Basic Principles of Point Estimation

Statistical Theory

Guillaume Dehaene
Ecole Polytechnique Fédérale de Lausanne

The Problem of Point Estimation

Point Estimation for Parametric Families

- Collection of r.v.'s (a random vector) $\mathbf{X} = (X_1, ..., X_n)$
- $\mathbf{X} \sim F_\theta \in \mathcal{F}$
- $\mathcal{F}$ a parametric class with parameter $\theta \in \Theta \subseteq \mathbb{R}^d$

The Problem of Point Estimation

1. Assume that F_θ is known up to the parameter θ which is unknown
2. Let $(x_1, ..., x_n)$ be a realization of $\mathbf{X} \sim F_\theta$ which is available to us
3. Estimate the value of θ that generated the sample given $(x_1, ..., x_n)$

So far considered aspects related to point estimation:

- Considered approximate distributions of $g(X_1, ..., X_n)$ as $n \uparrow \infty$
- Studied the information carried by $g(X_1, .., X_n)$ w.r.t θ
- Examined general parametric models

Today: How do we estimate θ in general? Some general recipes?

Point Estimators

Definition (Point Estimator)

Let $\{F_\theta\}$ be a parametric model with parameter space $\Theta \subseteq \mathbb{R}^d$ and let $\mathbf{X} = (X_1, \ldots, X_n) \sim F_{\theta_0}$ for some $\theta_0 \in \Theta$. A point estimator $\hat{\theta}$ of θ_0 is a statistic $T : \mathbb{R}^n \to \Theta$, whose primary purpose is to estimate θ_0

Therefore any statistic $T : \mathbb{R}^n \to \Theta$ is a candidate estimator!

$\hookrightarrow$ Harder to answer what a *good* estimator is!

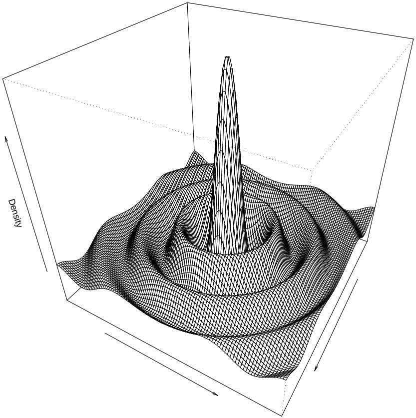
- Any estimator is of course a random variable
- Hence as a general principle, good should mean:

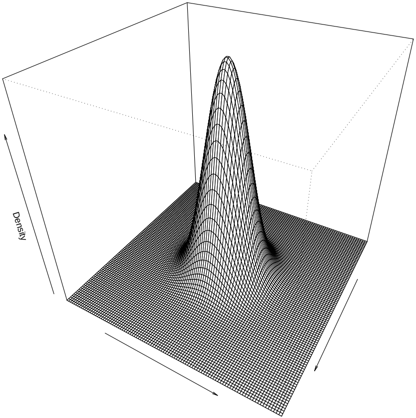
$$\text{dist}(\hat{\theta}) \text{ concentrated around } \theta$$

  $\hookrightarrow$ An ∞-dimensional description of quality.

- Look at some simpler measures of quality?

Concentration around a Parameter





Bias, Variance and Mean Squared Error

Bias and Mean Squared Error

Definition (Bias)

The *bias* of an estimator $\hat{\theta}$ of $\theta \in \Theta$ is defined to be

$$\text{bias}(\hat{\theta}) = \mathbb{E}_\theta[\hat{\theta}] - \theta$$

Describes how "off" we're from the target on average when employing $\hat{\theta}$.

Definition (Unbiasedness)

An estimator $\hat{\theta}$ of $\theta \in \Theta$ is *unbiased* if $\mathbb{E}_\theta[\hat{\theta}] = \theta$, i.e. $\text{bias}(\hat{\theta}) = 0$.

Will see that not too much weight should be placed on unbiasedness.

Definition (Mean Squared Error)

The *mean squared error* of an estimator $\hat{\theta}$ of $\theta \in \Theta \subseteq \mathbb{R}$ is defined to be

$$MSE(\hat{\theta}) = \mathbb{E}_\theta\left[(\hat{\theta} - \theta)^2\right]$$

Bias and Mean Squared Error

Bias and MSE combined provide a coarse but simple description of concentration around θ:

- Bias gives us an indication of the location of dist$(\hat{\theta})$ relative to θ (somehow assumes mean is good measure of location)
- MSE gives us a measure of spread/dispersion of dist$(\hat{\theta})$ around θ
- If $\hat{\theta}$ is unbiased for $\theta \in \mathbb{R}$ then $\text{Var}(\hat{\theta}) = MSE(\hat{\theta})$
- for $\Theta \subseteq \mathbb{R}^d$ have $MSE(\hat{\theta}) := \mathbb{E}\|\hat{\theta} - \theta\|^2$.

Example

Let $X_1, ..., X_n \stackrel{iid}{\sim} N(\mu, \sigma^2)$ and let $\hat{\mu} := \overline{X}$. Then

$$\mathbb{E}\hat{\mu} = \mu \quad \text{and} \quad MSE(\mu) = \text{Var}(\mu) = \frac{\sigma^2}{n}.$$

In this case bias and MSE give us a complete description of the concentration of dist$(\hat{\mu})$ around μ, since $\hat{\mu}$ is Gaussian and so completely determined by mean and variance.

The Bias-Variance Decomposition of MSE

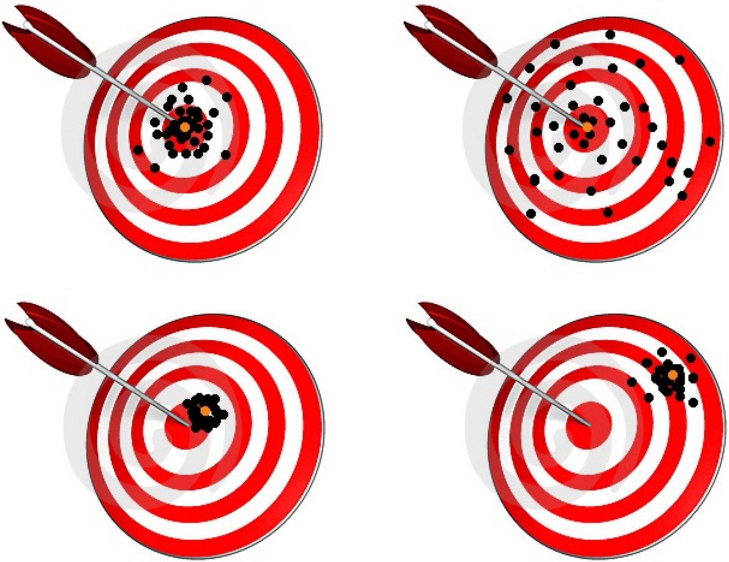
$$
\begin{aligned}
\mathbb{E}[\hat{\theta} - \theta]^2 &= \mathbb{E}[\hat{\theta} - \mathbb{E}\hat{\theta} + \mathbb{E}\hat{\theta} - \theta]^2 \\
&= \mathbb{E}\left\{(\hat{\theta} - \mathbb{E}\hat{\theta})^2 + (\mathbb{E}\hat{\theta} - \theta)^2 + 2(\hat{\theta} - \mathbb{E}\hat{\theta})(\mathbb{E}\hat{\theta} - \theta)\right\} \\
&= \mathbb{E}(\hat{\theta} - \mathbb{E}\hat{\theta})^2 + (\mathbb{E}\hat{\theta} - \theta)^2
\end{aligned}
$$

Bias-Variance Decomposition for $\Theta \subseteq \mathbb{R}$

$$MSE(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + \mathrm{bias}^2(\hat{\theta})$$

- A simple yet fundamental relationship
- Requiring a small MSE does not necessarily require unbiasedeness
- Unbiasedeness is a sensible property, but sometimes biased estimators perform better than unbiased ones
- Sometimes have bias/variance tradeoff
  (e.g. nonparametric regression)

Bias–Variance Tradeoff



Consistency

Can also consider quality of an estimator not for given sample size, but also as sample size increases.

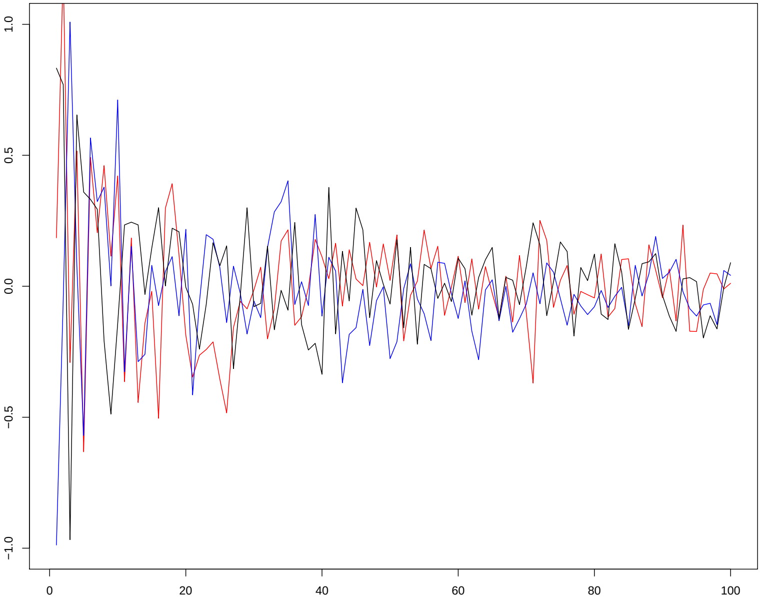
Consistency

A sequence of estimators $\{\hat{\theta}_n\}_{n\geq 1}$ of $\theta \in \Theta$ is said to be *consistent* if

$$\hat{\theta}_n \xrightarrow{P} \theta$$

- A consistent estimator becomes increasingly concentrated around the true value θ as sample size grows (usually have $\hat{\theta}_n$ being an estimator based on n iid values).
- Often considered as a "must have" property, but...
- A more detailed understanding of the "asymptotic quality" of $\hat{\theta}$ requires the study of dist$[\hat{\theta}_n]$ as $n \uparrow \infty$.

Consistency: $X_1, \ldots, X_n \sim \mathcal{N}(0,1)$, plot $\bar{X}_n$ for $n = 1, 2, \ldots$



The Plug-In Principle

Plug-In Estimators

Want to find general procedures for constructing estimators.

$\hookrightarrow$ An idea: $\theta \mapsto F_\theta$ is bijection under identifiability.

- Recall that more generally, a parameter is a function $\nu : \mathcal{F} \to \mathcal{N}$
- Under identifiability $\nu(F_\theta) = q(\theta)$, some q.

The Plug-In Principle

Let $\nu = q(\theta) = \nu(F_\theta)$ be a parameter of interest for a parametric model $\{F_\theta\}_{\theta\in\Theta}$. If we can construct an estimate $\hat{F}$ of F_θ on the basis of our sample $\mathbf{X}$, then we can use $\nu(\hat{F})$ as an estimator of $\nu(F_\theta)$. Such an estimator is called a *plug-in estimator*.

- Essentially we are "flipping" our point of view: viewing θ as a function of F_θ instead of F_θ as a function of θ.
- Note here that $\theta = \theta(F_\theta)$ if q is taken to be the identity.
- In practice such a principle is useful when we can explicitly describe the mapping $F_\theta \mapsto \nu(F_\theta)$.

Parameters as Functionals of F

Examples of "functional parameters":

- The mean: $\mu(F) := \int_{-\infty}^{+\infty} x dF(x)$
- The variance: $\sigma^2(F) := \int_{-\infty}^{+\infty} [x - \mu(F)]^2 dF(x)$
- The median: $\text{med}(F) := \inf\{x : F(x) \geq 1/2\}$
- An indirectly defined parameter $\theta(F)$ such that:

$$\int_{-\infty}^{+\infty} \psi(x - \theta(F)) dF(x) = 0$$

- The density (when it exists) at x_0: $\theta(F) := \left.\frac{d}{dx} F(x)\right|_{x=x_0}$

Plug-in Principle

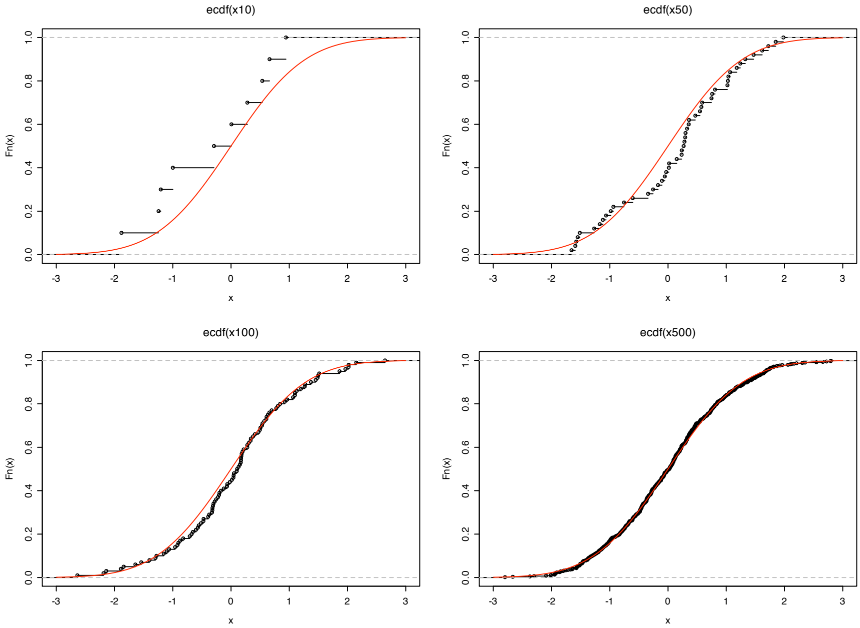
Converts problem of estimating θ into problem of estimating F. But how?

Consider the case when $\mathbf{X} = (X_1, .., X_n)$ has iid coordinates. We may define the empirical version of the distribution function $F_{X_i}(\cdot)$ as

$$\hat{F}_n(y) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i \leq y\}$$

- Places mass $1/n$ on each observation
- SLLN $\implies$ $\hat{F}_n(y) \xrightarrow{a.s.} F(y)\ \forall y \in \mathbb{R}$
  - $\hookrightarrow$ since $\mathbf{1}\{X_i \leq y\}$ are iid Bernoulli$(F(y))$ random variables

Suggests using $\nu(\hat{F}_n)$ as estimator of $\nu(F)$


ecdf(x10)
ecdf(x50)
ecdf(x100)
ecdf(x500)
Fn(x)
x
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3

The Empirical Distribution Function

Seems that we're actually doing better than just pointwise convergence...

Theorem (Glivenko-Cantelli)

Let $X_1, .., X_n$ be independent random variables, distributed according to F. Then, $\hat{F}_n(y) = n^{-1} \sum_{i=1}^{n} \mathbf{1}\{X_i \leq y\}$ converges uniformly to F with probability 1, i.e.

$$\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \xrightarrow{a.s.} 0$$

Proof.

Assume first that $F(y) = y\mathbf{1}\{y \geq 0\}$. (ie: $X_i \sim U([0,1])$).
Fix a regular finite partition $0 = x_1 \leq x_2 \leq \ldots \leq x_m = 1$ of [0,1] (so $x_{k+1} - x_k = (m-1)^{-1}$).
By monotonicity of $F, \hat{F}_n$

$$\sup_x |\hat{F}_n(x) - F(x)| < \max_k |\hat{F}_n(x_k) - F(x_{k+1})| + \max_k |\hat{F}_n(x_k) - F(x_{k-1})|$$

Adding and subtracting $F(x_k)$ within each term we can bound above by

$$2\max_k|\hat{F}_n(x_k)-F(x_k)| + \underbrace{\max_k|F(x_k)-F(x_{k+1})| + \max_k|F(x_k)-F(x_{k-1})|}_{=\max_k|x_k-x_{k+1}|+\max_k|x_k-x_{k-1}|=\frac{2}{m-1}}$$

by an application of the triangle inequality to each term. Letting $n\uparrow\infty$, the SSLN implies that the first term vanishes almost surely. Since m is arbitrary we have proven that, given any $\epsilon>0$,

$$\lim_{n\to\infty}\left[\sup_x|\hat{F}_n(x)-F(x)|\right]<\epsilon \quad a.s.$$

which gives the result when the cdf F is uniform.

For a general cdf F, we let $U_1,U_2,\ldots\overset{iid}{\sim}\mathcal{U}[0,1]$ and define

$$W_i:=F^{-1}(U_i)=\inf\{x:F(x)\geq U_i\}.$$

Observe that

$$W_i \leq x \iff U_i \leq F(x)$$

so that $W_i \stackrel{d}{=} X_i$. By Skorokhod's representation theorem, we may thus assume that

$$W_i = X_i \qquad \text{a.s.}$$

Letting $\hat{G}_n$ be the ecdf of $(U_1, ..., U_n)$ we note that

$$\hat{F}_n(y) = n^{-1} \sum_{i=1}^{n} \mathbf{1}\{W_i \leq y\} = n^{-1} \sum_{i=1}^{n} \mathbf{1}\{U_i \leq F(y)\} = \hat{G}_n(F(y)), \quad \text{a.s.}$$

in other words $$\hat{F}_n = \hat{G}_n \circ F, \text{ a.s.}$$

Now let $A = F(\mathbb{R}) \subseteq [0, 1]$ so that from the first part of the proof

$$\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| = \sup_{t \in A} |\hat{G}_n(t) - t| \leq \sup_{t \in [0,1]} |\hat{G}_n(t) - t| \stackrel{a.s.}{\to} 0$$

since obviously $A \subseteq [0, 1]$. $\square$

Example (Mean of a function)

Consider $\theta(F) = \int_{-\infty}^{+\infty} h(x)dF(x)$. A plug-in estimator based on the edf is

$$\hat{\theta} := \theta(\hat{F}_n) = \int_{-\infty}^{+\infty} h(x)d\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} h(X_i)$$

Example (Variance)

Consider now $\sigma^2(F) = \int_{-\infty}^{+\infty}(x-\mu(F))^2 dF(x)$. Plugging in $\hat{F}_n$ gives

$$\sigma^2(\hat{F}_n) = \int_{-\infty}^{+\infty} x^2 d\hat{F}_n(x) - \left(\int_{-\infty}^{+\infty} x d\hat{F}_n(x)\right)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2$$

Exercise

Show that $\sigma^2(\hat{F}_n)$ is a biased but consistent estimator for any F.

Example (Density Estimation)

Let $\theta(F) = f(x_0)$, where f is the density of F,

$$F(t) = \int_{-\infty}^{t} f(x)dx$$

If we tried to plug-in $\hat{F}_n$ then our estimator would require differentiation of $\hat{F}_n$ at x_0. Clearly, the edf plug-in estimator does not exist since $\hat{F}_n$ is a step function. We will need a "smoother" estimate of F to plug in, e.g.

$$\tilde{F}_n(x) := \int_{-\infty}^{\infty} G(x-y)d\hat{F}_n(y) = \frac{1}{n}\sum_{i=1}^{n} G(x - X_i)$$

for some continuous G concentrated at 0.

- Saw that plug-in estimates are usually easy to obtain via $\hat{F}_n$
- But such estimates are not necessarily as "innocent" as they seem.

The Moment Principle

The Method of Moments

Panaretos: "Perhaps the oldest estimation method (K. Pearson)"

Method of Moments

Let $X_1, \ldots, X_n$ be an iid sample from F_θ, $\theta \in \mathbb{R}^p$. The *method of moments* estimator $\hat{\theta}$ of θ is the solution w.r.t θ to the p random equations

$$\int_{-\infty}^{+\infty} x^{k_j} d\hat{F}_n(x) = \int_{-\infty}^{+\infty} x^{k_j} dF_\theta(x), \quad \{k_j\}_{j=1}^{p} \subset \mathbb{N}.$$

- In some sense this is a plug-in estimator - we estimate the theoretical moments by the sample moments in order to then estimate θ.
- Useful when exact functional form of $\theta(F)$ unavailable
- While the method was introduced by equating moments, it may be generalized to equating p theoretical functionals to their empirical analogues.

  $\hookrightarrow$ Choice of equations can be important

Motivational Diversion: The Moment Problem

Theorem

Suppose that F is a distribution determined by its moments. Let $\{F_n\}$ be a sequence of distributions such that $\int x^k dF_n(x) < \infty$ for all n and k. Then,

$$\lim_{n\to\infty} \int x^k dF_n(x) = \int x^k dF(x), \quad \forall\ k \geq 1 \implies F_n \xrightarrow{w} F.$$

BUT: Not all distributions are determined by their moments!

Lemma

The distribution of X is determined by its moments, provided that there exists an open neighbourhood A containing zero such that

$$M_X(u) = \mathbb{E}\left[e^{-\langle u, X\rangle}\right] < \infty, \quad \forall\ u \in A.$$

Example (Exponential Distribution)

Suppose $X_1, ..., X_n \stackrel{iid}{\sim} Exp(\lambda)$. Then, $\mathbb{E}[X_i^r] = \lambda^{-r}\Gamma(r+1)$. Hence, we may define a class of estimators of λ depending on r,

$$\hat{\lambda} = \left[\frac{1}{n\Gamma(r+1)}\sum_{i=1}^{n} X_i^r\right]^{-\frac{1}{r}}.$$

Tune value of r so as to get a "best estimator" (will see later...)

Example (Gamma Distribution)

Let $X_1, ..., X_n \stackrel{iid}{\sim} \text{Gamma}(\alpha, \lambda)$. The first two moment equations are:

$$\frac{\alpha}{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X} \quad \text{and} \quad \frac{\alpha}{\lambda^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

yielding estimates $\hat{\alpha} = \bar{X}^2/\hat{\sigma}^2$ and $\hat{\lambda} = \bar{X}/\hat{\sigma^2}$.

Example (Discrete Uniform Distribution)

Let $X_1, ..., X_n \stackrel{iid}{\sim} \mathcal{U}\{1, 2, ..., \theta\}$, for $\theta \in \mathbb{N}$. Using the first moment of the distribution we obtain the equation

$$\bar{X} = \frac{1}{2}(\theta + 1)$$

yielding the MoM estimator $\hat{\theta} = 2\bar{X} - 1$.

A nice feature of MoM estimators is that they generalize to non-iid data.
$\rightarrow$ if $\mathbf{X} = (X_1, ..., X_n)$ has distribution depending on $\theta \in \mathbb{R}^p$, one can choose statistics $T_1, ..., T_p$ whose expectations depend on θ:

$$\mathbb{E}_\theta T_k = g_k(\theta)$$

and then equate

$$T_k(\mathbf{X}) = g_k(\theta), \quad k = 1, ..., p.$$

$\rightarrow$ Important here that T_k is a reasonable estimator of $\mathbb{E}\, T_k$. (motivation)

Comments on Plug-In and MoM Estimators

- Usually easy to compute and can be valuable as preliminary estimates for algorithms that attempt to compute more efficient (but not easily computable) estimates.
- Can give a starting point to search for better estimators in situations where simple intuitive estimators are not available.
- Often these estimators are consistent, so they are likely to be close to the true parameter value for large sample size.
  - $\hookrightarrow$ Use empirical process theory for plug-ins
  - $\hookrightarrow$ Estimating equation theory for MoM's
- Can lead to biased estimators, or even completely ridiculous estimators (will see later)

Comments on Plug-In and MoM Estimators

- The estimate provided by an MoM estimator may $\notin \Theta$!
(exercise: show that this can happen with the binomial distribution, both n and p unknown).
- Will later discuss optimality in estimation, and appropriateness (or inappropriateness) will become clearer.
- Observation: many of these estimators do not depend solely on sufficient statistics
  - $\hookrightarrow$ Sufficiency seems to play an important role in optimality – and it does (more later)
- Will now see a method where estimator depends *only* on a sufficient statistic, when such a statistic exists.

The Likelihood Principle

The Likelihood Function

A central theme in statistics. Introduced by Ronald Fisher.

Definition (The Likelihood Function)

Let $\mathbf{X} = (X_1, ..., X_n)$ be random variables with joint density (or frequency function) $f(\mathbf{x};\theta)$, $\theta \in \Theta \subseteq \mathbb{R}^p$. The likelihood function $L(\theta)$ is the random function

$$L(\theta) = f(\mathbf{X};\theta)$$

$\hookrightarrow$ Notice that we consider L as a function of θ NOT of $\mathbf{X}$.

Interpretation: Most easily interpreted in the discrete case $\rightarrow$ How likely does the value θ make what we observed?

(can extend interpretation to continuous case by thinking of $L(\theta)$ as how likely θ makes something in a small neighbourhood of what we observed)

When $\mathbf{X}$ has iid coordinates with density $f(\cdot;\theta)$, then likelihood is:

$$L(\theta) = \prod_{i=1}^{n} f(X_i;\theta)$$

Maximum Likelihood Estimators

Definition (Maximum Likelihood Estimators)

Let $\mathbf{X} = (X_1, ..., X_n)$ be a random sample from F_θ, and suppose that $\hat{\theta}$ is such that

$$L(\hat{\theta}) \geq L(\theta), \quad \forall\ \theta \in \Theta.$$

Then $\hat{\theta}$ is called *a maximum likelihood estimator of* θ.

We call $\hat{\theta}$ *the* maximum likelihood estimator, when it is the unique maximum of $L(\theta)$,

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta).$$

Intuitively, a maximum likelihood estimator chooses that value of θ that is most compatible with our observation in the sense that *it makes what we observed most probable*. In not-so-mathematical terms, $\hat{\theta}$ is the value of θ that is most likely to have produced the data.

Comments on MLE's

Saw that MoMs and Plug-Ins often do not depend only on sufficient statistics.

$\hookrightarrow$ i.e. they also use "irrelevant" information

- If T is a sufficient statistic for θ then the Factorization theorem implies that

$$L(\theta) = g(T(\mathbf{X}); \theta)h(\mathbf{X}) \propto g(T(\mathbf{X}); \theta)$$

i.e. <u>any</u> MLE depends on data ONLY through the sufficient statistic

- MLE's are also invariant. If $g : \Theta \to \Theta'$ is a bijection, and if $\hat{\theta}$ is the MLE of θ, then $g(\hat{\theta})$ is the MLE of $g(\theta)$.

Comments on MLE's

- When the support of a distribution depends on a parameter, maximization is usually carried out by direct inspection.
- For a very broad class of statistical models, the likelihood can be maximized via differential calculus. If Θ is open, the support of the distribution does not depend on θ and the likelihood is differentiable, then the MLE satisfies the log-likelihood equations:

$$\nabla_\theta \log L(\theta) = 0$$

- Notice that maximizing $\log L(\theta)$ is equivalent to maximizing $L(\theta)$
- When Θ is not open, likelihood equations can be used, provided that we verify that the maximum does not occur on the boundary of Θ.

Example (Uniform Distribution)

Let $X_1, ..., X_n \overset{iid}{\sim} \mathcal{U}[0, \theta]$. The likelihood is

$$L(\theta) = \theta^{-n} \prod_{i=1}^{n} \mathbf{1}\{0 \leq X_i \leq \theta\} = \theta^{-n} \mathbf{1}\{\theta \geq X_{(n)}\}.$$

Hence if $\theta \leq X_{(n)}$ the likelihood is zero. In the domain $[X_{(n)}, \infty)$, the likelihood is a decreasing function of θ. Hence $\hat{\theta} = X_{(n)}$.

Example (Poisson Distribution)

Let $X_1, ..., X_n \overset{iid}{\sim}$ Poisson(λ). Then

$$L(\lambda) = \prod_{i=1}^{n} \left\{ \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} \right\} \implies \log L(\lambda) = -n\lambda + \log \lambda \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \log(x_i!)$$

Setting $\nabla_\lambda \log L(\lambda) = -n + \lambda^{-1} \sum x_i = 0$ we obtain $\hat{\lambda} = \bar{x}$ since $\nabla_\lambda^2 \log L(\lambda) = -\lambda^{-2} \sum x_i < 0$.