# 1 General problem

Model: $\{P_\theta : \theta \in \Theta\}$.
Observe $X \sim P_\theta, \theta \in \Theta$ unknown.
Estimate $\theta$. (Pick a plausible distribution from family. )
Or estimate $\tau = \tau(\theta)$.
Examples: $\theta = (\mu, \sigma^2), \tau(\theta) = \mu - \sigma$.
$\theta = (\beta_0, \beta_1, \sigma^2), \tau(\theta) = \beta_0 + \beta_1 z$.

# 2 Terminalogy

A **statistic** $W(X)$ is a function of the data $X$.
A **parameter** is a function $\tau(\theta)$. ("A" parameter $\tau(\theta)$ is a function of "the" parameter $\theta$.)
A **point estimator** of $\theta$ or $\tau(\theta)$ is a statistic $W(X)$ which is a single "value" (intended as an estimate of $\theta$ or $\tau(\theta)$).
We usually (but not always) require that a point estimator of $\theta$ should always belong to $\Theta$, and an estimator of $\tau(\theta)$ should always belong to $\tau(\Theta)$, the set of possible values of $\tau(\theta)$.
If the data $X$ is a random sample $(X_1, \ldots, X_n)$ from some population, these definitions correspond to those in elementary statistics:
A **statistic** is a characteristic of the sample.
A **parameter** is a characteristic of the population.
Notation: Point estimators of parameters $\theta$ or $\tau = \tau(\theta)$ are often designated $\hat{\theta} = \hat{\theta}(X)$ or $\hat{\tau} = \hat{\tau}(X)$.
**Examples of Parameters:**
Notation: $X = (X_1, \ldots, X_n)$ iid from the pdf (or pmf) $f(x \mid \theta)$.
$X$ is a single rv from $f(x \mid \theta)$.
For concreteness, think of $\theta = (\mu, \sigma^2)$ and $X \sim \mathrm{N}(\mu, \sigma^2)$.
Some parameters: $\tau(\theta) = \theta$
$\tau(\theta) = \mu$, or $\tau(\theta) = \mu^2$
$\tau(\theta) = \sigma^2$ or $\tau(\theta) = \sigma^4$
$\tau(\theta) = P_\theta(X \in A) = \int_A f(x \mid \theta) dx$
$\tau(\theta) = E_\theta X = \int x f(x \mid \theta) dx$ (general case)
$\tau(\theta) = E_\theta h(X) = \int h(x) f(x \mid \theta) dx$
$\tau(\theta) =$ median of $f(x \mid \theta)$.
$\tau(\theta) =$ interquartile range of $f(x \mid \theta)$.

$\tau(\theta) = 95^{th}$ percentile $f(x \mid \theta)$.
**Empirical Estimators:** It is often possible to estimate a population quantity by a natural sample analog.
**Examples:**

Table 1: default

| Parameter $\tau(\theta)$ | Estimate $\hat{\tau}(\mathbf{X})$ |
|---|---|
| $\underbrace{P_\theta(X \in A)}_{\text{(population proportion)}}$ | $\underbrace{n^{-1}\sum_{i=1}^{n} I(X_i \in A)}_{\text{sample proportion}}$ |
| $\underbrace{E_\theta X}_{\text{(population mean)}}$ | $\underbrace{n^{-1}\sum_{i=1}^{n} X_i}_{\text{(sample mean)}}$ |
| $\underbrace{E_\theta h(X)}_{\text{(population mean)}}$ | $\underbrace{n^{-1}\sum_{i=1}^{n} h(X_i)}_{\text{(sample mean)}}$ |
| population median | sample median |
| population IQR | sample IQR |
| population 95 th percentile | sample 95 th percentile |
| $\psi(F)$ | $\psi(\hat{F}_n)$ |

$F$ is the cdf of $f(x \mid \theta)$ (the population cdf) and $\hat{F}_n$ is the empirical cdf (defined later).

# 3 Intuitive approaches to estimation

## 3.1 Empirical Estimates (summary)

Estimate a population quantity by the natural sample analog. For example, estimate population mean by sample mean, population variance by sample variance, population quantile by sample quantile, etc.

## 3.2 Substitution principle (Plug-in Method)

Suppose $\alpha = \alpha(\theta)$ and $\beta = \beta(\theta)$ are two parameters related by $\alpha = h(\beta)$. If $\hat{\beta} = \hat{\beta}(\underset{\sim}{X})$ is a "reasonable" estimator of $\beta$, then $\hat{\alpha} = h(\hat{\beta})$ is a "reasonable" estimator of $\alpha$. More gener-

ally, if $\alpha, \beta_1, \beta_2, \ldots, \beta_k$ are parameters related by $\alpha = h(\beta_1, \beta_2, \ldots, \beta_k)$, and $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ are "reasonable" estimators of $\beta_1, \ldots, \beta_k$, then $\hat{\alpha} = h(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k)$ is a "reasonable" estimator of $\alpha$.
Example: $\tilde{X} = (X_1, X_2, \ldots X_n)$ iid $\mathrm{N}(\mu, \sigma^2), \theta = (\mu, \sigma^2)$. Estimate $\tau(\theta) = P(-1 < X < 1)$, where $X \sim \mathrm{N}(\mu, \sigma^2)$.

1. An empirical estimate: $\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} I(-1 < X_i < 1) =$ sample proportion.
2. A Plug-in estimate:

$$\begin{aligned}\tau(\theta) &= P(-1 < X < 1) = P\left(-\frac{1-\mu}{\sigma} < Z < \frac{1-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{1-\mu}{\sigma}\right) - \Phi(\frac{-1-\mu}{\sigma}\Big) \\ &= h(\mu, \sigma)\end{aligned}$$

where $\Phi$ is the cdf of $\mathrm{N}(0, 1)$.

Reasonable stimulates of $\mu, \sigma$ are

$$\hat{\mu} = \bar{x}, \quad \hat{\sigma} = s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

so that a plug-in estimate is given by

$$\hat{\tau} = h(\hat{\mu}, \hat{\sigma}) = \Phi\left(\frac{1-\bar{x}}{s}\right) - \Phi\left(\frac{-1-\bar{x}}{s}\right).$$

Which estimator is better? 1) or 2)? Intuition suggests 2) is better. Estimator 1) does not been use the assumption of normality. However, if it turns out that the normality assumption is false then 1) may end up giving the better estimate of $P(-1 < X < 1)$.
Example: $X_1, X_2, \ldots, X_n$ iid from a Cauchy location-scale family with pdf

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sigma} \cdot \frac{1}{\pi\left\{1 + \left(\frac{x-\mu}{\sigma}\right)^2\right\}}, \quad -\infty < x < \infty.$$

Estimate $\theta = (\mu, \sigma)$.
Note: This distribution does not have a finite mean. Thus $\bar{x}$ and $s^2$ are not useful here.
Useful Facts:

$$\begin{aligned}P(X < \mu) &= 0.5 \quad (\text{where } X \sim f(\cdot \mid \mu, \sigma)) \\ P(X < \mu - \sigma) &= 0.25 \\ P(X < \mu + \sigma) &= 0.75\end{aligned}$$

Notation: For $0 < p < 1$, let $\beta_p$ = population $p$th quantile, $Q_p$ = sample $p$th quantile.
Formal definitions: Let $F$ = population cdf: $F(t) = P(X \le t)$.
$\hat{F}$= sample cdf (empirical cdf): $\hat{F}(t) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le t)$.
Then

$$\begin{aligned} \beta_p &= \inf\{x : F(x) \ge p\} \\ Q_p &= \inf\{x : \hat{F}(x) \ge p\}. \end{aligned}$$

A "reasonable" estimate of $\beta_p$ is $\hat{\beta}_p = Q_p$. The "useful facts" say that for the Cauchy L-S family.

$$\beta_{0.5} = \mu, \quad \beta_{0.25} = \mu - \sigma, \quad \beta_{0.75} = \mu + \sigma$$

which implies

$$\mu = \beta_{0.5}, \quad \sigma = \frac{1}{2}(\beta_{0.75} - \beta_{0.25}).$$

Thus

$$\begin{aligned} \theta = (\mu, \sigma) &= h(\beta_{0.5}, \beta_{0.25}, \beta_{0.75}) \\ &= \left\{\beta_{0.5}, \frac{1}{2}(\beta_{0.75} - \beta_{0.25})\right\} \end{aligned}$$

so that a plug-in estimate is given by

$$\hat{\theta} = h(\hat{\beta}_{0.5}, \hat{\beta}_{0.25}, \hat{\beta}_{0.75}) = \left\{Q_{0.5}, \frac{1}{2}(Q_{0.75} - Q_{0.25})\right\}$$

# 4 Estimation by the Method of Moments (MOM)

MOM is basically fitting distribution by machining moments. MOM is a special case of the plug-in method.
Notation: $\mu_r = EX^r = r$th population moment ($\mu_r = \mu_r(\theta)$ is a parameter.)
$m_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r = r$th sample moment.
A "reasonable" estimate of $\mu_r$ is $\hat{\mu}_r = m_r$. Thus,
MOM: If parameter $\tau = \tau(\theta)$ can be written as a function of population moments $\tau = h(\mu_1, \mu_2, \ldots, \mu_k)$, then a "reasonable" estimate of $\tau$ is $\hat{\tau} = h(m_1, m_2, \ldots, m_k)$.

## 4.1 Parameter estimation by the Method of Moments

**Situation:** Suppose we have a model $X_1, X_2, \ldots, X_n$ iid $f(x \mid \theta)$, where $f(x \mid \theta)$ is the pdf (or pmf) of a family of distributions depending on a single parameter $\theta$. The value of $\theta$ is

unknown. We observe data $x_1, x_2, \ldots, x_n$. How do we estimate $\theta$?
**Notation:** Let $X$ denote a single observation from $f(x \mid \theta)$. Define

$$\begin{aligned} \mu &= \text{population mean} = EX \\ \hat{\mu} &= \bar{x} = \text{sample mean} = n \end{aligned}$$

Note that $\mu$ is a function of $\theta$, say $\mu = h(\theta)$.
**Method of Moments (MOM):** Estimate $\theta$ by that value $\hat{\theta}$ which makes the population mean $\mu$ equal to the sample mean $\bar{x}$.

**Formal Procedure:**

1. **Step 1:** Find $\mu$ as a function of $\theta$:

   $$\mu = EX = h(\theta).$$

   This is done either by looking up the family of distributions in the appendix or by doing the calculation

   $$EX = \int_{-\infty}^{\infty} x f(x \mid \theta), \quad \text{or} \quad \sum x f(x \mid \theta)$$

2. **Step 2:** Solve for $\theta$ as a function of $\mu$:

   $$\theta = g(\mu) \tag{1}$$

3. **Step 3:** Now plug in $\hat{\mu} = \bar{x}$ to obtain the MOM estimate:

   $$\hat{\theta} = g(\bar{x})$$

   Note: If $\mu$ does not depend on $\theta$ (for instance, if $\mu = 0$ for all $\theta$), then MOM is carried out using the second moment.
   Rationale: MOM works because the LLN guarantees that the sample mean $\bar{x}$ will be close to the population mean $\mu$ (with high probability) when the sample size $n$ is large. Since $g$ (in (5)) is a continuous function, $\bar{x} \approx \mu$ implies $g(\bar{x}) \approx g(\mu)$ which says that $\hat{\theta} \approx \theta$.

Example: MOM for Poisson($\lambda$) distribution

1. $\mu = EX = \sum_{x=0} x \frac{\lambda^x e^{-\lambda}}{x!} = \lambda$ Hence $\mu = \lambda$ ($\mu$ as a function of $\lambda$).
2. $\lambda = \mu$ (Solve for $\lambda$).
3. $\hat{\lambda} = \hat{\mu} = \bar{x}$ (Plug-in $\hat{\mu} = \bar{x}$ for $\mu$).

Conclusion: The MOM estimate of $\lambda$ is $\bar{x}$ ($\hat{\lambda} = \bar{x}$).

Example: Suppose you observe $X_1, X_2, \ldots, X_n \sim$ Geometric($p$). Find the MOM estimate of $p$.

1. Find $\mu$ as a function of $p$.

$$\mu = EX = \sum_{x=1}^{\infty} x \cdot p(1-p)^{x-1} = \frac{1}{p}$$

   $\mu = 1/p$.

2. Solve for $p$ as a function of $\mu$. $p = \frac{1}{\mu}$.

3. Plug in $\hat{\mu} = \bar{x}$ for $\mu$.

$$\hat{p} = \frac{1}{\hat{\mu}} = \frac{1}{\bar{x}}.$$

Conclusion: The MOM estimate of $p$ is $\frac{1}{\bar{x}}$ or $\hat{p} = \frac{1}{\bar{x}}$.

## 4.2 Parameter estimation by the Method of Moments in multi-parameter case

Suppose we have a model $X_1, X_2, \ldots, X_n$ iid $f(x \mid \theta)$, where $f(x \mid \theta)$ is the pdf (or pmf) of a family of distributions depending on a vector of parameters $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$. The vector of values $\theta$ is unknown. We observe data $x_1, x_2, \ldots, x_n$. How do we estimate $\theta$?
**Notation:** Let $X$ denote a single observation from $f(x \mid \theta)$. Define

$$\mu_k = (\text{population } k\text{-th moment}) = EX^k$$
$$\hat{\mu}_k = (\text{sample } k\text{-th moment}) = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

Note that $\mu_k$ is a function of $\theta$, say $\mu_k = h_k(\theta_1, \ldots, \theta_p)$.
**Method of Moments (MOM):** Estimate $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$ by those values $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_p)$ which make the population moments $(\mu_1, \mu_2, \ldots, \mu_p)$ equal to the sample moments $(\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_p) = (m_1, m_2, \ldots, m_p)$.
MOM for $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$

1. Find expressions for $\mu_1, \mu_2, \ldots, \mu_p$:

$$\begin{aligned} \mu_1 &= h_1(\theta_1, \ldots, \theta_p) \\ \mu_2 &= h_2(\theta_1, \ldots, \theta_p) \\ &\vdots \\ \mu_p &= h_p(\theta_1, \ldots, \theta_p) \end{aligned}$$

Look them up in appendix or evaluate using

$$\begin{aligned} \mu_k = EX^k &= \int_{-\infty}^{\infty} x^k f(x \mid \theta) dx \quad \text{(continuous)} \\ &= \sum x^k f(x \mid \theta) \quad \text{(discrete)} \end{aligned}$$

2. Solve this system of $p$ equations for $\theta_1, \theta_2, \ldots, \theta_p$:

$$\begin{aligned} \theta_1 &= g_1(\mu_1, \ldots, \mu_p) \\ \theta_2 &= g_2(\mu_1, \ldots, \mu_p) \\ &\cdots \\ \theta_p &= g_p(\mu_1, \ldots, \mu_p) \end{aligned}$$

3. Plug in $\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_p$ as estimates of $\mu_1, \ldots, \mu_p$:

$$\begin{aligned} \hat{\theta}_1 &= g_1(\hat{\mu}_1, \ldots, \hat{\mu}_p) \\ \hat{\theta}_2 &= g_2(\hat{\mu}_1, \ldots, \hat{\mu}_p) \\ &\cdots \\ \hat{\theta}_p &= g_p(\hat{\mu}_1, \ldots, \hat{\mu}_p) \end{aligned}$$

Special Case: $p = 2$

1. Find $\mu_1, \mu_2$:

$$\begin{aligned} \mu_1 &= h_1(\theta_1, \theta_2) \\ \mu_2 &= h_2(\theta_1, \theta_2). \end{aligned}$$

2. Solve for $\theta_1, \theta_2$:

$$\begin{aligned} \theta_1 &= g_1(\mu_1, \mu_2) \\ \theta_2 &= g_2(\mu_1, \mu_2). \end{aligned}$$

3. Plug in $\hat{\mu}_1, \hat{\mu}_2$ for $\mu_1, \mu_2$:

$$\begin{aligned} \hat{\theta}_1 &= g_1(\hat{\mu}_1, \hat{\mu}_2) \\ \hat{\theta}_2 &= g_2(\hat{\mu}_1, \hat{\mu}_2). \end{aligned}$$

## 5 Consistent Estimators

A sequence of estimators $W_n = W_n(X_1, X_2, \ldots, X_n)$ is a consistent sequence of estimators for the parameter $\tau = \tau(\theta)$ if, for every $\epsilon > 0$, and every $\theta \in \Theta$,

$$\lim_{n\to\infty} P_\theta(|W_n - \tau| < \epsilon) = 1 \tag{2}$$

The sequence is strongly consistent if we may replace (2) by

$$W_n \to \tau \quad \text{with probability 1 as } n \to \infty. \tag{3}$$

The sequence is consistent in 2nd mean (or in $L^2$) if we may replace (2) by

$$\lim_{n\to\infty} E_\theta(W_n - \tau)^2 = 0. \tag{4}$$

Let $X_1, X_2, X_3, \ldots,$ be iid.
Strong Law of Large Numbers: If $E|h(X_1)| < \infty$, then

$$\frac{1}{n}\sum_{i=1}^{n} h(X_i) \overset{wp1}{\to} Eh(X_1) \text{ as } n \to \infty.$$

Special case: If $\mu_r$ exists ( $E|X|^r < \infty$), then

$$m_r \overset{wp1}{\to} \mu_r.$$

Another Fact: Suppose the population $p$th quantile $\beta_p$ is unique (that is, there exists a unique value $x$ ( $= \beta_p$) such that $F(X) = p$, where $F$ is the population cdf), then $Q_p \overset{wp1}{\to} \beta_p$.

(Sections 5.5 and 10.1 discuss modes of convergence and consistency of estimates in greater detail.) The three types of consistency are (special cases of) 'convergence in probability', 'convergence almost surely, and 'convergence in $L_2$' (or in mean square), respectively.

Preservation of convergence by continuous functions:
If $W_n \to \tau$ in probability, and $g$ is a continuous function, then $g(W_n) \to g(\tau)$ in probability. Also true for functions of many variables:

If $U_n \to \xi$ and $W_n \to \tau$ in probability, and $g : \mathbb{R}^2 \to \mathbb{R}$ is a continuous function, then $g(U_n, W_n) \to g(\xi, \tau)$ in probability.
The previous facts remain true if "in probability" is everywhere replaced by "almost surely". Thus continuous functions of consistent estimates are consistent. As a consequence, it is typically true that estimators obtained by plug-in (substitution) are consistent. Method of Moments (MOM) estimators are consistent. Method of moments (MOM) estimators

are (typically) continuous functions of sample moments, which are consistent estimates of populations moments.

**Example:** If $X_1, X_2, X_3, \ldots$ are iid Geometric($p$), the MOM estimator of $p$ based on $X_1, \ldots, X_n$ is $1/\bar{x}_n$ where $\bar{x}_n = n^{-1}\sum_{i=1}^n X_i$.

WLLN implies $1/\bar{x}_n \to EX_1 = 1/p$ in probability (as $n \to \infty$). Thus $1/\bar{x}_n \to 1/(1/p) = p$ in probability. This holds for all $p \in (0, 1]$. Thus $1/\bar{x}_n$ is a consistent estimator of $p$.

Using the SLLN, the earlier statements remain true with in probability replaced by almost surely so that $1/\bar{x}_n$ is also a strongly consistent estimator of $p$.

What about consistency in 2nd mean (or in $L_2$)?

**Example:** Let $X_1, X_2, X_3, \ldots$ are iid N($\mu, \sigma^2$). The most commonly used estimate of $\sigma^2$ based on $X_1, \ldots, X_n$ is

$$s_n^2 = (n-1)^{-1}\sum_{i=1}^n (X_i - \bar{x}_n)^2$$

$s_n^2 \to \sigma^2$ in probability (for all $\mu$ and $\sigma^2$) $\to$ (*)

Thus, applying the continuous function $g(x) = \sqrt{x}$ to both sides: $s_n \to \sigma$ in probability (for all $\mu$ and $\sigma^2$). (These results don't require normality, but hold for any population with a finite second moment.)

**Proof of (*):** Show that $E(s_n^2 - \sigma^2)^2 = \mathrm{Var}(s_n^2) \to 0$. Alternatively, apply LLN to the identity:

$$s_n^2 = (n-1)^{-1}\left(\sum_{i=1}^n X_i^2 - n\bar{x}_n^2\right) = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{x}_n^2\right)$$

Example: MOM for Beta($\alpha, \beta$) distributions with pdf

$$f(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, 0 < x < 1$$

$\alpha, \beta > 0$, where $B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$. 2 parameters $\alpha, \beta$ implies to use two moments.

$$EX = \mu_1 = \frac{\alpha}{\alpha+\beta} \quad (5)$$

$$EX^2 = \mu_2 = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}. \quad (6)$$

Solve for $\alpha$ and $\beta$ in terms of $\mu_1$ and $\mu_2$.

$$R \equiv \frac{\mu_2}{\mu_1} = \frac{\alpha+1}{\alpha+\beta+1} = \frac{\frac{\alpha}{\alpha+\beta} + \frac{1}{\alpha+\beta}}{1 + \frac{1}{\alpha+\beta}}$$

$$= \frac{\mu_1 + \delta}{1+\delta}$$

where $\delta \equiv \frac{1}{\alpha+\beta}$.

$$\begin{aligned} R &= \frac{\mu_1+\delta}{1+\delta} \implies R + R\delta = \mu_1 + \delta \implies \\ R - \mu_1 &= \delta(1-R) \implies \delta = \frac{R-\mu_1}{1-R}. \end{aligned}$$

Note that

$$\begin{aligned} \mu_1 &= \frac{\alpha}{\alpha+\beta} = \alpha\delta \implies \alpha = \mu_1/\delta. \\ \beta &= (\alpha+\beta) - \alpha = \frac{1}{\delta} - \frac{\mu_1}{\delta} \implies \frac{(1-\mu_1)}{\delta} = \beta. \end{aligned}$$

Hence

$$\frac{1}{\delta} = \frac{1-R}{R-\mu_1} = \frac{1-\mu_2/\mu_1}{\mu_2/\mu_1 - \mu_1} = \frac{\mu_1-\mu_2}{\mu_2-\mu_1^2}.$$

In summary, $\alpha = \mu_1\xi, \beta = (1-\mu_1)\xi$ where

$$\xi \equiv \frac{1}{\delta} = \frac{\mu_1-\mu_2}{\mu_2-\mu_1^2}$$

so the MOM estimates are

$$\hat{\alpha} = m_1\hat{\xi}, \quad \hat{\beta} = (1-m_1)\hat{\xi}, \quad \hat{\xi} \equiv \frac{m_1-m_2}{m_2-m_1^2}.$$

# 6 Maximum Likelihood Estimation

Assume $\mathbf{X} \sim P_\theta, \theta \in \Theta$, with joint pdf (or pmf) $f(\mathbf{x} \mid \theta)$. Suppose we observe $\mathbf{X} = \mathbf{x}$. The Likelihood function is

$$L(\theta \mid \mathbf{x}) = f(\mathbf{x} \mid \theta)$$

as a function of $\theta$ (with the data $\mathbf{x}$ held fixed). The likelihood function $L(\theta \mid \mathbf{x})$ and joint pdf $f(\mathbf{x} \mid \theta)$ are the same except that $f(\mathbf{x} \mid \theta)$ is generally viewed as a function of $\mathbf{x}$ with $\theta$ held fixed, and $L(\theta \mid \mathbf{x})$ as a function of $\theta$ with $\mathbf{x}$ held fixed. $f(\mathbf{x} \mid \theta)$ is a density in $\mathbf{x}$ for each fixed $\theta$. But $L(\theta \mid \mathbf{x})$ is not a density (or mass function) in $\theta$ for fixed $\mathbf{x}$ (except by coincidence).

## 6.1 The Maximum Likelihood Estimator (MLE)

A point estimator $\hat{\theta} = \hat{\theta}(\mathbf{x})$ is a MLE for $\theta$ if

$$L(\hat{\theta} \mid \mathbf{x}) = \sup_{\theta} L(\theta \mid \mathbf{x}),$$

that is, $\hat{\theta}$ maximizes the likelihood. In most cases, the maximum is achieved at a unique value, and we can refer to "the" MLE, and write

$$\hat{\theta}(\mathbf{x}) = \text{argmax}_{\theta} L(\theta|\mathbf{x}).$$

(But there are cases where the likelihood has flat spots and the MLE is not unique.)

## 6.2 Motivation for MLE's

Note: We often write $L(\theta \mid \mathbf{x}) = L(\theta)$, suppressing $\mathbf{x}$, which is kept fixed at the observed data. Suppose $\mathbf{x} \in \mathbb{R}^n$.
Discrete Case: If $f(\cdot \mid \theta)$ is a mass function ($X$ is discrete), then

$$L(\theta) = f(\mathbf{x} \mid \theta) = P_{\theta}(\mathbf{X} = \mathbf{x}).$$

$L(\theta)$ is the probability of getting the observed data $\mathbf{x}$ when the parameter value is $\theta$.
Continuous Case: When $f(\cdot \mid \theta)$ is a continuous density $P_{\theta}(\mathbf{X} = \mathbf{x}) = 0$, but if $B \subset \mathbb{R}^n$ is a very, very small ball (or cube) centered at the observed data $\mathbf{x}$, then

$$P_{\theta}(\mathbf{X} \in B) \approx f(\mathbf{x} \mid \theta) \times \text{Volume}(B) \propto L(\theta).$$

$L(\theta)$ is proportional to the probability the random data $\mathbf{X}$ will be close to the observed data $\mathbf{x}$ when the parameter value is $\theta$. Thus, the MLE $\hat{\theta}$ is the value of $\theta$ which makes the observed data $\mathbf{x}$ "most probable".

To find $\hat{\theta}$, we maximize $L(\theta)$. This is usually done by calculus (finding a stationary point), but **not** always. If the parameter space $\Theta$ contains endpoints or boundary points, the maximum can be achieved at a boundary point without being a stationary point. If $L(\theta)$ is not "smooth" (continuous and everywhere differentiable), the maximum does not have to be achieved at a stationary point.
**Cautionary Example:** Suppose $X_1, \ldots, X_n$ are iid Uniform$(0, \theta)$ and $\Theta = (0, \infty)$. Given data $\mathbf{x} = (x_1, \ldots, x_n)$, find the MLE for $\theta$.

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} \theta^{-1} I(0 < x_i < \theta) = \theta^{-n} I(0 \leq x_{(1)}) I(x_{(n)} \leq \theta) \\ &= \begin{cases} \theta^{-n} & \text{for} \quad \theta \geq x_{(n)} \\ 0 & \text{for} \quad 0 < \theta < x_{(n)} \end{cases} \end{aligned}$$

which is maximized at $\theta = x_{(n)}$, which is a point of discontinuity and certainly not a stationary point. Thus, the MLE is $\hat{\theta} = x_{(n)}$.
Notes: $L(\theta) = 0$ for $\theta < x_{(n)}$ is just saying that these values of $\theta$ are absolutely ruled out by the data (which is obvious). A strange property of the MLE in this example (not typical):

$$P_\theta(\hat{\theta} < \theta) = 1$$

The MLE is biased; it is always less than the true value.
**A Similar Example:** Let $X_1, \ldots, X_n$ be iid Uniform$(\alpha, \beta)$ and $\Theta = \{(\alpha, \beta) : \alpha < \beta\}$. Given data $\mathbf{x} = (x_1, \ldots, x_n)$, find the MLE for $\theta = (\alpha, \beta)$.

$$\begin{aligned} L(\alpha, \beta) &= \prod_{i=1}^{n} (\beta - \alpha)^{-1} I(\alpha < x_i < \beta) = (\beta - \alpha)^{-n} I(\alpha \leq x_{(1)}) I(x_{(n)} \leq \beta) \\ &= \begin{cases} (\beta - \alpha)^{-n} & \text{for} \quad \alpha \leq x_{(1)}, x_{(n)} \leq \beta \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

which is maximized by making $\beta - \alpha$ as small as possible without entering "0 otherwise" region. Clearly, the maximum is achieved at $(\alpha, \beta) = (x_{(1)}, x_{(n)})$. Thus the MLE is $\theta = (\hat{\alpha}, \hat{\beta}) = (x_{(1)}, x_{(n)})$. Again, $P_{\alpha,\beta}(\alpha < \hat{\alpha}, \hat{\beta} < \beta) = 1$.