3. Estimation

data $x_1, \ldots, x_n$; also use $\mathbf{x}$

assume realizations of random variables $X_1, \ldots, X_n$ with joint p.d.f. $f(x_1, \ldots, x_n; \theta)$

Examples

1. i.i.d. Poisson (θ)

2. i.i.d. $\mathcal{N}(\mu, \sigma^2)$, here $\theta = (\mu, \sigma^2)$

3. Have n cells; X_j is the number of balls that land in cell j, then have Multinomial$(n, p_1, \ldots, p_n)$; with $\sum_{j=1}^{n} p_j = 1$

A function $t = t(\mathbf{x})$ of the observations but not of the parameter is called a *statistic*

If t is a statistic then

$$T = t(\mathbf{X}) = t(X_1, \ldots, X_n)$$

is called an *estimator* - random

$t(\mathbf{x})$, when used for estimation, is also called an *estimate* - not a random variable

Example $t(\mathbf{x}) = \bar{x}$ estimate, $\bar{X}$ estimator

Interval estimation

Goal: Find $a(\mathbf{X}), b(\mathbf{X})$ such that

$\mathbf{P}(a(\mathbf{X}) < \theta < b(\mathbf{X}))$ is large and $b(\mathbf{X}) - a(\mathbf{X})$ is small

If

$$\mathbf{P}(a(\mathbf{X}) < \theta < b(\mathbf{X})) = 1 - \alpha$$

then $(a(\mathbf{x}), b(\mathbf{x}))$ is a $100(1-\alpha)\%$ *confidence interval* for θ

Note: $\mathbf{X}$ is random, θ is not random!

Example:

$X_1, \ldots, X_n$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

distribution is independent of θ, so

$$\mathbf{P}\left(-t_{n-1}(\alpha/2) < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{n-1}(\alpha/2)\right) = 1 - \alpha,$$

and so

$$\begin{aligned} \mathbf{P}\Bigg\{\bar{X} - t_{n-1}(\alpha/2)\frac{S}{\sqrt{n}} &< \mu \\ &< \bar{X} + t_{n-1}(\alpha/2)\frac{S}{\sqrt{n}}\Bigg\} \\ = \quad & 1 - \alpha \end{aligned}$$

Choose

$$a(\mathbf{X}) = \bar{X} - t_{n-1}(\alpha/2)\frac{S}{\sqrt{n}}$$

$$b(\mathbf{X}) = \bar{X} + t_{n-1}(\alpha/2)\frac{S}{\sqrt{n}}$$

Substitute $\mathbf{X}$ by $\mathbf{x}$: $100(1-\alpha)\%$ confidence interval for μ

A function $g(\mathbf{x}, \theta)$ of both sample and parameter whose distribution does not depend on the parameter is called a *pivot*

In above example: $g(\mathbf{x}, (\mu, \sigma^2)) = \frac{\bar{x}-\mu}{s/\sqrt{n}}$

Above example:

$X_1, \ldots, X_n$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} :$$

pivot!

Can find c_1, c_2 such that distribution is independent of θ, so

$$\mathbf{P}\left(c_1 < \frac{(n-1)S^2}{\sigma^2} < c_2\right) = 1 - \alpha,$$

and so $100(1-\alpha)\%$ confidence interval for σ^2 is

$$\left(\frac{(n-1)s^2}{c_2}, \frac{(n-1)s^2}{c_1}\right)$$

equal tail convention:

$$\mathbf{P}(\chi^2_{n-1} < c_1) = \mathbf{P}(\chi^2_{n-1} > c_2)$$

Example: kitchen timer

Recorded times

$$293.7, 296.2, 296.4, 294.0, 297.3,$$

$$293.7, 294.3, 291.3, 295.1, 296.1$$

$\bar{x} = 294.81, \quad s^2 = 3.1232$

90% c.i. for μ: $t_0(.95) = 1.833$ from table, so

$$(293.8, 295.8)$$

(does not contain 300)

90% c.i. for σ^2: $c_1 = 3.325, c_2 = 28.109$ from table, so

$$(1.66, 8.45)$$

If $\sigma^2 = 8.45$: 90% c.i. for μ would be $(293.0, 296.6)$, still does not contain 300

Example: Two samples $X_1, \ldots, X_n$ i.i.d. $\mathcal{N}(\mu_1, \sigma^2)$, $Y_1, \ldots, Y_m$ i.i.d. $\mathcal{N}(\mu_2, \sigma^2)$, independent; then

$$\bar{X} - \bar{Y} \sim \mathcal{N}\left(\mu_1 - \mu_2, \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)\right)$$

and

$$\frac{(n-1)S_X^2 + (m-1)S_Y^2}{\sigma^2} \sim \chi^2_{n+m-2}$$

put

$$S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$$

then

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}$$

(and independent of $\bar{X} - \bar{Y}$), so

$$\frac{\bar{x} - \bar{y} - (\mu_1 - \mu_2)}{s\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

is a pivot

Example: Fastest recorded driving speed

males: $n = 17, \bar{x} = 102.1, s_X^2 = 17.05, s_X = 4.13$

females $m = 21, \bar{y} = 85.7, s_Y^2 = 9.39, s_Y = 3.06$

$\bar{x} - \bar{y} = 16.35,\ s_p^2 = 13.5,\ n + m - 2 = 36$

95 % confidence interval for $\mu_1 - \mu_2$ is $(7.5, 25.2)$, does not contain 0

Same variance? Can test for it.

Rule of thumb: Assume equal variances if neither sample standard deviation is more than twice that of the other standard deviation.

Pivot not always available.

Point estimation

Put $\widehat{\theta} = T(\mathbf{X})$ estimator; *Criteria:*

- Unbiased: $\mathbf{E}_\theta(\widehat{\theta}) = \int \widehat{\theta} f(\mathbf{x}, \theta) d\mathbf{x} = \theta$

  (in discrete case replace integral by sum)

  The *bias* is defined as

  $$bias(\widehat{\theta}) = E_\theta(\widehat{\theta} - \theta)$$

  (Often omit subscript θ)

- small variance

- related to a known distribution

The Methods of Moments (M.O.M.)

Assume $X_1, \ldots, X_n$ i.i.d. with p.d.f. $f(x, \theta)$

1. For a function h put

$$\bar{H} = \frac{1}{n} \sum_{i=1}^{n} h(X_i)$$

2. Then

$$\begin{aligned} E\bar{H} &= \frac{1}{n} \sum_{i=1}^{n} Eh(X_i) \\ &= \int h(x) f(x, \theta) dx =: k(\theta) \end{aligned}$$

(in discrete case replace integral by sum)

3. Solve

$$k(\theta) = \bar{H}$$

for θ; the solution $\tilde{\theta}$ is called the *method of moments estimator* (m.o.m. estimator)

Example

Random sample from $U(0, \theta)$

1. $h(x) = x$, $\bar{H} = \bar{X}$

2. $k(\theta) = E(X) = \frac{\theta}{2}$

3. Solve equation:

$$\tilde{\theta} = 2\bar{X}$$

Example: $\mathcal{N}(\mu, \sigma^2)$ with $\theta = (\mu, \sigma^2)$

1. $h_1(x) = x, h_2(x) = x^2$

2. $E\bar{H}_1 = \mu, E\bar{H}_2 = \mu^2 + \sigma^2$

3. Solve equations:

$$\tilde{\mu} = \bar{X}$$

$$\tilde{\mu}^2 + \tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

so

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Note that $\mathbf{E}\tilde{\sigma}^2 = \frac{n-1}{n}\sigma^2$, so this is not unbiased

Often: $h(x) = x^k$ for some k; $\mu_k = EX^k$ is called the *kth moment*

Maximum-Likelihood Estimators (M.L.E.)

Recall that the *likelihood* of the parameter θ given $\mathbf{x}$ is

$$L(\theta; \mathbf{x}) = f(x_1, \ldots, x_n; \theta)$$

Maximum-likelihood-estimator: $\widehat{\theta}$ such that $L(\theta, \mathbf{x})$ is maximized

In random sample of size n:

$$L(\theta, \mathbf{x}) = \prod_{i=1}^{n} f(x_i, \theta)$$

Example: $U([0,\theta])$ again

$$f(x,\theta) = \frac{1}{\theta}\mathbf{1}(0 \le x \le \theta)$$

$$L(\theta;\mathbf{x}) = \theta^{-n}\prod_{i=1}^{n}\mathbf{1}(0 \le x_i \le \theta)$$

so

$$L(\theta;\mathbf{x}) = \theta^{-n}\mathbf{1}(0 \le \min_i x_i; \max_i x_i \le \theta)$$

increases in $\max_i x_i$, so $\widehat{\theta} = \max_i x_i$

Variant: $U(0,\theta)$

$$f(x,\theta) = \frac{1}{\theta}\mathbf{1}(0 < x < \theta)$$

Note: $\widehat{\theta} = \max_i x_i$ but $L(\widehat{\theta},\mathbf{x}) = 0$, so m.l.e. does not exist

Variant: $U([\theta - \frac{1}{2}, \theta + \frac{1}{2}])$;

$$f(x, \theta) = \mathbf{1}(\theta - 1/2 \leq x \leq \theta + 1/2)$$

$$\begin{aligned} L(\theta; \mathbf{x}) &= \prod_{i=1}^{n} \mathbf{1}(\theta - 1/2 \leq x_i \leq \theta + 1/2) \\ &= \mathbf{1}(\theta \leq \min_i x_i + 1/2; \theta > \max_i x_i - 1/2) \end{aligned}$$

is maximal for *any*

$$\theta \in [\min_i x_i + 1/2; \max_i x_i - 1/2)]$$

So m.l.e. not unique

Special case: exponential family

An exponential family has a p.d.f. of the form

$$f(x,\theta) = e^{\sum_{i=1}^{k} p_i(\theta)h_i(x) - K(\theta) + b(x)}, \quad x \in A,$$

where A does not depend on θ

The quantities $\eta_i = p_i(\theta)$ are called the *natural parameters* of the family

Examples:

1. $Be(\theta)$: $f(x,\theta) = \theta^x(1-\theta)^{1-x}$, $x = 0, 1$

$$\begin{aligned} f(x,\theta) &= \theta^x(1-\theta)^{1-x} \\ &= \left(\frac{\theta}{1-\theta}\right)^x (1-\theta) \\ &= \exp\left\{x \log\left(\frac{\theta}{1-\theta}\right) + \log(1-\theta)\right\} \end{aligned}$$

choose $k = 1$,

$$b(x) = 0, h(x) = x, K(\theta) = -\log(1-\theta),$$

$p(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$

2. Exp(θ): $f(x,\theta) = \theta e^{-\theta x}$ for $x > 0$; choose $b(x) = 0, h(x) = -x, K(\theta) = -\log\theta$

3. χ^2_p, when p is considered as a continuous function

Also: Poisson, Normal, and many more

Not: Cauchy, F, t: cannot be written in required form

Not: $U(0, \theta)$: support depends on parameters

Fact: If X has an exponential family distribution, so does any one-to-one function of X

For $X_1, \ldots, X_n$ i.i.d. from 1-dim exponential family, θ natural parameter:

$$L(\theta, \mathbf{x}) = e^{\theta \sum h(x_j) - nK(\theta) + \sum b(x_j)}$$

so

$$l(\theta) = \log L(\theta) = \theta \sum h(x_j) - nK(\theta) + \sum b(x_j)$$

and

$$l'(\theta) = \sum h(x_j) - nk(\theta)$$

with $k(\theta) = K'(\theta)$; put equal to zero: shows that the m.l.e. in this case is also a m.o.m.

Fact: If the distribution comes from an exponential family and if it has a m.l.e. then the m.l.e. is unique.

Invariance: Let η be a function, then the m.l.e. of $\eta(\theta)$ is $\eta(\widehat{\theta})$.

Proof: Let

$$L^*(\tau, \mathbf{x}) = \sup_{\{\theta:\eta(\theta)=\tau\}} L(\theta, \mathbf{x})$$

then the m.l.e. of $\eta(\theta)$ is the value τ^* that maximizes L^*

The maxima of L^* and L coincide:

$$\begin{aligned} L^*(\tau^*, \mathbf{x}) &= \sup_{\tau} \sup_{\{\theta:\eta(\theta)=\tau\}} L(\theta, \mathbf{x}) \\ &= \sup_{\theta} L(\theta, \mathbf{x}) \\ &= L(\widehat{\theta}, \mathbf{x}) \end{aligned}$$

Furthermore

$$\begin{aligned} L(\widehat{\theta}, \mathbf{x}) &= \sup_{\{\theta:\eta(\theta)=\eta(\widehat{\theta})\}} L(\theta, \mathbf{x}) \\ &= L^*(\eta(\widehat{\theta}), \mathbf{x}) \end{aligned}$$

This finishes the proof.

Possible problems with m.l.e.:

- The m.l.e. may not exist

- It may not be unique

- It may not have a closed form expression.

Iteration procedure for m.l.e.

Assume $\theta \in \mathbf{R}$

Newton-Raphson: Let

$$U = -\frac{\partial l}{\partial \theta}$$

(*score function*) and

$$J = -\frac{\partial^2 l}{\partial \theta^2}$$

(*observed information*) then $\widehat{\theta}$ solves $U(\theta) = 0$

Taylor:

$$U(\widehat{\theta}) \approx U(\theta^*) - J(\theta^*)(\widehat{\theta} - \theta^*)$$

Hence

$$\widehat{\theta} \approx \theta^* + J^{-1}(\theta^*)U(\theta^*)$$

Given an initial guess $\theta^{(0)}$ for $\widehat{\theta}$, e.g. obtained by m.o.m., update estimate by

$$\widehat{\theta}^{(k+1)} \approx \widehat{\theta}^{(k)} + J^{-1}(\widehat{\theta}^{(k)})U(\widehat{\theta}^{(k)})$$

Terminate when

$$\| \widehat{\theta}^{(k+1)} - \widehat{\theta}^{(k)} \| < \epsilon$$

For higher-dimensional θ: use gradient vector $\underline{U}$, Hessian matrix J:

$$\underline{\widehat{\theta}}^{(k+1)} \approx \underline{\widehat{\theta}}^{(k)} + J^{-1}(\underline{\widehat{\theta}}^{(k)})\underline{U}(\underline{\widehat{\theta}}^{(k)})$$

Example: $Binomial(n, \theta)$, observe x

$$
\begin{aligned}
l(\theta) &= x \ln(\theta) + (n-x) \ln(1-\theta) + \log \binom{n}{x} \\
U(\theta) &= -\frac{x}{\theta} + \frac{n-x}{1-\theta} \\
J(\theta) &= \frac{x}{\theta^2} + \frac{n-x}{(1-\theta)^2}
\end{aligned}
$$

Assume $n = 5, x = 2, \epsilon = 0.01$ (in practice rather $\epsilon = 10^{-5}$); guess $\hat{\theta}^{(0)} = 0.55$

$$
\begin{aligned}
U(\hat{\theta}^{(0)}) &\approx 3.03 \\
\hat{\theta}^{(1)} &\approx \hat{\theta}^{(0)} + J^{-1}(\hat{\theta}^{(0)}) U(\hat{\theta}^{(0)}) \approx 0.40857 \\
U(\hat{\theta}^{(1)}) &\approx 0.1774 \\
\hat{\theta}^{(2)} &\approx \hat{\theta}^{(1)} + J^{-1}(\hat{\theta}^{(1)}) U(\hat{\theta}^{(1)}) \approx 0.39994
\end{aligned}
$$

Now $|\hat{\theta}^{(2)} - \hat{\theta}^{(1)}| < 0.01$ so stop

Compare: analytically, $\hat{\theta} = \frac{x}{n} = 0.4$

Modification: *Fisher Scoring Method*

Replace $J(\widehat{\theta}^{(k)})$ by $I(\widehat{\theta}^{(k)})$, where

$$I(\theta) = E_\theta \left\{ -\frac{\partial^2 l(\theta, \mathbf{x})}{\partial^2 \theta} \right\}$$

Large sample distribution

Assume here $X_1, X_2, \ldots$ i.i.d.

Recall CLT: If $X_1, X_2, \ldots$ i.i.d. mean μ, variance σ^2, then in distribution

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \to \mathcal{N}(0, 1)$$

M.o.m. asymptotics: $\widehat{\theta}$ solution of $\bar{H} = k(\theta)$

Theorem 1 *Assume that $k''(\theta)$ exists. In large samples, in distribution*

$$\widehat{\theta} \approx \mathcal{N}\left(\theta, \frac{\sigma_h^2(\theta)}{n(k'(\theta))^2}\right),$$

where

$$\sigma_h^2(\theta) = \int (h(x))^2 f(x, \theta) dx - (k(\theta))^2$$

Proof. *Recall* $\bar{H} = \frac{1}{n}\sum_{i=1}^{n} h(X_i)$, *and* $\mathbf{E}\bar{H} = k(\theta)$

CLT yields, in distribution

$$\frac{\bar{H} - k(\theta)}{\sigma_h(\theta)/\sqrt{n}} \to \mathcal{N}(0,1)$$

and (Taylor)

$$\bar{H} = k(\hat{\theta}) \approx k(\theta) - (\hat{\theta} - \theta)k'(\theta)$$

so

$$\frac{(\hat{\theta} - \theta)k'(\theta)}{\sigma_h(\theta)/\sqrt{n}} \approx \mathcal{N}(0,1)$$

so

$$(\hat{\theta} - \theta)k'(\theta) \approx \mathcal{N}\left(0, \frac{1}{n}\sigma_h^2(\theta)\right)$$

and

$$\hat{\theta} \approx \mathcal{N}\left(\theta, \frac{\sigma_h^2(\theta)}{n(k'(\theta))^2}\right),$$

which is the assertion.

Example $X \sim f(x,\theta) = \theta e^{-\theta}$, for $x > 0$ choose $h(x) = x$, then $\mathbf{E}\bar{H} = \mathbf{E}(X) = k(\theta) = \theta^{-1}$, $k'(\theta) = \theta^{-2}$,

$$\sigma_h^2(\theta) = Var(X) = \theta^{-2}$$

$$\widehat{\theta} = \frac{1}{\bar{X}} \approx \mathcal{N}\left(\theta, \frac{\theta^2}{n}\right)$$

M.I.e. asymptotics

Definition The function

$$\begin{aligned} I(\theta) &= E_\theta \left\{ \frac{\partial^2 l(\theta, x)}{\partial \theta^2} \right\} \\ &= \int \frac{\partial^2 \log f(x, \theta)}{\partial \theta^2} f(x, \theta) dx \end{aligned}$$

is called *Fisher's information per observation*

$$\begin{aligned} I_n(\theta) &= E_\theta \left\{ \frac{\partial^2 l(\theta, \mathbf{x})}{\partial \theta^2} \right\} \\ &= \int \frac{\partial^2 \log f(\mathbf{x}, \theta)}{\partial \theta^2} f(\mathbf{x}, \theta) d\mathbf{x} \end{aligned}$$

is called *Fisher's information* for a sample of size n

Regularity condition **R**:

(A) The range of values x does not depend on θ

(B) the first 3 partial derivatives w.r.t. θ of f are integrable w.r.t. x

Theorem 2 *Under (R), Fisher's information for a sample of size* n

$$\begin{aligned} I_n(\theta) &= -\int \frac{\partial^2 \log f(\mathbf{x},\theta)}{\partial^2 \theta} f(\mathbf{x},\theta) d\mathbf{x} \\ &= \int \left(\frac{\partial \log f(\mathbf{x},\theta)}{\partial \theta} \right)^2 f(\mathbf{x},\theta) d\mathbf{x} \end{aligned}$$

Proof. We have

$$\int f(\mathbf{x}, \theta) d\mathbf{x} = 1$$

for all θ, so differentiate

$$\begin{aligned} 0 &= \int \frac{\partial f(\mathbf{x}, \theta)}{\partial \theta} d\mathbf{x} \\ &= \int \left(\frac{\partial \log f(\mathbf{x}, \theta)}{\partial \theta} \right) f(\mathbf{x}, \theta) d\mathbf{x} \end{aligned}$$

as $\frac{\partial \log f}{\partial \theta} = \frac{f'}{f}$; differentiate again:

$$\begin{aligned} 0 &= \int \frac{\partial^2 \log f(\mathbf{x}, \theta)}{\partial^2 \theta} f(\mathbf{x}, \theta) d\mathbf{x} \\ &\quad + \int \left(\frac{\partial \log f(\mathbf{x}, \theta)}{\partial \theta} \right) \frac{\partial f(\mathbf{x}, \theta)}{\partial \theta} d\mathbf{x} \\ &= \int \frac{\partial^2 \log f(\mathbf{x}, \theta)}{\partial^2 \theta} f(\mathbf{x}, \theta) d\mathbf{x} \\ &\quad + \int \left(\frac{\partial \log f(\mathbf{x}, \theta)}{\partial \theta} \right)^2 f(\mathbf{x}, \theta) d\mathbf{x} \end{aligned}$$

where we applied $\frac{\partial \log f}{\partial \theta} = \frac{f'}{f}$ again; this finishes the proof.

Theorem 3 *Under (R), random sample, as* $n \to \infty$,

$$\hat{\theta} \approx \mathcal{N}\left(\theta, \frac{1}{I_n(\theta)}\right) = \mathcal{N}\left(\theta, \frac{1}{nI(\theta)}\right)$$

Sketch of Proof. Under random sample,

$$l(\theta, \mathbf{X}) = \sum_{i=1}^{n} \log f(X_i, \theta)$$

Assume

$$\frac{\partial l}{\partial \theta}|_{\theta=\hat{\theta}} = 0$$

Taylor:

$$\begin{aligned} \frac{\partial l}{\partial \theta} &= \frac{\partial l}{\partial \theta}|_{\theta=\hat{\theta}} \\ &\quad +(\theta - \hat{\theta})\frac{\partial^2 l}{\partial \theta^2}|_{\theta=\theta^*}, \end{aligned}$$

where θ^* is between θ and $\hat{\theta}$; so

$$\frac{1}{\sqrt{n}}\frac{\partial l}{\partial \theta} = \sqrt{n}(\theta - \hat{\theta})\frac{\partial^2 l}{n\partial \theta^2}|_{\theta=\theta^*}, \qquad (1)$$

Note

$$\frac{1}{\sqrt{n}}\frac{\partial l}{\partial \theta} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log f(X_i,\theta)}{\partial \theta}$$

and

$$\mathbf{E}\frac{\partial \log f(X_i,\theta)}{\partial \theta} = \int \frac{\partial \log f(x,\theta)}{\partial \theta} f(x,\theta)dx = 0$$

and

$$\mathbf{E}\left\{\left(\frac{\partial \log f(X_i,\theta)}{\partial \theta}\right)^2\right\} = I(\theta)$$

so by C.L.T.

$$\frac{1}{\sqrt{n}}\frac{\partial l}{\partial \theta} \approx \mathcal{N}(0, I(\theta))$$

Also

$$\mathbf{E}\frac{\partial^2 \log f(X,\theta)}{\partial \theta^2} = -I(\theta),$$

so by the law of large numbers

$$\frac{\partial^2 l(\theta;\mathbf{X})}{n\partial \theta^2} \to -I(\theta) \qquad (2)$$

in probability.

Assume $\theta^* \to \theta$ in probability, then (2) also true for θ^*

Collect; (1) gives

$$\begin{aligned}
\sqrt{n}(\theta - \widehat{\theta}) &= \frac{1}{\sqrt{n}}\frac{\partial l}{\partial \theta}\left\{\frac{\partial^2 l}{n\partial\theta^2}|_{\theta=\theta^*,}\right\}^{-1} \\
&\approx -\frac{1}{I(\theta)}\mathcal{N}(0, I(\theta)) \text{ in distribution} \\
&= N\left(0, \frac{1}{I(\theta)}\right) \text{ in distribution}
\end{aligned}$$

so, in distribution,

$$\widehat{\theta} \approx N\left(\theta, \frac{1}{nI(\theta)}\right).$$

This finishes the proof.

Example. $X_1, X_2, \ldots$ i.i.d. Bernoulli (θ), $\hat{\theta} = \bar{X}$

$$L(\theta, \mathbf{x}) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$$

and

$$l(\theta, \mathbf{x}) = \sum x_i(\log(\theta) - \log(1-\theta)) + n\log(1-\theta)$$

so

$$\begin{aligned} \frac{\partial l}{\partial \theta} &= \frac{1}{\theta}\sum x_i + \frac{1}{1-\theta}\sum x_i - \frac{n}{1-\theta} \\ &= \frac{n}{1-\theta}\left(\frac{\bar{x}}{\theta} - 1\right) \end{aligned}$$

and

$$\left(\frac{\partial l}{\partial \theta}\right)^2 = \frac{n^2}{(1-\theta)^2}\left(\frac{(\bar{x})^2}{\theta^2} - 2\frac{\bar{x}}{\theta} + 1\right)$$

and $\mathbf{E}\bar{X} = \theta$,

$$\mathbf{E}(\bar{X}^2) = Var(\bar{X}) + \theta^2 = \frac{\theta(1-\theta)}{n} + \theta^2,$$

so

$$\begin{aligned} I_n(\theta) &= \mathbf{E}\left\{\left(\frac{\partial l(\theta,\mathbf{X})}{\partial\theta}\right)^2\right\} \\ &= \frac{n^2}{(1-\theta)^2}\left(\frac{\theta(1-\theta)}{n\theta^2}+\frac{\theta^2}{\theta^2}-2\frac{\theta}{\theta}+1\right) \\ &= \frac{n}{\theta(1-\theta)}. \end{aligned}$$

Note: $I_n(\theta) = (Var(\bar{X}))^{-1}$

Also

$$\frac{\partial^2 l}{\partial\theta^2} = \frac{n}{(1-\theta)2}\left(\frac{\bar{x}}{\theta}-1\right)-\frac{n}{\theta^2(1-\theta)}\bar{x}$$

so

$$\begin{aligned} E\left(-\frac{\partial^2 l}{\partial\theta^2}(\mathbf{X})\right) &= -\frac{n}{(1-\theta)^2}\left(\frac{\theta}{\theta}-1\right)+\frac{n\theta}{\theta^2(1-\theta)} \\ &= \frac{n}{\theta(1-\theta)} \end{aligned}$$

gives the same answer, as it should.

Confidence intervals from point estimators

If $\widehat{\theta} \approx \mathbf{N}(0, \sigma^2(\theta))$ then

$$\frac{\sqrt{n}(\widehat{\theta} - \theta)}{\sigma(\theta)} \to \mathcal{N}(0,1) \text{ in distribution}$$

thus have *approximate pivot*

As θ unknown: replace $\frac{\sigma(\theta)}{\sqrt{n}}$ by $\frac{\sigma(\widehat{\theta})}{\sqrt{n}}$, the so-called *standard error*

Then

$$\begin{aligned} 1-\alpha \;\; \approx \;\; \mathbf{P}(\widehat{\theta} - z_{\alpha/2}\frac{\sigma(\widehat{\theta})}{\sqrt{n}} &< \theta \\ &< \widehat{\theta} + z_{\alpha/2}\frac{\sigma(\widehat{\theta})}{\sqrt{n}}) \end{aligned}$$

gives an approximate 100 $(1-\alpha)$% confidence interval

Example. Random sample, $f(x,\theta) = \theta e^{-\theta x}$, for $x > 0$

Then $\widehat{\theta} = \frac{1}{\overline{X}}$ and $I(\theta) = \frac{1}{\theta^2}$, so

$$\widehat{\theta} \approx \mathcal{N}\left(\theta, \frac{\theta^2}{n}\right)$$

and also

$$\widehat{\theta} \approx \mathcal{N}\left(\theta, \frac{\widehat{\theta}^2}{n}\right)$$

gives approximate 95% confidence interval

$$\frac{1}{\overline{x}} \pm 1.96 \frac{1}{\sqrt{n}\overline{x}} = \frac{1}{\overline{x}}\left(1 \pm \frac{1.96}{\sqrt{n}}\right)$$

Warning: left limit can be less than 0

Rule of Thumb: approximation good if $n \geq 30$; if underlying distribution is symmetric: already reasonably good for $n \geq 20$

Small sample properties

$T = t(\mathbf{X})$ estimator; $\mathbf{X}$ density $f(\mathbf{x}, \theta)$

Bias $b(\theta) = \mathbf{E}_\theta(T) - \theta$

Mean-square error (M.S.E.)

$$\mathbf{E}_\theta\{(T-\theta)^2\} = Var_\theta(T) + b(\theta)^2$$

Would like: small bias and small m.s.e.

Example: $X_1, X_2, \ldots$ i.i.d. $U([0, \theta])$;

$$\widehat{\theta} = X_{(n)}, \quad \widetilde{\theta} = \frac{n+1}{n} X_{(n)}$$

$\mathbf{E}\widehat{\theta} = \frac{n}{n+1}\theta$ not unbiased, $\mathbf{E}\widetilde{\theta} = \theta$, but

$$Var(\widehat{\theta}) = \frac{n}{(n+1)^2(n+2)}\theta^2$$

versus

$$Var(\widetilde{\theta}) = \frac{1}{n(n+2)}\theta^2 > Var(\widehat{\theta})$$

Theorem 4 Cramer-Rao lower bound (Information inequality

Assumptions: $X_1, \ldots, X_n$ *sample, density* $f(\mathbf{x}, \theta)$, T *unbiased, and (R)*

Then

$$Var(T(\mathbf{X})) \geq \frac{1}{I_n(\theta)}.$$

Note: Under regularity (R), MLE is asymptotically unbiased, achieves Cramer-Rao lower bound asymptotically

Proof. T unbiased, so

$$\theta = \int t(\mathbf{x}) f(\mathbf{x}, \theta) d\mathbf{x}$$

Differentiate both sides w.r.t. θ:

$$\begin{aligned} 1 &= \int t(\mathbf{x}) \frac{\partial}{\partial \theta} f(\mathbf{x}, \theta) d\mathbf{x} \\ &= \int t(\mathbf{x}) f(\mathbf{x}, \theta) \frac{\partial}{\partial \theta} \log f(\mathbf{x}, \theta) d\mathbf{x} \\ &= \mathbf{E}\left(T(\mathbf{X}) \frac{\partial}{\partial \theta} \log f(\mathbf{X}, \theta) \right) \end{aligned}$$

Cauchy-Schwarz inequality: (stated in statistical terms) For any two random variables,

$$(Cov(U, V))^2 \leq Var(U) Var(V)$$

Put

$$U = T(\mathbf{X}), \quad V = \frac{\partial}{\partial \theta} \log f(\mathbf{X}, \theta)$$

then

$$\begin{aligned}
\mathbf{E}V &= \int \left(\frac{\partial}{\partial \theta} \log f(\mathbf{x}, \theta) \right) f(\mathbf{x}, \theta) d\mathbf{x} \\
&= \int \frac{\partial}{\partial \theta} f(\mathbf{x}, \theta) d\mathbf{x} \\
&= \frac{\partial}{\partial \theta} \int f(\mathbf{x}, \theta) d\mathbf{x} \\
&= \frac{\partial}{\partial \theta} 1 = 0,
\end{aligned}$$

so

$$Cov(U, V) = \mathbf{E}(UV) - \mathbf{E}(U)\mathbf{E}(V) = \mathbf{E}(UV) = 1$$

and

$$Var(V) = \mathbf{E} \left(\frac{\partial}{\partial \theta} \log f(\mathbf{X}, \theta) \right)^2 = I_n(\theta),$$

so

$$VarT(\mathbf{X}) \geq \frac{Cov(U, V)^2}{Var(V)} = \frac{1}{I_n(\theta)}$$

This finishes the proof.

Extension: $T(\mathbf{X})$ biased, bias $b(\theta)$, then

$$VarT(\mathbf{X}) \geq \frac{\left(1 + \frac{db}{d\theta}\right)^2}{I_n(\theta)}$$

Proof: As before, only,

$$E(T(\mathbf{X})) = \theta + b(\theta) = \int t(\mathbf{x}) f(\mathbf{x}, \theta) d\mathbf{x}$$

Differentiate both sides with respect to θ; gives

$$1 + \frac{db}{d\theta}$$

instead of 1; and then

$$Cov(U, V) = 1 + \frac{db}{d\theta}.$$