1 Point Estimation

Definition 1 (Point Estimator)**.** A *point estimator* is any function $W(X_1, \ldots, X_n)$ of a sample; that is, any statistic is a point estimator. An *estimate* is a realised value $w(x_1, \ldots, x_n)$.

Note that there is the implicit restriction that the estimator is not a function of the parameter you are trying to estimate.

1.1 Methods of Finding Estimators

1.1.1 Method of Moments

The method of moments is performed using a random sample $X_1, \ldots, X_n$ from a population with unknown parameters $\theta_1, \ldots, \theta_k$ by computing the first k empirical moments $m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$ and matching them with the corresponding moments of the population $\mu'_k = \mathbb{E}[X^t]$.

$$\begin{aligned} m_1 &= \mu'_1(\theta_1, \ldots, \theta_k) \\ &\vdots \\ m_k &= \mu'_k(\theta_1, \ldots, \theta_k) \end{aligned}$$

From this you get k simultaneous equations that you can use to solve for the parameters of the population.

1.1.2 Maximum Likelihood Estimators

Definition 2 (Maximum Likelihood Estimator)**.** For each sample point $\boldsymbol{x}$, let $\hat{\theta}(\boldsymbol{x})$ be a parameter value at which the likelihood $L(\theta|\boldsymbol{x})$ attains its maximum as a function of θ, with $\boldsymbol{x}$ held fixed. A *maximum likelihood estimator* (MLE) of the parameter θ based on a sample $\boldsymbol{X}$ is $\hat{\theta}(\boldsymbol{X})$.

Suppose we want to find the MLE for some function of the parameter $\tau(\theta)$.

Definition 3 (Induced Likelihood)**.** Given some function of the parameter $\tau(\theta)$, we define the *induced likelihood function* L^* by

$$L^*(\eta|\boldsymbol{x}) = \sup_{\theta:\tau(\theta)=\eta} L(\theta|\boldsymbol{x}).$$

The value $\hat{\eta}$ that maximises $L^*(\eta|\boldsymbol{x})$ will be called the MLE of $\eta = \tau(\theta)$. It can be seen that the maxima of L^* and L coincide.

Theorem 1 (Invariance Property of MLEs)**.** IF $\hat{\theta}$ is the MLE of θ, then for any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$.

Remark 1.

1) The MLE can be an unstable function of the data.
2) When verifying maxima for multidimensional problems try to avoid going down the hessian route, which can be tedious.

1.1.3 Bayes Estimators

If we denote the prior distribution by $\pi(\theta)$ and the sampling distribution by $f(\boldsymbol{x}|\theta)$, then the posterior distribution of θ given the sample $\boldsymbol{x}$ is

$$\pi(\theta|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\theta)\pi(\theta)}{m(\boldsymbol{x})},$$

where $m(\boldsymbol{x})$ is the marginal distribution of the sample $\boldsymbol{X}$

$$m(\boldsymbol{x}) = \int f(\boldsymbol{x}|\theta)\pi(\theta)\mathrm{d}\theta.$$

Definition 4 (Conjugate Priors)**.** Let $\mathcal{F}$ denote the class of pdfs/pmfs $f(\boldsymbol{x}|\theta)$ (indexed by θ). A class Π of prior distribution is a *conjugate family* for $\mathcal{F}$ if, $\forall f \in \mathcal{F}$, $\forall$ priors $\in \mathcal{F}$ and $\forall \boldsymbol{x} \in \mathcal{X}$, the posterior distribution is in Π.

Note that this relation is not said to be symmetric.

Remark 2 (Some Examples)**.**

- $\mathcal{N}$ is self-conjugate as a family.
- Beta distribution is conjugate to binomial.

1.2 Methods of Evaluating Estimators

Definition 5 (Mean Squared Error)**.** The *mean squared error* (MSE) of an estimator W of a parameter θ is defined by $\mathbb{E}[(W-\theta)^2]$.

Lemma 1 (Bias-Variance Decomposition)**.**

$$\mathbb{E}[(W-\theta)^2] = \mathrm{Var}[W] + (\mathbb{E}[W]-\theta)^2 = \mathrm{Var}[W] + \mathrm{Bias}[W]^2$$

Definition 6 (Bias)**.** The *bias* of a point estimator W of a parameter θ is given by

$$\mathrm{Bias}[W] = \mathbb{E}[W] - \theta.$$

An estimator whose bias is 0 is called an *unbiased estimator* and has $\mathbb{E}[W] = \theta\ \forall\theta$.

Remark 3.

- Clearly, if an estimator is unbiased, then its MSE is equal to its variance.
- The MSE makes sense for location parameters but not so much for scale parameters, since it is symmetric and scale parameters have a natural floor at 0.
- The MSE may be a function of the thing you're trying to estimate. So which estimator you choose as being the 'best' may depend on the range you expect the parameter to lie within.

1.2.1 Best Unbiased Estimators

Definition 7 (Best Unbiased Estimator)**.** An estimator W^* is *best unbiased estimator* of of $\tau(\theta)$ if it satisfies $\mathbb{E}[W^*] = \tau(\theta)\ \forall\theta$, and for any other estimator with $\mathbb{E}[W] = \tau(\theta)\ \forall\theta$ we have $\mathrm{Var}[W*] \leq \mathrm{Var}[W]\ \forall\theta$. We also call W^* the *uniform minimum variance unbiased estimator* (UMVUE) of $\tau(\theta)$.

Remark 4. Suppose that we are trying to estimate θ and consider the class of estimators

$$\mathcal{C}_\tau = \{W : \mathbb{E}[W] = \tau(\theta)\}.$$

All estimators in this class have the same bias, so we can compare their MSEs by comparing their variances alone. (So the best estimator in this class is just the minimum variance one.)

This means that the considerations of this chapter can be applied to classes like $\mathcal{C}_\tau$, even if $\tau(\theta) \neq \theta$.

The best unbiased estimator, if it exists, could be hard to find. The following lower bound at least gives a stopping criterion to our search.

Theorem 2 (Cramér-Rao Inequality)**.** Let $X_1, \ldots, X_n$ be a sample with pdf $f(\boldsymbol{x}|\theta)$ and let $W(\boldsymbol{X}) = W(X_1, \ldots, X_n)$ be any estimator with finite variance satisfying

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}[W(\boldsymbol{X})] = \int_{\mathcal{X}} \frac{\partial}{\partial\theta}\left(W(\boldsymbol{x})f(\boldsymbol{x}|\theta)\right)\mathrm{d}\boldsymbol{x}.$$

Then

$$\mathrm{Var}[W(\boldsymbol{X})] \geq \frac{\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}[W(\boldsymbol{X})]\right)^2}{\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta)\right)^2\right]}.$$

The proof of this considers the correlation between the gradient of the log likelihood and the statistic. Note that the sample in the above theorem is not necessarily iid.

Remark 5. This hold for discrete distributions too, replacing integrals with sums.

Corollary 1. If the random sample $\mathcal{X}_1, \ldots, X_n$ is iid then the result becomes

$$\mathrm{Var}[W(\boldsymbol{X})] \geq \frac{\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}[W(\boldsymbol{X})]\right)^2}{n\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right]}.$$

Definition 8 (Fisher Information)**.** The quantity $\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta)\right)^2\right]$ is called the *Fisher Information.*

Lemma 2. If $f(x|\theta)$ satisfies

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right] = \int \frac{\partial}{\partial\theta}\left[\left(\frac{\partial}{\partial\theta}\log f(x|\theta)\right)f(x|\theta)\right]\mathrm{d}x$$

(as is true for an exponential family), then the Fisher information can be written

$$\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right] = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\right].$$

Remark 6. Even if the Cramér-Rao bound is applicable, it may not be sharp – there may not be an estimator that attains this bound.

Lemma 3 (Attainment)**.** Let $X_1, \ldots, X_n$ be iid $X \sim f(x|\theta)$, where $f(x|\theta)$ satisfies the conditions of the Cramér-Rao Theorem. Let $L(\theta|\boldsymbol{x}) = \prod_{i=1}^n f(x_i|\theta)$ denote the likelihood function. If $W(\boldsymbol{X})$ is any unbiased estimator of $\tau(\theta)$ then $W(\boldsymbol{X})$ attains the Cramér-Rao lower bound if and only if $\exists a(\theta)$ such that

$$a(\theta)[W(\boldsymbol{X}) - \tau(\theta)] = \frac{\partial}{\partial\theta}\log L(\theta|\boldsymbol{x}).$$

1.2.2 Sufficiency and Unbiasedness

Theorem 3 (Rao-Blackwell)**.** Let W be any unbiased estimator of $\tau(\theta)$ and let T be a sufficient statistic for θ. Define $\phi(T) = \mathbb{E}[W|T]$. Then $\mathbb{E}[\phi(T)] = \tau(\theta)$ and $\text{Var}[\phi(T)] \leq \text{Var}[W]\ \forall\theta$. That is, $\phi(T)$ is a *uniformly better unbiased estimator* of $\tau(\theta)$.

Remark 7.

- Conditioning on any unbiased estimator on a sufficient statistic will result in a uniform improvement, so we need consider only functions of a sufficient statistic when looking for best unbiased estimators.
- The proof doesn't require that the statistic we condition on is sufficient, but if it isn't then the resulting quantity will probably depend on the parameter we are trying to estimate.

Theorem 4. If W is a best unbiased estimator of $\tau(\theta)$ then W is unique.

The following theorem is mostly useful to show that a given estimator *isn't* best unbiased.

Theorem 5. If $\mathbb{E}[W] = \tau(\theta)$, then W is the best unbiased estimator of $\tau(\theta)$ if and only if W is uncorrelated with all unbiased estimators of 0.

The proof of this is by considering $\phi_a = W + aU$ where $\mathbb{E}[U] = 0$, then considering the variance.

Remark 8 (Unbiased estimator of 0)**.** Note that an unbiased estimator of 0 is simply noise (one should estimate 0 with 0). If an estimator can be improved by adding noise, then it is probably defective.

We are now in a position such that, if we can characterise all of the unbiased estimators of 0 then we can check if a given estimator is best unbiased. In general this is not easy and requires conditions on the distribution. However, if a distribution is complete then it admits no unbiased estimators of 0 other than 0 itself (by definition), so we will be done.

Note that due to the Rao-Blackwell theorem, only the distribution of the sufficient statistic needs to be complete (not the underlying population distribution).

Theorem 6. Let T be a complete sufficient statistic for a parameter θ and let $\phi(T)$ be any estimator based only on T. Then $\phi(T)$ is the unique best unbiased estimator of its expected value.

Theorem 7 (Lehmann-Scheffé)**.** Unbiased estimators based on complete sufficient statistics are unique.