Statistics - Point Estimation

Andrew Andreas

June 4, 2024

1 Introduction

This is a set of brief notes on point estimation in statistics. These may evolve over time as I look to expand their depth.

Please report any errors to andrew.andreas0@gmail.com

2 Motivation

Assume that a random sample of n observations $x_1, x_2, ..., x_n$ is available. The model for this situation is that the observed sample has arisen as a realisation of a set of n random variables $X_1, X_2, ..., X_n$ which are mutually independent and which each follow the same probability distribution. The form of the probability distribution is assumed known (normal, exponential, Poisson, whatever), but the distribution depends on the value(s) of any unknown parameter(s). The aim, therefore, is to use the values of $x_1, x_2, ..., x_n$ to estimate the parameter(s).

Therefore the topic of these notes relates to the concept of point estimation, which is the estimation of the value(s) of a parameter(s) of a statistical model by a single value for each parameter.

A point estimate is the realisation of the point estimator, which is a function of the random variables $X_1, X_2, ..., X_n$.

2.1 Maximum Likelihood Estimation

2.1.1 Likelihood Function

Before considering the maximum likelihood estimator, it is worth briefly revisiting the concept of a likelihood function.

Definition: Likelihood Given a sample of observations $x_1, x_2, ..., x_n$ that are the realisations of $X_1, X_2, ..., X_n$ and a probability model for this population $f(x|\theta)$, which is either a pdf or pmf, then the probabilities of $X_1, X_2, ..., X_n$ are each specified by the parameters, θ, of $f(x|\theta)$.

The likelihood seeks to reverse this relationship - given the observations $x_1, x_2, ..., x_n$, what is the most likely parameter θ of $f(x|\theta)$ that gives rise to these fixed observations.

Stated differently, the unknown quantity is now the parameter, θ and we have a sample of fixed observations. In the usual setting concerning a pdf or pmf, the model is specified by a known parameter(s) and this is a function of the random variables $X_1, X_2, ..., X_n$.

Definition: Likelihood Function, $\mathbf{L}(\theta)$ A likelihood function, $L(\theta) = f(x|\theta)$ is a function of the parameter, θ and measures the relative likelihood of a parameter, given the fixed observations.

In practice we are usually dealing with n observations so that the likelihood function is denoted $L(\theta) = f(x_1, x_2, ..., x_n|\theta) = f(\mathbf{x}|\theta)$ but quite often the data either arises independently or we assume the data are independent. As such we assume each observation arises from a univariate distribution, $f(x|\theta)$ meaning that we can rewrite the likelihood as a product of n univariate distributions

$$
\begin{aligned}
L(\theta) =& f(x_1, x_2, ..., x_n|\theta) && (1)\\
=& f(x_1|\theta) \times f(x_2|\theta) \times ...f(x_n|\theta) && (2)\\
=& \prod_{i=1}^{n} f(x_i|\theta) && (3)
\end{aligned}
$$

Definition: Maximum Likelihood Estimate The maximum likelihood estimate (MLE) is the value of θ that maximizes the likelihood function.

To determine the value of θ which maximizes $L(\theta)$ we can compute the gradient to find $\hat{\theta}$, which is then a candidate MLE and follow this by confirming whether this is indeed a maximum by computing the Hessian.

Example - Maximising a Binomial Likelihood This example concerns a coin tossed 100 times resulting in 54 heads; what is the maximum likelihood estimate of p?

The pmf of the binomial distribution is given below

$$
f(x|\theta) = \binom{100}{x} p^x (1-p)^{(100-x)} \tag{4}
$$

now if we were to substitute in the number of successes ($x = 54$), i.e., our fixed observations, this equation becomes

$$
L(\theta) = f(x|\theta) = \binom{100}{54} p^{54} (1-p)^{46} \tag{5}
$$

which is clear that our function now is a function of the unknown parameter, p, so we can go ahead and take derivatives, $\frac{dL(p)}{dp}$, and equate this to 0 and solve for $\hat{p}$.

Note I have skipped the steps of the derivative here but the result in equation (7) can be checked by use of the product rule

$$
\begin{aligned}
\frac{dL(p)}{dp} &= Cp^{54}(1-p)^{46} && (6)\\
&= Cp^{53}(1-p)^{45}(54 - 100p) = 0 && (7)\\
&= (54 - 100p) = 0 && (8)\\
\hat{p} &= 54/100 && (9)
\end{aligned}
$$

Taking second derivatives and evaluating at $\hat{p} = 0.54$ will reveal that $\frac{d^2L(p)}{dp} < 0$ and hence $\hat{p} = 0.54$ is a maximum

I have included a link here which illustrates this example numerically in python.

Note that we typically transform the likelihood function to the log-likelihood by taking its logarithm because i) it simplifies the calculations and ii) because the logarithmic function is monotonic increasing.

The first point can be seen when we consider the properties of logarithms, namely $log(ab) = log(a) + log(b)$ which is relevant since we have the product of a large number of univariate densities. This can cause problems when the densities are small values (< 1), since

$$
\lim_{n\to\infty} \prod_{i=1}^{n} f(x_i|\theta) = 0 \tag{10}
$$

The log-likelihood circumvents this issue as the product becomes a sum, which also turns out to be easier to handle when dealing with derivatives

$$l(\theta) = log\Big(\prod_{i=1}^{n} f(x|\theta)\Big) \tag{11}$$

$$= \sum_{i=1}^{n} log\Big(f(x_i|\theta)\Big) \tag{12}$$

The monotonic increasing nature of the logarithm function means that for $x_2 > x_1, x_i \in \mathcal{R}$ then $log(x_2) > log(x_1)$ and it ensures that the value of θ that maximises $L(\theta)$ also maximises $l(\theta)$.

A useful result of MLE is that the MLE of a function of a parameter is equivalent to the function of the MLE of the parameter

$$\hat{\tau} = h(\hat{\theta}) \tag{13}$$

with a simple example being that if we have an MLE for the standard deviation, σ, then the MLE for the variance is just the square of this, i.e., $h() = ()^2$

3 Properties of Estimators

3.1 Sampling Distributions

It is important to consider the fact that when we have a sample of data from a population $x_1, x_2, ..., x_n$ that this specific sample is just one possible realisation of the random variables $X_1, X_2, ..., X_n$. If we were to sample from the population again then the realisations would be $x_1^*, x_2^*, ..., x_n^*$.

As we know, our estimators are a function of the data and are therefore random variables themselves. To see this, consider that x_i and x_i^* are the realisations of the random variable X_i and thus two distinct possible estimators are

$$\hat{\theta} = t(x_1, x_2, ..., x_n) \tag{14}$$

$$\overline{\theta} = t(x_1^*, x_2^*, ..., x_n^*) \tag{15}$$

Since random variables are just functions $X : \Omega \to \Gamma$ which denotes a mapping from the sample space, Ω, to the target space Γ, which represents the set of possible outcomes, x.

With this in mind we can see that there is a range of possible values x_i can take and this is defined by the random variable X_i such that the parameter is a function of the random variables $\tilde{\theta} = t(X_1, X_2, ..., X_n)$. The distribution over the random variables induces a distribution over the estimators which is known as the **sampling distribution** of $\tilde{\theta}$.

Example.
Consider random variables that arise from the normal distribution, $X_i \sim N(\theta, \sigma^2)$ and we want to estimate the population mean, θ.

Consider as an estimate the sample mean $\tilde{\theta} = \overline{X} = \sum_{i=1}^{n} \frac{X_i}{n}$ then its sampling distribution is characterised by two parameters, the mean and variance which we can find considering the mean and variance of the estimator and using knowledge of the fact that the distribution of the sum of normal random variables is also normal

$$
\begin{aligned}
\mathbb{E}\big[\overline{X}\big] &= \mathbb{E}\big[\sum_{i=1}^{n} \frac{X_i}{n}\big] && (16)\\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big[X_i\big] && (17)\\
&= \frac{1}{n}\sum_{i=1}^{n}\theta && (18)\\
&= \theta && (19)
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{V}\big[\overline{X}\big] &= \mathbb{V}\big[\sum_{i=1}^{n} \frac{X_i}{n}\big] && (20)\\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\big[X_i\big] && (21)\\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 && (22)\\
&= \frac{\sigma^2}{n} && (23)
\end{aligned}
$$

Thus we have determined the sampling distribution of the sample mean, $\overline{X} \sim N\big(\theta, \frac{\sigma^2}{n}\big)$

3.2 Bias of an Estimator

Definition: Bias The bias of an estimator is defined as $B(\tilde{\theta}) = \theta - \tilde{\theta}$.

The sample mean is an unbiased estimator of the population mean, regardless of the distribution, provided that the population mean is finite. Equations 16-19 showed this as $\mathbb{E}\big[X_i\big] = \theta$ regardless of the distribution.

3.2.1 Unbiased estimator of the population variance

The sample variance is defined as

$$
S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\big(X_i - \overline{X}\big)^2 \tag{24}
$$

Using the trick of adding 0 into the squared term in the form of $-\mu + \mu$

$$
\begin{aligned}
\sum_{i=1}^{n}\big(X_i - \overline{X}\big)^2 &= \sum_{i=1}^{n}\big(X_i - \mu + \mu - \overline{X}\big)^2 && (25)\\
&= \sum_{i=1}^{n}\big((X_i - \mu) + (\mu - \overline{X})\big)^2 && (26)\\
&= \sum_{i=1}^{n}\big(X_i - \mu)^2 - n\big(\overline{X} - \mu\big)^2 && (27)
\end{aligned}
$$

Taking expectations

$$\begin{aligned}
\mathbb{E}\big[S^2\big] &= \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n}\big(X_i-\mu\big)^2 - n\big(\overline{X}-\mu\big)^2\right] && (28)\\
&= \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}\big[\big(X_i-\mu\big)^2\big] - n\mathbb{E}\big[\big(\overline{X}-\mu\big)^2\big] && (29)\\
&= \frac{1}{n-1}\Big[\mathbb{V}(X) - n\mathbb{V}(\overline{X})\Big] && (30)\\
&= \frac{1}{n-1}\Big[n\sigma^2 - n\frac{\sigma^2}{n}\Big] && (31)\\
&= \frac{1}{n-1}\Big[n\sigma^2 - \sigma^2\Big] && (32)\\
&= \frac{1}{n-1}\Big[(n-1)\sigma^2\Big] && (33)\\
&= \sigma^2 && (34)
\end{aligned}$$

So the sample mean and sample variance are both unbiased estimators of the population mean and variance. However, by just considering the bias of an estimator we neglect a crucial element of its sampling distribution - the variation of the parameter's sampling distribution. A larger variance results in higher uncertainty around the true value of the parameter.

It turns out then that the best estimator's are those with minimum variance and are termed **minimum variance unbiased estimators** or **MVUEs**

3.3 Cramer-Rao Lower Bound

The Cramer-Rao Lower Bound (**CRLB**) is an inequality that defines the lower bound for the variance of an unbiased estimator

$$\mathbb{V}(\tilde{\theta}) \geq \frac{1}{\mathbb{E}_X\left[\{l'(\theta)\}^2\right]} \tag{35}$$

Another more practical form of the **CRLB** is presented below. First we must consider the likelihood, but note that here we are considering the log-likelihood *prior* to making any observations, so our notation is in terms of the random variables, X_i rather than x_i. What this means is we are considering the log-likelihood function over all possible values of x.

It is also helpful to define

$$\begin{aligned}
\phi &= l'(\theta) && (36)\\
&= l'(\theta|X_1, X_2, ..., X_n) && (37)\\
&= \frac{d}{d\theta}\left[\sum_{i=1}^{n} log\big[f(X_i|\theta)\big]\right] && (38)\\
&= \sum_{i=1}^{n}\frac{d}{d\theta}log\big[f(X_i|\theta)\big] && (39)\\
&= \sum_{i=1}^{n} U_i && (40)
\end{aligned}$$

Using the definition of the variance and rewriting this as

$$\begin{aligned}
\mathbb{E}_X\Big[\{l'(\theta|X_1, X_2, ..., X_n)\}^2\Big] &= \mathbb{E}_X\left[\phi^2\right] && (41)\\
&= \mathbb{V}_X(\phi) + \{\mathbb{E}_X(\phi)\}^2 && (42)
\end{aligned}$$

If we consider the second term $\{\mathbb{E}_X(\phi)\}^2$, focusing on the expectation of the derivative of the log-likelihood for independent random variables X_i

$$\begin{aligned}
\mathbb{E}_X(\phi) &= \sum_{i=1}^{n} \mathbb{E}_X(U_i) && (43)\\
&= \sum_{i=1}^{n} \mathbb{E}_X\left[\frac{d}{d\theta} log\big[f(X_i|\theta)\big]\right] && (44)\\
&= \sum_{i=1}^{n} \int \frac{d}{d\theta} log\big[f(x|\theta)\big] f(x|\theta)dx && (45)\\
&= \sum_{i=1}^{n} \int \frac{1}{f(x|\theta)} f(x|\theta) \frac{d}{d\theta} f(x|\theta)dx && (46)\\
&= \sum_{i=1}^{n} \int \frac{d}{d\theta} f(x|\theta)dx = \sum_{i=1}^{n} \frac{d}{d\theta} \int f(x|\theta)dx && (47)\\
&= \sum_{i=1}^{n} \frac{d}{d\theta}(1) = 0 && (48)
\end{aligned}$$

So we know the second term $\{\mathbb{E}_X(\phi)\}^2 = 0$ and therefore we just consider this as the variance of the estimator. If we denote $\mathbb{V}_X(\phi) = \sigma_U^2$ then we can rewrite this as

$$\begin{aligned}
\mathbb{V}_X(\phi) &= \sum_{i=1}^{n} \mathbb{V}_X\big[U_i\big] && (49)\\
&= \sum_{i=1}^{n} \sigma_U^2 && (50)\\
&= n\sigma_U^2 && (51)
\end{aligned}$$

Since $\mathbb{V}\big[U_i\big] = \mathbb{E}\big[U_i^2\big] + \mathbb{E}\big[U_i\big]^2$ and since we know $\mathbb{E}\big[U_i\big] = 0$, then the last term vanishes and we can rewrite the **CRLB** as

$$\begin{aligned}
\mathbb{V}(\tilde{\theta}) &\geq \frac{1}{n\mathbb{E}\big[U_i^2\big]} && (52)\\
&= \frac{1}{n\mathbb{E}\big[U_i^2\big]} && (53)\\
&= \frac{1}{n\mathbb{E}\left[\left(\frac{d}{d\theta} log\big[f(X_i|\theta)\big]\right)^2\right]} && (54)\\
& && (55)
\end{aligned}$$

meaning that that the **CRLB** is proportional to $\frac{1}{n}$ which tells us that taking larger samples will minimise the lower bound of an unbiased estimator.

3.4 Efficiency of an unbiased estimator

There is not always an unbiased estimator that achieves the **CRLB** but we can determine the estimator's efficiency, which is defined as the ratio of the CRLB to the variance of the estimator

$$\begin{aligned}
\text{efficiency}(\tilde{\theta}) &= \frac{\textbf{CRLB}}{\mathbb{V}[\tilde{\theta}]} && (56)\\
&= \frac{1}{n\mathbb{E}\left[\left(\frac{d}{d\theta} log\big[f(X_i|\theta)\big]\right)^2\right]\mathbb{V}[\tilde{\theta}]} && (57)
\end{aligned}$$

A fully efficient estimator achieves the **CRLB** and hence has efficiency of 1

3.5 Mean squared error

Another way to think about the efficacy of an estimator is through the it's **MSE**, $M(\tilde{\theta})$

$$M(\tilde{\theta}) = \mathbb{E}\big\{(\tilde{\theta} - \theta)^2\big\} \tag{58}$$

The **MSE** can be written in terms of the bias and variance

$$\begin{align} M(\tilde{\theta}) &= \mathbb{E}\big\{(\tilde{\theta} - \theta)^2\big\} \tag{59}\\ &= \mathbb{E}\big\{(\tilde{\theta} - \mathbb{E}[\tilde{\theta}] + \mathbb{E}[\tilde{\theta}] - \theta)^2\big\} \tag{60}\\ &= \mathbb{E}\big\{(\tilde{\theta} - \mathbb{E}[\tilde{\theta}]) + (\mathbb{E}[\tilde{\theta}] - \theta)^2\big\} \tag{61}\\ &= \mathbb{E}\big\{(\tilde{\theta} - \mathbb{E}[\tilde{\theta}])^2 + (\tilde{\theta} - \mathbb{E}[\tilde{\theta}])(\mathbb{E}[\tilde{\theta}] - \theta) + (\mathbb{E}[\tilde{\theta}] - \theta)^2\big\} \tag{62}\\ & \tag{63} \end{align}$$

Due to the linearity of expectation and the fact that only $\tilde{\theta}$ is a random variable, the middle term vanishes

$$\begin{align} \mathbb{E}\Big[(\tilde{\theta} - \mathbb{E}[\tilde{\theta}])(\mathbb{E}[\tilde{\theta}] - \theta)\Big] &= \ \mathbb{E}\big[(\tilde{\theta} - \mathbb{E}[\tilde{\theta}])\big](\mathbb{E}[\tilde{\theta}] - \theta) \tag{64}\\ &= \Big[\mathbb{E}\big[\tilde{\theta}\big] - \mathbb{E}\big[\mathbb{E}[\tilde{\theta}]\big]\Big](\mathbb{E}[\tilde{\theta}] - \theta) \tag{65}\\ &= \Big[\mathbb{E}\big[\tilde{\theta}\big] - \big[\mathbb{E}[\tilde{\theta}]\big]\Big](\mathbb{E}[\tilde{\theta}] - \theta) \tag{66}\\ &= 0 \times (\mathbb{E}[\tilde{\theta}] - \theta) \tag{67}\\ &= 0 \tag{68} \end{align}$$

So we are left with $M(\tilde{\theta}) = \mathbb{E}\Big\{(\tilde{\theta} - \mathbb{E}[\tilde{\theta}])^2\Big\} + \mathbb{E}\Big\{(\mathbb{E}[\tilde{\theta}] - \theta)^2\Big\}$ and recall that only $\tilde{\theta}$ is a random variable so that we can rewrite this as

$$M(\tilde{\theta}) = \mathbb{V}\big[\theta\big] + \big\{\mathrm{B}(\theta)\big\}^2 \tag{69}$$

3.6 Large sample properties

3.6.1 Asymptotic Unbiasedness

Despite the fact we may not always require an estimator to be unbiased, we will almost always require the estimator to be asymptotically unbiased

Definition. Asymptotic Unbiasedness An estimator is said to be asymptotically unbiased if $\lim_{n\to\infty} B(\tilde{\theta}) = 0$

3.6.2 Asymptotic Variance

A good estimator will have an asymptotic variance of 0, i.e., $\lim_{n\to\infty} \mathbb{V}\big[\tilde{\theta}\big] = 0$

A consequence of the these two desirable large sample properties is that estimators which satisfy these requirements have an asymptotic **MSE** of 0.

3.7 Consistency

An estimator is said to be consistent if

$$P\big(|\tilde{\theta} - \theta| > \epsilon\big) \to 0 \ \text{ as } n \to \infty \text{ for any } \epsilon > 0 \tag{70}$$

If an estimator is consistent, then it is said to **converge in probability** to θ.

3.8 Sufficiency

Definition. Statistic A statistic, $T(\mathbf{X})$ is a function of the observations in a sample. In contrast to an estimator, it does not need to estimate a parameter.

Definition. Sufficiency A statistic, $T(\mathbf{X})$ is said to be sufficient for a parameter θ if it contains all information needed to estimate the parameter θ. An alternate interpretation of the definition is *no other statistic that can be calculated from the same sample provides any additional information as to the value of the parameter.*

Fisher-Neyman Factorisation Theorem $T(\mathbf{X})$ is a sufficient statistic for θ if and only if there exist two functions g and h such that the likelihood can be written

$$L(\theta) = g\big(T(\mathbf{x}), \theta\big) \times h(\mathbf{x}) \tag{71}$$

$$\tag{72}$$

where h is a function of $\mathbf{x}$ only (including constants) but is not dependent of θ. One can also take the log-likelihood and it becomes

$$\begin{aligned} l(\theta) &= log\big(g(T(\mathbf{x}), \theta)\big) + log\big(h(\mathbf{x})\big) && (73) \\ &= c\big(g(T(\mathbf{x}), \theta)\big) + d(\mathbf{x}) && (74) \end{aligned}$$

Example - The Poisson Likelihood

The Poisson distribution gives the probability of an event happening a k times within a given interval of time or space. The pmf of the Poisson distribution is given by $\frac{\lambda^k e^{-\lambda}}{k!}$.

The log-likelihood of the Poisson distribution with respect to the parameter λ is

$$l(\lambda) = -n\lambda + n\overline{k}log\lambda - \sum_{i=1}^{n} log(k!) \tag{75}$$

$$\tag{76}$$

If we compare the $c\big(g(T(\mathbf{x}), \theta)\big) + d(\mathbf{x})$ to the above, we can see that the final term is independent of the parameter λ whilst the first two terms are dependent on λ. As such we can say that

$$\begin{aligned} d(\mathbf{x}) &= -\sum_{i=1}^{n} log(k!) && (77) \\ c\big(g(T(\mathbf{k}) &= -n\lambda + n\overline{k}log\lambda && (78) \end{aligned}$$

Therefore, we can confirm that $T(X) = \overline{X}$ is a sufficient statistic for λ

Example - The Exponential Likelihood
The exponential distribution is often used to model the time lapse between events

Suppose $X_1, X_2, ..., X_n$ are a set of independent random variables arising from the same exponential distribution with parameter λ. The pdf of this distribution is

$$\begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & x \leq 0 \end{cases} \tag{79}$$

The likelihood function for the exponential distribution is given by $l(\lambda) = nlog\lambda - n\lambda\overline{x}$
If we consider $\overline{X}$ and compare it to $c\big(g(T(\mathbf{x}), \theta)\big) + d(\mathbf{x})$, then we can see that

$$\begin{aligned} c\big(g(T(\mathbf{x}), \theta)\big) &= nlog\lambda - n\lambda\overline{x} && (80) \\ d(\mathbf{x}) &= 0 && (81) \end{aligned}$$

Then if we were to consider $\frac{1}{\overline{X}}$ as a statistic, we can see that

$$c\big(g(T(\mathbf{x}), \theta)\big) = nlog\lambda - \frac{n\lambda}{\frac{1}{\overline{x}}} \tag{82}$$

This demonstrates that a sufficient statistic is not unique. More generally, if $T(\mathbf{X})$ is a sufficient statistic for θ and u is any one-to-one function of T, then $u\big(T(\mathbf{X})\big)$ is also sufficient for θ.

4 Properties of the MLE

The MLE, $\hat{\theta}$ has the following properties, are provided without proof:

- $\hat{\theta}$ is not necessarily an unbiased estimator
- If an unbiased estimator achieves the **CRLB**, then this is both the MLE and MVUE
- $\hat{\theta}$ is an asymptotically unbiased estimator of θ
- The variance of $\hat{\theta}$ tends to zero as n $\rightarrow \infty$. This means that asymptotically the variance coincides with the the value of the **CRLB**
- $\hat{\theta}$ is a consistent estimator of θ
- if $T(\mathbf{X})$ is a sufficient statistic for θ, then $\hat{\theta}$ can be expressed as $\hat{\theta} = u\Big(T(\mathbf{X})\Big)$ for some function of u