Habib University | Dr. Muntazir Abidi
Dhanani School of Science and Engineering | Integrated Sciences and Mathematics (iSciM)

AstroStatistics Spring 2022
Exercise Sheet 3
Issued : 30 March 2022 **Due : 14 April 2022**

Useful Notes :

— **Unbiased estimators :** An estimator of a parameter θ is a function $T = T(X)$ which we use to estimate θ from an observation of X. T is said to be unbiased if

$$\mathrm{E}(T) = \theta \tag{1}$$

— **Maximum likelihood estimation :** Suppose that the random variable X has probability density function $f(x|)$. Given the observed value x of X, the likelihood of θ is defined by

$$\mathcal{L}(\theta) = f(x|\theta) \tag{2}$$

Thus we are considering the density as a function of θ, for a fixed x. In the case of multiple observations we write the likelihood function as

$$\mathcal{L}(\theta) = f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) \tag{3}$$

Thus the maximum likelihood estimate $\hat{\theta}(x)$ of θ is defined as the value of θ that maximizes the likelihood ; we call it the maximum likelihood estimator (MLE) of θ.

— **Sufficient statistics :** If MLE exists is always a function of the sufficient statistics. The sufficient statistics summarises all information which is relevant to do inference about θ. It is important to know the following theorem called the factorization criterion :

Theorem : *The statistics T is sufficient for θ if and only if $f(x|\theta)$ can be expressed as*

$$f(x|\theta) = g(T(X), \theta)h(x). \tag{4}$$

Problem 1

Gaussian 1D problem. The surface temperature on Mars is measured by a probe 10 times, yielding the following data (units of K) :

$$\text{Data} : \{191.9, \quad 201.6, \quad 206.1, \quad 200.4, \quad 203.2, \quad 201.6, \quad 196.5, \quad 199.5, \quad 194.1, \quad 202.4\} \tag{5}$$

1. Assume that each measurement is independently Normally distributed with known variance $\sigma^2 = 25$ K2. What is the likelihood function for the whole data set ?
2. Find the Maximum Likelihood Estimate (MLE) for the surface temperature, $\hat{T}_{\mathrm{ML}}$, and express your result to 4 significant figures accuracy.
3. Determine symmetric confidence intervals at 68.3%, 95.4%, and 99% around $\hat{T}_{\mathrm{ML}}$ (4 significant figures accuracy)
4. How many measurements would you need to make if you wanted to have a 1σ confidence interval around the mean of length less than 1 K (on each side) ?

Problem 2

You flip a coin $n = 10$ times and you obtain 8 heads.

1. What is the likelihood function for this measurement ? Identify explicitly what are the data and what is the free parameter you are trying to estimate.
2. What is the Maximum Likelihood Estimate for the probability of obtaining heads in one flip, p ?
3. Approximate the likelihood function as a Gaussian around its peak and derive the 1σ confidence interval for p. How would you report your result for p ?
4. With how many σ confidence can you exclude the hypothesis that the coin is fair ? (*Hint : compute the distance between the MLE for p and $p = 1/2$ and express the result in number of σ*).
5. You now flip the coin 1000 times and obtain 800 heads. What is the MLE for p now and what is the 1σ confidence interval for p ? With how many σ confidence can you exclude the hypothesis that the coin is fair now ?

Problem 3

An experiment counting particles emitted by a radioactive decay measures r particles per unit time interval. The counts are Poisson distributed.

1. If λ is the average number of counts per per unit time interval, write down the appropriate probability distribution function for r.
2. Now we seek to determine λ by repeatedly measuring for M times the number of counts per unit time interval. This series of measurements yields a sequence of counts $r = \hat{r}_1, \hat{r}_2, \hat{r}_3, \ldots, \hat{r}_M$. Each measurement is assumed to be independent. Derive the joint likelihood function for λ, $\mathcal{L}(\lambda) = P(\hat{r}|\lambda)$, given the measured sequence of counts $\hat{r}$
3. Use the Maximum Likelihood Principle applied to the the log likelihood $\ln \mathcal{L}(\lambda)$ to show that the Maximum Likelihood estimator for the average rate λ is just the average of the measured counts, $\hat{r}$, i.e
$$\hat{\lambda}_{\mathrm{ML}} = \frac{1}{M} \sum_{i=1}^{M} \hat{r}_i \tag{6}$$
4. By considering the Taylor expansion of $\ln \mathcal{L}(\lambda)$ to second order around $\hat{\lambda}_{\mathrm{ML}}$, derive the Gaussian approximation for the likelihood $\mathcal{L}(\lambda)$ around the Maximum Likelihood point, and show that it can be written as
$$\mathcal{L}(\lambda) \approx L_0 \exp\Big(-\frac{1}{2} \frac{M}{\hat{\lambda}_{\mathrm{ML}}} (\lambda - \hat{\lambda}_{\mathrm{ML}})^2 \Big) \tag{7}$$
where L_0 is a normalisation constant.
5. Compare with the equivalent expression for M Gaussian-distributed measurements to show that the variance σ^2 of the Poisson distribution is given by $\sigma^2 = \lambda$.

Problem 4

Suppose you measure the flux F of photons from a laser source using 4 different instruments and you obtain the following results (units of 104 photons/cm^2) :

$$\text{Data} : \{34.7 \pm 5.0, \quad 28.9 \pm 2.0, \quad 27.1 \pm 3.0, \quad 30.6 \pm 4.0.\} \tag{8}$$

1. Write down the likelihood for each measurement, and explain why a Gaussian approximation is justified in this case
2. Write down the joint likelihood for the combination of the 4 measurements.
3. Find the MLE of the photon flux, $\hat{F}_{\mathrm{ML}}$, and show that it is given by :

$$\hat{F}_{\mathrm{ML}} = \sum_i \frac{\hat{n}_i}{\hat{\sigma}_i^2/\bar{\sigma}^2} \tag{9}$$

   where

$$\frac{1}{\bar{\sigma}^2} \equiv \sum_i \frac{1}{\hat{\sigma}_i^2} \tag{10}$$

4. Compute $\hat{F}_{\mathrm{ML}}$ from the data above and compare it with the sample mean.
5. Find the 1σ confidence interval for your MLE for the mean, and show that it is given by :

$$\left(\sum_i \frac{1}{\hat{\sigma}_i^2}\right)^{-1/2} \tag{11}$$

   Evaluate the confidence interval for the above data. How would you summarize your measurement of the flux F ?

Problem 5

Suppose $X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$, where μ and σ^2 are unknown and the be estimated from the data.

1. Find the maximum likelihood estimator (MLE) for μ and σ. The MLEs are defined as $\hat{\mu}_{\mathrm{ML}}$ and $\hat{\sigma}^2_{\mathrm{ML}}$
2. Show that $\hat{\mu}_{\mathrm{ML}}$ is unbiased but $\hat{\sigma}^2_{\mathrm{ML}}$ is not unbiased. Find the unbiased estimator for the variance.
3. Can you give an example of the estimator of the variance which is neither the MLE nor unbiased ? (Hint : *the estimator which minimizes the mean squared error. more Hint : estimator of the form* $\lambda \sum_{i=1}^n (X_i - \bar{X})^2$, for what value of λ it is neither the MLE nor unbiased.)

Problem 6

In each of cases (a)–(c) write down the likelihood of θ and show that the stated $T(X)$ is a sufficient statistic for θ.

In each case also find a MLE of θ and show that it is a function of $T(X)$. Find the distribution of $T(X)$ and determine whether or not the MLE is an unbiased estimator of θ. If it is not, verify that it is asymptotically unbiased, and find some other estimator which is unbiased.

1. $X_1, \ldots, X_n$ are independent Poisson random variables, with X_i having mean $i\theta$, where $\theta > 0$. $T(X) = \sum_{i=1}^n X_i$.
2. $X_1, \ldots, X_n$ are independent normal random variables, with $X_i \sim N(\theta, \sigma_i^2)$, and $\sigma_i^2, i = 1, \ldots, n$, known. $T(X) = \sum_{i=1}^n X_i/\sigma_i^2$.
3. $X_1, \ldots, X_n$ are $n > 2$ independent and exponentially distributed random variables, with parameter θ, i.e., with density $f(x|\theta) = \theta e^{-\theta x}, x > 0$. $T(X) = \sum_{i=1}^n X_i$

Hint : *In case (a),* $T(X) \sim P(\frac{1}{2}n(n+1)\theta)$. *In case (b),* $T(X) \sim N(\theta \sum_i \sigma_i^{-2}, \sum_i \sigma_i^{-2})$. *In case (c),* $T(X) \sim gamma(n, \theta)$. *Do you understand why ?*