GR5204: Statistical Inference – Lecture 1*
Estimation

Bodhisattva Sen

October 23, 2017

1 Motivation

Statistical inference is concerned with making *probabilistic statements* about *random variables* encountered in the analysis of data.

Examples: means, median, variances ...

Example 1.1. *A company sells a certain kind of electronic component. The company is interested in knowing about how long a component is likely to last on average.*

They can collect data on many such components that have been used under typical conditions.

They choose to use the family of exponential distributions[1] *to model the length of time (in years) from when a component is put into service until it fails.*

The company believes that, if they knew the failure rate θ, then $\boldsymbol{X}_n = (X_1, X_2, \ldots, X_n)$ would be n i.i.d random variables having the exponential distribution with parameter θ. We may ask the following questions:

1. *Can we* **estimate** *θ from this data? If so, what is a reasonable estimator?*
2. *Can we quantify the uncertainty in the estimation procedure, i.e., can we construct* **confidence interval** *for θ?*

*Notes for Chapter 7 of DeGroot and Schervish adapted from Giovanni Motta's and Martin Lindquist's notes for STAT W4109/W4105.

[1] X has an exponential distribution with (failure) rate $\theta > 0$, i.e., $X \sim \text{Exp}(\theta)$, if the p.d.f of X is given by

$$f_\theta(x) = \theta e^{-\theta x} \mathbf{1}_{[0,\infty)}(x), \quad \text{for } x \in \mathbb{R}.$$

The mean (or expected value) of X is given by $\mathbb{E}(X) = \theta^{-1}$, and the variance of X is $\text{Var}(X) = \theta^{-2}$.

1.1 Recap: Some results from probability

Definition 1 (Sample mean)**.** *Suppose that $X_1, X_2, \ldots, X_n$ are n i.i.d r.v's with (unknown) mean $\mu \in \mathbb{R}$ (i.e., $\mathbb{E}(X_1) = \mu$) and variance $\sigma^2 < \infty$. A natural "estimator" of μ is the sample mean (or average) defined as*

$$\bar{X}_n := \frac{1}{n}(X_1 + \ldots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Lemma 1.2. $\mathbb{E}(\bar{X}_n) = \mu$ *and* $\mathrm{Var}(\bar{X}_n) = \sigma^2/n$.

Proof. Observe that

$$\mathbb{E}(\bar{X}_n) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \frac{1}{n}\cdot n\mu = \mu.$$

Also,

$$\mathrm{Var}(\bar{X}_n) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}.$$

□

Theorem 1.3 (Weak law of large numbers)**.** *Suppose that $X_1, X_2, \ldots, X_n$ are n i.i.d r.v's with finite mean μ. Then for any $\epsilon > 0$, we have*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}(X)\right| > \epsilon\right) \to 0 \qquad \text{as } n \to \infty.$$

This says that if we take the sample average of n i.i.d r.v's the sample average will be close to the true population average.

Definition 2 (Convergence in probability)**.** *In the above, we say that the sample mean $\frac{1}{n}\sum_{i=1}^{n} X_i$ converges in probability to the true (population) mean.*

More generally, we say that the sequence of r.v's $\{Z_n\}_{n=1}^{\infty}$ converges to Z **in probability**, *and write*

$$Z_n \xrightarrow{\mathbb{P}} Z,$$

if for every $\epsilon > 0$,

$$\mathbb{P}(|Z_n - Z| > \epsilon) \to 0 \qquad \text{as } n \to \infty.$$

This is equivalent to saying that for every $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|Z_n - Z| \leq \epsilon) = 1.$$

Definition 3 (Convergence in distribution)**.** *We say a sequence of r.v's* $\{Z_n\}_{i=1}^n$ *with c.d.f's* $F_n(\cdot)$ **converges in distribution** *to* F *if*

$$\lim_{n\to\infty} F_n(u) = F(u)$$

for all u *such that* F *is continuous*[2] *at* u *(here* F *is itself a c.d.f).*

The second fundamental result in probability theory, after the law of large numbers (LLN), is the Central limit theorem (CLT), stated below. The CLT gives us the approximate (asymptotic) distribution of $\bar{X}_n$

Theorem 1.4 (Central limit theorem)**.** *If* $X_1, X_2, \ldots$ *are i.i.d with mean zero and variance* 1*, then*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \xrightarrow{d} N(0,1),$$

where $N(0,1)$ *is the standard normal distribution. More generally, the usual rescaling tell us that, for* $X_1, X_2, \ldots$ *are i.i.d with mean* μ *and variance* $\sigma^2 < \infty$

$$\sqrt{n}(\bar{X}_n - \mu) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu) \xrightarrow{d} N(0, \sigma^2).$$

Another class of useful results we will use very much in this course go by the name "continuous mapping theorem". Here are two such results.

Theorem 1.5. *If* $Z_n \xrightarrow{\mathbb{P}} b$ *and if* g *is a function that is continuous at* b*, then*

$$g(Z_n) \xrightarrow{\mathbb{P}} g(b).$$

Theorem 1.6. *If* $Z_n \xrightarrow{d} Z$ *and if* g *is a function that is continuous, then*

$$g(Z_n) \xrightarrow{d} g(Z).$$

[2]Explain why do we need to restrict our attention to continuity points of F. (Hint: think of the following sequence of distributions: $F_n(u) = I(u \geq 1/n)$, where the "indicator" function of a set A is one if $x \in A$ and zero otherwise.)

It's worth emphasizing that convergence in distribution — because it only looks at the c.d.f. — is in fact **weaker** than convergence in probability. For example, if p_X is symmetric, then the sequence $X, -X, X, -X, \ldots$ trivially converges in distribution to X, but obviously doesn't converge in probability.

Also, if $U \sim \text{Unif}(0,1)$, then the sequence

$$U, 1-U, U, 1-U, \ldots$$

converge in distribution to a uniform distribution. But obviously they do not converge in probability.

1.2 Back to Example 1.1

In the first example we have the following results:

- by the LLN, the sample mean $\bar{X}_n$ converges in probability to the expectation $1/\theta$ (failure rate), i.e.,
$$\bar{X}_n \xrightarrow{\mathbb{P}} \frac{1}{\theta};$$
- by the continuous mapping theorem (see Theorem 1.5) $\bar{X}_n^{-1}$ converges in probability to θ, i.e.,
$$\bar{X}_n^{-1} \xrightarrow{\mathbb{P}} \theta;$$
- by the CLT, we know that
$$\sqrt{n}(\bar{X}_n - \theta^{-1}) \xrightarrow{d} N(0, \theta^{-2})$$
where $\mathrm{Var}(X_1) = \theta^{-2}$;
- But how does one find an approximation to the distribution of $\bar{X}_n^{-1}$?

1.3 Delta method

The first thing to note is that if $\{Z_n\}_{i=1}^n$ converges in distribution (or probability) to a constant θ, then $g(Z_n) \xrightarrow{d} g(\theta)$, for any continuous function $g(\cdot)$.

We can also "zoom in" to look at the asymptotic distribution (not just the limit point) of the sequence of r.v's $\{g(Z_n)\}_{i=1}^n$, whenever g is sufficiently smooth.

Theorem 1.7. *Let $Z_1, Z_2, \ldots, Z_n$ be a sequence of r.v's and let Z be a r.v with a continuous c.d.f F^*. Let $\theta \in \mathbb{R}$, and let $a_1, a_2, \ldots,$ be a sequence such that $a_n \to \infty$. Suppose that*
$$a_n(Z_n - \theta) \xrightarrow{d} F^*.$$
Let g be a function with a continuous derivative such that $g'(\theta) \neq 0$. Then
$$a_n \frac{g(Z_n) - g(\theta)}{g'(\theta)} \xrightarrow{d} F^*.$$

Proof. We will only give an outline of the proof (think $a_n = n^{1/2}$, if Z_n as the sample mean). As $a_n \to \infty$, Z_n must get close to θ with high probability as $n \to \infty$.

As $g(\cdot)$ is continuous, $g(Z_n)$ will be close to $g(\theta)$ with high probability.

Let's say $g(\cdot)$ has a Taylor expansion around θ, i.e.,
$$g(Z_n) \approx g(\theta) + g'(\theta)(Z_n - \theta),$$

where we have ignored all terms involving $(Z_n - \theta)^2$ and higher powers.

Then if

$$a_n(Z_n - \theta) \xrightarrow{d} Z,$$

for some limit distribution F^* and a sequence of constants $a_n \to \infty$, then

$$a_n \frac{g(Z_n) - g(\theta)}{g'(\theta)} \approx a_n(Z_n - \theta) \xrightarrow{d} F^*.$$

□

In other words, limit distributions are passed through functions in a pretty simple way. This is called the **delta method** (I suppose because of the deltas and epsilons involved in this kind of limiting argument), and we'll be using it a lot.

The main application is when we've already proven a CLT for Z_n,

$$\frac{\sqrt{n}(Z_n - \mu)}{\sigma} \xrightarrow{d} N(0, 1),$$

in which case

$$\sqrt{n}(g(Z_n) - g(\mu)) \xrightarrow{d} N(0, \sigma^2(g'(\mu))^2).$$

Exercise 1: Assume $n^{1/2}Z_n \xrightarrow{d} N(0, 1)$. What is the asymptotic distribution of

1. $g(Z_n) = (Z_n - 1)^2$?
2. What about $g(Z_n) = Z_n^2$? Does anything go wrong when applying the delta method in this case? Can you fix this problem?

1.4 Back to Example 1.1

By the delta method, we can show that

$$\sqrt{n}(\bar{X}_n^{-1} - \theta) \xrightarrow{d} N(0, (\theta^2)^2\theta^{-2}),$$

where we have considered $g(x) = \frac{1}{x}$; $g'(x) = -\frac{1}{x^2}$ (observe that g is continuous on $(0, \infty)$). Note that the variance of X_1 is $\text{Var}(X_1) = \theta^{-2}$.

2 Statistical Inference

Definition 4 (Statistical model)**.** *A statistical model is*

- *an identification of random variables of interest,*
- *a specification of a joint distribution or a family of possible joint distributions for the observable random variables,*
- *the identification of any parameters of those distributions that are assumed unknown,*
- *(Bayesian approach, if desired) a specification for a (joint) distribution for the unknown parameter(s).*

Definition 5 (Statistical Inference)**.** *A statistical inference is a procedure that produces a probabilistic statement about some or all parts of a statistical model.*

Definition 6 (Parameter space)**.** *In a problem of statistical inference, a characteristic or combination of characteristics that* **determine the joint distribution** *for the random variables of interest is called a* **parameter** *of the distribution.*

The set Ω *of* **all possible values of a parameter** θ *or of a vector of parameters* $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ *is called the parameter space.*

Examples:

- The family of *binomial* distributions has parameters n and p.
- The family of *normal* distributions is parameterized by the mean μ and variance σ^2 of each distribution (so $\boldsymbol{\theta} = (\mu, \sigma^2)$ can be considered a pair of parameters, and $\Omega = \mathbb{R} \times \mathbb{R}^+$).
- The family of *exponential* distributions is parameterized by the rate parameter θ (the failure rate must be positive: Ω will be the set of all positive numbers).

The parameter space Ω must contain all possible values of the parameters in a given problem.

Example 2.1. *Suppose that n patients are going to be given a treatment for a condition and that we will observe for each patient whether or not they recover from the condition.*

For each patient $i = 1, 2, \ldots$, let $X_i = 1$ if patient i recovers, and let $X_i = 0$ if not. As a collection of possible distributions for $X_1, X_2, \ldots$, we could choose to say that the X_i are i.i.d. having the Bernoulli distribution with parameter p, for $0 \leq p \leq 1$.

In this case, the parameter p is known to lie in the closed interval $[0, 1]$, and this interval could be taken as the parameter space. Notice also that by the LLN, p is the limit as $n \to \infty$ of the proportion of the first n patients who recover.

Definition 7 (Statistic)**.** *Suppose that the observable random variables of interest are $X_1, \ldots, X_n$. Let φ be a real-valued function of n real variables. Then the random variable $T = \varphi(X_1, \ldots, X_n)$ is called a* **statistic***.*

Examples:

- the sample mean $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$;
- the maximum $X_{(n)}$ of the values $X_1, \ldots, X_n$;
- the sample variance $S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$ of the values $X_1, \ldots, X_n$.

Definition 8 (Estimator/Estimate)**.** *Let $X_1, \ldots, X_n$ be observable data whose joint distribution is indexed by a parameter θ taking values in a subset Ω of the real line.*

An **estimator** *$\widehat{\theta}_n$ of the parameter θ is a real-valued function $\widehat{\theta}_n = \varphi(X_1, \ldots, X_n)$.*

If $\{X_1 = x_1, \ldots, X_n = x_n\}$ is observed, then $\varphi(x_1, \ldots, x_n)$ is called the **estimate** *of θ.*

Definition 9 (Estimator/Estimate)**.** *Let $X_1, \ldots, X_n$ be observable data whose joint distribution is indexed by a parameter $\boldsymbol{\theta}$ taking values in a subset Ω of k-dimensional space, i.e., $\Omega \subset \mathbb{R}^k$.*

Let $\boldsymbol{h} : \Omega \to \mathbb{R}^d$, be a function from Ω into d-dimensional space. Define $\boldsymbol{\psi} = \boldsymbol{h}(\boldsymbol{\theta})$.

An estimator of $\boldsymbol{\psi}$ is a function $\boldsymbol{g}(X_1, \ldots, X_n)$ that takes values in d-dimensional space. If $\{X_1 = x_1, \ldots, X_n = x_n\}$ are observed, then $\boldsymbol{g}(x_1, \ldots, x_n)$ is called the estimate of $\boldsymbol{\psi}$.

When $\boldsymbol{h}$ in Definition 9 is the identity function $\boldsymbol{h}(\boldsymbol{\theta}) = \boldsymbol{\theta}$, then $\boldsymbol{\psi} = \boldsymbol{\theta}$ and we are estimating the original parameter $\boldsymbol{\theta}$. When $g(\boldsymbol{\theta})$ is one coordinate of $\boldsymbol{\theta}$, then the ψ that we are estimating is just that one coordinate.

Definition 10 (Consistent (in probability) estimator)**.** *A sequence of estimators* $\widehat{\boldsymbol{\theta}}_n$ *that* **converges in probability** *to the unknown value of the parameter* $\boldsymbol{\theta}$ *being estimated is called a* **consistent sequence of estimators**, *i.e.,* $\widehat{\theta}_n$ *is consistent if and only if for every* $\epsilon > 0$,

$$\mathbb{P}(|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}| > \epsilon) \to 0, \qquad \text{as } \ n \to \infty.$$

In this Chapter we shall discuss three types of estimators:

- **Method of moments** estimators,
- **Maximum likelihood** estimators, and
- **Bayes** estimators.

3 Method of Moments estimators

The *method of moments* (MOM) is an intuitive method for estimating parameters when other, more attractive, methods may be too difficult (to implement/compute).

Definition 11 (Method of moments estimator)**.** *Assume that* $X_1, \ldots, X_n$ *form a random sample from a distribution that is indexed by a* k*-dimensional parameter* $\boldsymbol{\theta}$ *and that has at least* k *finite moments. For* $j = 1, \ldots, k$, *let*

$$\mu_j(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}(X_1^j).$$

Suppose that the function $\boldsymbol{\mu}(\boldsymbol{\theta}) = (\mu_1(\boldsymbol{\theta}), \ldots, \mu_k(\boldsymbol{\theta}))$ *is a one-to-one function of* $\boldsymbol{\theta}$. *Let* $M(\mu_1, \ldots, \mu_k)$ *denote the inverse function, that is, for all* $\boldsymbol{\theta}$,

$$\boldsymbol{\theta} = M(\mu_1, \ldots, \mu_k).$$

Define the ***sample moments*** *as*

$$m_j := \frac{1}{n} \sum_{i=1}^{n} X_i^j \quad \text{for } j = 1, \ldots, k.$$

The method of moments estimator of $\boldsymbol{\theta}$ *is* $M(m_1, \ldots, m_k)$.

The usual way of implementing the method of moments is to set up the k equations $m_j = \mu_j(\boldsymbol{\theta})$ and then solve for $\boldsymbol{\theta}$.

Example 3.1. *Let $X_1, X_2, \ldots, X_n$ be from a $N(\mu, \sigma^2)$ distribution. Thus $\boldsymbol{\theta} = (\mu, \sigma^2)$. What is the MOM estimator of $\boldsymbol{\theta}$?*

Solution: *Consider $\mu_1 = \mathbb{E}(X_1)$ and $\mu_2 = \mathbb{E}(X_1^2)$. Clearly, the parameter $\boldsymbol{\theta}$ 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

$$\hat{\mu} = m_1 = \bar{X} \, ,$$

and

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

Example 3.2. *Suppose that $X_1, X_2, \ldots, X_n$ are i.i.d Gamma(α, β), $\alpha, \beta > 0$. Thus, $\boldsymbol{\theta} = (\alpha, \beta) \in \Omega := \mathbb{R}_+ \times \mathbb{R}_+$. The first two moments of this distribution are:*

$$\mu_1(\boldsymbol{\theta}) = \frac{\alpha}{\beta}, \qquad \mu_2(\boldsymbol{\theta}) = \frac{\alpha(\alpha+1)}{\beta^2},$$

which implies that

$$\alpha = \frac{\mu_1^2}{\mu_2 - \mu_1^2}, \qquad \beta = \frac{\mu_1}{\mu_2 - \mu_1^2}.$$

The MOM says that we replace the right-hand sides of these equations by the sample moments and then solve for α and β. In this case, we get

$$\hat{\alpha} = \frac{m_1^2}{m_2 - m_1^2}, \qquad \hat{\beta} = \frac{m_1}{m_2 - m_1^2}.$$

MOM can thus be thought of as "plug-in" estimates; to get an estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta} = M(\mu_1, \mu_2, \ldots, \mu_k)$, we plug-in estimates of the μ_i's, which are the m_i's, to get $\hat{\boldsymbol{\theta}}$.

Result: If M is continuous, then the fact that m_i converges in probability to μ_i, for every $i = 1, \ldots, k$, entails that

$$\hat{\boldsymbol{\theta}} = M(m_1, m_2, \ldots, m_k) \xrightarrow{\mathbb{P}} M(\mu_1, \mu_2, \ldots, \mu_k) = \boldsymbol{\theta}.$$

Thus MOM estimators are consistent under mild assumptions.

Proof. LLN: the sample moments converge in probability to the population moments $\mu_1(\boldsymbol{\theta}), \ldots, \mu_k(\boldsymbol{\theta})$.

The generalization of the continuous mapping theorem (Theorem 6.2.5 in the book) to functions of k variables implies that $M(\cdot)$ evaluated at the sample moments converges in probability to $\boldsymbol{\theta}$, i.e., the MOM estimator converges in probability to $\boldsymbol{\theta}$. □

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

Example 3.3. *Let $X_1, X_2, \ldots, X_n$ be the indicators of n Bernoulli trials with success probability θ. We are going to find a MOM estimator of θ.*

Solution: *Note that θ is the probability of success and satisifes,*

$$\theta = \mathbb{E}(X_1)\ ,\ \theta = \mathbb{E}(X_1^2)\,.$$

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

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

and

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

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.*

Note that

$$Var_\theta(\bar{X}) = \frac{\theta\,(1-\theta)}{n}\,,$$

and a MOM estimate of $Var_\theta(\bar{X})$ is obtained as

$$\widehat{Var_\theta(\bar{X})}_{MOM} = \frac{\bar{X}\,(1-\bar{X})}{n}\,.$$

Example 3.4. *Let $X_1, X_2, \ldots, X_n$ be i.i.d. Poisson(λ). Find the MOM estimator of λ.*

Solution: *We know that,*

$$\mathbb{E}(X_1) = \mu_1 = \lambda$$

and $Var(X_1) = \mu_2 - \mu_1^2 = \lambda$. *Thus*

$$\mu_2 = \lambda + \lambda^2.$$

Now, a MOM estimate of λ *is clearly given by* $\hat{\lambda} = m_1 = \bar{X}$*; thus a MOM estimate of* $\mu_2 = \lambda^2 + \lambda$ *is given by* $\bar{X}^2 + \bar{X}$.

On the other hand, the obvious MOM estimate of m_2 *is*

$$m_2 = \frac{1}{n}\sum_{j=1}^{n} X_j^2.$$

However these two estimates are not necessarily equal; in other words, it is not necessarily the case that $\bar{X}^2 + \bar{X} = (1/n)\sum_{j=1}^{n} X_j^2$.

This illustrates one of the disadvantages of MOM estimates — they may not be uniquely defined.

Example 3.5. *Consider* n *systems with failure times* $X_1, X_2, \ldots, X_n$ *assumed to be i.i.d Exp*(λ), $\lambda > 0$. *Find the MOM estimators of* λ.

Solution: *It is not difficult to show that*

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

Therefore

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

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

$$\hat{\lambda}_{MOM} = \frac{1}{m_1}\,,$$

and the estimate based on the second moment is

$$\hat{\lambda}_{MOM} = \sqrt{\frac{2}{m_2}}\,.$$

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.

4 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.

Definition 12 (Likelihood function)**.** *The likelihood function for the sample* $\boldsymbol{X}_n = (X_1, X_2, \ldots, X_n)$ *is*

$$L_n(\theta, \boldsymbol{X}_n) = \prod_{i=1}^{n} f(X_i, \theta).$$

This is simply the joint density (or mass function) but we now think of this as a function of θ *for a fixed* $\boldsymbol{X}_n$*; namely the* $\boldsymbol{X}_n$ *that is realized.*

Heutistics: Suppose for the moment that X_i's are discrete, so that f is actually a p.m.f. Then $L_n(\boldsymbol{X}_n, \theta)$ is exactly the probability that the observed data is realized or "happens".

We now seek to obtain that $\theta \in \Omega$ for which $L_n(\boldsymbol{X}_n, \theta)$ 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 **maximum likelihood estimate** (MLE).

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.

Sometimes, in the case of multiple maximizers, we choose one which is more desirable according to some "sensible" criterion.

Example 4.1. *Suppose that $X_1, \ldots, X_n$ are i.i.d Poisson(θ), $\theta > 0$. Find the MLE of θ.*

Solution: *In this case, it is easy to see that*

$$L_n(\theta, \boldsymbol{X}_n) = \prod_{i=1}^{n} \frac{e^{-\theta}\, \theta^{X_i}}{X_i\,!} = C(\boldsymbol{X}_n)\, e^{-n\,\theta}\, \theta^{\sum_{i=1}^{n} X_i}\,.$$

To maximize this expression, we set

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

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

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

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).

Example 4.2. *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,*

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

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

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

We split the maximization problem into the following 3 cases.

(i) $\bar{X}_n = 1$; 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) $\bar{X}_n = 0$; 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) $0 < \bar{X}_n < 1$*; in this case it is easy to see that the function* $\log L_n(\theta, \boldsymbol{X}_n)$ *goes to* $-\infty$ *as* θ *approaches 0 or 1, so that for purposes of maximization we can restrict to* $0 < \theta < 1$*. To maximize* $\log L_n(\theta, \boldsymbol{X}_n)$*, we solve the equation,*

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

which yields,

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

This gives $\theta = \bar{X}$*.*

It can be checked by computing the second derivative at $\bar{X}_n$ *or noticing that* $\log\, L_n(\theta, \boldsymbol{X}_n)$ *is concave in* θ *that the function attains a maximum at* $\bar{X}_n$*. Thus, the MLE,* $\hat{\theta}_n = \bar{X}_n$ *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 4.3. *Suppose* $X_1, X_2, \ldots, X_n$ *are i.i.d Unif*$([0, \theta])$ *random variables, where* $\theta > 0$*. We want to obtain the MLE of* θ*.*

Solution: *The likelihood function is given by,*

$$\begin{aligned} L_n(\theta, \boldsymbol{X}_n) &= \prod_{i=1}^{n} \frac{1}{\theta} I_{[0,\theta]}(X_i) \\ &= \frac{1}{\theta^n} \prod_{i=1}^{n} I_{[X_i,\infty)}(\theta) \\ &= \frac{1}{\theta^n} I_{[\max_{i=1,\ldots,n} X_i, \infty)}(\theta). \end{aligned}$$

It is then clear that $L_n(\theta, \boldsymbol{X}_n)$ *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*

$$\hat{\theta}_n = \max_{i=1,\ldots,n} X_i.$$

Here, differentiation will not help you to get the MLE because the likelihood function is not differentiable at the point where it hits the maximum.

Example 4.4. *Suppose that* $X_1, X_2, \ldots, X_n$ *are i.i.d* $N(\mu, \sigma^2)$*. We want to find the MLEs of the mean* μ *and the variance* σ^2*.*

Solution: *We write down the likelihood function first. This is,*

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

It is easy to see that,

$$\begin{aligned}\log L_n(\mu, \sigma^2, \boldsymbol{X}_n) &= -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 + \textit{constant} \\ &= -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 - \frac{n}{2\sigma^2}(\bar{X}_n - \mu)^2.\end{aligned}$$

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

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

showing that we can choose $\mu_{MLE} = \bar{X}_n$.

It then remains to maximize $\log L_n(\bar{X}_n, \sigma^2, \boldsymbol{X}_n)$ with respect to σ^2 to find σ^2_{MLE}.

Now,

$$\log L_n(\bar{X}_n, \sigma^2, \boldsymbol{X}_n) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

Differentiating the left-side w.r.t σ^2 gives,

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

where $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$. The above equation leads to,

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

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 4.5. *We now tweak the above situation a bit. Suppose now that we restrict the parameter space, so that μ has to be non-negative, i.e., $\mu \geq 0$.*

Thus we seek to maximize $\log L_n(\mu, \sigma^2, \boldsymbol{X}_n)$ but subject to the constraint that $\mu \geq 0$ and $\sigma^2 > 0$.

Solution: *Clearly, the MLE is $(\bar{X}_n, \hat{\sigma}^2)$ if $\bar{X}_n \geq 0$.*

In case, that $\bar{X}_n < 0$ we proceed thus. For fixed σ^2, the function $\log\, L_n(\mu, \sigma^2, \boldsymbol{X}_n)$, as a function of μ, attains a maximum at $\bar{X}_n$ and then falls off as a parabola on either side. The $\mu \geq 0$ for which the function $\log\, L_n(\mu, \sigma^2, \boldsymbol{X}_n)$ is the largest is 0; thus $\mu_{MLE} = 0$ and $\log\, L_n(\hat{\mu}, \sigma^2, \boldsymbol{X}_n)$ is then given by,

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

Proceeding as before (by differentiation) it is shown that

$$\sigma^2_{MLE} = \frac{1}{n} \sum_{i=1}^{n} X_i^2.$$

Thus the MLEs can be written as,

$$(\mu_{MLE}, \sigma^2_{MLE}) = I_{(-\infty,0)}(\bar{X}) \left(0, \frac{1}{n} \sum_{i=1}^{n} X_i^2 \right) + I_{[0,+\infty)}(\bar{X})(\bar{X}, \hat{\sigma}^2).$$

Example 5 (non-uniqueness of MLE): Suppose that $X_1, \ldots, X_n$ form a random sample from the uniform distribution on the interval $[\theta, \theta+1]$, where $\theta \in \mathbb{R}$ is unknown. We want to find the MLE of θ.

Solution: The likelihood has the form

$$L_n(\theta) = \prod_{i=1}^{n} I_{[\theta,\theta+1]}(X_i).$$

The condition that $\theta \leq X_i$, for all $i = 1, \ldots, n$, is equivalent to the condition that $\theta \leq \min\{X_1, \ldots, X_n\} = X_{(1)}$.

Similarly, the condition that $X_i \leq \theta + 1$, for all $i = 1, \ldots, n$, is equivalent to the condition that $\theta \geq \max\{X_1, \ldots, X_n\} - 1 = X_{(n)} - 1$. Thus the likelihood can be written as

$$L_n(\theta) = I_{[X_{(n)}-1, X_{(1)}]}(\theta).$$

Hence it is possible to select as an MLE any value of θ in the interval $[X_{(n)} - 1, X_{(1)}]$, and thus the MLE is not unique.

Example 4.6. *Consider a random variable X that can come with equal probability either from a $N(0, 1)$ or from $N(\mu, \sigma^2)$, where both μ and σ are unknown.*

Thus, the p.d.f. $f(\cdot, \mu, \sigma^2)$ of X is given by

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

Suppose now that $X_1, \ldots, X_n$ form a random sample from this distribution. As usual, the likelihood function

$$L_n(\mu, \sigma^2) = \prod_{i=1}^{n} f(X_i, \mu, \sigma^2).$$

We want to find the MLE of $\theta = (\mu, \sigma^2)$.

Let X_k denote one of the observed values. If we let $\mu = X_k$ and let $\sigma^2 \to 0$ then the factor $f(X_k, \mu, \sigma^2)$ will grow large without bound, while each factor $f(X_i, \mu, \sigma^2)$, for $i \neq k$, will approach the value

$$\frac{1}{2\sqrt{2\pi}}e^{-X_i^2/2}.$$

Hence, when $\mu = X_k$ and $\sigma^2 \to 0$, we find that $L_n(\mu, \sigma^2) \to \infty$.

Note that 0 is not a permissible estimate of σ^2, because we know in advance that $\sigma > 0$. Since the likelihood function can be made arbitrarily large by choosing $\mu = X_k$ and choosing σ^2 arbitrarily close to 0, it follows that the MLE does not exist.

4.1 Properties of MLEs

4.1.1 Invariance

Theorem 4.7 (Invariance property of MLEs)**.** *If $\widehat{\theta}_n$ is the MLE of θ and if h is any function, then $h\big(\widehat{\theta}_n\big)$ is the MLE of $h(\theta)$.*

See Theorem 7.6.2 and Example 7.6.3 in the text book.

Thus if $X_1, \ldots, X_n$ be i.i.d N(μ, σ^2), then the MLE of μ^2 is $\bar{X}_n^2$.

4.1.2 Consistency

Consider an estimation problem in which a random sample is to be taken from a distribution involving a parameter θ.

Then, under certain conditions, which are typically satisfied in practical problems, the sequence of MLEs is *consistent*, i.e.,

$$\widehat{\theta}_n \xrightarrow{\mathbb{P}} \theta, \qquad \text{as } n \to \infty.$$

4.2 Computational methods for approximating MLEs

Example: Suppose that $X_1, \ldots, X_n$ are i.i.d from a Gamma distribution for which the p.d.f is as follows:

$$f(x, \alpha) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x}, \qquad \text{for } x > 0.$$

The likelihood function is

$$L_n(\alpha) = \frac{1}{\Gamma(\alpha)^n} \left(\prod_{i=1}^{n} X_i \right)^{\alpha-1} e^{-\sum_{i=1}^{n} X_i},$$

and thus the log-likelihood is

$$\ell_n(\alpha) = \log L_n(\alpha) = -n \log \Gamma(\alpha) + (\alpha - 1) \sum_{i=1}^{n} \log(X_i) - \sum_{i=1}^{n} X_i,$$

The MLE of α will be the value of α that satisfies the equation

$$\begin{aligned} \frac{\partial}{\partial \alpha} \ell_n(\alpha) &= -n \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \sum_{i=1}^{n} \log(X_i) = 0 \\ \text{i.e.,} \quad \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} &= \frac{1}{n} \sum_{i=1}^{n} \log(X_i). \end{aligned}$$

4.2.1 Newton's Method

Let $f(x)$ be a real-valued function of a real variable, and suppose that we wish to solve the equation

$$f(x) = 0.$$

Let x_0 be an initial guess at the solution.

Newton's method replaces the initial guess with the updated guess

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

The rationale behind the Newton's method is: approximate the curve by a line tangent to the curve passing through the point $(x_0, f(x_0))$. The approximating line crosses the horizontal axis at the revised guess x_1. [Draw a figure!]

Typically, one replaces the initial guess with the revised guess and iterates Newton's method until the results stabilize (see e.g., http://en.wikipedia.org/wiki/Newton's_method).

4.2.2 The EM Algorithm

Read Section 7.6 of the text-book. I will cover this later, if time permits.

5 Principles of estimation

Setup: Our data $X_1, X_2, \ldots X_n$ are i.i.d observations from the distribution P_θ where $\theta \in \Omega$, the parameter space (Ω is assumed to be the k-dimensional Euclidean space). We assume identifiability of the parameter, i.e. $\theta_1 \neq \theta_2 \Rightarrow P_{\theta_1} \neq P_{\theta_2}$.

Estimation problem: Consider now, the problem of estimating $g(\theta)$ where g is some function of θ.

In many cases $g(\theta) = \theta$ itself.

Generally $g(\theta)$ will describe some important aspect of the distribution P_θ.

Our estimator of $g(\theta)$ will be some function of our observed data $\boldsymbol{X}_n = (X_1, X_2, \ldots, X_n)$.

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 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 $|T_n - g(\theta)|$ 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 $\epsilon > 0$,

$$\mathbb{P}_\theta \left(|T_n - g(\theta)| > \epsilon \right) \to 0.$$

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?

A way out of this difficulty is to obtain an average measure of the error, or in other words, average out $|T_n - g(\theta)|$ 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 absolute error** and can be written compactly (using acronym) as:

$$\text{MAD} = \mathbb{E}_\theta \left(|T_n - g(\theta)| \right).$$

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** (MSE) and is

$$\text{MSE}(T_n, g(\theta)) = \mathbb{E}_\theta \left[(T_n - g(\theta))^2 \right].$$

Of course, this is meaningful, only if the above quantity is finite for all θ. Good estimators are those for which the MSE is generally not too high, whatever be the value of θ.

There is a standard decomposition of the MSE that helps us understand its components. We have,

$$\begin{aligned}
\text{MSE}(T_n, g(\theta)) &= \mathbb{E}_\theta \left[(T_n - g(\theta))^2 \right] \\
&= \mathbb{E}_\theta \left[(T_n - \mathbb{E}_\theta(T_n) + \mathbb{E}_\theta(T_n) - g(\theta))^2 \right] \\
&= \mathbb{E}_\theta \left[(T_n - \mathbb{E}_\theta(T_n))^2 \right] + (\mathbb{E}_\theta(T_n) - g(\theta))^2 \\
&\qquad\qquad + 2\, \mathbb{E}_\theta[(T_n - \mathbb{E}_\theta(T_n))(\mathbb{E}_\theta(T_n) - g(\theta))] \\
&= \text{Var}_\theta(T_n) + b(T_n, g(\theta))^2,
\end{aligned}$$

where $b(T_n, g(\theta)) = \mathbb{E}_\theta(T_n) - g(\theta)$ is the **bias** of T_n as an estimator of $g(\theta)$.

The cross product term in the above display vanishes since $\mathbb{E}_\theta(T_n) - g(\theta)$ is a constant and $\mathbb{E}_\theta(T_n - \mathbb{E}_\theta(T_n)) = 0$.

The bias measures, on an average, by how much T_n overestimate or underestimate $g(\theta)$. If we think of the expectation $\mathbb{E}_\theta(T_n)$ 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 MSEs. Under parameter value θ, T_n dominates S_n as an estimator if

$$\text{MSE}(T_n, \theta) \leq \text{MSE}(S_n, \theta) \qquad \text{for all } \theta \in \Omega.$$

In this situation we say that S_n is *inadmissible* in the presence of T_n.

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**, except 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 $\mathbb{E}_\theta(T_n) = g(\theta)$ for all possible values of θ; i.e.,

$$b(T_n, g(\theta)) = 0 \qquad \text{for all } \theta \in \Omega.$$

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 MSE 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 **minimum variance unbiased estimator** (MVUE). Here is a formal definition.

MVUE: We call S_n an MVUE of $g(\theta)$ if

$$(i) \;\; \mathbb{E}_\theta(S_n) = g(\theta) \qquad \text{for all } \theta \in \Omega$$

and (ii) if T_n is an unbiased estimate of $g(\theta)$, then $\text{Var}_\theta(S_n) \leq \text{Var}_\theta(T_n)$.

Here are a few examples to illustrate some of the various concepts discussed above.

(a) Consider $X_1, \ldots, X_n$ i.i.d $N(\mu, \sigma^2)$.

A natural unbiased estimator of $g_1(\theta) = \mu$ is $\bar{X}_n$, 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 $\bar{X}_n$. Recall that the variance of $\bar{X}_n$ is simply σ^2/n.

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

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

Out of these $\hat{\sigma}^2$ is not unbiased for σ^2 but s^2 is. 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 $g(\theta) = \mathbb{E}_\theta(X_1)$.

Then the sample mean $\bar{X}_n$ is *always* an unbiased estimate of $g(\theta)$. Whether it is MVUE or not depends on the underlying structure of the model.

(c) Suppose that $X_1, X_2, \ldots, X_n$ be i.i.d Ber(θ). It can be shown that $\bar{X}_n$ 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 favor 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 $T_n \equiv \bar{X}_n/(1 - \bar{X}_n)$. 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 $X_1, X_2, \ldots, X_n$ i.i.d Unif($[0, \theta]$). Here $\Omega = (0, \infty)$.

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 $\bar{X}_n$ is an unbiased estimate of $\theta/2$, the common mean of the X_i's; hence $2\,\bar{X}_n$ is an unbiased estimate of θ.

Show that $X_{(n)}$ in the sense of MSE outperforms $2\bar{X}_n$ by an order of magnitude.

The best unbiased estimator (MVUE) of θ is $(1 + n^{-1})X_{(n)}$.

Solution: We can show that

$$\begin{aligned} \mathrm{MSE}(2\bar{X}_n, \theta) &= \frac{\theta^2}{3n} = \mathrm{Var}(2\bar{X}_n) \\ \mathrm{MSE}((1 + n^{-1})X_{(n)}, \theta) &= \frac{\theta^2}{n(n+2)} = \mathrm{Var}((1 + n^{-1})X_{(n)}) \\ \mathrm{MSE}(X_{(n)}, \theta) &= \frac{\theta^2}{n(n+2)} \cdot \frac{n^2}{(n+1)^2} + \frac{\theta^2}{(n+1)^2}, \end{aligned}$$

where in the last equality we have two terms – the variance and the squared bias.

So far we have discussed some broad general principles and exemplified them somewhat.

6 Sufficient Statistics

In some problems, there may not be any MLE, or there may be more than one. Even when an MLE is unique, it may not be a suitable estimator (as in the Unif$(0, \theta)$ example, where the MLE always underestimates the value of θ).

In such problems, the search for a good estimator must be extended beyond the methods that have been introduced thus far.

In this section, we shall define the concept of a *sufficient statistic*, which can be used to simplify the search for a good estimator in many problems.

Suppose that in a specific estimation problem, two statisticians A and B must estimate the value of the parameter θ.

Statistician A can observe the values of the observations $X_1, X_2, \ldots, X_n$ in a random sample, and statistician B cannot observe the individual values of $X_1, X_2, \ldots, X_n$ but can learn the value of a certain statistic $T = \varphi(X_1, \ldots, X_n)$.

In this case, statistician A can choose any function of the observations $X_1, X_2, \ldots, X_n$ as an estimator of θ (including a function of T). But statistician B can use only a function of T. Hence, it follows that A will generally be able to find a better estimator than will B.

In some problems, however, B will be able to do just as well as A. In such a problem, the single function $T = \varphi(X_1, \ldots, X_n)$ will in some sense summarize all the information contained in the random sample about θ, and knowledge of the individual values of $X_1, \ldots, X_n$ will be irrelevant in the search for a good estimator of θ.

A statistic T having this property is called a *sufficient statistic*.

A statistic is *sufficient* with respect to a statistical model P_θ and its associated unknown parameter θ if it provides "all" the information on θ; .e.g., if "no other statistic that can be calculated from the same sample provides any additional information as to the value of the parameter".

Definition 13 (Sufficient statistic)**.** *Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution indexed by a parameter $\theta \in \Omega$. Let T be a statistic. Suppose that, for every $\theta \in \Omega$ and every possible value t of T, the conditional joint distribution of $X_1, X_2, \ldots, X_n$ given that $T = t$ (at θ) depends only on t but not on θ.*

That is, for each t, the conditional distribution of $X_1, X_2, \ldots, X_n$ given $T = t$ is the same for all θ. Then we say that T is a sufficient statistic for the parameter θ.

So, if T is sufficient, and one observed only T instead of $(X_1, \ldots, X_n)$, one could, at least in principle, simulate random variables $(X_1', \ldots, X_n')$ with the same joint distribution.

In this sense, T is sufficient for obtaining as much information about θ as one could get from $(X_1, \ldots, X_n)$.

We shall now present a simple method for finding a sufficient statistic that can be applied in many problems.

Theorem 6.1 (Factorization criterion)**.** *Let $X_1, X_2, \ldots, X_n$ form a random sample from either a continuous distribution or a discrete distribution for which the p.d.f or the p.m.f is $f(x, \theta)$, where the value of θ is unknown and belongs to a given parameter space Ω.*

A statistic $T = r(X_1, X_2, \ldots, X_n)$ is a sufficient statistic for θ if and only if the joint p.d.f or the joint p.m.f $f_n(\boldsymbol{x}, \theta)$ of $(X_1, X_2, \ldots, X_n)$ can be factored as follows for all

values of $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ *and all values of* $\theta \in \Omega$*:*

$$f_n(\boldsymbol{x}, \theta) = u(\boldsymbol{x})\nu(r(\boldsymbol{x}), \theta),$$

where

- u *and* ν *are both non-negative,*
- *the function* u *may depend on* $\boldsymbol{x}$ *but does not depend on* θ*,*
- *the function* ν *will depend on* θ *but depends on the observed value* $\boldsymbol{x}$ *only through the value of the statistic* $r(\boldsymbol{x})$*.*

Example: Suppose that $X_1, \ldots, X_n$ are i.i.d Poi(θ), $\theta > 0$. Thus, for every non-negative integers $x_1, \ldots, x_n$, the joint p.m.f $f_n(\mathbf{x}, \theta)$ of $(X_1, \ldots, X_n)$ is

$$f_n(\mathbf{x}, \theta) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} = \frac{1}{\prod_{i=1}^{n} x_i!} e^{-n\theta}\theta^{\sum_{i=1}^{n} x_i}.$$

Thus, we can take $u(\boldsymbol{x}) = 1/(\prod_{i=1}^{n} x_i!)$, $r(\boldsymbol{x}) = \sum_{i1=}^{n} x_i$, $\nu(t, \theta) = e^{-n\theta}\theta^t$. It follows that $T = \sum_{i=1}^{n} X_i$ is a sufficient statistic for θ.

Exercise: Suppose that $X_1, \ldots, X_n$ are i.i.d Gamma(α, β), $\alpha, \beta > 0$, where α is known, and β is unknown. The joint p.d.f is

$$f_n(\boldsymbol{x}, \beta) = \Big\{ [\Gamma(\alpha)]^n \big(\underbrace{\prod_{i=1}^{n} x_i}_{u(\boldsymbol{x})} \big)^{\alpha - 1} \Big\}^{-1} \times \Big\{ \underbrace{\beta^{n\alpha} \exp(-n\beta t)}_{\nu(t,\beta)} \Big\}, \quad \text{where } t = \sum_{i=1}^{n} x_i.$$

The sufficient statistics is $T_n = \sum_{i=1}^{n} X_i$.

Exercise: Suppose that $X_1, \ldots, X_n$ are i.i.d Gamma(α, β), $\alpha, \beta > 0$, where α is unknown, and β is known.

The joint p.d.f in this exercise is the same as that given in the previous exercise. However, since the unknown parameter is now α instead of β, the appropriate factorization is now

$$f_n(\boldsymbol{x}, \alpha) = \Big\{ \underbrace{\exp\big(-\beta \sum_{i=1}^{n} x_i \big)}_{u(\boldsymbol{x})} \Big\} \times \Big\{ \underbrace{\frac{\beta^{n\alpha}}{[\Gamma(\alpha)]^n} t^{\alpha - 1}}_{\nu(t,\alpha)} \Big\}, \quad \text{where } t = \prod_{i=1}^{n} x_i.$$

The sufficient statistics is $T_n = \prod_{i=1}^{n} X_i$.

Exercise: Suppose that $X_1, \ldots, X_n$ are i.i.d Unif$(0, \theta)$, $\theta > 0$ is the unknown parameter.

Show that $T = \max\{X_1, \ldots, X_n\}$ is the sufficient statistic.

7 Bayesian paradigm

Frequentist versus Bayesian statistics:

Frequentist:

- Data are a repeatable random sample – there is a frequency.
- *Parameters are fixed.*
- Underlying parameters remain constant during this repeatable process.

Bayesian:

- Parameters are unknown and described probabilistically
- *Analysis is done conditioning on the observed data; i.e., data is treated as fixed.*

7.1 Prior distribution

Definition 14 (Prior distribution)**.** *Suppose that one has a statistical model with parameter θ. If one treats θ as random, then the distribution that one assigns to θ before observing the data is called its* **prior distribution**.

Thus, now θ is random and will be denoted by Θ (note the change of notation).

We will assume that if the prior distribution of Θ is continuous, then its p.d.f is called the prior p.d.f of Θ.

Example: Let Θ denote the probability of obtaining a head when a certain coin is tossed.

- Case 1: Suppose that it is known that the coin either is fair or has a head on each side. Then Θ only takes two values, namely $1/2$ and 1. If the prior probability that the coin is fair is 0.8, then the prior p.m.f of Θ is $\xi(1/2) = 0.8$ and $\xi(1) = 0.2$.
- Case 2: Suppose that Θ can take any value between $(0, 1)$ with a prior distribution given by a Beta distribution with parameters $(1, 1)$.

Suppose that the observable data $X_1, X_2, \ldots, X_n$ are modeled as random sample from a distribution indexed by θ. Suppose $f(\cdot|\theta)$ denote the p.m.f/p.d.f of a single random variable under the distribution indexed by θ.

When we treat the unknown parameter Θ as random, then the joint distribution of the observable random variables (i.e., data) indexed by θ is understood as the **conditional distribution** of the data given $\Theta = \theta$.

Thus, in general we will have $X_1, \ldots, X_n|\Theta = \theta$ are i.i.d with p.d.f/p.m.f $f(\cdot|\theta)$, and that $\Theta \sim \xi$, i.e.,

$$f_n(\boldsymbol{x}|\theta) = f(x_1|\theta) \ldots f(x_n|\theta),$$

where f_n is the joint conditional distribution of $\boldsymbol{X} = (X_1, \ldots, X_n)$ given $\Theta = \theta$.

7.2 Posterior distribution

Definition 15 (Posterior distribution)**.** *Consider a statistical inference problem with parameter θ and random variables $X_1, \ldots, X_n$ to be observed. The conditional distribution of Θ given $X_1, \ldots, X_n$ is called the posterior distribution of θ.*

The conditional p.m.f/p.d.f of Θ given $X_1 = x_1, \ldots, X_n = x_n$ is called the posterior p.m.f/p.d.f of θ and is usually denoted by $\xi(\cdot|x_1, \ldots, x_n)$.

Theorem 7.1. *Suppose that the n random variables $X_1, \ldots, X_n$ form a random sample from a distribution for which the p.d.f/p.m.f is $f(\cdot|\theta)$. Suppose also that the value of the parameter θ is unknown and the prior p.d.f/p.m.f of θ is $\xi(\cdot)$. Then the posterior p.d.f/p.m.f of θ is*

$$\xi(\theta|\boldsymbol{x}) = \frac{f(x_1|\theta) \cdots f(x_n|\theta)\xi(\theta)}{g_n(\boldsymbol{x})}, \qquad \text{for } \theta \in \Omega,$$

where g_n is the marginal joint p.d.f/p.m.f of $X_1, \ldots, X_n$.

7.3 Sampling from a Bernoulli distribution

Theorem 7.2. *Suppose that $X_1, \ldots, X_n$ form a random sample from the Bernoulli distribution with mean $\theta > 0$, where $0 < \theta < 1$ is unknown. Suppose that the prior distribution of Θ is Beta(α, β), where $\alpha, \beta > 0$.*

Then the posterior distribution of Θ given $X_i = x_i$, for $i = 1, \ldots, n$, is Beta$(\alpha + \sum_{i=1}^n x_i, \beta + n - \sum_{i=1}^n x_i)$.

Proof. The joint p.m.f of the data is

$$f_n(\boldsymbol{x}|\theta) = f(x_1|\theta)\cdots f(x_n|\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}.$$

Therefore the posterior density of $\Theta|X_1 = x_1, \ldots, X_n = x_n$ is given by

$$\begin{aligned} \xi(\theta|\boldsymbol{x}) &\propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \cdot \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} \\ &= \theta^{\sum_{i=1}^n x_i+\alpha-1}(1-\theta)^{\beta+n-\sum_{i=1}^n x_i-1}, \end{aligned}$$

for $\theta \in (0,1)$. Thus, $\Theta|X_1 = x_1, \ldots, X_n = x_n \sim Beta(\alpha + \sum_{i=1}^n x_i, \beta + n - \sum_{i=1}^n x_i)$. □

7.4 Sampling from a Poisson distribution

Theorem 7.3. *Suppose that $X_1, \ldots, X_n$ form a random sample from the Poisson distribution with mean $\theta > 0$, where θ is unknown. Suppose that the prior distribution of Θ is $Gamma(\alpha, \beta)$, where $\alpha, \beta > 0$.*

Then the posterior distribution of Θ given $X_i = x_i$, for $i = 1, \ldots, n$, is $Gamma(\alpha + \sum_{i=1}^n x_i, \beta + n)$.

Definition: Let $X_1, X_2, \ldots,$ be conditionally i.i.d given $\Theta = \theta$ with common p.m.f/p.d.f $f(\cdot|\theta)$, where $\theta \in \Omega$.

Let Ψ be a family of possible distributions over the parameter space Ω. Suppose that no matter which prior distribution ξ we choose from Ψ, no matter how many observations $\boldsymbol{X} = (X_1, \ldots, X_n)$ we observe, and no matter what are their observed values $\boldsymbol{x} = (x_1, \ldots, x_n)$, the posterior distribution $\xi(\cdot|\boldsymbol{x})$ is a member of Ψ.

Then Ψ is called a *conjugate family of prior distributions* for samples from the distributions $f(\cdot|\theta)$.

7.5 Sampling from an Exponential distribution

Example: Suppose that the distribution of the lifetime of fluorescent tubes of a certain type is the exponential distribution with parameter θ. Suppose that $X_1, \ldots, X_n$ is a random sample of lamps of this type.

Also suppose that $\Theta \sim \text{Gamma}(\alpha, \beta)$, for known α, β.

Then

$$f_n(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} \theta e^{-\theta x_i} = \theta^n e^{-\theta \sum_{i=1}^n x_i}.$$

Then the posterior distribution of Θ given the data is

$$\xi(\theta|\boldsymbol{x}) \propto \theta^n e^{-\theta \sum_{i=1}^n x_i} \cdot \theta^{\alpha-1} e^{-\beta\theta} = \theta^{n+\alpha-1} e^{-(\beta+\sum_{i=1}^n x_i)\theta}.$$

Therefore, $\Theta|\boldsymbol{X}_n = \boldsymbol{x} \sim \text{Gamma}(\alpha + n, \beta + \sum_{i=1}^n x_i)$.

8 Bayes Estimators

An estimator of a parameter is some function of the data that we hope is close to the parameter, i.e., $\hat{\theta} \approx \theta$.

Let $X_1, \ldots, X_n$ be data whose joint distribution is indexed by a parameter $\theta \in \Omega$.

Let $\delta(X_1, \ldots, X_n)$ be an estimator of θ.

Definition: A *loss function* is a real-valued function of two variables, $L(\theta, a)$, where $\theta \in \Omega$ and $a \in \mathbb{R}$.

The interpretation is that the statistician loses $L(\theta, a)$ if the parameter equals θ and the estimate equals a.

Example: (Squared error loss) $L(\theta, a) = (\theta - a)^2$.

(Absolute error loss) $L(\theta, a) = |\theta - a|$.

Suppose that $\xi(\cdot)$ is a prior p.d.f/p.m.f of $\theta \in \Omega$. Consider the problem of estimating θ without being able to observe the data. If the statistician chooses a particular estimate a, then her expected loss will be

$$\mathbb{E}[L(\theta, a)] = \int_\Omega L(\theta, a)\xi(\theta) d\theta.$$

It is sensible that the statistician wishes to choose an estimate a for which the expected loss is *minimum*.

Definition: Suppose now that the statistician can observe the value $\boldsymbol{x}$ of a the data $\boldsymbol{X}_n$ before estimating θ, and let $\xi(\cdot|\boldsymbol{x})$ denote the posterior p.d.f of $\theta \in \Omega$. For each estimate a that the statistician might use, her expected loss in this case will be

$$\mathbb{E}[L(\theta, a)|\boldsymbol{x}] = \int_\Omega L(\theta, a)\xi(\theta|\boldsymbol{x}) d\theta. \tag{1}$$

Hence, the statistician should now choose an estimate a for which the above expectation is minimum.

For each possible value $\boldsymbol{x}$ of $\boldsymbol{X}_n$, let $\delta^*(\boldsymbol{x})$ denote a value of the estimate a for which the expected loss (1) is minimum. Then the function $\delta^*(\boldsymbol{X}_n)$ is called the *Bayes estimator* of θ.

Once $\boldsymbol{X}_n = \boldsymbol{x}$ is observed, $\delta^*(\boldsymbol{x})$ is called the Bayes estimate of θ.

Thus, a Bayes estimator is an estimator that is chosen to minimize the posterior mean of some measure of how far the estimator is from the parameter.

Corollary 8.1. *Let $\theta \in \Omega \subset \mathbb{R}$. Suppose that the squared error loss function is used and the posterior mean of Θ, i.e., $\mathbb{E}(\Theta|\boldsymbol{X}_n)$ is finite. Then the Bayes estimator of θ is $\delta^*(X_n) = \mathbb{E}(\Theta|\boldsymbol{X}_n)$.*

Example 1: (Bernoulli distribution with Beta prior)

Suppose that $X_1, \ldots, X_n$ form a random sample from the Bernoulli distribution with mean $\theta > 0$, where $0 < \theta < 1$ is unknown. Suppose that the prior distribution of Θ is Beta(α, β), where $\alpha, \beta > 0$.

Recall that $\Theta|X_1 = x_1, \ldots, X_n = x_n \sim \text{Beta}(\alpha + \sum_{i=1}^n x_i, \beta + n - \sum_{i=1}^n x_i)$. Thus,

$$\delta^*(\boldsymbol{X}) = \frac{\alpha + \sum_{i=1}^n X_i}{\alpha + \beta + n}.$$

8.1 Sampling from a normal distribution

Theorem 8.2. *Suppose that $X_1, \ldots, X_n$ form a random sample from $N(\mu, \sigma^2)$, where μ is unknown and the value of the variance $\sigma^2 > 0$ is known. Suppose that $\Theta \sim N(\mu_0, v_0^2)$. Then*

$$\Theta|X_1 = x_1, \ldots, X_n = x_n \sim N(\mu_1, v_1^2),$$

where

$$\mu_1 = \frac{\sigma^2\mu_0 + nv_0^2\bar{x}_n}{\sigma^2 + nv_0^2} \qquad \text{and} \qquad v_1^2 = \frac{\sigma^2 v_0^2}{\sigma^2 + nv_0^2}.$$

Proof. The joint density has the form

$$f_n(\boldsymbol{x}|\theta) \propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \theta)^2\right].$$

The method of completing the squares tells us that

$$\sum_{i=1}^n (x_i - \theta)^2 = n(\theta - \bar{x}_n)^2 + \sum_{i=1}^n (x_i - \bar{x}_n)^2.$$

Thus, by omitting the factor that involves $x_1, \ldots, x_n$ but does depend on θ, we may rewrite $f_n(\boldsymbol{x}|\theta)$ as

$$f_n(\boldsymbol{x}|\theta) \propto \exp\left[-\frac{n}{2\sigma^2}(\theta - \bar{x}_n)^2\right].$$

Since the prior density has the form

$$\xi(\theta) \propto \exp\left[-\frac{1}{2v_0^2}(\theta - \mu_0)^2\right],$$

it follows that the posterior p.d.f $\xi(\theta|\boldsymbol{x})$ satisfies

$$\xi(\theta|\boldsymbol{x}) \propto \exp\left[-\frac{n}{2\sigma^2}(\theta - \bar{x}_n)^2 - \frac{1}{2v_0^2}(\theta - \mu_0)^2\right].$$

Completing the squares again establishes the following identity:

$$\frac{n}{\sigma^2}(\theta - \bar{x}_n)^2 + \frac{1}{v_0^2}(\theta - \mu_0)^2 = \frac{1}{v_1^2}(\theta - \mu_1)^2 + \frac{n}{\sigma^2 + nv_0^2}(\bar{x}_n - \mu_0)^2.$$

The last term on the right side does not involve on θ. Thus,

$$\xi(\theta|\boldsymbol{x}) \propto \exp\left[-\frac{1}{2v_1^2}(\theta - \mu_1)^2\right].$$

□

Thus,

$$\delta^*(\boldsymbol{X}) = \frac{\sigma^2\mu_0 + nv_0^2\bar{X}_n}{\sigma^2 + nv_0^2}.$$

Corollary 8.3. *Let $\theta \in \Omega \subset \mathbb{R}$. Suppose that the absolute error loss function is used. Then the Bayes estimator of θ $\delta^*(X_n)$ equals the median of the posterior distribution of Θ.*