Estimation (Chapter 7)

General Problem:

Model: $\{P_\theta : \theta \in \Theta\}$

Observe $X \sim P_\theta$, $\theta \in \Theta$, θ unknown.

Estimate θ. (Pick a plausible distn from family.)

Or estimate $\tau = \tau(\theta)$.

Example: $\theta = (\mu, \sigma^2)$, $\tau(\theta) = \mu - \sigma$.

Terminology

A **point estimator** is a statistic $W(X)$ (a function of the data). We usually (but not always) require that $W(X) \in \Theta$.

A **parameter** is a function $\tau(\theta)$. ("A" parameter $\tau(\theta)$ is a function of "the" parameter θ.)

If the data X is a random sample $(X_1, \ldots, X_n)$, 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 θ or $\tau = \tau(\theta)$ are often designated $\widehat{\theta} = \widehat{\theta}(X)$ or $\widehat{\tau} = \widehat{\tau}(X)$.

Examples of Parameters:

Notation:

$\boldsymbol{X} = (X_1, \ldots, X_n)$ 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 $X \sim N(\mu, \sigma^2)$.

Some parameters:

$\tau(\theta) = \theta$

$\tau(\theta) = \mu$ or $\tau(\theta) = \mu^2$

$\tau(\theta) = \sigma^2$ or $\tau(\theta) = \sigma^4$

$\tau(\theta) = P_\theta(X \in A) = \int_A f(x \mid \theta)$

$\tau(\theta) = E_\theta X = \int x f(x \mid \theta)\, dx$ (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 $f(x \mid \theta)$

$\tau(\theta) = 95^{\text{th}}$ percentile of $f(x \mid \theta)$

Empirical Estimators:

It is often possible to estimate a population quantity by a natural sample analog.

Examples:

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

†: F is the cdf of $f(x\,|\,\theta)$ (the population cdf) and $\widehat{F}_n$ is the empirical cdf (defined later).

Intuitive approaches to estimation

Empirical Estimates (summary)

Estimate a population quantity by the natural sample analog.

For example, estimate

population mean by sample mean,
pop. variance by sample variance,
pop. quantile by sample quantile,
etc.

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 β, then $\hat{\alpha} = h(\hat{\beta})$ is a "reasonable" estimator of α.

More generally, 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 α.

Example:

$$\underbrace{X_1, X_2, \ldots, X_n}_{\underset{\sim}{X}} \text{ iid } N(\underbrace{\mu, \sigma^2}_{\theta})$$

Estimate $\tau(\theta) = P(-1 < X < 1)$

where $X \sim N(\mu, \sigma^2)$.

(A) An empirical estimate

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} I(-1 < X_i < 1)$$

= the sample proportion

$$Z = \frac{X-\mu}{\sigma}$$

(B) A Plug-in estimate

$$\tau(\theta) = P(-1 < X < 1) = P\left(\frac{-1-\mu}{\sigma} < Z < \frac{1-\mu}{\sigma}\right)$$

$$= \Phi\left(\frac{+1-\mu}{\sigma}\right) - \Phi\left(\frac{-1-\mu}{\sigma}\right)$$

where Φ is cdf of $N(0,1)$

$$= h(\mu, \sigma)$$

Reasonable estimates of μ, σ 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? (A) or (B)?

Intuition suggests (B) is better.
Estimator (A) does <u>not</u> even use the assumption of normality.
However, if it turns out that the normality assumption is <u>false</u>, then (A) may end up giving the better estimate of $P(-1<X<1)$.

<u>Example</u>: $X_1, X_2, \ldots, X_n$ iid from a Cauchy location-scale family with

pdf $f(x|\underbrace{\mu,\sigma}_{\theta}) = \frac{1}{\sigma} \cdot \frac{1}{\pi\left(1+\left(\frac{x-\mu}{\sigma}\right)^2\right)}$

for $-\infty < x < \infty$.

Estimate $\theta = (\mu, \sigma)$.

Note: This distribution does <u>not</u> have a finite mean. Thus $\bar{x}$ and s^2 are <u>not</u> useful here.

Useful Facts:

$P(X<\mu) = .5$ (where $X \sim f(\cdot|\mu,\sigma)$)

$P(X<\mu-\sigma) = .25$

$P(X<\mu+\sigma) = .75$

Notation For $0<p<1$,

Let β_p = population p^{th} quantile,

Q_p = sample p^{th} quantile.

Formal definitions:

Let F = pop. cdf : $F(t) = P(X \le t)$

$\hat{F}$ = sample cdf : $\hat{F}(t) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le t)$ (empirical cdf)

Then

$\beta_p = \inf\{x : F(x) \ge p\}$

$Q_p = \inf\{x : \hat{F}(x) \ge p\}$.

A "reasonable" estimate of β_p is $\hat{\beta}_p = Q_p$.

The "useful facts" say that for the Cauchy L-S family

$$\left.\begin{aligned} \beta_{.5} &= \mu \\ \beta_{.25} &= \mu - \sigma \\ \beta_{.75} &= \mu + \sigma \end{aligned}\right\} \Rightarrow \begin{aligned} \mu &= \beta_{.5} \\ \sigma &= \tfrac{1}{2}(\beta_{.75} - \beta_{.25}) \end{aligned}$$

Thus

$$\theta = (\mu, \sigma) = h(\beta_{.5}, \beta_{.25}, \beta_{.75}) = (\beta_{.5}, \tfrac{1}{2}(\beta_{.75} - \beta_{.25}))$$

so that a plug-in estimate is given by

$$\hat{\theta} = h(\hat{\beta}_{.5}, \hat{\beta}_{.25}, \hat{\beta}_{.75}) = (Q_{.5}, \tfrac{1}{2}(Q_{.75} - Q_{.25})).$$

Estimation by the Method of Moments (MOM)

(Fitting distns by matching moments)

MOM is a special case of the Plug-in method.

Notation: (delete primes used in text)

$\mu_r = E X^r = r^{th}$ 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 μ_r is $\hat{\mu}_r = m_r$.

Thus ...

MOM: If parameter $\tau = \tau(\theta)$ can be written as a function of pop. moments

$$\tau = h(\mu_1, \mu_2, \ldots, \mu_K),$$

then a "reasonable" estimate of τ is

$$\hat{\tau} = h(m_1, m_2, \ldots, m_K).$$

Parameter estimation by the Method of Moments

Situation:

Suppose we have a model

$$X_1, X_2, \ldots, X_n \text{ iid } f(x|\theta)$$

where $f(x|\theta)$ is the pdf (or pmf) of a family of distributions depending on a single parameter θ.

The value of θ is unknown.

We observe data $x_1, x_2, \ldots, x_n$.

How do we estimate θ?

Notation:

Let X denote a single observation from $f(x|\theta)$.

Define

$$\mu = \text{population mean} = EX$$

$$\hat{\mu} = \bar{x} = \text{sample mean} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Note that μ is a function of θ, say $\mu = g(\theta)$.

Method of Moments (MOM):

Estimate θ by that value $\hat{\theta}$ which makes the population mean μ equal to the sample mean $\bar{x}$.

Formal Procedure:

Step 1: Find μ as a function of θ:

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

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

$$EX = \int_{-\infty}^{\infty} x f(x|\theta)\,dx \quad \text{or} \quad \sum_{\text{all } x} x f(x|\theta)\,.$$

Step 2: Solve for θ as a function of μ:

$$\theta = g(\mu)\,. \qquad (\dagger)$$

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

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

Note: If μ does not depend on θ (for instance, if $\mu = 0$ for all θ), 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 μ (with high probability) when the sample size n is large.

Since g (in $\dagger$) 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(λ) distribution

① $\mu = EX = \underbrace{\sum_{x=0}^{\infty} x \cdot \frac{\lambda^x e^{-\lambda}}{x!}}_{\text{not needed in this case}} = \lambda$ ← from appendix

$\mu = \lambda$ (μ as function of λ)

② $\lambda = \mu$ (Solve for λ)

③ $\hat{\lambda} = \hat{\mu} = \bar{x}$ (plug in $\hat{\mu} = \bar{x}$ for μ)

Conclusion: The MOM estimate of λ is $\bar{x}$.

($\hat{\lambda} = \bar{x}$)

Example: Suppose you observe

$X_1, X_2, \ldots, X_n$ iid Geometric(p).

Find the MOM estimate of p.

---

① Find μ as a function of p.

$$\mu = EX = \underbrace{\sum_{x=1}^{\infty} x \cdot p(1-p)^{x-1}}_{\text{not needed in this case}} = \frac{1}{p}$$

↑ from apperdix

$$\mu = \frac{1}{p}$$

② Solve for p as a function of μ.

$$p = \frac{1}{\mu}$$

③ Plug in $\hat{\mu} = \bar{x}$ for μ

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

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

$$\left(\hat{p} = \frac{1}{\bar{x}}\right)$$

Parameter estimation by the Method of Moments

Situation: (multi-parameter case)

Suppose we have a model

$$X_1, X_2, \ldots, X_n \text{ iid } f(x|\theta)$$

where $f(x|\theta)$ 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 θ is unknown.

We observe data $x_1, x_2, \ldots, x_n$.

How do we estimate θ?

Notation:

Let X denote a single observation from $f(x|\theta)$.

Define

$$\mu_k = \text{(population } k\text{-th moment)} = EX^k$$

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

(Special case: $\mu = \mu_1$ and $\bar{x} = \hat{\mu} = \hat{\mu}_1$.)

Note that μ_k is a function of θ, 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 $(\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_p)$.

MOM for $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$

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

$$\mu_1 = h_1(\theta_1, \ldots, \theta_p)$$
$$\mu_2 = h_2(\theta_1, \ldots, \theta_p)$$
$$\vdots$$
$$\mu_p = h_p(\theta_1, \ldots, \theta_p)$$

Look them up in appendix or evaluate using $\mu_k = EX^k = \int_{-\infty}^{\infty} x^k f(x|\theta)\,dx$ (continuous)

or $\sum_{\text{all } x} x^k f(x|\theta)$ (discrete)

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

$$\theta_1 = g_1(\mu_1, \ldots, \mu_p)$$
$$\theta_2 = g_2(\mu_1, \ldots, \mu_p)$$
$$\vdots$$
$$\theta_p = g_p(\mu_1, \ldots, \mu_p)$$

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

$$\hat{\theta}_1 = g_1(\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_p)$$
$$\hat{\theta}_2 = g_2(\hat{\mu}_1, \ldots, \hat{\mu}_p)$$
$$\vdots$$
$$\hat{\theta}_p = g_p(\hat{\mu}_1, \ldots, \hat{\mu}_p).$$

Special Case: $p=2$

① Find μ_1, μ_2:

$$\mu_1 = h_1(\theta_1, \theta_2)$$
$$\mu_2 = h_2(\theta_1, \theta_2)$$

② Solve for θ_1, θ_2:

$$\theta_1 = g_1(\mu_1, \mu_2)$$
$$\theta_2 = g_2(\mu_1, \mu_2)$$

③ Plug in $\hat{\mu}_1, \hat{\mu}_2$ for μ_1, μ_2:

$$\hat{\theta}_1 = g_1(\hat{\mu}_1, \hat{\mu}_2)$$
$$\hat{\theta}_2 = g_2(\hat{\mu}_1, \hat{\mu}_2)$$

Consistent Estimators

A sequence of estimators

$$W_n = W_n(X_1, X_2, \ldots, X_n)$$

is a <u>consistent</u> sequence of estimators for the parameter $\tau = \tau(\theta)$ if, for every $\epsilon > 0$ and every $\theta \in \Theta$,

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

The sequence is <u>strongly consistent</u> if we may replace (*) by

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

The sequence is <u>consistent in 2nd mean</u> (or in L^2) if we may replace (*) by

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

Let $X_1, X_2, X_3, \ldots$ be iid.

Strong Law of Large Numbers

If $E|h(X_1)| < \infty$, then

$$\frac{1}{n}\sum_{i=1}^{n} h(X_i) \xrightarrow{wp1} E\,h(X_1)$$

as $n \to \infty$.

Special Case:

If μ_r exists ($E|X|^r < \infty$),

then $m_r \xrightarrow{wp1} \mu_r$.

Another Fact:

Suppose the population p^{th} quantile β_p is unique (that is, there exists a unique value x ($=\beta_p$) such that $F(x) = p$ where F is the pop. cdf),

then $Q_p \xrightarrow{wp1} \beta_p$.

(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 $\bar{x}_n \to EX_1 = 1/p$ in probability (as $n \to \infty$).

Thus $1/\bar{x}_n \to 1/(1/p) = p$ in probability.

This holds for all p. 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 σ^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 μ and σ^2). (†)

Thus, applying the continuous function $g(x) = \sqrt{x}$ to both sides: $s_n \to \sigma$ in probability (for all μ and σ^2).

(These results don't require normality, but hold for any population with a finite second moment.)

Proof of (†):

Show that $E(s_n^2 - \sigma^2)^2 = \text{Var}(s_n^2) \to 0$.

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(α, β) distn with pdf $f(x) = \dfrac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$, $0<x<1$ $(\alpha, \beta > 0)$

where $B(\alpha,\beta) = \dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.

---

2 params $\alpha, \beta \Longrightarrow$ use two moments.

$$EX = \mu_1 = \frac{\alpha}{\alpha+\beta}$$

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

} Solve for α, β in terms of μ_1, μ_2

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

$$= \frac{\mu_1 + \delta}{1+\delta} \quad \text{where} \quad \delta \equiv \frac{1}{\alpha+\beta}$$

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

$$\Longrightarrow R - \mu_1 = \delta(1-R) \Longrightarrow \delta = \frac{R-\mu_1}{1-R}$$

Note that $\mu_1 = \frac{\alpha}{\alpha+\beta} = \alpha\delta$

$$\Rightarrow \boxed{\alpha = \mu_1/\delta}$$

$$\beta = (\alpha+\beta) - \alpha = \frac{1}{\delta} - \frac{\mu_1}{\delta}$$

$$= \boxed{\frac{1}{\delta}(1-\mu_1) = \beta}$$

$$\frac{1}{\delta} = \frac{1-R}{R-\mu_1} = \frac{1-(\mu_2/\mu_1)}{\frac{\mu_2}{\mu_1} - \mu_1}$$

$$= \frac{\mu_1 - \mu_2}{\mu_2 - \mu_1^2}$$

In summary:

$$\alpha = \mu_1 \xi, \quad \beta = (1-\mu_1)\xi$$

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

so the MOM estimates are

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