, this is outside of the parameter space $\Theta = [0, \infty)$. If $\bar{x}_n < 0$, then the log-likelihood $\ell_n : [0, \infty) \to \mathbb{R}$ equals

$$\ell_n(\theta) = \log \left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-(x_i - \theta)^2/2} \right) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^{n} x_i^2 + n\bar{x}_n \theta - \frac{n\theta^2}{2}.$$

Differentiating, $\ell_n'(\theta) = n(\bar{x}_n - \theta) < 0$ if $\theta > 0$ and $\bar{x}_n < 0$. Thus ℓ_n is a decreasing function of θ for $\theta \geq 0$ if $\bar{x}_n < 0$, and is thus maximized at $\theta = 0$. The MLE is therefore

$$\hat{\theta}_{ML} = \bar{X}_n 1\{\bar{X}_n \geq 0\} = \begin{cases} \bar{X}_n & \text{if } \bar{X}_n \geq 0, \\ 0 & \text{if } \bar{X}_n < 0. \end{cases}$$

(b) The likelihood equals

$$f_\theta(x) := \frac{1}{(2\pi\theta)^{n/2}} \exp \left(-\frac{1}{2\theta} \sum_{j=1}^{n} (x_j - \theta)^2 \right).$$

At $x_1 = \cdots = x_n = 0$ the pdf tends to infinity as $\theta \to 0$ and so the MLE does not exist for $X_1 = \cdots = X_n = 0$ (note $0 \notin \Theta$). To determine the MLE if there is one j with $X_j \neq 0$, we write

$$\begin{aligned} \ell_n(\theta) &= -\frac{n}{2} \log(2\pi\theta) - \frac{1}{2} \sum_{j=1}^{n} \frac{(x_j - \theta)^2}{\theta}, \\ \ell_n'(\theta) &= -\frac{n}{2\theta} + \sum_{j=1}^{n} \frac{x_j - \theta}{\theta} + \frac{1}{2} \sum_{j=1}^{n} \frac{(x_j - \theta)^2}{\theta^2} \\ &= -\frac{n}{2\theta^2}\theta + \frac{1}{2\theta^2} \sum_{j=1}^{n} (x_j + \theta)(x_j - \theta) \\ &= -\frac{n}{2\theta^2} \left(\theta^2 + \theta - \overline{x^2} \right), \end{aligned}$$

where $\overline{x^2} = \frac{1}{n} \sum_{j=1}^{n} x_j^2$. The zeros are

$$\theta_\pm = -\frac{1}{2} \pm \sqrt{\left(\frac{1}{2}\right)^2 + \overline{x^2}}.$$

For $x \neq 0$ we have $\theta_- < 0$ and $\theta_+ > 0$. Since $\ell_n'(\theta) > 0$ on $(0, \theta_+)$ and $\ell_n'(\theta) < 0$ on (θ_+, ∞), $\ell : (0, \infty) \to \mathbb{R}$ has a global maximum at θ_+ and thus θ_+ is the MLE.

8. The log-likelihood and score equal

$$\ell_n(\theta) = -n \log 2 - \sum_{j=1}^{n} |x_j - \theta| \qquad \text{and} \qquad \ell_n'(\theta) = \sum_{j=1}^{n} \text{sign}(x_j - \theta),$$

where the score function is defined for $\theta \neq x_1, \ldots, x_n$. We shall directly maximize $-\sum_{j=1}^n |x_j - \theta|$. Without loss of generality, reorder the observations so that $X_1 < X_2 < \cdots < X_n$, and assume these take distinct values, which happens with probability one. When $\ell_n(\theta)$ is differentiable, we have $\ell_n'(\theta) = \sum_{j=1}^n \mathrm{sign}(x_j - \theta) \neq 0$ since n is odd. So the maximizer must occur at one of the points where the function is non-differentiable, i.e. one of $X_1, \ldots, X_n$. We see that $\ell_n(\theta)$ is continuous everywhere, piecewise linear and decreasing for $\theta < X_{(n+1)/2}$ and increasing for $\theta > X_{(n+1)/2}$. Therefore, the maximizer is given by the sample median $X_{(n+1)/2}$.

Let $X_{(1)}, \ldots, X_{(n)}$ be the order statistics. We again study $\ell_n'(\theta) = \sum_{j=1}^n \mathrm{sign}(x_j - \theta)$. If n is even and $X_{(n/2)} = X_{(n/2+1)}$, then the MLE is unique and equal to the sample median. If n is even and $X_{(n/2)} \neq X_{(n/2+1)}$, then any point in $[X_{(n/2)}, X_{(n/2+1)}]$ is an MLE. Strictly speaking, the Fisher information does not exist since the likelihood is not differentiable. However, note that by symmetry of the Laplace distribution about θ, $E_\theta \mathrm{sign}(X_j - \theta) = 0$. Using this and the independence, one can therefore calculate

$$E_\theta[\ell_n'(\theta)^2] = E_\theta\left[\left(\sum_{j=1}^n \mathrm{sign}(X_j - \theta)\right)^2\right] = E\left[\sum_{j=1}^n (\mathrm{sign}(X_j - \theta))^2\right] = n.$$

9. This is the standard linear regression model with Gaussian errors, so the MLE is $(X^TX)^{-1}X^TY$ (see Statistical Modelling I notes). Since $Y = X\theta + Z$ for $Z \sim N_n(0, I_n)$, the MLE equals in distribution $(X^TX)^{-1}X^T(X\theta + Z) = \theta + (X^TX)^{-1}X^TZ$. Using that if $W \sim N(0, \Sigma)$ then $AW \sim N(0, A\Sigma A^T)$, the MLE has distribution $N_p(\theta, (X^TX)^{-1})$.

The log-likelihood in this model is

$$\ell(\theta) = \log\left(\frac{1}{(2\pi)^{n/2}} \exp\left(-\tfrac{1}{2}\|Y - X\theta\|^2\right)\right) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\|Y - X\theta\|^2,$$

so that $\nabla_\theta \ell(\theta) = X^T(Y - X\theta)$. Using that $Y - X\theta = Z \sim N_n(0, I_n)$ under E_θ, we can compute the Fisher information as

$$\begin{aligned} I(\theta) = E_\theta[\nabla_\theta \ell(\theta) \nabla_\theta \ell(\theta)^T] &= E_\theta[X^T(Y - X\theta)(Y - X\theta)^TX] \\ &= X^TE_\theta[ZZ^T]X = X^TI_nX = X^TX. \end{aligned}$$

The covariance of the MLE therefore equals the inverse Fisher information. For $p = n$ and $X = I$, the MLE is Y and the Fisher information is I.

10. The asymptotic distributions are given in the following table.

(a)	(b)	(c)	(d)	(e)	(f)	(g)
$N(0, \theta(1-\theta))$	$N(0,1)$	$N(0, 2\theta^2)$	see below	$N(0,\theta)$	$N(0,\theta^2)$	$N(0,\theta^2)$

The asymptotic distribution in (d) is given by

$$N\left(0, \begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix}\right).$$

In (a), (b), (c), (e) and (f) the MLE is a mean and the asymptotic distributions follows by the CLT.

For (g), set $f_\lambda(x) = \lambda e^{-\lambda x} = (1/\theta)e^{-x/\theta}$ using the parametrization from (f). We want to derive the asymptotic distribution of $\sqrt{n}(\hat{\lambda} - \lambda) = \sqrt{n}(1/\hat{\theta} - 1/\theta)$, where θ and $\hat{\theta}$ are as in (f). From (f) we know $\sqrt{n}(\hat{\theta} - \theta) \to N(0, \theta^2)$. A Taylor expansion of $f(x) = 1/x$ at θ yields

$$\frac{1}{x} = \frac{1}{\theta} + f'(\theta)(x - \theta) + \frac{1}{2}f''(t)(x - \theta)^2,$$

for some t between x and θ. Letting $x = \hat{\theta}$ yields

$$\sqrt{n}\left(\frac{1}{\hat{\theta}} - \frac{1}{\theta}\right) = -\frac{\sqrt{n}(\hat{\theta}-\theta)}{\theta^2} + \frac{1}{t^3}\sqrt{n}(\hat{\theta}-\theta)(\hat{\theta}-\theta).$$

By the WLLN, we now that $\hat{\theta} \to^p \theta$, and hence also $t \to^p \theta$. From (f), we know that $\sqrt{n}(\hat{\theta} - \theta) \to N(0, \theta^2)$. So by Slutsky's lemma, the last term converges in distribution to zero. For $f_\lambda(x) = \lambda e^{-\lambda x}$, we obtain

$$\sqrt{n}\left(\hat{\lambda} - \lambda\right) = \sqrt{n}\left(\frac{1}{\hat{\theta}} - \frac{1}{\theta}\right) = -\frac{\sqrt{n}(\hat{\theta}-\theta)}{\theta^2} \to N(0, \theta^{-2}) = N(0, \lambda^2).$$

[This is roughly the proof of the delta method, which we will see later in the course].

In (d) we have by the CLT $\sqrt{n}(\overline{X} - \mu) \to^d N(0, \sigma^2)$. For the estimator of the variance in (d), observe that

$$\begin{aligned}\sqrt{n}\left(\tfrac{1}{n}\textstyle\sum_{j=1}^n (X_j - \overline{X})^2 - \sigma^2 \right) &= \sqrt{n}\left(\tfrac{1}{n}\textstyle\sum_{j=1}^n ((X_j - \mu) - (\overline{X} - \mu))^2 - \sigma^2 \right) \\ &= \sqrt{n}\left(\tfrac{1}{n}\textstyle\sum_{j=1}^n (X_j - \mu)^2 - (\overline{X} - \mu)^2 - \sigma^2 \right)\end{aligned}$$

By the CLT for the i.i.d. random variables $(X_j - \mu)^2$ and using normality of the X_j, we have

$$\sqrt{n}\left(\tfrac{1}{n}\textstyle\sum_{j=1}^n (X_j - \mu)^2 - \sigma^2 \right) \to^d N(0, 2\sigma^4).$$

By the WLLN $(\overline{X} - \mu) \to^d 0$ and as seen above by the CLT $\sqrt{n}(\overline{X} - \mu) \to^d N(0, \sigma^2)$, so that by Slutsky's lemma $\sqrt{n}(\overline{X} - \mu)^2 \to^d 0$. We conclude using Slutsky's lemma

$$\sqrt{n}\left(\tfrac{1}{n}\textstyle\sum_{j=1}^n (X_j - \overline{X})^2 - \sigma^2 \right) \to^d N(0, 2\sigma^4).$$

It is well known that for $X_j \sim^{i.i.d.} N(\mu, \sigma^2)$, the two random variables $\overline{X}$ and $\frac{1}{n}\sum_{j=1}^n (X_j - \overline{X})^2$ are independent. Collecting the above results, we obtain in (d)

$$\sqrt{n}(\hat{\theta}_{MLE} - \theta) \to^d N\left(0, \begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix}\right).$$

11. (i) (a) We fix $\theta_2 = b$. The log-likelihood equals

$$\begin{aligned}\ell(\theta_1) &= -\log(2\pi|\Sigma|^{\frac{1}{2}}) - \frac{1}{2}(x_1 - \theta_1,\ x_2 - b)\Sigma^{-1}\begin{pmatrix} x_1 - \theta_1 \\ x_2 - b \end{pmatrix} \\ \ell'(\theta_1) &= (1,\ 0)\Sigma^{-1}\begin{pmatrix} x_1 - \theta_1 \\ x_2 - b \end{pmatrix} \\ \ell''(\theta_1) &= -(1,\ 0)\Sigma^{-1}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = -(\Sigma^{-1})_{11},\end{aligned}$$

and hence $I(\theta_1) = (\Sigma^{-1})_{11}$. The Cramér–Rao lower bound is thus $I(\theta_1)^{-1} = (\Sigma^{-1})_{11}^{-1}$.
(b) In this case,

$$\begin{aligned}\ell(\theta) &= -\log(2\pi|\Sigma|^{\frac{1}{2}}) - \frac{1}{2}(x_1 - \theta_1,\ x_2 - \theta_2)\Sigma^{-1}\begin{pmatrix} x_1 - \theta_1 \\ x_2 - \theta_2 \end{pmatrix} \\ \nabla\ell(\theta) &= \left((1,\ 0)\Sigma^{-1}\begin{pmatrix} x_1 - \theta_1 \\ x_2 - \theta_2 \end{pmatrix}, (0,\ 1)\Sigma^{-1}\begin{pmatrix} x_1 - \theta_1 \\ x_2 - \theta_2 \end{pmatrix}\right).\end{aligned}$$

Differentiating again,

$$\frac{\partial^2 \ell}{\partial\theta_i \partial\theta_j} = -e_i^T \Sigma^{-1} e_j = -(\Sigma^{-1})_{ij},$$

where $e_1 = (1, 0)$ and $e_2 = (0, 1)$ are the usual basis vectors. Thus the Fisher information is $I(\theta) = \Sigma^{-1}$. We now want to compute the Cramer-Rao bound for the one-dimensional quantity $g(\theta) = \theta_1$. We know from a remark in the notes that this equals

$$\nabla_\theta g(\theta)^T I(\theta)^{-1} \nabla_\theta g(\theta) = (1,\ 0) I(\theta)^{-1} (1,\ 0)^\top = (1,\ 0)\Sigma(1,\ 0)^\top = \Sigma_{11}.$$

(ii) For a diagonal matrix Σ, we have $(\Sigma^{-1})_{11} = \Sigma_{11}^{-1}$, so that $(\Sigma^{-1})_{11}^{-1} = \Sigma_{11}$.

(iii) For a 2×2 covariance matrix Σ,

$$\Sigma = \begin{pmatrix} a & b \\ b & c \end{pmatrix}, \qquad \Sigma^{-1} = \frac{1}{ac - b^2} \begin{pmatrix} c & -b \\ -b & a \end{pmatrix},$$

and $c > 0$ since Σ is positive definite.

$$(\Sigma^{-1})_{11}^{-1} = \frac{ac - b^2}{c} = a - \frac{b^2}{c} \leq a = \Sigma_{11}.$$

If $X_1 - \theta_1$ and $X_2 - \theta_2$ are uncorrelated, then the Cramér–Rao lower bounds are the same regardless of whether θ_2 is known or unknown. However, if $X_1 - \theta_1$ and $X_2 - \theta_2$ are correlated and θ_2 is known, then the Fisher information is larger than for unknown θ_2. The error in the second component $X_2 - \theta_2$ contains additional information for estimation of θ_1 through the correlation of $X_1 - \theta_1$ and $X_2 - \theta_2$.