Statistical Inference

Jayati Kaushik

May 27, 2022

Contents

Chapter 1

Parameter estimation

Let X be a random variable with values in $\mathcal{X}$ and assume that the distribution $\mathcal{L}(X)$ is known upto some paramter θ; i.e.

$$\mathcal{L}(X) = \mathcal{P}_\theta^X = \{P_\theta : \theta \in \Theta\} \quad \Theta = \quad paramter \quad space$$

Here, we assume that $\Theta \subset \mathbb{R}^d$ for $d \geq 1$ and $X = (X_1, X_2, \ldots, X_n)$ with independent random variables X_i. Based on the realization $x = (x_1, x_2, \ldots x_n)$ we want to find the value of θ.

Defintion 1.0.1. *1. A statistic T is a function on $\mathcal{X}$ which does not depend on the unknown parameter theta.*

2. A statistic $T(X_1, X_2, \ldots, X_n) : \mathcal{X} \to (\Theta) = \{g(\theta) : \theta \in \Theta\}$ is called an estimator. We will write

$$\hat{g(\theta)}_n := T(X_1, X_2, \ldots X_n)$$

for the estimator w.r.t sample size n.

3. The value $T(X_1, X_2, \ldots, X_n)$ taken for the observations $x_1, x_2, \ldots, x_n$ is called an estimate.

Remark:

Since the distribution of X is determined by the parameter θ, we will write $P_\theta(X \in B)$ and $E_\theta[g(X)]$ for the probabilities and expectations calculated under the assumption that θ is the true parameter of the distribution of X. **Example** 0

Let $\mathcal{X} = \{0, 1\}^n$ and $X = (X_1, X_2, \ldots, X_n)$ with $X_1, X_2, \ldots, X_n$ iid with

$$P(X_i = 1) = p \quad \text{and} \quad P(X_i = 0) = 1 - p$$

p can be interpreted as the probability of success in Bernoulli experiment. Then $\theta = p$, $\Theta = [0, 1]$ and $\mathcal{L}(X) \in \{P_\theta : \theta \in \Theta\}$ where

$$P_\theta(\{(x_1, x_2, \ldots, x_n)\}) = \prod_{i=1}^{n} P(X_i = x_i) = \theta^s (1 - \theta)^{n-s}$$

with $s = \sum_{i=1}^{n} x_i$
Goal: Estimate $\theta = p$
Guess: $\hat{p} = \frac{s}{n}$. NOTE: The random variabel $S = \sum_{i=1}^{n} X_i$ has a binomial distribution with parameter $p, \mathcal{L}(S) = B(n, p)$.

1.1 Properties of Estimators

In this section we attempt to study some properties of a "good" estimator. $T(X_1, X_2, \ldots, X_n) \equiv 5$ is an estimator. But this is not a "good" estimator; it does not make sense in any applicatins. So, we try to introduce several properties of "good" estimators, as well as measures of quality an estimator.

1.1.1 Highest Concentrations and MSE

Defintion 1.1.1. 1. *If T_1 and T_2 are 2 estimators of parameter θ and $P(\theta - \delta < T_1 < \theta + \delta) \geq P(\theta - \delta < T_2 < \theta + \delta)$ $\forall \theta \in \Theta$ and $\delta > 0$ Then T_1 is preferable to T_2 as an estimator of θ and T_1 is said to be more concentrated around θ than T_2.*

2. *The highest concentration criteria states that the estimator that has the highest concentration is the best.*

Let T_2 be any constant, say c. Then in definition 1.1.1 1, R.H.S. is unity for all θ for which $\delta > |c - \theta|$ Then for the best estimator, L.H.S. should also be unity for θ for which $\delta > |c - \theta|$. Hence a best estimator according to this criteria does not exist.

Defintion 1.1.2. *The Mean Squared Error of an estimator T is defined as $E_\theta[T - \theta]^2$*

Defintion 1.1.3. *If T_1 and T_2 are 2 estimators of θ, then T_1 is better than T_2 if*

$$E_\theta[T_1 - \theta]^2 \leq E_\theta[T_2 - \theta]^2 \forall \theta \in \Theta$$

If the above is true $\forall \quad T_2 \neq T_1$ then T_1 is a best minimum MSE estimator of θ.

Comparision of MSE estimators is 2-dimensional. If T_1 is a better estimator of θ accorinding to definition 1.1.1 it is also better according to definition 1.1.3. As before, let T_2 be a constant c. THen from the definition 1.1.3 at $\theta = c$, R.H.S in definition 1.1.3 is 0. This implies $E_\theta[T_1 - \theta]^2 \leq 0$ at $\theta = c$. This is only possible if $T_1 = c$. Hence, a best estimator according to this criteria is also not possible. Hence, this criterion is commonly used for comparing two estimators, but not for finding a best estimator.

1.1.2 Consistency

A reasonable criterion is that the quality of the estimator should improve with increse in sample size. This leads us to the following definition.

Defintion 1.1.4. *A sequence $T_n = T_n(X_1, X_2, \dots X_n), n \in \mathbb{N}$, of estimators for $g(\theta)$ is called consistent if for $\theta \in \Theta$*

$$P_\theta(||T_n(X_1, \dots, X_n) - g(\theta)|| \geq \epsilon) \xrightarrow{n \to \infty} 0 \quad \forall \epsilon > 0$$

That means that $T_n(X_1, \dots X_n) \xrightarrow{p} g(\theta)$ if $\mathcal{L}(X_1, \dots, X_n) = P_{\theta,n}$ for all $n \geq 1$

Example 1

Let $X_1, X_2 \dots$ be iid real valued random variables with $E(X_i) = \mu$. Then, the law of large numbers implies that $\hat{\mu} = \bar{X}_n$ is a consitent estimator. Consequently,

$$\hat{p} = \frac{s}{n}$$

in example 1 is a consistent estimator.

1.1.3 Unbiasedness

Defintion 1.1.5. 1. *The bias of an estimator $\hat{\theta}$ is given by*

$$bias_\theta(\hat{\theta}) = E_\theta(\hat{\theta}) - \theta$$

2. *An estimator with $bias_\theta(\hat{\theta}) = 0$, i.e.$E_\theta(\hat{\theta}) = \theta$, for all $\theta \in \Theta$ is called unbiased.*

On average, an unbiased estimator estimates the correct parameter value, i.e.,the estimator is centred correctly.

Theorem 1.1.1. *Let $X_1, \dots, X_n$ be an iid with mean μ and variance σ^2. Then $\bar{X}_n$ is an unbiased estimator for μ,*

$$\hat{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \quad \text{and} \quad \hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

are unbiased estimators for σ^2.

Proof. Left as an excercise. □

Example 2

An unbiased estimator does not necessarily have a lower MSE than a biased one. For instance, if $\mathcal{L}(X_i) = N(\mu, \sigma^2)$ then

$$MSE(\frac{n-1}{n}\hat{S}_n^2) < MSE(\hat{S}_n^2)$$

But $\frac{n-1}{n}\hat{S}_n^2$ is biased and underestimates the true value.

Theorem 1.1.2. *The MSE of $\hat{\theta}$ for $\theta \in \Theta \subset \mathbb{R}$ can be expressed as*

$$MSE_\theta(\hat{\theta}) = E_\theta[(\hat{\theta} - \theta)^2] = var_\theta\hat{\theta} + (bias_\theta\hat{\theta})^2.$$

Proof.

$$E_\theta[(\hat\theta-\theta)^2] = E_\theta(\hat\theta^2) - [E_\theta(\hat\theta)]^2 + [E_\theta(\hat\theta)]^2 - 2E_\theta(\hat\theta\theta) + \theta^2 = var_\theta\hat\theta + (bias_\theta\hat\theta)^2$$

□

The MSE of an unbiased estimator consequently reduces to it's variance. Consistency of a sequence of unbiased estimators can therefore be proven by showing that their variance tends to zero.

We know from definition 1.1.1 and definition 1.1.3 that no uniformly consistent estimator exists. So we restrict our class of estimators and look for the best estimator within the that. Hence, we arrive at the following definition.

Defintion 1.1.6. *An estimator $T_n^*(X_1,\dots,X_n)$ is called best unbiased estimator of $g(\theta)\in\mathbb{R}$ if it is unbiased and satisfies*

$$var_\theta T_n^*(X_1,\dots,X_n) \le var_\theta T_n(X_1,\dots,X_n) \quad \forall\theta\in\Theta$$

for all unbiased estimators T_n. T_n^* is also called uniform minimum variance unbiased estimator (UMVUE) of $g(\theta)$.

Theorem 1.1.3. *If T_n is a best unbiased estimator of θ it is almost surely unique.*

With this theorem we can not talk about the best estimator, as if it was unique.

1.1.4 Sufficiency and Completeness

When using a statistic T to make inference on a parameter θ, two samples x and y are considered equal if $T(x)=T(y)$. Hence, T can be regarded as a means of data reduction. This is not always reasonalbe ($e.g., T\equiv 0$). The task now is to reduce data without losing any information about the parameter θ that we want to estimate.

Defintion 1.1.7. *Let X be a sample from a family $\mathcal{P}_\theta^X$. A statistic S is called sufficient for $\theta\in\Theta$ if*

$$P_\theta(X\in B|S(X)=t)$$

does not depend on the unkown parameter θ for all t with $P_\theta(S(X)=t)\neq 0$.

Example 3
In Example 1, $S(X)=\sum_{i=1}^n X_i$ is a sufficient statistic for $\theta=p$:

$$\begin{aligned} P_p(X_1=x_1,\dots,X_n=x+n|S(X)=k) = &\ \frac{P_p(X_1=x_1,\dots,X_n=x_n,S(X)=k)}{P_p(S(X)=k)} \\ = &\ \begin{cases} 0 & \sum_{i=1}^n x_i\neq k \\ \frac{p^k(1-p)^{n-k}}{\binom{n}{k}p^k(1-p)^{n-k}} = \frac{1}{\binom{n}{k}} & \text{otherwise} \end{cases} \end{aligned}$$

for all $x_1,\dots,x_n\in\{0,1\}$ and $0\le k\le n$.

Theorem 1.1.4. *Rao Blackwell Theorem*
Let S be a sufficient statistic. For any unbiased estimator $T(X)$ of $g(\theta)$ there exists another unbiased estimator $\tilde{T}(S(X))$ with

$$var_\theta \tilde{T}(S(X)) \leq var_\theta T(X)$$

Such an estimator is given by

$$\tilde{T}(t) = E_\theta[T(X)|S(X) = t]$$

An unbiased estimator that only depends on the information contained in S is uniformly at least as good as that of T

Proof. $var_\theta(X) = E_\theta[(T(X) - g(\theta))^2]$
$= E_\theta[(T(X) - \tilde{T}(S(X)) + \tilde{T}(S(X)) - g(\theta))^2]$
$= E_\theta[(T(X) - \tilde{T}(S(X)))^2] + var_\theta \tilde{T}(S(X)) + 2E_\theta[(T(X) - \tilde{T}(S(X)))(\tilde{T}(S(X)) - g(\theta))]$
$\geq var \tilde{T}(S(X))$ □

Defintion 1.1.8. *A statistic S is called complete if for all functions g with $E_\theta[g(S(X))] = 0$ for all $\theta \in \Theta$ we have*

$$P_\theta(g(S(X)) = 0) = 1 \quad \forall \theta \in \Theta$$

Example 4
For binomial distribution, if parameter space is restricted to $\Theta = (0, 1)$, then the statistic $S(X) = \sum_{i=1}^{n} X_i$ is complete.

Theorem 1.1.5. *Lehmann - Scheffé*
Let S be a sufficeint and complete statistic for a family of distributions. If there exists an unbiased estimator T of $g(\theta)$ then $\tilde{T}(S(X))$ with $\tilde{T}(s) = E_\theta[T(X)|S(X) = s]$ is the almost surely unique best unbiased estimator.

1.2 Finding Estimators

1.2.1 Method of Moments

1.2.2 Maximum Likelihood Estimators

Defintion 1.2.1. *Let x be a realization of X with values in $\mathcal{X} \in \{P_\theta : \theta \in \Theta\}$. If $P_\theta(X = x) > 0$ the likelihood function is defined as*

$$L(\theta|x) = P_\theta(X = x) \quad x \in \mathcal{X}, \theta \in \Theta$$

If the estimaotr $\hat{\theta}$ satisfies

$$L(\hat{\theta}(X)|X) = max_{\theta \in \Theta} L(\theta|X)$$

it is called maximum liklihood estimator(MLE) of θ

NOTE:
It is often convinient to use the log-likelihood function

$$l(\theta|x) = logL(\theta|x) \quad x \in \mathcal{X}, \theta \in \Theta$$

Example 5
Let $X-1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$-distributed. Then,

$$\begin{aligned} L(\mu, \sigma^2|X_1 \ldots, X_2) = & \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} exp(-\frac{(X_i-\mu)^2}{2\sigma^2}) \\ = & (2\pi\sigma^2)^{\frac{-n}{2}} exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(X_i-\mu)^2) \end{aligned}$$

Differentiating w.r.t σ^2 and μ yeilds

$$\begin{aligned} \frac{\partial l}{\partial \mu}(\mu, \sigma^2|X_1 \ldots, X_2) = & \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 \\ \frac{\partial l}{\partial \sigma^2}(\mu, \sigma^2|X_1 \ldots, X_2) = & -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(X_i-\mu)^2 \end{aligned}$$

Setting these to zero, we obtain

$$\begin{aligned} \hat{\mu} = & \bar{X}_n \\ \hat{\sigma^2} = & \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2 \end{aligned}$$

NOTE:

- We know that $\hat{\sigma^2}$ is a biased estimator of σ^2. Nevertheless, we would prefer this estimator because

  $$MSE(\hat{\sigma^2}) \leq MSE(\hat{S}^2)$$

  .

- In general ML Estimators are not unique.
- It can be shown that MLEs are consitent estimators.
- MLEs are also min max estimaotrs, i.e., they minimize maximum risk.
- MLEs are also invariant.

1.2.3 EM Algorithm