Parametric Inference

Moulinath Banerjee

University of Michigan

April 14, 2004

1 General Discussion

The object of statistical inference is to glean information about an underlying population based on a sample collected from it. The actual population is assumed to be described by some probability distribution. Statistical inference is concerned with learning about the distribution or at least some characteristics of the distribution that are of scientific interest.

In parametric statistical inference, which we will be primarily concerned with in this course, the underlying distribution of the population is taken to be parametrized by a Euclidean parameter. In other words, there exists a subset Θ of k-dimensional Euclidean space such that the class of dsitributions $\mathcal{P}$ of the underlying population can be written as $\{P_\theta : \theta \in \Theta\}$. You can think of the θ's as labels for the class of distributions under consideration. More precisely this will be our set-up: Our data $X_1, X_2, \ldots X_n$ are i.i.d. observations from the distribution P_θ where $\theta \in \Theta$, the parameter space. We assume identifiability of the parameter, i.e. $\theta_1 \neq \theta_2 \Rightarrow P_{\theta_1} \neq P_{\theta_2}$. In general, we will also assume that X_1 has a density $f(x, \theta)$ (this can either be a probability mass function or an ordinary probability density function). Here x is a typical value assumed by the random variable. Thus, $f(x, \theta)$ for a discrete random variable X_1 just gives us the probability that X_1 assumes the value x when the underlying parameter is indeed θ. For a continuous random variable, $f(x, \theta)$ gives us the density function of the random variable X_1 at the point x when θ is the underlying parameter. Thus $f(x, \theta)\, dx$ where dx is a very small number is approximately the probability that X_1 lives in the interval $[x, x + dx]$ under parameter value θ.

We will be interested in estimating θ, or more generally, a function of θ, say $g(\theta)$.

Let us consider a few examples that will enable us to understand these notions better.

(1) Let $X_1, X_2, \ldots, X_n$ be the outcomes of n independent flips of the same coin. Here, we code $X_i = 1$ if the i'th toss produces H and 0 otherwise. The parameter of interest is θ, the probability of H turning up in a single toss. This can be any number between 0 and 1. The

X_i's are i.i.d. and the common distribution P_θ is the Bernoulli(θ) distribution which has probability mass function:

$$f(x,\theta) = \theta^x \, (1-\theta)^{1-x} \quad , x \in \{0,1\} \, .$$

Check that this is indeed a valid expression for the p.m.f. Here the parameter space, i.e. the set of all possible values for θ is the closed interval $[0,1]$.

(2) Let $X_1, X_2, \ldots, X_n$ denote the failure times of n different bulbs. We can think of the X_i's as independent and identically distributed random variables from an exponential distribution with an unknown parameter θ which we want to estimate. If $F(x,\theta)$ denotes the distribution function of X_1 under parameter value θ, then

$$F(x,\theta) = P_{\theta \text{ is true}} \, (X_1 \leq x) = 1 - e^{-\theta \, x} \, .$$

The common density function is given by,

$$f(x,\theta) = \theta \, e^{-x \, \theta} \, .$$

Here the parameter space for θ is $(0,\infty)$.

Note that θ is very naturally related to the mean of the distribution. We have $E_{theta}(X_1) = 1/\theta$. The expression $E_\theta(X_1)$ should be read as the *expected value of* X_1 *when the true parameter is* θ. In general, whenever I write an expression with θ as a subscript, interpret that as *under the scenario that the true underlying parameter is* θ.

(3) Let $X_1, X_2, \ldots, X_n$ be the number of customers that arrive at n different identical counters in unit time. Then the X_i's can be thought of as i.i.d. random variables with a (common) Poisson distribution with mean θ. Once again θ which is also the parameter that completely specifices the Poisson distribution varies in the set $(0,\infty)$ which therefore is the parameter space Θ. The probability mass function is:

$$f(x,\theta) = \frac{e^{-\theta} \, \theta^x}{x \, !} \, .$$

(4) Let $X_1, X_2, \ldots, X_n$ be i.i.d. observations from a Normal distribution with mean μ and variance σ^2. The mean and the variance completely specify the normal distribution. We can take the $\theta = (\mu, \sigma^2)$. Thus, we have a two dimensional parameter and Θ, the set of all possible values of θ is the set in $\mathbb{R}^2$ given by $(-\infty,\infty) \times (0,\infty)$. The density function $f(x,\theta)$ is then given by:

$$f(x,\theta) = \frac{1}{\sqrt{2\,\pi}\,\sigma} \exp\left[-\frac{(x-\mu)^2}{2\,\sigma^2}\right] \, .$$

Note that each different value of θ gives you a different normal curve. if you fix μ, the first component of θ and vary σ^2 you get a family of normal (density) curves all centered at μ but with varying spread. A smaller value of σ^2 corresponds to a curve that is more peaked about μ and also more tightly packed around it. If you fix σ^2 and vary μ you get a family of curves that are all translates of a fixed curve (say, the one centered at 0).

Consider now, the problem of estimating $g(\theta)$ where g is some function of θ. (In class, I've used $\Psi(\theta)$ for $g(\theta)$.) In many cases $g(\theta) = \theta$ itself; for example, we could be interested in estimating θ, the probability of H in Example 1 above. Generally $g(\theta)$ will describe some important aspect of the distribution P_θ. In Example 1, $g(\theta) = \theta$ describes the probability of the coin landing heads; in Example 3, $g(\theta) = 1/\theta$ is the expected value of the lifetime of a bulb. Our estimate of $g(\theta)$ will be some function of our observed data $X = (X_1, X_2, \ldots, X_n)$. We will generically denote an estimate of $g(\theta)$ by $T_n(X_1, X_2, \ldots, X_n)$ and will write T_n for brevity. Thus T_n is some function of the observed data and is therefore a random variable itself.

Let's quickly look at an example. In Example 1, a natural estimate of θ, as we have discussed before, is $\overline{X}_n$, the mean of the X_i's. This is simply the sample proportion of Heads in n tosses of the coin. Thus

$$T_n(X_1, X_2, \ldots, X_n) = \frac{X_1 + X_2 + \ldots + X_n}{n}\,.$$

By the WLLN $\overline{X}_n$ converges in probability to θ and is therefore a reasonable estimator, at least in this sense. Of course, this is not the only estimator of θ that one can propose (but this is indeed the best estimator in more ways than one). One could also propose the proportion of heads in the first m tosses of the coin as an estimator, m being the floor of $n/2$. This will also converge in probability to θ as $n \to \infty$, but its variance will always be larger than that of $\overline{X}_n$.

In general there will be several different estimators of $g(\theta)$ which may all seem reasonable from different perspectives – the question then becomes one of finding the most optimal one. This requires an objective measure of the performance of the estimator. If T_n estimates $g(\theta)$ a criterion that naturally suggests itself is the distance of T_n from $g(\theta)$. Good estimators are those for which $\mid T_n - g(\theta) \mid$ is generally small. Since T_n is a random variable no deterministic statement can be made about the absolute deviation; however what we can expect of a good estimator is a high chance of remaining close to $g(\theta)$. Also as n, the sample size, increases we get hold of more information and hence expect to be able to do a better job of estimating $g(\theta)$. These notions when coupled together give rise to the consistency requirement for a sequence of estimators T_n; as n increases, T_n ought to converge in probability to $g(\theta)$ (under the probability distribution P_θ). In other words, for any $\epsilon > 0$,

$$P_\theta\left(\mid T_n - g(\theta) \mid > \epsilon\right) \to_P 0\,.$$

The above is clearly a large sample property; what it says is that with probability increasing to 1 (as the sample size grows), T_n estimates $g(\theta)$ to any pre-determined level of accuracy. However, the consistency condition alone, does not tell us anything about how well we are performing for any particular sample size, or the rate at which the above probability is going to 0.

For a fixed sample size n, how do we measure the performance of an estimator T_n? We have seen that $\mid T_n - g(\theta) \mid$ is itself random and therefore cannot even be computed as a function of θ before the experiment is carried out. A way out of this difficulty is to obtain an average measure of the error, or in other words, average out $\mid T_n - g(\theta) \mid$ over all possible realizations of T_n. The resulting quantity is then still a function of θ but no longer random It is called the mean

absloute error and can be written compactly (using acronym) as:

$$M.A.D. = E_\theta \mid T_n - g(\theta) \mid .$$

However, it is more common to avoid absolute deviations and work with the square of the deviation, integrated out as before over the distribution of T_n. This is called the mean-squared error (M.S.E.) and is

$$M.S.E.(T_n, g(\theta)) = E_\theta \ (T_n - g(\theta))^2 .$$

Of course, this is meaningful, only if the above quantity is finite for all θ. Good estimators are those for which the M.S.E. is generally not too high, whatever be the value of θ. There is a standard decomposition of the M.S.E. that helps us understand its components. We have,

$$\begin{aligned} M.S.E, (T_n, g(\theta)) &= E_\theta \ (T_n - g(\theta))^2 \\ &= E_\theta \ (T_n - E_\theta(T_n) + E_\theta(T_n) - g(\theta))^2 \\ &= E_\theta \ (T_n - E_\theta(T_n))^2 + (E_\theta(T_n) - g(\theta))^2 + 2\, E_\theta[(T_n - E_\theta(T_n))\,(E_\theta(T_n) - g(\theta))] \\ &= \mathrm{Var}_\theta \ (T_n) + b(T_n, \theta)^2 , \end{aligned}$$

where $b(T_n, g(\theta)) = E_\theta(T_n) - g(\theta)$ is the bias of T_n as an estimator of $g(\theta)$ (the cross product term in the above display vanishes since $E_\theta(T_n) - g(\theta)$ is a constant and $E_\theta(T_n - E_\theta(T_n)) = 0$). It measures, on an average, by how much T_n overestimates or underestimates $g(\theta)$. If we think of the expectation $E_\theta(T_n)$ as the center of the distribution of T_n, then the bias measures by how much the center deviates from the target. The variance of T_n, of course, measures how closely T_n is clustered around its center. Ideally one would like to minimize both simultaneously, but unfortunately this is rarely possible. Two estimators T_n and S_n can be compared on the basis of their mean squared errors. Under parameter value θ, T_n dominates S_n as an estimator if $M.S.E.(T_n, \theta) \leq M.S.E.(S_n, \theta)$. We say that S_n is inadmissible in the presence of T_n if

$$M.S.E.(T_n, \theta) \leq M.S.E.(S_n, \theta) \ \ \forall \ \ \theta .$$

The use of the term "inadmissible" hardly needs explanation. If, for all possible values of the parameter, we incur less error using T_n instead of S_n as an estimate of $g(\theta)$, then clearly there is no point in considering S_n as an estimator at all. Continuing along this line of thought, is there an estimate that improves all others? In other words, is there an estimator that makes every other estimator inadmissible ? The answer is no, execpt in certain pathological situations.

As we have noted before, it is generally not possible to find a universally best estimator. One way to try to construct optimal estimators is to restrict oneself to a subclass of estimators and try to find the best possible estimator in this subclass. One arrives at subclasses of estimators by constraining them to meet some desirable requirements. One such requirement is that of *unbiasedness*. Below, we provide a formal definition.

Unbiased estimator: An estimator T_n of $g(\theta)$ is said to be unbiased if $E_\theta(T_n) = g(\theta)$ for all possible values of θ; i.e. $b(T_n, g(\theta)) = 0$.

Thus, unbiased estimators, on an average, hit the target value. This seems to be a reasonable constraint to impose on an estimator and indeed produces meaningful estimates in a variety of situations. Note that for an unbiased estimator T_n, the M.S.E under θ is simply the variance of T_n under θ. In a large class of models, it is possible to find an unbiased estimator of $g(\theta)$ that has the smallest possible variance among all possible unbiased estimators. Such an estimate is called an MVUE (minimum variance unbiased estimator). Here is a formal definition.

MVUE: We call S_n an MVUE of $g(\theta)$ if (i) $E_\theta(S_n) = g(\theta)$ for all values of θ and (ii) if T_n is an unbiased estimate of $g(\theta)$, then $\mathrm{Var}_\theta(S_n) \leq \mathrm{Var}_\theta(T_n)$.

We will not discuss the most general method of constructing best unbiased estimators in this course. Later on we will provide a sufficient condition for checking whether an unbiased estimator is the best possible one in the sense that it has the minimnum variance. But before we proceed further, here are a few examples to illustrate some of the various concepts discussed above.

(a) Consider Example (4) above. A natural unbiased estimator of $g_1(\theta) = \mu$ is $\overline{X}_n$, the sample mean. It is also consistent for μ by the WLLN. It can be shown that this is also the MVUE of μ. In other words, *any* other unbiased estimate of μ will have a larger variance than $\overline{X}_n$. Recall that the variance of $\overline{X}_n$ is simply σ^2/n.

Consider now, the estimation of σ^2. Two estimates of this that we have considered in the past are

$$(i)\ \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \ \text{ and } \ (ii)\ s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\,.$$

Out of these $\hat{\sigma}^2$ is not unbiased for σ^2 but s^2 is, as you will show in a homework exercise. In fact s^2 is also the MVUE of σ^2.

(b) Let $X_1, X_2, \ldots, X_n$ be i.i.d. from some underlying density function or mass function $f(x, \theta)$. Let $g(\theta) = E_\theta(X_1)$. Then the sample mean $\overline{X}_n$ is *always* an unbiased estimate of $g(\theta)$. Whether it is MVUE or not depends on the underlying structure of the model.

(c) In Example 1 above, $\overline{X}_n$ is the MVUE of θ. Now define $g(\theta) = \theta/(1-\theta)$. This is a quantity of interest because it is precisely the odds in favour of Heads. It can be shown that there is *no unbiased estimator* of $g(\theta)$ in this model. However an intuitively appealing estimate of $g(\theta)$ is $T_n \equiv \overline{X}_n/(1-\overline{X}_n)$. It is *not unbiased* for $g(\theta)$; however it does converge in probability to $g(\theta)$. This example illustrates an important point – unbiased estimators may not always exist. Hence imposing unbiasedness as a constraint may not be meaningful in all situations.

(d) Unbiased estimators are not always better than biased estimators. Remember, it is the MSE that gauges the performance of the estimator and a biased estimator may actually outperform an unbiased one owing to a significantly smaller variance. Consider the situation

where $X_1, X_2, \ldots, X_n$ are i.i.d. Uniform(0,θ). Here $\Theta = (0, \infty)$. A natural estimate of θ is the maximum of the X_i's, which we denote by $X_{(n)}$. Another estimate of θ is obtained by observing that $\overline{X}$ is an unbiased estimate of $\theta/2$, the common mean of the X_i's; hence $2\,\overline{X}_n$ is an unbiased estimate of θ. You will show in the homework that $X_{(n)}$ in the sense of M.S.E outperforms $2\,\overline{X}$ by an order of magnitude. The best unbiased estimator (MVUE) of θ is $(1 + n^{-1})\, X_{(n)}$.

So far we have discussed some broad general principles and exemplified them somewhat. Our next goal is to actually describe some common methods of constructing estimators in parametric models. We will discuss two important approaches in the next section.

2 Construction of Estimates

. We discuss (a) Estimation by the Method of Moments and (b) Maximum Likelihood Estimation.

2.1 Method of Moments Estimation

As before, our set-up is the following: $X_1, X_2, \ldots, X_n$ are i.i.d. observations from some density or mass function $f(x, \theta)$. Suppose we know how the first k moments of X_1 look like, as a function of θ. Thus, let $\mu_i(\theta) = E_\theta(X_1^i)$ for $i = 1, 2, \ldots, k$ (we tacitly assume that these moments exist). Thus we can construct the function,

$$\theta \mapsto (\mu_1(\theta), \mu_2(\theta), \ldots, \mu_k(\theta))\,.$$

Now, suppose that this map is invertible, in the sense that we can express θ as a function of the first k moments. Let this inverse map be denoted by h (note that implies that the distribution of X_1 is completely determined by its first k moments). Now, the μ_i's are also known as population moments. Define, the i'th sample moment, $\hat{\mu}_i$ by,

$$\hat{\mu}_i = \frac{1}{n} \sum_{j=1}^{n} X_j^i\,.$$

Then the MOM principle, says, that we should estimate θ by $\hat{\theta}$ where

$$\hat{\theta} = h(\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_k)\,;$$

in other words $\hat{\theta}$ is the unique value of the parameter such that,

$$(\mu_1(\hat{\theta}), \mu_2(\hat{\theta}), \ldots, \mu_k(\hat{\theta}))) = (\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_k)\,.$$

Method of Moments can thus be thought of as "plug-in" estimates; to get an estimate $\hat{\theta}$ of $\theta = h(\mu_1, \mu_2, \ldots, \mu_k)$, we plug in estimates of the μ_i's which are the $\hat{\mu}_i$'s, to get $\hat{\theta}$. If h is continuous, then the fact that the $\hat{\mu}_i$ converges in probability to μ_i for every i, entails that $\hat{\theta} = h(\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_k)$ converges in probability to $h(\mu_1, \mu_2, \ldots, \mu_k) = \theta$; thus MOM estimates are

consistent under mild assumptions.

In general, we might be interested in estimating $g(\theta)$ where g is some (known) function of θ; in such a case, the MOM estimate of $g(\theta)$ is $g(\hat{\theta})$ where $\hat{\theta}$ is the MOM estimate of θ.

Example 1: Let $X_1, X_2, \ldots, X_n$ be from a $N(\mu, \sigma^2)$ distribution. Thus $\theta = (\mu, \sigma^2)$. We want to obtain MOM estimates of θ. Consider $\mu_1 = E(X_1)$ and $\mu_2 = E(X_1^2)$. Clearly, the paramter θ can be expressed as a function of the first two population moments, since

$$\mu = \mu_1 \, , \; \sigma^2 = \mu_2 - \mu_1^2 \, .$$

To get MOM estimates of μ and σ^2 we are going to plug in the sample moments. Thus

$$\hat{\mu} = \hat{\mu_1} = \overline{X} \, ,$$

and

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{j=1}^{n} X_j^2 \; - \overline{X}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2 \, .$$

Example 2: Let $X_1, X_2, \ldots, X_n$ be the indicators of n Bernoulli trials with success probability θ. We are going to find MOMs of θ. Note that θ is the probability of success and satisifes,

$$\theta = E(X_1) \, , \; \theta = E(X_1^2) \, .$$

Thus we can get MOMs of θ based on both the first and the second moments. Thus,

$$\hat{\theta}_{MOM} = \overline{X} \, ,$$

and

$$\hat{\theta}_{MOM} = \frac{1}{n} \sum_{j=1}^{n} X_j^2 = \sum_{j=1}^{n} X_j = \overline{X} \, .$$

Thus the MOM estimates obtained in these two different ways coincide. Note that

$$\mathrm{Var}_\theta(\overline{X}) = \frac{\theta \, (1 - \theta)}{n} \, ,$$

and a MOM estimate of $\mathrm{Var}_\theta(\overline{X})$ is obtained as

$$\widehat{\mathrm{Var}_\theta(\overline{X})}_{MOM} = \frac{\overline{X} \, (1 - \overline{X})}{n} \, .$$

Here, the MOM estimate based on the second moment μ_2 coincides with the MOM estimate based on μ_1 because of the 0-1 nature of the X_i's (which entails that $X_i^2 = X_i$). However, this is not necessarily the case; the MOM estimate of a certain parameter may not be unique as illustrated

by the following example.

Example 3: Let $X_1, X_2, \ldots, X_n$ be i.i.d. Poisson(λ). We know that, $E(X_1) = \mu_1 = \lambda$ and $\text{Var}(X_1) = \mu_2 - \mu_1^2 = \lambda$. Thus $\mu_2 = \lambda + \lambda^2$. Now, a MOM estimate of λ is clearly given by $\hat{\lambda} = \hat{\mu}_1 = \overline{X}$; thus a MOM estimate of $\mu_2 = \lambda^2 + \lambda$ is given by $\overline{X}^2 + \overline{X}$. On the other hand, the obvious MOM estimate of μ_2 is $\hat{\mu}_2 = (1/n) \sum_{j=1}^n X_j^2$. However these two estimates are not necessarily equal; in other words, it is not necessarily the case that $\overline{X}^2 + \overline{X} = (1/n) \sum_{j=1}^n X_j^2$. This illustrates one of the disadvantages of MOM estimates - they may not be uniquely defined.

Example 4: Consider n systems with failure times $X_1, X_2, \ldots, X_n$ assumed to be independent and identically distributed as exp(λ). It is not difficult to show that

$$E(X_1) = \frac{1}{\lambda}\,,\; E(X_1^2) = \frac{2}{\lambda^2}\,.$$

Therefore

$$\lambda = \frac{1}{\mu_1} = \sqrt{\frac{2}{\mu_2}}\,.$$

The above equations lead to two different MOM estimates for λ; the estimate based on the first moment is

$$\hat{\lambda}_{MOM} = \frac{1}{\hat{\mu_1}}\,,$$

and the estimate based on the second moment is

$$\hat{\lambda}_{MOM} = \sqrt{\frac{2}{\hat{\mu_2}}}\,.$$

Once again, note the non-uniqueness of the estimates. We finish up this section by some key observations about method of moments estimates.

- (i) The MOM principle generally leads to procedures that are easy to compute and which are therefore valuable as preliminary estimates.
- (ii) For large sample sizes, these estimates are likely to be close to the value being estimated (consistency).
- (iii) The prime disadvantage is that they do not provide a unique estimate and this has been illustrated before with examples.

2.2 Method of Maximum Likelihood.

As before we have i.i.d. observations $X_1, X_2, \ldots, X_n$ with common probability density or mass function $f(x, \theta)$ and θ is a Euclidean parameter indexing the class of distributions being considered. The goal is to estimate θ or some $\Psi(\theta)$ where Ψ is some known function of θ. Define the likelihood function for the sample $\underline{X} = (X_1, X_2, \ldots, X_n)$ as

$$L_n(\theta, \underline{X}) = \Pi_{i=1}^n f(X_i, \theta)\,.$$

This is simply the joint density but we now think of this as a function of θ for a fixed $\underline{X}$; namely the $\underline{X}$ that is realized. Suppose for the moment that X_i's are discrete, so that f is actually a frequency function. Then $L_n(\underline{X}, \theta)$ is exactly the probability that the observed data is realized or "happens". We now seek to obtain that $\theta \in \Theta$ for which $L_n(\underline{X}, \theta)$ is maximized. Call this $\hat{\theta}_n$ (assume that it exists). Thus $\hat{\theta}_n$ is that value of the parameter that maximizes the likelihood function, or in other words, makes the observed data most likely. It makes sense to pick $\hat{\theta}_n$ as a guess for θ. When the X_i's are continuous and $f(x, \theta)$ is in fact a density we do the same thing – maximize the likelihood function as before and prescribe the maximizer as an estimate of θ. For obvious reasons, $\hat{\theta}_n$ is called an MLE (maximum likelihood estimate). Note that $\hat{\theta}_n$ is itself a deterministic function of $(X_1, X_2, \ldots, X_n)$ and is therefore a random variable. Of course there is nothing that guarantees that $\hat{\theta}_n$ is unique, even if it exists. The uniqueness problem is handled by imposing structural requirements on f that guarantee uniqueness, or with probability increasing to 1, uniqueness in the long run. Sometimes, in the case of multiple maximizers, we choose one which is more desirable according to some "sensible"criterion. We will not worry about these issues at this stage; rather we will focus on doing some MLE computations.

Example 1: Let $X_1, X_2, \ldots, X_n$ be i.i.d. Ber(θ) where $0 \leq \theta \leq 1$. We want to find the MLE of θ. Now,

$$\begin{aligned} L_n(\theta, \underline{X}) &= \Pi_{i=1}^n \, \theta^{X_i} \, (1-\theta)^{1-X_i} \\ &= \theta^{\sum X_i} \, (1-\theta)^{n - \sum X_i} \\ &= \theta^{\, n \, \overline{X}_n} \, (1-\theta)^{\, n \, (1 - \overline{X}_n)} \, . \end{aligned}$$

Maximizing $L_n(\theta, \underline{X})$ is equivalent to maximizing $\log \, L_n(\theta, \underline{X})$. Now,

$$\log \, L_n(\theta, \underline{X}) = n \, \overline{X}_n \, \log \, \theta + n \, (1 - \overline{X}_n) \, \log(1-\theta) \, .$$

We split the maximization problem into the following 3 cases.

- (i) $\overline{X}_n = 1$; this means that we observed a success in every trial. It is not difficult to see that in this case the MLE $\hat{\theta}_n = 1$ which is compatible with intuition.
- (ii) $\overline{X}_n = 0$; this means that we observed a failure in every trial. It is not difficult to see that in this case the MLE $\hat{\theta}_n = 0$, also compatible with intuition.
- (iii) $0 < \overline{X}_n < 1$; in this case it is easy to see that the function $\log \, L_n(\theta, \underline{X})$ goes to $-\infty$ as θ approaches 0 or 1, so that for purposes of maximization we can restrict to $0 < \theta < 1$. To maximize $\log \, L_n(\theta, \underline{X})$, we solve the equation,

  $$\frac{\partial}{\partial \, \theta} \, \log \, L_n(\theta, \underline{X}) = 0 \, ,$$

  which yields,

  $$\frac{\overline{X}_n}{\theta} - \frac{1 - \overline{X}_n}{1 - \theta} = 0 \, .$$

This gives $\theta = \overline{X}$. It can be checked by computing the second derivative at $\overline{X}_n$ or noticing that $\log\, L_n(\theta, \underline{X})$ is concave in θ that the function attains a maximum at $\overline{X}_n$. Thus, the MLE, $\hat{\theta}_n = \overline{X}_n$ and this is just the sample proportion of 1's.

Thus, in every case, the MLE is the sample proportion of 1's. Note that this is also the MOM estimate of θ.

Example 2: Suppose that $X_1, X_2, \ldots, X_n$ are i.i.d. Poisson(θ). In this case, it is easy to see that

$$L_n(\theta, X) = \Pi_{i=1}^n \frac{e^{-\theta}\, \theta^{X_i}}{X_i\,!} = C(\underline{X})\, e^{-n\,\theta}\, \theta^{\sum\, X_i}\,.$$

To maximize this expression, we set

$$\frac{\partial}{\partial\, \theta}\, \log\, L_n(\theta, \underline{X}) = 0\,.$$

This yields that

$$\frac{\partial}{\partial\, \theta}\, \left[-n\, \theta + \left(\sum_{i=1}^n\, X_i \right)\, \log\, \theta \right] = 0;$$

i.e.

$$-n + \frac{\sum_{i=1}^n\, X_i}{\theta} = 0\,,$$

showing that

$$\hat{\theta}_n = \overline{X}\,.$$

It can be checked (by computing the second derivative at $\hat{\theta}_n$) that the stationary point indeed gives (a unique) maximum (or by noting that the log-likelihood is a (strictly) concave function). By the CLT, note that

$$\sqrt{n}\, \left(\overline{X}_n - \theta \right) \rightarrow_d N(0, \theta)\,.$$

The above result can be used to construct (approximate) confidence intervals for θ.

Example 3: Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. $U(0, \theta)$ random variables, where $\theta > 0$. We want to obtain the MLE of θ. It is clear that in this case, knowing the largest value makes the other values completely non-informative about θ. So the MLE inuititvely should be based on the largest value. (Intrinsic to this intuitive observation is the notion of sufficiency, which has to do with data-reduction; one condenses the data, to make life simpler (store less information) but at the same time not lose any information about the parameter.)

The likelihood function is given by,

$$\begin{aligned} L_n(\theta, X) &= \Pi_{i=1}^n \frac{1}{\theta}\, I(X_i \leq \theta) \\ &= \frac{1}{\theta^n}\, \Pi_{i=1}^n\, I(\theta \geq X_i) \\ &= \frac{1}{\theta^n}\, I(\theta \geq \max X_i)\,. \end{aligned}$$

It is then clear that $L_n(\theta, X)$ is constant and equals $1/\theta^n$ for $\theta \geq \max X_i$ and is 0 otherwise. By plotting the graph of this function, you can see that $\hat{\theta}_n = \max X_i$. Here, differentiation will not help you to get the MLE because the likelihood function is not differentiable at the point where it hits a maximum..

Example 4: Suppose that $X_1, X_2, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$. We want to find the MLE s of the mean μ and the variance σ^2. We write down the likelihood function first. This is,

$$L_n(\mu, \sigma^2, \underline{X}) = \frac{1}{(2\,\pi)^{n/2}}\,\sigma^{-n}\exp\left(-\frac{1}{2\,\sigma^2}\,\sum(X_i-\mu)^2\right).$$

It is easy to see that,

$$\begin{aligned}\log\,L_n(\mu,\sigma^2,\underline{X}) &= (-n/2)\,\log\,\sigma^2 - (1/2\,\sigma^2)\,\sum(X_i-\mu)^2 + \text{constant}\\ &= (-n/2)\,\log\,\sigma^2 - (1/2\,\sigma^2)\,\sum(X_i-\overline{X})^2 - (n/2\,\sigma^2)\,(\overline{X}-\mu)^2\,.\end{aligned}$$

To maximize the above expression wrt. μ and σ^2 we proceed as follows. For any (μ, σ^2) we have,

$$\log\,L_n(\mu,\sigma^2,\underline{X}) \leq \log\,L_n(\overline{X}_n,\sigma^2,\underline{X})\,,$$

showing that we can choose $\mu_{MLE} = \overline{X}$. It then remains to maximize $\log\,L_n(\overline{X}_n, \sigma^2, \underline{X})$ with respect to σ^2 to find σ^2_{MLE}. Now,

$$\log\,L_n(\overline{X}_n,\sigma^2,\underline{X}) = (-n/2)\,\log\,\sigma^2 - (1/2\,\sigma^2)\,\sum(X_i-\overline{X})^2\,.$$

Differentiating the left-side wrt. σ^2 gives,

$$(-n/2)\,(1/\sigma^2) + (1/2)\,(1/\sigma^2)^2\,n\,\hat{\sigma}^2 = 0\,,$$

where $\hat{\sigma}^2 = (1/n)\,\sum(X_i-\overline{X})^2$ is the sample variance. The above equation leads to,

$$\sigma^2_{MLE} = \hat{\sigma}^2 = (1/n)\,\sum(X_i-\overline{X})^2\,.$$

The fact that this actually gives a global maximizer follows from the fact that the second derivative at $\hat{\sigma}^2$ is negative. Note that, once again, the MOM estimates coincide with the MLEs.

Example 4 (continued): We now tweak the above situation a bit. Suppose now that we restrict the parameter space, so that μ has to be non-negative. Thus we seek to maximize $\log\,L_n(\mu, \sigma^2, \underline{X})$ but subject to the constraint that $\mu \geq 0$ and $\sigma^2 > 0$. Clearly, the MLE is $(\overline{X}_n, S^2)$ if $\overline{X}_n \geq 0$. In case, that $\overline{X}_n < 0$ we proceed thus. For fixed σ^2, the function $\log\,L_n(\mu, \sigma^2, \underline{X})$, as a function of μ, attains a maximum at $\overline{X}_n$ and then falls off as a parabola on either side. The $\mu \geq 0$ for which the function $\log\,L_n(\mu, \sigma^2, \underline{X})$ is the largest is 0; thus $\mu_{MLE} = 0$ and $\log\,L_n(\hat{\mu}, \sigma^2, \underline{X})$ is then given by,

$$\begin{aligned}\log\,L_n(\hat{\mu},\sigma^2,\underline{X}) &= (-n/2)\,\log\,\sigma^2 - (1/2\sigma^2)\,\sum(X_i-\overline{X})^2 - (1/2\sigma^2)\,n\,\overline{X}^2\\ &= (-n/2)\,\log\,\sigma^2 - \sum X_i^2/2\sigma^2\,.\end{aligned}$$

Proceeding as before (by differentiation) it is shown that

$$\sigma^2_{MLE} = \frac{1}{n} \sum X_i^2 \,.$$

Thus the MLEs can be written as,

$$(\mu_{MLE}, \sigma^2_{MLE}) = 1\,(\overline{X} < 0)\,(0, \frac{1}{n} \sum X_i^2) + 1\,(\overline{X} \geq 0)\,(\overline{X}, S^2)\,.$$

3 The Information Inequality and its Role in Asymptotic Theory

We saw in the above section that for a variety of different models one could differentiate the log–likelihood function with respect to the parameter θ and set this equal to 0 to obtain the MLE of θ. In these examples, the log–likelihood as a function of θ looks like an inverted bowl and hence solving for the stationary point gives us the unique maximizer of the log–likelihood. We start this section by introducing some notation. Let $l(x,\theta) = \log\, f(x,\theta)$. Also let $\dot{l}(x,\theta) = \partial/\partial\,\theta\; l(x,\theta)$. As before, $\underline{X}$ denotes the vector $(X_1, X_2, \ldots, X_n)$ and $\underline{x}$ denotes a particular value $(x_1, x_2, \ldots, x_n)$ assumed by the random vector $\underline{X}$. We denote by $p_n(\underline{x},\theta)$ the value of the density of $\underline{X}$ at the point $\underline{x}$. Thus $p_n(\underline{x},\theta) = \Pi_{i=1}^n\, f(x_i,\theta)$. Thus,

$$L_n(\theta, X) = \Pi_{i=1}^n\, f(X_i,\theta) = p_n(\underline{X},\theta)$$

and

$$l_n(\underline{X},\theta) = \log\, L_n(\theta,\underline{X}) = \sum_{i=1}^n\, l(X_i,\theta)\,.$$

Differentiating with respect to θ yields

$$\dot{l}_n(\underline{X},\theta) = \frac{\partial}{\partial\,\theta}\, \log\, p_n(\underline{X},\theta) = \sum_{i=1}^n\, \dot{l}(X_i,\theta)\,.$$

We call $\dot{l}(x,\theta)$ the score function and $\dot{l}_n(\underline{X},\theta) = 0$ the score equation. If differentiation is permissible for the purpose of obtaining the MLE, then $\hat{\theta}_n$ the MLE solves the equation

$$\dot{l}_n(\underline{X},\theta) \equiv \sum_{i=1}^n\, \dot{l}(X_i,\theta) = 0\,.$$

In this section, our first goal is to find a (nontrivial) lower bound on the variance of unbiased estimators of $\Psi(\theta)$ where Ψ is some differentiable function of θ. If we can indeed find such a bound (albeit under some regularity conditions) and there is an unbiased estimator of $\Psi(\theta)$ that attains this lower bound, we can conclude that it is the MVUE of $\Psi(\theta)$.

We now impose the following restrictions (regularity conditions) on the model.

(A.1) The set $A_\theta = \{x : f(x,\theta) > 0\}$ actually does not depend on θ and is subsequently denoted by A.

(A.2) If $W(\underline{X})$ is a statistic such that $E_\theta(\mid W(\underline{X}) \mid) < \infty$ for all θ, then,

$$\frac{\partial}{\partial\theta} E_\theta(W(\underline{X})) = \frac{\partial}{\partial\theta} \int_A W(\underline{x})\, p_n(\underline{x},\theta)\, dx = \int_A W(\underline{x})\, \frac{\partial}{\partial\theta}\, p_n(\underline{x},\theta)\, dx\,.$$

(A.3) The quantity $\partial/\partial\,\theta \log f(x,\theta)$ exists for all $x \in A$ and all $\theta \in \Theta$ as a well defined finite quantity.

The first condition says that the set of possible values of the data vector on which the distribution of $\underline{X}$ is supported does not vary with θ; this therefore rules out families of distribution like the uniform. The second assumption is a "smoothness assumption" on the family of densities and is generally happily satisfied for most parametric models we encounter in statistics and in particular, for exponential families of distributions, that form a very broad class of parametric models (and are extensively used in statistical modelling). There are various types of simple sufficient conditions that one can impose on $f(x,\theta)$ to make the intechange of integration and differentiation possible - we shall however not bother about this for the moment.

We define the information about the parameter θ in the model, namely $I(\theta)$, by

$$I(\theta) = E_\theta\, \left[\dot{l}(X,\theta)\right]^2\,,$$

provided it exists as a finite quantity for every θ.

We then have the following theorem.

Theorem 3.1 *All notation being as above,*

$$Var_\theta(T(\underline{X})) \geq \frac{(\psi^{'}(\theta))^2}{n\, I(\theta)}\,,$$

provided, the assumptions A.1, A.2 and A.3 hold and $I(\theta)$ exists as a finite quantity for all θ.

The above inequality is the celebrated Cramer-Rao inequality (or the information inequality) and is one of the most well known inequalities in statistics and has important ramifications in even more advanced forms of inference. Notice that if we take $\psi(\theta) = \theta$ then $(I_n(\theta))^{-1} \equiv (n\, I(\theta))^{-1}$ (the definition of $I_n(\theta)$ and its relationship to $I(\theta)$ is explained below) gives us a lower bound on unbiased estimators of θ in the model. If $I(\theta)$ is small, the lower bound is large, so unbiased estimators are doing a poor job in general – in other words, the data is not that informative about θ (within the context of unbiased estimation). On the other hand, if $I(\theta)$ is big, the lower bound is small, and so if we have a best unbiased estimator of θ that actually atttains this lower bound, we are doing a good job. That is why $I(\theta)$ is referred to as the information about θ.

Proof of Theorem 3.1: Let ρ_θ denote the correlation between $T(\underline{X})$ and $\dot{l}_n(\underline{X},\theta)$. Then $\rho_\theta^2 \leq 1$ which implies that

$$\mathrm{Cov}_\theta^2 \left(T(\underline{X}), \dot{l}_n(\underline{X},\theta)\right) \leq \mathrm{Var}_\theta(T(\underline{X})) \,.\, \mathrm{Var}_\theta(\dot{l}_n(\underline{X},\theta)) \,. \tag{3.1}$$

Now,

$$1 = \int p_n(\underline{x},\theta)\, d\underline{x}\,;$$

on differentiating both sides of the above identity with respect to θ and using $A.2$ with $W(\underline{x}) \equiv 1$ we obtain,

$$0 = \int \frac{\partial}{\partial\,\theta}\, p_n(\underline{x},\theta)\, d\underline{x} = \int \left(\frac{\partial}{\partial\,\theta}\, p_n(\underline{x},\theta)\right) \frac{1}{p_n(\underline{x},\theta)}\, p_n(\underline{x},\theta)\, d\underline{x} = \int \left(\frac{\partial}{\partial\,\theta}\, \log\, p_n(\underline{x},\theta)\right) p_n(\underline{x},\theta)\, d\underline{x}.$$

The last expression in the above display is precisely $E_\theta(\dot{l}_n(\underline{X},\theta))$ which therefore is equal to 0. Now note that,

$$E_\theta(\dot{l}_n(\underline{X},\theta)) = E_\theta\left(\sum_{i=1}^n \dot{l}(X_i,\theta)\right) = n\, E_\theta\left(\dot{l}(X_1,\theta)\right),$$

since the $\dot{l}(X_i,\theta)$'s are i.i.d. Thus, we have $E_\theta\left(\dot{l}(X_1,\theta)\right) = 0$. This implies that $I(\theta) = \mathrm{Var}_\theta(\dot{l}(X_1,\theta))$. Further,

$$\begin{aligned}
I_n(\theta) &\equiv E_\theta(\dot{l}_n^2(\underline{X},\theta)) \\
&= \mathrm{Var}_\theta(\dot{l}_n(\underline{X},\theta)) \\
&= \mathrm{Var}_\theta\left(\sum_{i=1}^n \dot{l}(X_i,\theta)\right) \\
&= \sum_{i=1}^n \mathrm{Var}_\theta(\dot{l}(X_i,\theta)) \\
&= n\, I(\theta)\,.
\end{aligned}$$

We will refer to $I_n(\theta)$ as the information based on n observations. Since $E_\theta(\dot{l}_n(\underline{X},\theta)) = 0$, it follows that

$$\begin{aligned}
\mathrm{Cov}_\theta\left(T(\underline{X}), \dot{l}_n(\underline{X},\theta)\right) &= \int T(\underline{x})\, \dot{l}_n(\underline{x},\theta)\, p_n(\underline{x},\theta)\, d\underline{x} \\
&= \int T(\underline{x}) \frac{\partial}{\partial\,\theta}\, p_n(\underline{x},\theta)\, d\underline{x} \\
&= \frac{\partial}{\partial\,\theta} \int T(\underline{x})\, p_n(\underline{x},\theta)\, d\underline{x} \;\; (\text{by } A.2) \\
&= \frac{\partial}{\partial\,\theta}\, \psi(\theta) \\
&= \psi'(\theta)\,.
\end{aligned}$$

Using the above in conjunction in (3.1) we get,

$$\psi^{'}(\theta)^2 \leq \mathrm{Var}_\theta(T(\underline{X}))\, I_n(\theta)$$

which is equivalent to what we set out to prove. □

There is an alternative expression for the information $I(\theta)$ in terms of the second derivative of the log-likelihood with respect to θ. If $\ddot{l}(x,\theta) \equiv (\partial^2/\partial\,\theta^2)\,\log\, f(x,\theta)$ exists for all $x \in A$ and for all $\theta \in \Theta$ then, we have the following identity:

$$I(\theta) = E_\theta\left(\dot{l}(X,\theta)^2\right) = -E_\theta\left(\ddot{l}(X,\theta)\right)\,,$$

provided we can differentiate twice under the integral sign; more concretely, if

$$\int \frac{\partial^2}{\partial\,\theta^2}\, f(x,\theta)\, dx = \frac{\partial^2}{\partial\,\theta^2} \int f(x,\theta)\, dx = 0 \quad (\star).$$

To prove the above identity, first note that,

$$\dot{l}(x,\theta) = \frac{1}{f(x,\theta)}\,\frac{\partial}{\partial\,\theta}\, f(x,\theta)\,.$$

Now,

$$\begin{aligned}
\ddot{l}(x,\theta) &= \frac{\partial}{\partial\,\theta}\left(\dot{l}(x,\theta)\right) \\
&= \frac{\partial}{\partial\,\theta}\left(\frac{1}{f(x,\theta)}\,\frac{\partial}{\partial\,\theta}\, f(x,\theta)\right) \\
&= \frac{\partial^2}{\partial\,\theta^2}\, f(x,\theta)\,\frac{1}{f(x,\theta)} - \frac{1}{f^2(x,\theta)}\left(\frac{\partial}{\partial\,\theta}\, f(x,\theta)\right)^2 \\
&= \frac{\partial^2}{\partial\,\theta^2}\, f(x,\theta)\,\frac{1}{f(x,\theta)} - \dot{l}(x,\theta)^2\,.
\end{aligned}$$

Thus,

$$\begin{aligned}
E_\theta(\ddot{l}(X,\theta)) &= \int \ddot{l}(x,\theta)\, f(x,\theta)\, dx \\
&= \int \frac{\partial^2}{\partial\,\theta^2}\, f(x,\theta)\, dx - E_\theta(\dot{l}^2(X,\theta)) \\
&= 0 - E_\theta(\dot{l}^2(X,\theta))\,,
\end{aligned}$$

where the first term on the right side vanishes by virtue of $(\star)$. This establishes the desired equality. It follows that,

$$I_n(\theta) = E_\theta(-\ddot{l}_n(\underline{X},\theta))\,,$$

where $\ddot{l}_n(\underline{X},\theta)$ is the second partial derivative of $l_n(\underline{X},\theta)$ with respect to θ. To see this, note that,

$$\ddot{l}_n(\underline{X},\theta) = \frac{\partial^2}{\partial\,\theta^2}\left(\sum_{i=1}^{n} l(X_i,\theta)\right) = \sum_{i=1}^{n} \ddot{l}(X_i,\theta)\,,$$

so that

$$E_\theta(\ddot{l}_n(\underline{X},\theta)) = \sum_{i=1}^{n} E_\theta(\ddot{l}(X_i,\theta)) = n\,E_\theta(\ddot{l}(X_1,\theta)) = -n\,I(\theta)\,.$$

We now look at some applications of the Cramer-Rao inequality.

Example 1: Let $X_1, X_2, \ldots, X_n$ be i.i.d. Poi(θ). Then $E(X_1) = \theta$ and $\mathrm{Var}(X_1) = \theta$. Let us first write down the joint likelihood of the data. We have,

$$p_n(\underline{x},\theta) = \Pi_{i=1}^{n}\,\frac{e^{-\theta}\,\theta^{x_i}}{x_i!} = e^{-n\,\theta}\,\theta^{\sum x_i}(\Pi\,x_i!)^{-1}\,.$$

Thus,

$$\begin{aligned}\dot{l}_n(\underline{x},\theta) &= \frac{\partial}{\partial\,\theta}\left(-n\,\theta + \left(\sum x_i\right)\,\log\,\theta - \log\,\Pi\,x_i\right)\\ &= -n + \sum x_i\,/\theta\,.\end{aligned}$$

Thus the information about θ based on n observations is given by,

$$I_n(\theta) = \mathrm{Var}_\theta\left(-n + \sum X_i\,/\theta\right) = \mathrm{Var}_\theta\left(\sum X_i\,/\theta\right) = n\,\theta\,/\theta^2 = n/\theta\,.$$

The assumptions needed for the Cramer Rao inequality to hold are all satisfied for this model, and it follows that for any unbiased estimator $T(\underline{X})$ of $\psi(\theta) = \theta$ we have,

$$\mathrm{Var}_\theta(T(\underline{X})) \geq 1/I_n(\theta) = \theta/n\,;$$

since $\overline{X}_n$ is unbiased for θ and has variance θ/n we conclude that $\overline{X}_n$ is the Best Unbiased Estimator of θ.

Example 2: Let $X_1, X_2, \ldots, X_n$ be i.i.d. $N(0, V = \sigma^2)$. Consider once again, the joint density of the n observations:

$$p_n(\underline{x},V) = \frac{1}{(2\,\pi\,V)^{n/2}}\,\exp\left(-\frac{1}{2V}\,\sum x_i^2\right)\,.$$

Now,

$$\begin{aligned}\dot{l}_n(\underline{x},V) &= \frac{\partial}{\partial\,V}\left(-(n/2)\,\log\,2\pi - (n/2)\,\log\,V - \frac{1}{2V}\,\sum x_i^2\right)\\ &= -\frac{n}{2V} + \frac{1}{2\,V^2}\,\sum x_i^2\,.\end{aligned}$$

Differentiating yet again we obtain,

$$\ddot{l}_n(\underline{x}, V) = \frac{n}{2\,V^2} - \frac{1}{V^3}\sum x_i^2\,.$$

Then, the information for V based on n observations is,

$$I_n(V) = -E_V\left(\frac{n}{2\,V^2} - \frac{1}{V^3}\sum X_i^2\right) = \frac{n}{2\,V^2} + \frac{1}{V^3}\,nV = \frac{n}{2\,V^2}\,.$$

Now consider the problem of estimating $\psi(V) = V$. For any unbiased estimator $S(X)$ of V, the Cramer-Rao inequality tells us that

$$\mathrm{Var}_V\left(S(X)\right) \geq I_n(V)^{-1} = 2\,V^2/n\,.$$

Consider, $\sum X_i^2/n$ as an estimator of V. This is clearly unbiased for V and the variance is given by,

$$\mathrm{Var}_V\left(\sum X_i^2/n\right) = (1/n)\,\mathrm{Var}\,X_1^2 = (V^2/n)\,\mathrm{Var}(X_1^2/V) = 2\,V^2/n\,,$$

since X_1^2/V follows χ_1^2 which has variance 2. It follows that $\sum X_i^2/n$ is the best unbiased estimator of V in this model.

3.1 Large Sample Properties of the MLE and Variance Stabilizing Transformations

In this section we study some of the large sample properties of the MLE in standard parametric models and how these can be used to construct confidence sets for θ or a function of θ. The Method of Variance Stabilizing Transformations will also be discussed as a means of obtaining more precise confidence sets.

We will see in this section that in the long run MLE's are the best possible estimators in a variety of different models. We will stick to models satisfying the restrictions (A1, A2 and A3) imposed in the last section. Hence our results will not apply to the Uniform distribution (or ones similar to the Uniform). Let us throw our minds back to the Cramer-Rao inequality. When does an unbiased estimator $T(\underline{X})$ of $\Psi(\theta)$ attain the bound given by this inequality? This requires:

$$\mathrm{Var}_\theta(T(\underline{X})) = \frac{(\Psi'(\theta))^2}{n\,I(\theta)}\,.$$

But this is equivalent to the assertion that the correlation between $T(\underline{X})$ and $\dot{l}_n(\underline{X}, \theta)$ is equal to 1 or -1. This means that $T(\underline{X})$ can be expressed as a linear function of $\dot{l}_n(\underline{X}, \theta)$. In fact, this is a necessary and sufficient condition for the information bound to be attained by the variance of $T(\underline{X})$. It turns out that this is generally difficult to achieve. Thus, there will be many different functions of θ, for which best unbiased estimators will exist but whose variance will not hit the information bound. The example below will illustrate this point.

Example: Let $X_1, X_2, \ldots, X_n$ be i.i.d.Bernoulli(θ). Then, the joint density is:

$$p_n(\underline{X}, \theta) = \Pi_{i=1}^n \, \theta^{X_i} \, (1-\theta)^{1-X_i}) \, .$$

We have,

$$f(x, \theta) = \theta^x \, (1-\theta)^{1-x} \, .$$

Thus,

$$l(x, \theta) = x \, \log(\theta) + (1-x) \, \log(1-\theta) \, ,$$

$$\dot{l}(x, \theta) = \frac{x}{\theta} - \frac{1-x}{1-\theta}$$

and

$$\ddot{l}(x, \theta) = -\frac{x}{\theta^2} - \frac{1-x}{(1-\theta)^2} \, .$$

Thus,

$$\dot{l}_n(\underline{X}, \theta) = \sum_{i=1}^n \dot{l}(X_i, \theta) = \frac{\sum_{i=1}^n X_i}{\theta} - \frac{n - \sum_{i=1}^n X_i}{1-\theta} \, .$$

Recall that the MLE solves $\dot{l}_n(\underline{X}, \theta) = 0$. Check that in this situation, this gives you precisely $\overline{X}_n$ as your MLE. Let us compute the information $I(\theta)$. We have,

$$I(\theta) = -E_\theta(\ddot{l}(X_1, \theta)) = E_\theta \left(\frac{X_1}{\theta^2} + \frac{1-X_1}{(1-\theta)^2} \right) = \frac{1}{\theta} + \frac{1}{1-\theta} = \frac{1}{\theta \, (1-\theta)} \, .$$

Thus,

$$I_n(\theta) = n \, I(\theta) = \frac{n}{\theta \, (1-\theta)} \, .$$

Consider unbiased estimation of $\Psi(\theta) = \theta$ based on $\underline{X}$. Let $T(X)$ be an unbiased estimator of θ. Then, by the information inequality,

$$\mathrm{Var}_\theta(T(X)) \geq \frac{\theta \, (1-\theta)}{n} \, .$$

Note that the variance of $\overline{X}$ is precisely $\theta \, (1-\theta)/n$, so that it is the MVUE of θ. Note that,

$$\dot{l}_n(X, \theta) = \frac{n \, \overline{X}}{\theta} - \frac{n \, (1-\overline{X})}{1-\theta} = \left(\frac{n}{\theta} + \frac{n}{1-\theta} \right) \overline{X} - \frac{n}{1-\theta} \, .$$

Thus, $\overline{X}_n$ is indeed linear in $\dot{l}_n(\underline{X}, \theta)$.

Consider now estimating a different function of θ, say $h(\theta) = \theta^2$. This is the probability of getting two consecutive heads. Suppose we try to find an unbiased estimator of this function. Then $S(X) = X_1 \, X_2$ is an unbiased estimator ($E_\theta(X_1 \, X_2) = E_\theta(X_1) \, E_\theta \, (X_2) = \theta^2$), but then so is $X_i \, X_j$ for any $i \neq j$. We can find the best unbiased estimator of θ^2 in this model by using

techniques beyond the scope of this course – it can be shown that any estimator $T(X)$ that can be written as a function of $\overline{X}$ and is unbiased for θ^2 is an MVUE (and indeed there is one such). Verify that,

$$T^*(X) = \frac{n\,\overline{X}^2 - \overline{X}}{n-1}$$

is unbiased for θ^2 and is therefore an (in fact *the*) MVUE. However, the variance of $T^*(X)$ does not attain the information bound for estimating $h(\theta)$ which is $4\,\theta^3\,(1-\theta)/n$ (verify). This can be checked by direct (somewhat tedious) computation or by noting that $T^*(X)$ is not a linear function of $\dot{l}_n(\underline{X}, \theta)$.

The question then is whether we can propose an estimator of θ^2 that does achieve the bound, at least approximately, in the long run. It turns out that this is actually possible. Since the MLE of θ is $\overline{X}$, the MLE of $h(\theta)$ is proposed as the plug-in value $h(\overline{X}) = \overline{X}^2$. This is *not an unbiased estimator of* $h(\theta)$ in finite samples, but has excellent behavior in the long run. In fact,

$$\sqrt{n}\,(h(\overline{X}_n) - h(\theta)) \rightarrow_d N(0, 4\,\theta^3\,(1-\theta))\,.$$

Thus for large values of n, $h(\overline{X})$ behaves approximately like a normal random variable with mean $h(\theta)$ and variance $4\,\theta^3\,(1-\theta)/n$. In this sense, $h(\overline{X}_n)$ is *asympotitcally (in the long run) unbiased* and *asympotically efficient* (in the sense that it has minimum variance).

This is not an isolated phenomenon but happens repeatedly. To this end, here is a general proposition.

Proposition 1: Suppose T_n is an estimator of $\Psi(\theta)$ (based on i.i.d. observations, $X_1, X_2, \ldots, X_n$ from P_θ) that satisfies:

$$\sqrt{n}\,(T_n - \Psi(\theta)) \rightarrow_d N(0, \sigma^2(\theta))\,.$$

Here $\sigma^2(\theta)$ is the limiting variance and depends on the underlying parameter θ. Then, for a continuously differentiable function h such that $h^{'}(\Psi(\theta)) \neq 0$, we have:

$$\sqrt{n}\,(h(T_n) - h(\Psi(\theta))) \rightarrow_d N(0, h^{'}(\Psi(\theta))^2\,\sigma^2(\theta))\,.$$

This is a powerful proposition. It's proof relies on using a Taylor expansion and is postponed for the moment (we will discuss heuristics in class if time permits). We will see crucial applications of the above shortly.

Here is another important proposition that establishes the limiting behavior of the MLE.

Proposition 2: If $\hat{\theta}_n$ is the MLE of θ obtained by solving $\sum_{i=1}^n \dot{l}(X_i, \theta) = 0$, then the following representation for the MLE is valid:

$$\sqrt{n}(\hat{\theta} - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I(\theta)^{-1}\,\dot{l}(X_i, \theta) + r_n\,,$$

where r_n converges to 0 in probability. It follows by a direct application of the CLT that,

$$\sqrt{n}(\hat{\theta} - \theta) \rightarrow_d N(0, I(\theta)^{-1}) .$$

We can now deduce the limiting behavior of the MLE of $g(\theta)$ given by $g(\hat{\theta})$ for any smooth function g such that $g^{'}(theta) \neq 0$. Combining Proposition 2 with Proposition 1 yields (take $\Psi(\theta) = \theta, T_n = \hat{\theta}$ and $h = g$)

$$\sqrt{n}(g(\hat{\theta}) - g(\theta)) \rightarrow_d N(0, g^{'}(\theta)^2 \, I(\theta)^{-1}) .$$

Thus, for large n,

$$g(\hat{\theta}) \sim_{appx} N(g(\theta), g^{'}(\theta)^2 \, (n \, I(\theta))^{-1}) .$$

Thus $g(\hat{\theta})$ is unbiased for $g(\theta)$ in the long run and its variance is approximately the information bound for unbiased estimators of $g(\theta)$.

Indeed, the MLE $\hat{\theta}$ is the best possible estimator available. Not only does its long term distribution center around θ, the quantity of interest, its distribution is also less spread out than that of any "reasonable" estimator of θ. If S_n is a "reasonable" estimator of θ, with

$$\sqrt{n} \, (S_n - \theta) \rightarrow_d N(0, \xi^2(\theta)) ,$$

then $\xi^2(\theta) \geq I(\theta)^{-1}$.

Exercise: Verify, in the Bernoulli example above in this section, that

$$\sqrt{n} \, (h(\overline{X}_n) - h(\theta)) \rightarrow_d N(0, 4 \, \theta^3 \, (1 - \theta)) .$$

Constructing confidence sets for θ: Since,

$$\sqrt{n}(\hat{\theta} - \theta) \rightarrow_d N(0, I(\theta)^{-1}) ,$$

it follows that

$$\sqrt{n \, I(\theta)}(\hat{\theta} - \theta) \rightarrow_d N(0, 1) .$$

Thus, the left side acts as an approximate pivot for θ. We have,

$$P_\theta \left[-z_{\alpha/2} \leq \sqrt{n \, I(\theta)}(\hat{\theta} - \theta) \leq z_{\alpha/2} \right] =_{appx} 1 - \alpha .$$

An approximate level $1 - \alpha$ confidence set for θ is obtained as

$$\{ \theta : -z_{\alpha/2} \leq \sqrt{n \, I(\theta)}(\hat{\theta} - \theta) \leq z_{\alpha/2} \} .$$

To find the above confidence set, one needs to solve for all values of θ satisfying the inequalities in the above display; this can however be a potentially complicated exercise depending on the functional form for $I(\theta)$. However, if the sample size n is large $I(\hat{\theta})$ can be expected to be close to $I(\theta)$ with high probability and hence the following is also valid:

$$P_\theta \left[-z_{\alpha/2} \leq \sqrt{n \, I(\hat{\theta})}(\hat{\theta} - \theta) \leq z_{\alpha/2} \right] =_{appx} 1 - \alpha . \; (\star\star)$$

This immediately gives an approximate level $1-\alpha$ C.I. for θ as:

$$\left[\hat{\theta} - \frac{1}{\sqrt{n\, I(\hat{\theta})}}\, z_{\alpha/2}, \hat{\theta} + \frac{1}{\sqrt{n\, I(\hat{\theta})}}\, z_{\alpha/2}\right].$$

Let's see what this implies for the Bernoulli example discussed above. Recall that $I(\theta) = (\theta\,(1-\theta))^{-1}$ and $\hat{\theta} = \overline{X}$. The approximate level $1-\alpha$ C.I. is then given by,

$$\left[\overline{X} - \sqrt{\frac{\overline{X}\,(1-\overline{X})}{n}}\, z_{\alpha/2}, \overline{X} + \sqrt{\frac{\overline{X}\,(1-\overline{X})}{n}}\, z_{\alpha/2}\right].$$

Exercise: Find explicitly

$$\left\{\theta : -z_{\alpha/2} \leq \sqrt{n\, I(\theta)}(\hat{\theta} - \theta) \leq z_{\alpha/2}\right\}$$

in the following cases (a) $X_1, X_2, \ldots, X_n$ are i.i.d. Bernoulli(θ). (b) $X_1, X_2, \ldots, X_n$ are i.i.d. Poisson(θ).

You will see that this involves solving for the roots of a quadratic equation. As in the Bernoulli Example, one can also get an approximate C.I. for θ in the Poisson setting on using $(\star\star)$. Verify that this yields the following level $1-\alpha$ C.I. for θ:

$$\left[\overline{X} - \sqrt{\frac{\overline{X}}{n}}\, z_{\alpha/2}, \overline{X} + \sqrt{\frac{\overline{X}}{n}}\, z_{\alpha/2}\right].$$

The recipe $(\star\star)$ is somewhat unsatisfactory because it involves one more level of approximation in that $I(\theta)$ is replaced by $I(\hat{\theta})$ (note that there is already one level of approximation in that the pivots being considered are only approximately $N(0,1)$ by the CLT). This introduces more variability in the approximate C.I's which one would ideally like to avoid. While solving the inequalities

$$-z_{\alpha/2} \leq \sqrt{n\, I(\theta)}(\hat{\theta} - \theta) \leq z_{\alpha/2}$$

is one way of avoiding this, a somewhat more elegant and easily implementable method is provided by the idea of *variance stabilizing transformations*.

Variance Stabilizing Transformations: Using Proposition 1, we conclude that for any differentiable function $v(\theta)$ with non-vanishing derivative $v'(\theta)$,

$$\sqrt{n}(v(\hat{\theta}) - v(\theta)) \to_d N(0, v'(\theta)^2/I(\theta)).$$

The idea is now to choose $v(\theta)$ in such a way that ensures that $v'(\theta)^2/I(\theta)$ is a constant and therefore does not depend on θ. Without loss of generality suppose that this constant value equals 1. Then we must have,

$$v'(\theta)^2 = I(\theta).$$

It suffices to take $v^{'}(\theta) = \sqrt{I(\theta)}$. Thus,

$$v(\theta) = \int \sqrt{I(\theta)}\, d\,\theta\,.$$

Since $v'(\theta) = I(\theta) > 0$ this implies that v is a strictly increasing function on its domain. Thus v^{-1} is well-defined and also strictly increasing. Furthemore,

$$\sqrt{n}(v(\hat{\theta}) - v(\theta)) \rightarrow_d N(0,1)\,.$$

It follows that an approximate level $1-\alpha$ C.I. for $v(\theta)$ is given by

$$\left[v(\hat{\theta}) - \frac{z_{\alpha/2}}{\sqrt{n}}\,,\, v(\hat{\theta}) + \frac{z_{\alpha/2}}{\sqrt{n}}\right]\,.$$

Using the fact that v^{-1} is strictly increasing we construct our level $1-\alpha$ C.I. for θ as

$$\left[v^{-1}\left(v(\hat{\theta}) - \frac{z_{\alpha/2}}{\sqrt{n}}\right)\,,\, v^{-1}\left(v(\hat{\theta}) + \frac{z_{\alpha/2}}{\sqrt{n}}\right)\right]\,.$$

Examples: (1) Suppose that the X_i's are i.i.d.Bernoulli(θ). Then,

$$v(\theta) = \int \frac{1}{\sqrt{\theta\,(1-\theta)}}\, d\theta\,.$$

Check that $v(\theta) = 2\sin^{-1}(\sqrt{\theta})$, compute the inverse transformation and find a 95% confidence interval for θ.

(2) Suppose that the X_i's are i.i.d. Poisson(θ). Then,

$$v(\theta) = \int \frac{1}{\sqrt{\theta}}\, d\theta\,.$$

Check that $v(\theta) = 2\sqrt{\theta}$, find v^{-1} and the resulting 95% C.I. for θ.

The notes "On Variance Stabilizing Transformations" (especially Page 3) will be helpful in this context.

(3) Use the method of variance stabilizing transformations to find an approximate level $1-\alpha$ C.I. for θ, based on i.i.d. observations $X_1, X_2, \ldots, X_n$ from an Exponential(θ) distribution.

4 Exercises

(1) Suppose that $X_1, X_2, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$. Show that

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

is unbiased for σ^2. Also, compute the bias of

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\,.$$

(2) Let $X_1, X_2, \ldots, X_n$ be i.i.d. observations from the Uniform$(0, \theta)$ distribution. Consider the following two estimators of θ.

$$(a)\ X_{(n)}\ \ (b) = (1 + 1/n)\, X_{(n)}\ \text{ and }\ (c) 2\,\overline{X}_n\,.$$

Compare the three estimators in terms of their MSE. The second estimator is the MVUE of θ.

(3) Let $X_1, X_2, \ldots, X_n$ be i.i.d exp(λ).

(a) Compute the MLE $\hat{\lambda}$ of λ and show that $\hat{\lambda}^{-1}$ is an unbiased estimator of λ^{-1}. Show also that $\hat{\lambda}$ coincides with a MOM estimate of λ.

(b) Show that for large n, $\hat{\lambda}^{-1}$ is approximately normally distributed with certain parameters. Use the central limit theorem and identify these parameters. Suggest an approximate 95% confidence interval for λ.

(4) . Let $X_1, X_2, \ldots, X_n$ be i.i.d. $U(a, b)$ where $0 < a < b$.

(i) Write down the likelihood function for the parameters (a, b).

(ii) Given the data, sketch the region in the x-y plane (the parameter a is plotted on the x-axis and the parameter b is plotted on the y-axis) where the likelihood function is non-zero.

(iii) Find the MLE's of a and b.

(5) . Let $X_1, X_2, \ldots, X_n$ be i.i.d. Geometric(θ). Thus $P(X_1 = k) = (1 - \theta)^{k-1}\,\theta\,.$

(i) Find a MOM estimate for θ based on the first population moment μ.

(ii) Find the MLE of θ and show that it coincides with the MOM estimate in (i).

(6) . (Censored geometric model). If time is measured in discrete periods, a model that is often used for the time to failure, X, of an item is,

$$P_\theta(X = k) = \theta^{k-1}\,(1 - \theta)\ \ k = 1, 2, \ldots, .$$

The above is just a geometric model, with θ being the probability that an item does not fail in one period. Suppose (for cost efficiency, say), we only record the time of failure, if

failure occurs on or before time r and otherwise record that the item has lived at least $r+1$ periods. Thus if X is the time to failure of an item, we just observe $\min(X, r+1)$. Let $Y = \min(X, r+1)$. Let $f(k,\theta) = P_\theta(Y = k)$ denote the probability mass function of the random vector Y. Note that k is between 1 and $r+1$.

(i) Compute $f(k,\theta)$.

(ii) Suppose we observe $Y_1, Y_2, \ldots, Y_n$ which is an i.i.d sample from the distribution of Y. Let M denote the number of indices i such that $Y_i = r+1$. Show that the likelihood function $l(Y_1, Y_2, \ldots, Y_n, \theta)$ can be written as

$$L_n(\theta, Y_1, Y_2, \ldots, Y_n) = \theta^{\sum_{i=1}^n Y_i - n} (1-\theta)^{n-M},$$

and hence find the MLE of θ.

(iii) **Difficult (not for Homework Purposes):** Show that the MLE $\hat{\theta}_n$ computed in (ii) converges in probability to θ.

(7) Let $X_1, X_2, \ldots, X_n$ be i.i.d. from the density,

$$f(x,\theta) = \frac{x}{\theta^2} e^{-\frac{x^2}{2\theta^2}} \quad , \quad x > 0 \, , \, \theta > 0 \, .$$

(a) Show that $E(X_1^2) = 2\theta^2$ and use this to compute a MOM estimate of θ. (Hint: What is the distribution of X_1^2?)

(b) Compute the MLE of θ and show that it coincides with the MOM in (a).

(c) By the CLT, $\sum_{i=1}^n X_i^2/n$ is approximately normal with parameters $(\mu(\theta), \sigma^2(\theta))$, where $\mu(\theta)$ and $\sigma^2(\theta)$ are functions of θ. Find $\mu(\theta)$ and $\sigma^2(\theta)$ and subsequently find a level $1-\alpha$ confidence interval for θ.

MORE PROBLEMS:

(1) Problems from Chapter 8 of Rice's book: Problem 19 (a) and (b), Problem 21, Problem 29, Problem 39 (a) and (b) and (c), Problem 42, Problem 43, Problem 44 (except Part (d)), Problem 49.

(2) Let $X_1, X_2, \ldots, X_n$ be i.i.d. $N(\mu_x, \sigma^2)$ and $Y_1, Y_2, \ldots, Y_m$ be i.i.d. $N(\mu_y, \sigma_2)$. Assume that the X's and Y's are independent. Find the MLE's of μ_x, μ_y, σ^2 and construct a level $1-\alpha$ C.I. for $\mu_x - \mu_y$ based on these estimates.

(3) Let X follow Binomial(n,θ) with $0 < \theta < 1$. Thus,

$$p(\underline{X}, \theta) = \binom{n}{X} \theta^X (1-\theta)^{n-X} \, .$$

Define

$$T(\underline{X}) = 1 \text{ if } X = n\,,$$

and

$$T(\underline{X}) = 0\,, \quad \text{otherwise}\,.$$

Thus $T(\underline{X})$ is a Bernoulli random variable. Let,

$$\Psi(\theta) = E_\theta(T(\underline{X})) = P_\theta(T(\underline{X}) = 1)\,.$$

Then clearly $T(\underline{X})$ is an unbiased estimator of $\Psi(\theta)$.

(a) Show that the information bound obtained from the Cramer-Rao inequality for unbiased estimators of $\Psi(\theta)$ is $n\,\theta^{2n-1}\,(1-\theta)$.
(b) Show that the variance of $T(X)$ is simply $\theta^n\,(1-\theta^n)$ and that this is strictly larger than the bound obtained from the Cramer-Rao inequality for $0 < \theta < 1$ – i,e.

$$\theta^n\,(1-\theta^n) > n\,\theta^{2n-1}\,(1-\theta)\,.$$

Note: It can be shown that $T(\underline{X})$ is the best unbiased estimator of $\Psi(\theta)$. This shows that the bound obtained in the Cramer-Rao inequality is not sharp in the sense that best unbiased estimators may not achieve this bound.

(4) Consider $X_1, X_2, \ldots,$ be an i.i.d. sequence from the density,

$$f(x,\theta) = (\theta+1)\,x^\theta,\ 0 < x < 1\,.$$

Let $\mu_m = E(X_1^m)$ and consider the first n observations, $X_1, X_2, \ldots, X_n$, from this sequence.
(i) Show that $\mu_m = (\theta+1)/(\theta+m+1)$ and use this result to show that the MOM estimate of θ is

$$\hat{\theta}_{MOM} = \frac{(m+1)\,\hat{\mu}_m - 1}{1-\hat{\mu}_m}\,,$$

where

$$\hat{\mu}_m = \frac{1}{n}\sum_{i=1}^{n} X_i^m\,.$$

(ii) Show that for fixed m, $\hat{\theta}_{MOM}$ converges in probability to θ.
(iii) Show that for a fixed data set $X_1, X_2, \ldots, X_n$,

$$\lim_{m\to\infty} \hat{\theta}_{MOM} = -1\,.$$

This shows that using a very high moment to estimate θ will give you a pretty bad estimate. (**Hint:** Note that each X_i is < 1 and therefore so is their maximum $X_{(n)}$ and this is larger than $(X_1^m + X_2^m + \ldots + X_n^m)/n$. As $m \to \infty$ what happens to $X_{(n)}^m$? What happens to $m\,X_{(n)}^m$

?)

(iv) We know that the MLE of θ in this model is

$$\hat{\theta}_{MLE} = -\frac{n}{\sum_{i=1}^{n} \log X_i} - 1\,.$$

Show that $\hat{\theta}_{MLE}$ converges in probability to θ.

Hint: Let $Y_i = -\log X_i$. Show that Y_i is an exponential random variable and use the law of large numbers to deduce that $\overline{Y}_n = -\sum_{i=1}^{n} \log X_i/n$ converges in probability to $1/(\theta+1)$. Proceed from there.

(5) Let $X_1, X_2, \ldots, X_n$ be i.i.d. Geometric(θ). Thus, each X_i has probability mass function,

$$f(x,\theta) = (1-\theta)^{x-1}\,\theta\,,\ \ \theta > 0\,,\ \ x = 1, 2, 3, \ldots$$

(a) Compute the information about θ, i.e. $I(\theta)$ based on one observation.

(b) Identify the limit distribution of $\sqrt{n}(\hat{\theta}_n - \theta)$ and use this result along with Proposition 1 to to show that

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

Is there a simpler/more direct way of arriving at this result?

(6) Let X follow a Poi(θ)distribution. Define a statistic $T(X)$ as follows:

$$T(X) = 1 \ \text{ if } \ X = 0\,,$$

and

$$T(X) = 0\,, \quad \text{otherwise}\,.$$

(i) Compute the information bound for the variance of $T(X)$ obtained from the information inequality.

(ii) Compute the variance of $T(X)$ and show that the bound obtained from the information inequality is strictly less than the variance. (This shows that the bound is not sharp. Nevertheless, it can be shown that $T(X)$ is the best unbiased estimator of its expectation. The above example illustrates that best unbiased estimators may not always achieve the lower bound given by the information inequality) (**Hint:** For $x > 0$, $e^x - 1 > x$.)

(7) Let $X_1, X_2, \ldots, X_n$ be a sample from a population with density

$$f(x,\theta) = \theta\,(\theta+1)\,x^{\theta-1}\,(1-x)\,,\ 0 < x < 1,\ \ ,\theta > 0\,.$$

(i) Show that

$$\hat{\theta} = \frac{2\,\overline{X}_n}{1-\overline{X}_n}\,,$$

is a MOM estimator of θ.

(ii) Is $\hat{\theta}$ consistent for θ?

(iii) Identify the limit distribution of $\overline{X}_n$; i.e. find constants $a(\theta)$ and $b^2(\theta)$ such that

$$\sqrt{n}(\overline{X}_n - a(\theta)) \to_d N(0, b^2(\theta))\,.$$

(iv) Show that

$$\sqrt{n}\,(\hat{\theta} - \theta) \to_d N(0, \xi^2(\theta))$$

where $\xi^2(\theta) = \theta\,(\theta+2)^2/(2\,(\theta+3))$.

(v) Show that the MLE of θ is obtained by solving the equation

$$\theta\,(\theta+1)\,V_n + (2\,\theta+1) = 0\,,$$

where $V_n = n^{-1}\sum_{i=1}^n \log\,X_i$.

(vi) Let $Y_i = -\log\,X_i$, for $i = 1, 2, \ldots, n$. Show that $Y_1, Y_2, \ldots, Y_n$ are i.i.d. with density function,

$$g(y,\theta) = \theta\,(\theta+1)\,e^{-\theta\,y}\,(1-e^{-y})\,,\;\; 0 < y < \infty\,.$$

(vii) Let $\mu_1 = E(Y_1)$. Compute μ_1 in terms of θ and then show that,

$$\theta = \frac{\sqrt{\mu_1^2+4} - (\mu_1 - 2)}{2\,\mu_1}\,.$$

(Remember that $\mu_1 > 0$.) Find a MOM estimate of θ using the above equation.

(viii) Call the above MOM estimate $\theta^\star$. Show that $\theta^\star$ is precisely the MLE that one gets by solving the equation in (e). Also show that $\theta^\star$ is the MLE of θ based on the transformed data $Y_1, Y_2, \ldots, Y_n$.

(ix) Find a minimal sufficient statistic for θ.

(x) Calculate the limit distribution of $\theta^\star$.

(8) (a) Suppose that I toss a coin 100 times and that p, the chance of getting H in a single toss of the coin is 10^{-6}. If S denotes the total number of heads obtained, can I use the CLT to study the distribution of S? Explain briefly.

(b) Let $X_1, X_2, \ldots, X_n$ be an i..i.d sample from the density

$$f(x, \lambda, \theta) = \lambda \exp(-\lambda(x - \theta))\, 1(x \geq \theta)\,, \quad \theta > 0, \lambda > 0\,.$$

Show that the likelihood function for the data is,

$$L(\theta, \lambda) = \lambda^n \exp\left(-\lambda \left(\sum_{i=1}^{n} X_i - n\,\theta\right)\right) 1(\min X_i \geq \theta)\,.$$

Compute the MLE's of θ and λ.